KTH Computer Science
and Communication

Human interaction as a model for spoken dialogue system behaviour

Anna Hjalmarsson

Doctoral Thesis
Stockholm, Sweden 2010

Cover: Peter Sellers in Being There (1979). © Getty Images 2010.
"*Chance, a simple gardener, has never left the estate until his employer dies. His simple TV-informed utterances are mistaken for profundity.*"
www.imdb.com (August 3, 2010)

TRITA-CSC-A 2010:10
ISSN-1653-5723
ISRN–KTH/CSC/A--10/10-SE
ISBN 978-91-7415-703-1

KTH School of Computer Science and Communication
SE-100 44 Stockholm
SWEDEN

Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges till offentlig granskning för avläggande av teknologie doktorsexamen i tal- och musikkommunikation med inriktning på talkommunikation fredagen den 3 september 2010 klockan 10.00 i F3, Kungl Tekniska högskolan, Lindstedtsvägen 26, Stockholm.

Tryck: Universitetsservice US AB

Abstract

This thesis is a step towards the long-term and high-reaching objective of building dialogue systems whose behaviour is similar to a human dialogue partner. The aim is not to build a machine with the same conversational skills as a human being, but rather to build a machine that is human enough to encourage users to interact with it accordingly. The behaviours in focus are *cue phrases*, *hesitations* and *turn-taking cues*. These behaviours serve several important communicative functions such as providing feedback and managing turn-taking. Thus, if dialogue systems could use interactional cues similar to those of humans, these systems could be more intuitive to talk to. A major part of this work has been to collect, identify and analyze the target behaviours in human-human interaction in order to gain a better understanding of these phenomena. Another part has been to reproduce these behaviours in a dialogue system context and explore listeners' perceptions of these phenomena in empirical experiments.

The thesis is divided into two parts. The first part serves as an overall background. The issues and motivations of humanlike dialogue systems are discussed. This part also includes an overview of research on human language production and spoken language generation in dialogue systems.

The next part presents the data collections, data analyses and empirical experiments that this thesis is concerned with. The first study presented is a listening test that explores human behaviour as a

model for dialogue systems. The results show that a version based on human behaviour is rated as more *humanlike*, *polite* and *intelligent* than a constrained version with less variability. Next, the DEAL dialogue system is introduced. DEAL is used as a platform for the research presented in this thesis. The domain of the system is a trade domain and the target audience are second language learners of Swedish who want to practice conversation. Furthermore, a data collection of human-human dialogues in the DEAL domain is presented. Analyses of cue phrases in these data are provided as well as an experimental study of turn-taking cues. The results from the turn-taking experiment indicate that turn-taking cues realized with a diphone synthesis affect the expectations of a turn change similar to the corresponding human version.

Finally, an experimental study that explores the use of talkspurt-initial cue phrases in an incremental version of DEAL is presented. The results show that the incremental version had shorter response times and was rated as more efficient, more polite and better at indicating when to speak than a non-incremental implementation of the same system.

Acknowledgements

First and foremost, I would like to thank my primary supervisor Rolf Carlson whose gentleness and thoroughness has provided me with great support over the years. Rolf amazing capacity to provide wise and insightful advice and he has always been available when I needed him. I am also very grateful for the help and support provided by my secondary supervisor, Joakim Gustafson. Joakim has an ability to view things from innovative perspectives and has contributed with many new and stimulating ideas (and references).

Much of the work presented in this thesis is done in cooperation with others and I have really enjoyed working with all of you! Thanks to, Preben Wik and Jenny Brusk for exciting work on the DEAL dialogue system. Many thanks to Gabriel Skantze for providing help with and tools for dialogue system development as well as stimulating work within the GENDIAL project. Thanks to Jens Edlund and Mattias Heldner for cooperations as well as stimulating discussions related to the vision of humanlike dialogue systems. Another person that has provided me with much help and valuable comments is Julia Hirschberg. Julia contributed with a lot of inspiration and enthusiasm during her year as a visiting professor at KTH and I also had the pleasure of working with her and Noémie Elhadad during my six months stay at Columbia University. I would further like to thank Jenny Klarenfjord for her help with the data collection and the annotations of the DEAL corpus.

The final version of my thesis was completed over the summer and I am very grateful to Rebecca Hincks who spent parts of her vacation proof-reading. I also wish to thank the anonymous reviewers that have provided me with helpful comments that have greatly helped to improve the overall quality of this thesis. Furthermore, the studies presented in this thesis are of an empirical nature and could not have been done without my patient subjects who had to put up with peculiar and sometimes tedious experiments.

Thanks also to my roommates Preben and Giampiero for being great company and answering all kinds of questions. I would further like to thank Rolf Carlson and Björn Granström for heading the speech group and Anders Askenfelt for heading the Department of Speech, Music, and Hearing. Thanks also to all the people who work here for creating such a friendly and stimulating place to work in.

A source of great inspiration throughout my years as a PhD student is all the work that has been done off-work. Thanks to all the people with whom I discussed my research with during conferences, GSLT retreats and after work beer at Östra Station. I would also like to thank Anick with whom I shared PhD student experiences while running in Lill-Jansskogen.

There is also a group of people who has contributed with a great deal of support by taking my mind off work: My supportive and loving family, Göran, Lena, Martin and Eva and all my amazing friends.

Finally, I would like to thank the most important person in my life, Tom. I am endlessly grateful to all your support you have given me especially during the last couple of months. And thanks also for all the help with the graphic design of this thesis. It would not have looked even half as impressive without you!

This thesis was carried out at KTH with the support of the Swedish Graduate School of Technology (GSLT) and the Centre for Speech Technology (CTT). The Ragnar and Astrid Signeul Foundation has also contributed with some of the travelling expenses.

Contents

Publications and contributors

The majority of the work presented in this thesis has already been published in journals and conference proceedings. Some of this work has partly been done in collaboration with others. The publication list presented below specifies the details of the collaborations.

Chapter 2

Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. Speech Communication, 50(8-9), 630-645.

Chapter 2 is partly based on the discussion presented in the publication referenced above. In this publication, Anna Hjalmarsson contributed to the research discussed in "Off-line data manipulations" and manuscript authoring.

Chapter 5

Hjalmarsson, A., & Edlund, J. (2008). Humanlikeness in utterance generation: effects of variability. In Perception in Multimodal Dialogue Systems - Proceedings of the 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, PIT 2008, Kloster Irsee, Germany, June 16-18, 2008. (pp. 252-255). Berlin/Heidelberg: Springer.

Anna Hjalmarsson contributed to the main part of this work. Jens Edlund contributed to manuscript authoring.

Chapter 6

Hjalmarsson, A., Wik, P., & Brusk, J. (2007). Dealing with DEAL: a dialogue system for conversation training. In Proceedings of SIGdial (pp. 132-135). Antwerp, Belgium.

Anna Hjalmarsson, Preben Wik and Jenny Brusk all contributed to the ideas behind the system. Anna Hjalmarsson was responsible for implementing the dialogue modules set up in Higgins (Pickering, Galatea and Ovidius). Anna Hjalmarsson and Preben Wik were responsible for setting up the dialogue manager. Preben Wik developed the graphical user interface.

Wik, P., & Hjalmarsson, A. (2009). Embodied conversational agents in computer assisted language learning. Speech Communication, 51(10), 1024-1037.

Anna Hjalmarsson was responsible for manuscript authoring the DEAL section.

Chapter 7

Hjalmarsson, A. (2008). Speaking without knowing what to say... or when to end. In Proceedings of SIGdial 2008. Columbus, Ohio, USA.

Hjalmarsson, A. (in press), The vocal intensity of turn-initial cue phrases and filled pauses in dialogue, In Proceedings of SIGdial, Tokyo, Japan.

Chapter 8

Hjalmarsson, A. (in press), The additive effect of turn-taking cues in human and synthetic voice, Submitted to Speech Communication.

Hjalmarsson, A. (2009). On cue - additive effects of turn-regulating phenomena in dialogue. In Proceedings of Diaholmia. Stockholm, Sweden.

Chapter 9

Skantze, G., & Hjalmarsson, A, (in press), Towards Incremental Speech Production in Dialogue Systems, In Proceedings of SIGdial. Tokyo, Japan.

Anna Hjalmarsson and Gabriel Skantze both contributed to the user experiment, data analysis and manuscript authoring. Gabriel Skantze was responsible for the incremental implementation of DEAL.

Part I. Introduction and background

1. Introduction

1.1. Motivation

Throughout the past decades, researchers in the fields of artificial intelligence and speech technology have been engaged in the challenge to build machines that interact with users through spoken language. Employing speech in computer interfaces is motivated by the ease whit which humans communicate. Spoken face-to-face conversation is the primary setting of human interaction. It is an efficient and robust way to interact that leaves the hand and the eyes of the speaker free to perform other tasks. Speech is also an efficient way to express complex meaning, engage in social relationships and solve problems. Still, the latter types of features are rarely explored in today's spoken dialogue systems. In fact, interacting with most of today's dialogue systems is more similar to filling out web-based forms than engaging in conversation. The work presented in this thesis focuses on dialogue systems that to a larger extent follow the principles of human communication. While conversational dialogue systems are not necessarily superior to any other type of speech interface, they have appealing potential, allowing researchers and system designers to explore the full potential of spoken language. For example:

- Humans are experienced speakers and this experience is something dialogue system designers may benefit from. A dialogue system with conversational capabilities may encour-

age users to transfer some of the knowledge gained from their long experience in human communication. Dialogue systems perceived as similar to human conversational partners may also be more intuitive and engaging to talk to (for discussion see Edlund et al., 2008).

- Dialogue systems that engage in humanlike dialogue open up new and interesting areas of research and new domains such as computer games and tutoring systems where the interaction itself is entertaining and meaningful to the user (c.f. Iuppa & Borst, 2007; Gustafson et al., 2004).

- Dialogue systems that engage in humanlike dialogue can be used in controlled experiments in order to further explore human behaviour and the underlying cognitive processes of human language processing (c.f. Schlangen, 2009; Edlund & Beskow, 2009).

Human language processing, however, is very complex, and building a machine with the full conversational powers of a human being is not realistic. Instead, in this thesis, it is anticipated that dialogue systems that to some extent display human behaviour will affect users *willing of suspension of disbelief* to perceive these systems as humanlike conversational partners. That is, as long as the interaction is intuitive, the users will be prepared to accept the system's limited capabilities in order to exploit or enjoy the functionalities of the system.

1.2. Thesis focus

Dialogue system designers and research have focused on how to build systems that in a prompt, accurate and syntactically well-formed way provide the user with a particular piece of information. In order to reduce speech recognition errors and misunderstanding, the system makes explicit requests, limiting the users' ability to interact freely. This style of interaction encourages users to interact with dialogue

systems in a way that is more similar to how we interact with other machine interfaces than to how we normally interact through speech in everyday face-to-face conversation.

The aim of this thesis is to explore the potentials of *humanlike* spoken language generation in dialogue systems. More specifically, the work presented here investigates the behavioural patterns and listeners' perceptions of a number of verbal behaviours that serve communicative functions on the interaction level of dialogue. One part of this work has been to collect, identify and analyze the target behaviours in human-human interaction in order to gain a better understanding of these phenomena. The other part has been to reproduce these behaviours in a dialogue system context and explore listeners' perceptions of these behaviours in empirical experiments.

This thesis also presents DEAL, a spoken dialogue system for conversation training. DEAL is a game with a spoken language interface designed for second language learners. The system is intended as a multidisciplinary research platform where challenges and potential benefits of combining elements from computer games, dialogue systems and language learning can be explored. The DEAL domain is a flea market where a talking animated agent is the owner of a shop where used objects are sold.

1.2.1. Human speech production

It is often argued that interlocutors produce speech *incrementally* and on-line as the dialogue progresses, using information from several different sources (c.f. Kempen & Hoenkamp, 1982; Kilger & Finkler, 1995 and Levelt, 1989). Hence, while speaking, processes at all levels – semantic, syntactic, phonologic and articulatory – work in parallel to render the message under construction. This is an efficient processing strategy since speakers may employ the time devoted to articulating the first part of a message to plan the rest. However, speech production in dialogue is limited by time restrictions, and interlocutors need to somehow connect different segments of speech

and try to maintain a coherent dialogue structure. Thus, to accommodate the current dialogue context, speakers incrementally modify the message they are about to convey.

The focus of this thesis is how to employ human dialogue behaviours in spoken dialogue systems to make users' interaction with these systems more intuitive. The thesis primarily explores three different conversational behaviours, namely *cue phrases*, *hesitations*, and *turn-taking cues*.

1.2.2. Cue phrases

Central to the concept of incremental speech production is that articulation can be initiated before the speaker has a complete plan of what to say. Speakers often initiate new turns with *cue phrases* (Gravano, 2009). Cue phrases or so-called *discourse markers* are a class of linguistic devices used to signal pragmatic and semantic relations between different segments of speech. Examples in English are: "oh", "well", "now", "then", "however", "you know", "I mean", "because", "and", "but" and "or" (c.f. Schourup, 1999). Though these lexical expressions have relatively little propositional impact at the local speech segment level, cue phrases serve significant pragmatic functions that help our addressees segment and structure the dialogue at different communicative levels. For example, cue phrases are often used to give feedback, indicate a change of topic or signal turn-management functions.

In this thesis, cue phrases are first manually annotated and analyzed in a corpus of dyadic face-to-face conversations. The aim of these analyses is to identify standardized expressions that can be employed in spoken dialogue systems to connect segments of speech incrementally. The focus is on turn-initial cue phrases and how these elements can be employed to signal that the system claims the floor.

1.2.3. Hesitations

To produce speech incrementally in smaller segments without a complete plan of what to say, speakers occasionally need to alter or refine previous speech segments in order to adjust to new incoming information. Behaviours associated with such modifications are often referred to as *disfluencies*. As the term suggests, disfluencies are often regarded as irregularities in what is otherwise described as a smooth flow of speech. Yet, psycholinguistic research suggests that disfluencies do not necessarily affect comprehension negatively (c.f. Brennan & Schober, 2001). If anything, interlocutors appear to make use of these phenomena in order to coordinate the interaction with their conversational partners. Though these findings suggest that such behaviours could signal important communicative functions, they are rarely generated in today's dialogue systems.

1.2.4. Turn-taking cues

One crucial aspect in dialogue systems is to control *turn-taking*, that is to regulate the flow of dialogue contributions between the system and the user. However, very few dialogue systems use sophisticated methods to manage turn-taking. The systems are generally poor both at detecting users' end of turns and at generating appropriate turn-management behaviour to help users discriminate momentary pauses from ends of turns. Humans do not generate speech in regular constant pace of vocalized segments, but in streams of fragments in varying sizes (Butterworth, 1975). If the motivation is to produce speech in a similar fashion, an important aspect is to help users discriminate momentary pauses from ends of turns in order for them to identify appropriate places to speak. Duncan (c.f. Duncan, 1972, Duncan & Fiske, 1977) suggests that speakers attend to various lexical and non-lexical behavioural cues or signals in the message of the preceding speaker. If dialogue systems could use similar strategies to communicate appropriate places for users to take the turn, turn-taking in such systems will be more intuitive.

1.3. Thesis goals

Natural language has been studied from many perspectives and for many different purposes. The majority of this research has focused on written texts, monologues or task-oriented dialogue. However, natural language in its most frequent setting, spontaneous conversation, is still relatively unexplored. The work presented in this thesis explores human dialogue behaviours used to maintain dialogue at the interactional level and how these behaviours are perceived in the context of a dialogue system. The overall goal is to investigate the potential benefits and possibilities of producing speech in dialogue systems in a more humanlike manner. More specifically, the motivation is to explore a set of human behaviours to signal communicative functions in dialogue systems. It is further proposed that these behaviours can be employed in dialogue system capable of incremental processing. The motivation for doing this is to provide the system with a set of communicative signals that can be used to make the system's processes more transparent to the user.

1.4. Thesis overview

The thesis is divided into two parts.

1.4.1. Part I

Part I serves as an overall introduction and background of this thesis. Chapter 2 discusses the issues and motivations of humanlike spoken dialogue systems. Chapter 3 presents a background to human language production. The focus of this chapter is on a set of conversational behaviours that serve important pragmatic functions in conversation. Chapter 4 presents a background to spoken language generation in dialogue systems. Some critical aspects of how generation in such systems can be made more humanlike are discussed.

1.4.2. Part II

Part II starts with a presentation of a listening test that explores how a dialogue system with human behaviour is perceived compared to a system with constrained human behaviour. Chapter 6 presents DEAL, a dialogue system for second language learning of Swedish. The target audience is language learners who want to practise Swedish through conversation. A dialogue system with human conversation skills is therefore desirable. Chapter 7 presents a data collection of human face-to-face dialogues in the DEAL domain. This chapter also describes the annotation and analyses of a set of conversational behaviours in this corpus. The focus is on how cue phrases, hesitations and turn-taking cues can be used to accommodate incremental speech production under time constraints. Chapter 8 presents an experimental study of turn-taking cues. Chapter 1 presents an experimental study with an incremental version of DEAL. An overall discussion on the contributions of this thesis and future work is provided in Chapter 1. Finally, there is a reference list with all referenced publications and a collection of appendixes (A-G).

2. Designing humanlike spoken dialogue systems

This chapter is concerned with the motivations and implications of using human spontaneous conversation as a model for spoken dialogue systems. A basic assumption behind this work is that human conversational behaviour can be generated in spoken dialogue systems in such a manner that they are perceived as having similar communicative functions as they do in human-human conversation.

2.1. Conceptual metaphors in interface design

Metaphors are devices for seeing one thing in terms of something else. Lakoff & Johnson (1980) challenged the traditional view of metaphors as something poetic by proposing that metaphors have a fundamental role in human cognition. They further argue that not only are metaphors extremely common, but they also shape conversation as well as how we think and act. Lakoff & Johnson (1980) even go as far as claiming: "*our conceptual system…is fundamentally metaphoric in nature*". To perceive something through a *conceptual metaphor* is to understand an idea or domain in terms something else.

When designing software applications, it is essential to make the capabilities as well as the limitations of the system visible to the user. In order to visualize the system's functionalities, system developers

sometimes use conceptual metaphors as themes for applications. A well-known example is the desktop in Windows and Mac operating systems. Here, the user interface is designed graphically as a desktop where the user can store objects such as documents or folders.

The choice of metaphor in conceptual design should not be taken lightly since it creates certain expectations from the users. There are many examples of interfaces designed according to farfetched metaphors. The metaphors used in these systems confuse rather than help their users if the gaps between the source domains and the target domains are too wide. As a result, the use of metaphors in interaction design is hardly without controversy. Norman (1998) states two principles when designing for people: (1) provide a good conceptual model and (2) make things visible. A good conceptual model is a model that allows us to predict the effects of our actions. A basic assumption is that a metaphor changes how we perceive and interact with its subject (Burke, 1969). A user can learn how to interact with an interface without the use of a metaphor, however, a system consistent with a wisely chosen metaphor will likely be adopted faster and without much training or explicit instructions. This idea of cross mapping (see Figure 1) between a source domain and a target domain is central in conceptual design.

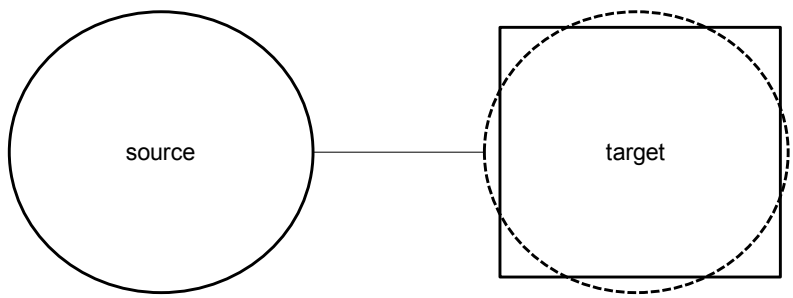


Figure 1. Understanding the unfamiliar through the familiar – schematic illustration adapted by permission from an original by Jeffrey J. Morgan

2.2. Conceptual metaphors in speech interfaces

Building applications that allow users to interact with machines through conversational speech has been a challenging goal in both artificial intelligence and speech technology for some time. At present, there are many different speech applications available to the public. Examples of typical domains include in-car systems, timetable information, and customer service call-routing. This variety of tasks and domains allows for many potential design metaphors. However, this thesis argues that users mainly perceive dialogue systems through either a *human metaphor* or an *interface metaphor*. In support of this argument, Riccardi & Gorin (2000) noted that there was a bimodal distribution in the length of users' response to a greeting prompt. One of these was similar to the length of responses to greetings in human-human interaction, and was brief and fragmental. Riccardi & Gorin (2000) call the latter type "menu-speak".

2.2.1. The human metaphor

A dialogue system perceived through a *human metaphor* is system perceived as an interlocutor, a dialogue partner with humanlike conversational abilities. Caporeal & Heyes (1997) claim that humans are inclined to *anthropomorphise* – that is, to personify inanimate objects and ascribe them human features. Furthermore, based on a series of experiments, Reeves & Nass (1996) argue that also people with much technical experience tend to personify a wide range of artefacts. There is also a long tradition of describing computer processes in terms of human activities. Operating systems "read" and "write" to disk and software applications use "dialogue" boxes. Regardless whether people have a tendency to personify all artefacts or not, the idea that users personify and allot dialogue systems human capabilities is not farfetched. Spoken language is primarily a human activity. It is through human interaction that we learn to speak and this is also how spoken language is most frequently used. There are dia-

logue systems that already employ human features in order to encourage users to perceive these systems in the view of a human metaphor. For example, a number of commercial chatbots (c.f. ALICE, http://www.alicebot.org/) and other types of spoken language interfaces use animated embodied humans, so-called embodied conversational agents (ECAs), (c.f. Cassell et al., 1999). There is also the Loebner prize[1], which is a formal instantiation of the Turing Test. The Loebner prize gold medal is awarded to chatbots that are indistinguishable from a human.

While there are good reasons to believe that users perceive dialogue systems through a human metaphor, it is also plausible that user perceive dialogue systems as some type of machine interface.

2.2.2. The interface metaphor

Saffer (2005) claims that metaphors degrade over time as we become more familiar with an application. For example, does anyone think of computers as desktops anymore? As we become familiar with a new application, this application likely becomes a new potential metaphor. Hence, it is plausible that users perceive dialogue systems in the light of their previous experiences with speech applications or other types of machine interfaces.

A majority of the commercial dialogue systems available to the public are systems developed to provide the user with a specific piece of information. Typical domains include timetable information, travel booking or customer service call-routing. The aim of these systems is to provide the user promptly, accurately and unobtrusively with a particular piece of information. Similar services are also frequently provided using web-based forms which use other input devices such as mouse or a keyboard. This has strongly influenced the design of these interfaces and they are likely perceived metaphorically as some type of machine interface. In this thesis, this metaphor will be referred to as the *interface metaphor*. Characteristic to the interface

[1] http://www.loebner.net/Prizef/-loebner-prize.html

metaphor is that speech is used in a way that is similar to the way a keyboard or mouse is used. Below is an example of how a dialogue consistent with an interface metaphor can unfold (see Example 1). The dialogue is a made-up example taken from Riccardi & Gorin (2000). The example is used to illustrate the following scenario: User U tries to speak freely on several occasions (U_1, U_3), but system S fails to recognize U's intentions. In order to address this problem, S makes clarification requests (S_2, S_4) that describe in a detailed manner how U should provide the information. To accommodate these guidelines, U responds in brief menu-speak style utterances (U_2, U_4).

S1. Please say collect, calling card or operator.
U1. I would like to reverse the charges to Nancy.
S2. Please say collect, calling card or operator.
U2. Collect, please.
S3. Please speak the telephone number now.
U3. The number is 1 2 3 4 5 6 7 8 9 0 area code 1 2 3.
S4. Invalid telephone number. Please speak the telephone number now.
U4. 1 2 3 1 2 3 4 5 6 7 8 9 0.

Example 1. Dialogue excerpt taken from Riccardi & Gorin (2000)

This example illustrates how the dialogue system prompts are designed to restrict the users' behaviour in order to facilitate automatic speech recognition and language understanding. Below some of the typical characteristics in which human-machine dialogues differ from human-human dialogues are listed.

- Spontaneous speech is characterized by high rates of repetitions, revisions, hesitations and false starts. These phenomena aggravate speech recognition and language understanding and dialogue systems are often designed to reduce these types of behaviour. For example, Karsenty (2002) suggests that dialogue systems should avoid open ended questions and instead use explicit requests in order to help users structure their responses and avoid long utterances. The aim of this approach is to avoid burdening the users with high planning demands which are associated with high disfluency rates (Oviatt, 1995). Oviatt also presents results that show that humans use fewer disfluencies when speaking to dialogue systems. Hence, it appears as dialogue system users are aware of the difficulties that these cause and adjust their behaviour accordingly.

- In spontaneous speech, speakers often use fragmental utterances rather than full expressions. The dialogue excerpt below is an example of a fragmental utterance:

  U1. How much is the red one?
  U2. And the blue?

Fragmental utterances are efficient, but in order to interpret this utterance, the entity that this refers to needs be recovered from context. In dialogue systems, such constructions can lead to errors if the system is unable to identify or misunderstands which entities these refers to. It has been shown that humans have higher frequencies of syntactically well-formed phrases and use more full sentences when they were led to believe that they were talking to a machine than when talking to another human (Fraser & Gilbert, 1991a).

- A majority of today's dialogue systems are designed to perform a single task and the dialogue leading up to the completion of this task leaves no room for side tracks such as in-between small talk.

- Human interlocutors often overlap each other. Whereas some of these overlaps are results of miscalculated speaker endings, many overlaps appear as intentional and well calculated (Jefferson, 1986). For example, overlaps occur frequently during laughter, greetings, turn-transitions and feedback. However, overlapping talk is problematic for speech recognition and is generally viewed as an error in dialogue systems. To avoid simultaneous talk, the system is typically designed to turn silent immediately if interrupted by the user. Furthermore, many dialogue systems use silences for end of turn detection and the system only overlaps the users when long pauses are mistaken for turn-endings.

- Compared to human speakers, dialogue systems have a very limited vocabulary. Users tend to imitate the system's vocabulary, and a basic design principle is to only generate words that the system is able to recognize. Additionally, Hauptmann & Rudnicky (1988) present results which show that humans use a reduced vocabulary when talking to a machine compared to when talking to a person.

Except for the technical difficulties associated with the behaviours listed above, there are other potential reasons for why these phenomena are not modelled in dialogue systems. For example: (1) the behaviours are regarded as irrelevant for task completion or/and (2) the behaviours are regarded as mistakes or errors.

This thesis does not argue that one metaphor is "better" than another. All depends on the domain and the criteria of the system, as well as the services that it provides. It is also possible that the inter-

face metaphor and the human metaphor attract different users. Tomko & Rosenfeld (2004) present a user study of the Speech Graffiti system. The system is an approach to standardizing user's language in to a small subset of keywords and expressions in order to reduce speech recognition errors and simplify language understanding. Before interacting with the system, the users need to learn the particular language that the system understands. The results from the user study showed that the subjects with the highest word error rates and the lowest user satisfaction scores were the same users who preferred a natural language interface – Movieline – to the Graffiti system. This group of subjects was further characterized as the subjects with least computer programming background. Thus, it is possible that system designers may want to allow users to choose between interfaces designed according to different metaphors based on their individual preferences.

From now on we move away from the interface metaphor and focus on the qualities and potentials of designing dialogue systems consistent with a human metaphor.

2.3. Motivations of increased humanlikeness in spoken dialogue systems

In the area of spoken dialogue systems, researchers and system developers try to build systems that benefit and exploit the features of spoken language. Yet, humans are experienced speakers and today's speech interfaces are far from being as skilled conversation partners as humans are. In order to cope with complexity, many dialogue systems try to restrict the user's behaviour, for example by only requesting one piece of information at a time. Such requests help to restrict the users vocabulary and syntax, which in turn simplifies automatic speech recognition and language interpretation (c.f. Kamm et al., 1997), but they also limit how the users can express themselves. Face-to-face conversation is a natural, efficient, and robust way for humans to interact. By means of conversation, speakers communi-

cate complex meaning, engage in social relationships, and discuss future events as well as things that happened in the past. Dialogue systems with more humanlike capabilities could encourage users to speak freely and this will allow system developers to exploit and profit from the features that are so exceptional to spoken language. Furthermore, humanlike dialogue systems open up for new and interesting approaches to dialogue system research. Below three potential approaches are listed:

- New areas of use – There are many aspects of conversational speech that is relatively unexplored in today's dialogue systems. If dialogue systems to a larger extent were capable of producing and understanding conversational speech, these systems could be used for more complex tasks such as negotiation, abstract reasoning and social interaction. Examples of new and challenging domains include computer games and tutoring systems where the interaction itself is entertaining and meaningful to the user (c.f. Iuppa & Borst, 2007; Gustafson et al., 2004).

- Conversational affordance – the term affordance refers to an object's or environment's intrinsic qualities to guide its use. A dialogue system with sophisticated conversational capabilities needs to encourage its users to explore these features. Conversational affordance refers to a system's ability to do so. For example, if a system uses prosodic or lexical cues to yield the turn to the user, it needs to be verified that these cues have a turn-yielding effect. In order to explore conversational affordance, criteria that can be used to evaluate whether the users responses match the human behaviours are needed.

- Dialogue systems as cognitive models – Schlangen (2009) proposes that human cognition can be explored by means of spoken dialogue systems. By modelling a certain aspect of human cognition in a dialogue system, we can make this phenomenon testable in an online interactive setting. According to Schlangen (2009) "...artificial agents embody a theory of communication, whose adequacy is evaluated through the reactions it provokes in a naturalistic setting". Thus, dialogue systems that engage in conversation with human speakers can be used in controlled experiments to further explore human cognition. The domain and the context, however, needs to be carefully considered in order to make the experiment ecologically valid and the constraints of the system clear to the users.

In conclusion, dialogue systems designed according to a human metaphor have appealing potential. Nevertheless, the design community has often argued against the use of anthropomorphism in user interfaces. According to Shneiderman (1995):

> *"Anthropomorphic terms and concepts have continually been rejected by consumers, yet some designers fail to learn the lesson."*

Shneiderman further argues that interfaces that invite anthropomorphism risk being an "empty promise" and perceived as having vague goals and functionalities. This argument is based on the design principle of *direct manipulation,* which claims that machine interfaces should allow their users to directly control and manipulate the interface in order to achieve their individual goals (for discussion see Shneiderman & Maes, 1997). An agent designed to invite anthropomorphism will likely be perceived as holding human capabilities. If these capabilities are not supported, there is a mismatch between the interface and its design metaphor which in turn gives the users false expectations. However, as pointed out by Don et al. (1992),

today's computers have new and interesting sectors of applications and whether to anthropomorphise an interface or not should depend on what functionalities it provides. According to Brennan (see Don et al., 1992):

> *"There are classes of things that are done better with speech and natural language than with direct manipulation. These things include delegating complex or redundant actions and doing anything that's not in the here and now… We should stop worrying about anthropomorphism and work on making systems capable of behaving as coherent interactive partners."*

Another objection to the use of anthropomorphism is the *uncanny valley* (Mori, 1970). According to the hypothesis of *uncanny valley,* users feel discomfort when the boundary between machines and humans is blurred. It is argued that users' positive reactions and feeling of empathy increase in line with a robot's behaviour becoming more human. Yet, at a certain point, the machine becomes *too* humanlike and the user's positive experience changes into a feeling of revulsion. Whether this dip in the proposed curve of increased user compassion exists for speech interfaces is, however, yet to be explored.

2.3.1. How human is humanlike?

To build a mechanical human is an overwhelming and probably impossible task. However, the aim of using a conceptual metaphor in system design is not to make users believe that they are actually interacting with the metaphor. Instead, to know a conceptual metaphor is to know the mappings between the source and its target. According to Cassell (2007), we should strive for *"a machine that acts human enough that we respond to it as we respond to another human"*. In line with Cassell's vision, the aim of implementing the conversational behaviours explored in this thesis is to guide users towards a human metaphor rather than some other metaphor.

An often-used concept within computer games and works of fiction is *willing suspension of disbelief* (e.g. Hayes-Roth, 2004). This phrase refers to a human motivation to overlook some of the non-realistic elements of a work of fiction in order to be entertained. This notion can also be considered relevant for dialogue systems. That is, in order to be perceived as coherent with a human metaphor, dialogue systems need to encourage users to suspend some of their disbeliefs and interact with these systems in a way that is similar to interacting with another human being. Yet in order to do this it is not necessary for them to believe that they are actually speaking to another person. The goal of humanlikeness is still high-reaching and this visionary goal needs to be approached in more practical terms. The rest of this chapter discusses how the research area of humanlike dialogue systems can be approached empirically, although it is still in its infancy.

2.3.2. Symmetry

Related to the vision of humanlikeness is the issue of how to match the system's conversational skills with the users' expectations. An oft-stated design principle in dialogue systems is the principle of *symmetry* – that is, that the system should be capable of understanding all the behaviours that it evokes. However, since a system with a full range of conversational capabilities will probably not occur in the near future, it is essential to match the system capabilities with its expectations without necessarily being able to understand all of the behaviours that it evokes. Thus, in order to be successful, the system does *not* need to match human competencies at all levels. Within limited domains, this thesis explores how to employ human behaviours in dialogue systems which are far from being as sophisticated information processors as humans are. Allen et al. refers to this assumption as the Practical dialogue hypothesis (Allen et al., 2001, p. 3): "*The conversational competence required for practical dialogues, while still complex, is significantly simpler to achieve than general human conversational competence*".

2.3.3. Level of analysis

The level of analysis approached in this thesis needs to be discussed. In his work on visual perception, Marr (1982) proposes that cognitive processes can be described on three different levels, the *computational*, *algorithmic*, and *implementational* level. These levels are described as follows (Marr, 1982, p. 470):

- Computational theory: What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?

- Representation and algorithm: How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?

- Hardware implementation: How can the representation and algorithm be realized physically?

The initial aim of this thesis is to identify mappings between a certain conversational behaviour and its pragmatic function in a specific context. The next step is to understand how these behaviours can be modelled computationally, and a later ambition is to implement these models in a dialogue system and evaluate these behaviours in an online interactive setting. Described in terms of Marr's levels of analysis, this thesis is concerned with all three levels of analysis. Some behaviour are analysed only at the computational level whereas others are implemented and evaluated in an on-line interactive setting.

2.4. Data collection methods in humanlike dialogue system development

In order to build more humanlike dialogue systems, methods to collect representative dialogue data, and methods to evaluate the effects of our models are needed. Software design is often done in an *itera-*

tive manner. Rather than developing the system in sequential steps where the design process starts with planning and ends with deployment, iterative system development is done in cycles incrementally. The software system is further evaluated repeatedly during the design process in order to take experiences from previous cycles into account. The next sections describe three methods that are used to collect dialogue data in this thesis, namely *Wizard-of-Oz simulations*, *human-human data manipulation* and *micro domains*.

2.4.1. Wizard-of-Oz simulations

An issue in iterative system development is how to collect representative data of user experience before the system is fully functional. To address this issue, system developers sometimes let users interact with simulations of a system. This method is called *prototyping* (Naumann & Jenkins, 1982). In prototyping, the sophistication of the system simulations ranges from simple mock-ups to fully functional systems.

In dialogue system development, prototyping is typically done through so-called *Wizard-of-Oz* (WoZ) simulations (for a thorough discussion see Wooffitt et al., 1997). In WoZ data collections, the system is simulated with the help of a person, a *Wizard*, who plays the role of the computer. In order to collect representative data, the experimental subjects are led to believe that they are interacting with a fully working speech interface. Yet, in reality the wizard controls the system (or parts of the system). How much of the system that is operated by the Wizard differs and depends on how far along the system development is as well as the aim of the study.

There are several variations of the WoZ paradigm. In *Wizard-as-component*, the Wizard operates a specific function of a system. In this setup, the Wizard needs a well-designed interface which enables him to focus on the task without distractions of other components. The Wizard-as-component setup is typically used when important system functionalities are missing in order to collect data on how users interact with the system. The purpose of the study can also be to study the Wizard's actions – that is, *Wizard-as-subject*. An example

of Wizard-as-subject is presented in Gustafson et al. (2008). In this study, the actions of real customer care operators were studied in order to explore how short feedback utterances can be used to elicit information from customers in a commercial call-routing system.

WoZ studies have many valuable features, but there are also methodological issues associated with this paradigm. One problem is how to setup experiments where the Wizard's behaviour is representative of the systems future behaviour. If the Wizard is allowed to interact freely, the data may not be representative of the end-users' behaviour since the technical and functional constraints of the system has not been considered (Dybkjær et al., 1993). To address this issue, the wizard's behaviour is often constrained, for example by instructing the Wizard to act more machinelike (Dahlbäck et al., 1993). However, research that aims for increased humanlikeness should try to do the opposite. With human behaviour as a gold standard, it is possible to exploit the fact that the Wizard is human and there is no need to instruct this person to "act like a computer would" (Allwood & Haglund, 1992). Regardless of whether the Wizard plays the role of the entire system or operates specific components, the Wizard should be encouraged truly to represent himself. At the same time, it is important to restrict the Wizard's behaviour in order to accommodate the limitations of the domain of and the capabilities of the dialogue system. For example, the operators acting as Wizards in Gustafson et al. (2008) were provided with a "prompt piano", a set of pre-recorded human feedback expressions that could be played back to the callers. In this way, the Wizards' choices of feedback expressions could be explored in a controlled manner within the limitations of the call-routing system's functionalities.

2.4.2. Human-human data manipulation

Spontaneous human-human interaction is a valuable resource that can be used in our efforts towards increased humanlikeness. In *human-human data manipulation*, recordings of dialogues are manipulated in some way in order to explore the effects of some behaviour

that would otherwise be difficult to study in a controlled experiment. Such manipulations can be done either *off-line* or *on-line*. Off-line data manipulation takes place after the dialogue data has been recorded. The effects of these manipulations can later be explored in perceptual experimentation. On-line data manipulations take place during the actual interaction.

An example of on-line data manipulation is presented in Fraser & Gilbert (1991a). In this study, a vocoder was used to transform the wizard's speech on-line, creating an illusion of a dialogue system. The advantage of on-line data manipulation is that the effects of the manipulations can be studied directly since they affect the ongoing conversation. Thus, the vocoder allowed Fraser & Gilbert (1991a) to directly assess the differences in behaviour between subjects speaking to a human operator with a human voice and subjects who were led to believe that they were speaking to a machine by means of the vocoder. A methodological issue in on-line data manipulations is how to control the effects of our manipulations. Thus, if an interlocutor's speech is manipulated, the addressee will likely alter his or her response in order to adjust to this manipulation. The series of behaviours that follows is difficult to predict and the experiment needs to be designed carefully in order to control for interfering variables.

In off-line data manipulations, the dialogues are manipulated in a post-recording step and the manipulated dialogues can later be used as stimuli in a perceptual experiment. An example of off-line data manipulation is presented in Schaffer (1983), who explores the role of intonation in turn-taking. In order to isolate the prosodic realization from the semantic influence, the dialogue segments used as stimuli were band-passed filtered to render them intelligible. In the perceptual experiment, the subjects listened to the segments with and without intelligible speech and judged whether the same speaker was going to continue or if there was going to be a change of speakers. The judgements were analyzed to explore whether the subjects could predict the outcome of the dialogues based on intonation only.

Off-line data manipulations provide researchers with a great deal of control, being able to do fine-tuned manipulations of the recordings in advance. Furthermore, it is possible to manipulate the target behaviour only, without the need to control for unexpected interfering variables. This, however, may also be problematic since the interaction is not affected by our manipulations, as it would be in its original setting. Thus, we risk ending up with dialogues that are no longer representative of the interaction that we try to model.

2.4.3. Micro-domains

Another approach that can be used to explore the effects of a particular dialogue phenomenon is to implement a system that operates within a limited domain. If such a *Micro-domain* system is designed carefully, providing the users with a good conceptual model, a system with limited resources can be tailored to elicit sophisticated interactional data. The most well-known example of a system that successfully operated within a Micro-domain is Eliza (Weizenbaum, 1966). The predictability of the dialogues in Eliza made it possible to create an illusion of a sophisticated system through very simple means..

2.4.4. Evaluation

The data collection methods described in previous sections are useful to collect data that can be used to model human behaviour in spoken dialogue systems. One concern that has been mentioned only very briefly is how to measure the effects of such models. Objective metrics such as the number of words, word error rate (WER) and task success are suitable for use in systems that aim to enable effective information transfer. However, these metrics do not address the dialogue qualities that are aimed for in this thesis. The objective here is related to what is often referred to as *naturalness* in dialogue. The term natural is, however, fuzzy and rarely defined. This thesis adopts Boyce & Gorin (1996) objective: "*Our goal is to design a dialog that is natural, which we define as being one that closely resembles a conversation two humans might have"*. One way to approach the evaluation of

this objective is to employ methods used in the area of *computer-directed* speech. Research in this area, explores differences between human-directed and computer-directed speech. This thesis aims to do the opposite. Namely, the aim is explore if human conversational behaviour in dialogue systems encourages users to speak in a way that is similar to how they speak to human interlocutors. Below, some of the empirical studies that have explored differences in human-directed and computer-directed speech are presented.

Hauptmann & Rudnicky (1988) studied users' behaviour when they were led to believe that they were interacting with an email application capable of natural language understanding. Three different conditions were explored, *speech-to-computer mode*, *speech-to-human mode* and *typing-to-computer mode*. Speech-to-computer mode was speech directed towards the computer, speech-to-human mode was speech directed towards an experimenter who was translating their utterances into typed commands, and typing-to-computer mode was typed natural language directed towards the computer. The results analysis shows several differences between the three experimental conditions. For example, the subjects used significantly more words in the spoken conditions than in the typed condition. However, there was no difference in words per task between speech-to-computer mode and speech-to-human mode. In addition, utterances were longer in the speech-to-computer mode than in the two other conditions.

In another study, Fraser & Gilbert (1991a) explored differences between human and machine directed speech in the domain of a flight information service. The service was managed by a speaker whose voice was manipulated through a vocoder in order to reduce the speaker's prosodic variation and to make the voice sound synthetic. A comparison between the dialogues with the synthesized service agent and a reference corpus collected with a human speaker without a vocoder showed that the human-human dialogues contained more words and more word forms. It was further shown that the human-human dialogues contained more "non-words". These

non-words were unfinished words and "noise words such as 'erm'". Syntactic analyses showed that the dialogues collected with the synthesized voice contained no ellipses and less overlapping talk and fewer relative clauses such as "the plane arrives in the afternoon" than the corpus collected with the human voice.

There are a few studies that have used similar methods to evaluate synchrony in human-machine dialogue rather than differences. For example, to explore the potentials of using humanlike fragmentary questions in dialogue systems, Skantze et al. (2006) manipulated the prosodic realization of one-word questions in a perceptual experiment. The results show that the prosodic realization affected how the participants responded to the system. Furthermore, manual annotations of the responses suggest that the participants interpreted the different prosodic realisations of the clarification requests according to their associated pragmatic meaning[2].

Finally, data collection paradigms such as on-line data manipulations or Micro-domains can be used to compare manipulated human-machine dialogues to a control group of unconstrained human-human dialogues in the same context. This allows us to explore the effects of our independent variables quantitatively. For example, Fraser & Gilbert (1991b), who compared differences in length of speaker turns and differences in the frequency of various linguistic phenomena such as anaphora and filler words (e.g. "eh" and "ehm").

2.5. Summary

This chapter have discussed the use of conceptual metaphors as themes for spoken dialogue systems. The focus has been on the potentials and possibilities of applying a *human metaphor*. A dialogue system viewed in the light of a human metaphor is a system perceived as an interlocutor, a dialogue partner with humanlike conversational abilities. It is argued that a human metaphor is plausible

[2] The pragmatic meaning of the different prosodic realizations were explored in a previous experiment (see Edlund et al., 2005).

since spoken language is primarily a human activity. Users may, however, also perceive dialogue systems through an *interface metaphor* since the style of interaction in these systems sometimes are more similar to web-based forms than spoken conversation. Hence, a system perceived through an interface metaphor is a system perceived as some type of machine interface and is typically responded to in short and command like utterances, so-called *menu-speak* style.

The vision of this thesis is to build humanlike dialogue systems is motivated by the possibilities to explore the full potential of spoken language in these systems. For example, dialogue systems can be employed to engage in complex reasoning in new types of domains such as computer games and tutoring applications. Furthermore, employing conversational behaviours in dialogue systems can evoke users to transfer knowledge from human conversation and this, in turn, can make these systems more intuitive to use. Finally, as proposed by Schlangen (2009), we can implement models of human speech processing in order to test these in an on-line interactive setting and by doing so further explore processes of human cognition. The objective of humanlike dialogue systems does not entail that a dialogue consistent with a human metaphor needs to have all the conversational capabilities of a human speaker. Instead, as put by Cassell (2007), this thesis aims for: *"a machine that acts human enough that we respond to it as we respond to another human"*.

The end of this chapter is concerned with different data collection methods and how to approach evaluation of humanlike dialogue systems. The data collation methods discussed are *Wizard-of-Oz simulations*, *human-human data manipulations*, and *micro domains*. These data collection paradigms allow us to collect data in both on-line and off-line settings. Wizard-of-Oz simulations or fully functional dialogue systems can be used to collect data in an on-line interactive setting. Dialogue data can also be manipulated off-line and studied in perceptual experiments. Finally, it is proposed that comparisons between human-machine dialogue and human-human dialogue in a similar setting can be used to determine whether our models are per-

ceived as intended that is, whether the system's behaviour encourage the users to act more humanlike.

The next chapter discusses the characteristics of human speech production in more detail.

3. Human speech production

In the previous chapter, the overall motivations for this thesis were discussed in general terms. The present chapter is concerned with the subject of our analyses – that is, spontaneous human speech production. The focus is on a set of *interactional cues* – that is, conversational behaviours that help interlocutors coordinate and maintain a coherent dialogue structure.

3.1. Models of human speech production

An influential model that provides a comprehensive overview of the underlying cognitive processes of speech production was presented by Levelt (1989). Similar proposals had been presented earlier by Garrett (1975), Levelt (1983), Kempen & Hoenkamp (1987), and others. According to Levelt, human speech production involves a set of cognitive, motoric and linguistic processes that are (relatively) automated and distributed over different components. The structure of the model is derived from research on reaction times for different language production tasks. The fact that Levelt's proposal is based on human speech processing operating in real-time makes it well suited to serve as an initial outline for speech production in dialogue systems. Figure 2 presents an overview of this model, introducing the subsystems and processing components that are involved in human speech production. According to Levelt there are three main components, namely, the *conceptualizer*, the *formulator*, and the *articulator*.

3.1.1. The Conceptualizer

The conceptualizer produces a first abstract conceptual representation of a message. This involves coming up with an intention to say something and selecting the relevant information to mediate this message. The product of the conceptualizer is a *preverbal message* that has a *propositional* representation. In order to produce this preverbal message, different types of knowledge are needed. First, our *Working Memory* is continually updated with our knowledge of what goes on in the conversation by attending to our own speech as well as what other speakers are saying. The processes of *conceptualizing* also require *encyclopedic knowledge* and *situational knowledge*. Encyclopedic knowledge is the knowledge we have acquired during a lifetime about the world and ourselves. Situational knowledge is the contextual knowledge we have about the dialogue including information about the other speakers and the surrounding environment. The conceptualizer also supervises our speech and adjusts our output to the current state of the dialogue. These *monitoring* aspects will be discussed in Section 3.4. The preverbal message produced by the conceptualizer is the input to the *Formulator*, which is the next component of Levelt's speech production system.

3.1.2. The Formulator

The formulator transforms the preverbal conceptual message into a linguistic representation. The formulator has two subcomponents: the *grammatical encoder* and the *phonological encoder*. The grammatical encoder retrieves lemmas whose meanings match parts of the preverbal message. The lemmas are stored in a mental lexicon that has information about the lexical item's meaning as well syntactic information. The grammatical encoder produces a *surface structure*, an ordered string of lemmas, which is the input of the phonological encoder. The phonological encoder makes use of the phonological and morphological information and produces an articulatory or phonetic plan for each lexical item and for the speech segment as a whole. This representation is internal and yet to be realized as *overt speech*.

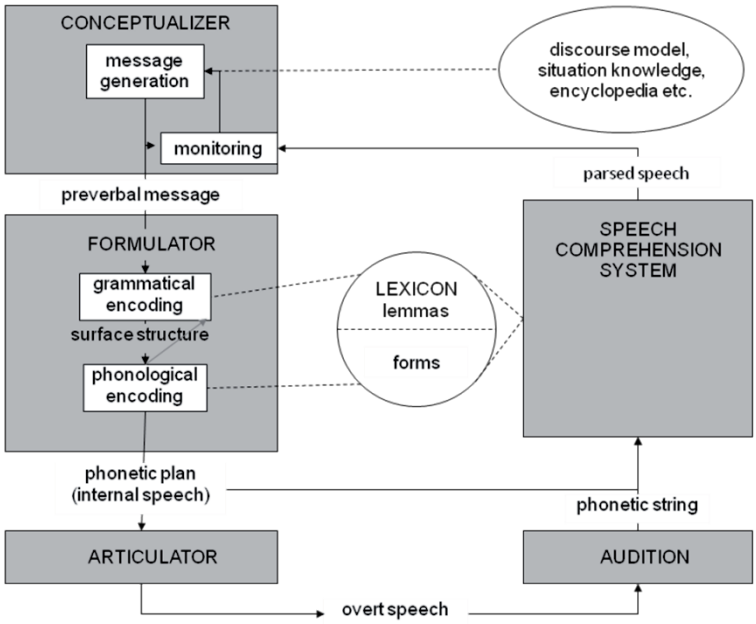


Figure 2. Blueprint of speech production, reproduced with permission from an original by Levelt (1989)

3.1.3. The Articulator

The articulator executes the phonetic plan by articulatory movements that transform the message into overt speech. These processes are not necessarily synchronized with the generation of the articulatory plan. In order to cope with these asynchronies, the articulatory plan may need to be stored temporarily in the *articulatory buffer*, which is where the articulator retrieves chunks of articulatory plans in order to execute them.

3.2. Incremental speech processing

Levelt further argues that the cognitive processes responsible for language production work partly in parallel. Kempen and Hoenkamp refer to this phenomenon as *incremental processing* (Kempen & Hoenkamp, 1982; Kempen & Hoenkamp, 1987). When speaking, processes at all levels (e.g. semantic, syntactic, phonologic and articulatory) work in parallel to render the utterance. Kempen & Hoenkamp (1987) postulates that the preverbal message has a *linearized* structure that is split up into messages that can be realized in a piecemeal fashion. This is an efficient processing strategy since subsequent processing levels do not need to wait until the entire utterance has been completed. Instead, as soon as one process level has encoded an utterance unit, the next level of processes can start to operate on this constituent (see Figure 3).

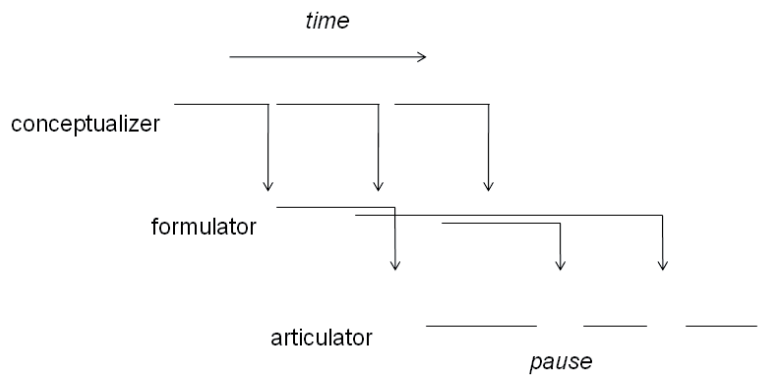


Figure 3. Parallelism in incremental speech production – schematic illustration reproduced with permission from an original by De Smedt, 1990

A central issue related to incremental speech production is what type of units the different levels of processes operate on and to what extent speakers plan ahead. Empirical studies have used reaction times to study the extent of conceptual (Butterworth, 1975), syntactic planning (Meyer, 1996), phonologic planning (Damian & Dumay,

2007), and phonetic planning (Wheeldon & Lahiri, 1997). However, there is disagreement as to how much of an utterance is planned before articulation starts. Damian & Dumay (2007) present results which suggest that the unit of phonological planning is larger than one word and that the size of the planning unit appears to be unaffected by time pressure. Furthermore, it has been shown that hesitation phenomena are more likely to occur before longer utterances (Shriberg, 1994). These findings and spoonerisms such as "cuff of coffee" (Fromkin, 1973) suggest that planning (at least) goes beyond the immediate phonological word[3]. Still, while speakers may not produce speech phone by phone or word by word, there are findings which show that planning affects articulation, suggesting that these processes occur almost simultaneously. For example, in a sentence production task that included arithmetic calculations, the duration of utterances was affected by problem difficulty (Ferreira & Swets, 2002). Moreover, Brysbaert et al. (1998) showed that the response time for answers to arithmetical problems was shorter when the problem was presented so that the first term in the calculation corresponded to the first phonological word uttered. Hence, when the problem was presented as tens plus units (20+4), French speakers (who say "twenty four") responded faster than Dutch speakers did. When the problem was presented in reverse order, that is, units plus tens (4+20), the Dutch speakers, who say numbers in reverse order (24 is read "four and twenty"), had shorter response times.

Whereas the intermediate levels of speech production and the units of processing are still under debate (c.f. Caramazza, 1997), the literature referenced above provide empirical evidence of incremental speech production. With this fundamental property of language production in mind, the next section focuses on the interactional aspects of dialogue.

[3] See Nespor (1986) for a definition of phonological words in terms of stress values.

3.3. Interaction control

Levelt's model of speech production has been very influential. Still, his approach has been criticised for not addressing the interactive and social aspects of dialogue. For example, in a discussion about Levelt's model O'Connell (1992) claims that: "*In short, the instrumental, social, conscious, and interactive components of speaking skills are systematically subordinated to the cognitive or are disregarded*".

Many researchers have stressed the view of human interaction as a joint project performed in close coordination (c.f. Clark, 1996). Garrod & Pickering (2004) argues that the production of dialogue contributions is facilitated by its contextual foundation and its success depends on a mutual understanding of the situation. Interlocutors' tendency to produce output that share characteristics with previously perceived input is often discussed in relation to conversation. Many different terms have been used to refer to this phenomenon, for example *alignment*, *accommodation* and *entrainment*. Alignment between speakers has been studied on many different levels of processing, e.g. syntactic, semantic and phonological, and it is argued to be a largely automated process (c.f. Pickering & Garrod, 2006; Dijksterhuis & Bargh, 2001).

According to Moore (2007), feedback is an essential part of dialogue (or any type of behaviour) and the processes of speech perception and speech production are tightly coupled. Moore further argues that when we speak, our future behaviour is affected by the feedback we receive. This idea stems from Powers (1973), who introduced the *perceptual control theory* (PCT). According to PCT the behaviour of organisms is affected by how we perceive the effects of our actions in relation to our intents. Thus, when we speak, we continually try to evaluate how our listener(s) perceives what we say and adjust our future behaviour to address this feedback. Hence, small variations in our conversational behaviour may also affect the response we receive from our dialogue partners. In the rest of this chapter, we will explore some of these conversational behaviours in more detail.

3.3.1. Are "disfluencies" disfluent?

The behaviours we are about to discuss in this chapter are typically referred to as "disfluencies". Examples of so-called disfluencies are mid-word interruptions, repetitions, restarts, mispronunciations, pauses and fillers. Previous studies have shown that disfluencies are very frequent in dialogue. There are about six disfluencies per 100 words (c.f. Fox Tree, 1995; Brennan & Schober, 2001). The term "disfluencies" suggest that these phenomena are irregularities or deviations from a flow of fluent speech. This view is salient in Chomskyan linguistics where disfluencies are regarded as deviations in human performance from some optimal delivery of an utterance (Chomsky, 1965). According to this tradition, disfluencies are not part of language and need to be excluded from studies of linguistic theory. Furthermore, Lickley & Bard (1996) presents findings which show that listeners have difficulties transcribing disfluent segments of speech. The subjects' task was to transcribe utterances incrementally word by word, as it was played to them. Analyses of the transcriptions show that compared to fluent speech, disfluent speech segments were often transcribed incorrectly and words were left out. Furthermore, when subjects were asked to give a verbatim transcription, the recall of disfluent word segments was worse than the recall of fluent speech segments (Bard & Lickley, 1997).

Another related view is that disfluencies indicate problems in speech production, but that these phenomena are relevant and need to be studied in order to better understand human language production (c.f. Goldman–Eisler, 1961; Levelt, 1989). Furthermore, these phenomena appear to provide listeners with important pragmatic information. For example, a study by Brennan & Williams (1995) shows that listeners rated responses to general knowledge questions as less certain when they contained filler words than when they did not. Another study by Brennan & Schober (2001) shows that fillers help listeners compensate for disruptions and delays in spontaneous speech. In four experiments, participants identified visual objects on a screen from verbal descriptions. Compared to fluent descriptions,

the participants identified the target object faster, but not less accurately when the description contained an error repair with a filler as an editing term. This effect was also salient compared to fluent descriptions where the disfluent segments had been replaced with a pause of equal length. These findings suggest that it is the disfluency per se and not the extra processing time that facilitates comprehension. Furthermore, mid-word interruptions resulted in fewer errors than between-word interruptions. The mid-word interruption appears to indicate that the word was "deleted" and this highlighting facilitates identification of the correct object. Arnold et al. (2003) present another study on comprehension of disfluent speech. In this study, an eye-tracker was used to explore the effect of disfluencies on comprehensions. The results show that when verbal descriptions of visual objects contained filled pauses, the participants looked more frequently at objects that had not been mentioned earlier in the experiment. These findings suggest that filled pauses signal new information. However, as argued by Corley & Stewart (2008), it is difficult to determine if this effect is a result of a tendency to predict the difficult referent or to rule out the easy one. In another experiment (Corley et al., 2007), participants were presented with words visually and asked to guess which words they had heard previously in the experiment. If the word was preceded by a disfluency when it was first introduced, it was more likely to be recognized the second time. Furthermore, Watanabe et al. (2008) showed that listeners predict hesitations to be followed by words with high complexity.

A third approach is to treat disfluencies as intentional devices used to control the interaction in dialogue (c.f. Clark, 1996). It is argued that speakers use repeats, repairs, fillers and prolonged syllables intentionally to synchronize their own internal processes with those of their addressees. Clark & Fox Tree (2002) present the filler-as-word hypothesis and argue that filled pauses are not automatic or unintentional consequences of speech processing, but linguistic signals controlled by speakers. Analyses of the London – Lund corpus consisting of face-to-face conversations in British English, suggest

that the filler "um" was followed by longer and more frequent delays than "uh". In line with these results, Clark & Fox Tree stipulate that "uh" is used to express what is expected to be a minor delay whereas "um" signal an expected major delay. However, O'Connell & Kowal (2005) have later questioned these results.

Regardless of whether disfluencies are intentional or not, the research presented above suggests that the term "disfluencies" is misleading. The pauses, repetitions and abrupt mid-word interruptions in spontaneous speech appear to play a prominent role in communication. Hence, from now on, we choose not to use the term "disfluency". Instead, we will focus on how these phenomena are perceived in the context of spontaneous conversation.

3.4. Self-monitoring

A characteristic of spontaneous conversation is that we do not know what lies ahead. Spontaneous speech is produced in real time and our resources for planning of future utterances are limited by time restrictions. We may have a vague plan of what to say next, but as the dialogue progresses, we obtain new information that forces us to revise and refine our plan as we go along. Both speakers and listeners continually update and adjust to the current pragmatic and semantic context of the dialogue. As we speak, in a stepwise fashion we refine, alter and revise our plans of what to say. Theories of such *self-monitoring* processes diverge. A central issue is how the monitoring devices are distributed. Levelt (1989) argues that self-monitoring is a part of the speech comprehension system and is performed by the same processes that attend to errors in other interlocutors' speech. This is the so-called perceptual loop theory. The perceptual loop theory assumes that speech monitoring can be performed in three stages: first during conceptualization, later on a pre-articulatory version of inner speech and finally on overt speech. Van Wijk & Kempen (1987), on the other hand, claim that monitoring devices are dispensed over the speech production system and these attend to in-

termediate results at different levels of processing. A problem with such distributed monitoring, according to Levelt, is the reduplication of knowledge. A monitoring device that controls the output of a certain process must contain similar information as the process itself.

Another theory of speech monitoring is the *spreading-activation theory* (c.f. Dell, 1986). In this connectionist model of speech processing, the modularized approach with processes distributed over different components is entirely abandoned. Instead, speech production is presented as a layered network of nodes. The layers contain different types of nodes with different responsibilities, including the functions of conceptualization, lexicalization and articulation. In order to produce speech, a probabilistic path is taken through the connected nodes, and processes for speech monitoring are intertwined in this network. Whenever a node is activated, there is a two-way activation or priming. Thus, for each connection, such as from a particular concept to a particular lexical item, there is also a bottom-up connection, that is, spreading activation in the opposite direction. This backward activation gives feedback to the previous node in order help confirm that the correct node was activated. If a node is incorrectly activated, the previous node is not activated as much as expected and the error is detected. This spreading-activation theory is argued to allow for immediate repairs of errors. For a more thorough discussion of speech monitoring models see Postma (2000).

The next section moves away from the cognitive models of speech processing and focuses on the consequences of speech monitoring.

3.5. Repairs

Different types of errors are common in fluent speech, and repairs are used to alter or refine erroneous or underspecified segments of speech. Examples of the various aspects that speakers attend to and repair are errors in lexical, syntactic and articulatory performance. Findings presented by Levelt (1983) show that 46% of the speech errors were corrected in a corpus of Dutch speaking subjects describ-

ing visual properties to each other. Whether the subjects did not bother to correct the other half or whether these errors were undetected is difficult to know. Furthermore, the percentage of repaired speech errors increases towards the end of a syntactic constituent. The provided explanation is that in the beginning of utterances speakers are busy planning the rest of the phrase, whereas towards the end, more resources are available for monitoring.

The general structure of repairs is presented as in Figure 4 (or as adaptations of this structure). The *reparandum* is the segment of speech to be repaired or substituted by the alteration. The *moment of interruption* is where the speaker interrupts herself. This is the point Shriberg (1994) refers to as *interruption point* and what Blackmer & Mitton (1991) calls *cut-off*. The editing phase is what Blackmer & Mitton refers to as *cut-off-to-repair time*. Thus, after the point of interruption there is a variable duration of time before the actual repair is initiated. This is the so-called *editing phase* or *interregnum*. This phase can be either silent or, as in the example in Figure 4, include an editing term. Common editing expressions in English include "sorry", "I mean", "er", and "oh". The repair can start immediately after the reparandum, after the editing phase or retrace to an earlier point. In Figure 4 the speaker retraces to an earlier point and starts the repair with "from", which is the preposition prior to the reparandum. This is what Levelt refers to as "span of retracing".

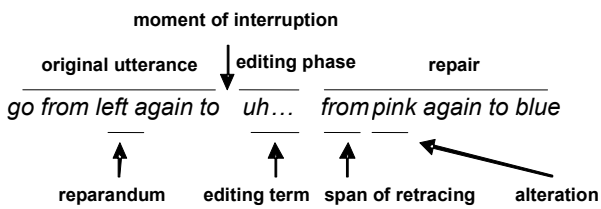


Figure 4. The structure of repairs (from Levelt, 1983)

Levelt further argues that errors are corrected as soon as they are detected. This is the so-called *main interruption rule*. The main interruption rule is based on the observation that speech appears to be interrupted immediately after an error has been detected and cut-offs can be located anywhere in the flow of speech, even at locations that may not appear as linguistically motivated. Early points of interruption in the course of a word suggest that there is pre-articulatory self-monitoring. For example in the following utterance: "*To the left side of the purple disk is a v- a horizontal line*" (from Levelt, 1989 p. 474). The speaker is about to say vertical instead of horizontal but detects this error early, and makes a repair before the word has been fully articulated. The duration after word onset until the repair is shorter than 200 milliseconds, which have been suggested as the average duration it takes from word onset to recognize a word in spontaneous speech (Marslen-Wilson & Tyler, 1980). Levelt refers to errors that occur *before* articulation as *covert repairs* and repairs that occur *after* articulation as *overt repairs*. The perceptual loop theory assumes that a repair of an error passes through the speech production system in sequential order starting from the conceptualizer. However, Blackmer & Mitton (1991) show that 12.4% of the cut-off-to-repair times were zero milliseconds. Since their data also include longer cut-off-to-repair times where planning may have occurred after the cut-off, planning for these immediate repairs must have occurred during fluent speech. The error segment and its repair cannot have been conceptualized as a single unit, but still the repair comes immediately after the error, without any delay. Thus, such immediate repairs cannot be explained by the perceptual loop theory. In line with these findings, Blackmer & Mitton argue that processes responsible for self-monitoring operate in an incremental fashion.

Repairs are often complex and it is difficult to develop comprehensive classification systems. As discussed above, the speech monitoring system appears to operate on different levels of processing. Many classification systems try to capture the underlying intention or problem behind the repair, which is a difficult undertaking. For

example, Blackmer & Mitton (1991) make a distinction between *conceptually based repairs* and *production-based repairs,* where the former are repairs originating in the conceptualizer, whereas the latter originate in the formulator. Furthermore, a common distinction adopted from Levelt is that between *error repairs* and *appropriate repairs*. Whereas error repairs include an erroneous word that needs to be "undone", appropriate repairs make the previous speech segment more precise or appropriate. Blackmer & Mitton (1991) make a further distinction between appropriateness repairs used to replace previous segments of speech, so-called *appropriateness replacements*, and appropriateness repairs used to add extended information to previous speech segments, so-called *appropriateness inserts*. Rather than trying to interpret the motivation behind a repair, Van Wijk & Kempen (1987) focus on their overt realization and use the terms *retracing repairs* and *non-retracing repairs*. Retracing repairs backtrack to an earlier point of the utterance and repeat this segment as it was originally realized or partly modified. Non-retracing repairs replace the reparandum without backtracking. Another repair categorization is so-called *restarts*, *false starts* or *fresh starts* where the speaker abandons a previously initiated segment of speech and starts over (c.f. Heeman & Allen, 1999). For many restarts, there is little or no connection between the abandoned segment and the new restart segment.

There are many different repair classification systems and to describe them all is beyond the scope of this overview: for further reading see for example: Blackmer & Mitton (1991), Maclay & Osgood (1959), Shriberg (1994), and Levelt (1983).

3.5.1. Prosodic realization of repairs

Acoustic analyses show that speakers appear to highlight or mark repairs prosodically. For example, Levelt & Cutler (1983) manually annotated intonation of lexical repairs in a corpus of Dutch speakers. Whether a repair was prosodically marked or not was decided based on the following target question: "*is the prosody of the trouble word roughly the same as the prosody of its correction, or is it different?*"

(Levelt & Cutler, 1983 p. 208). The analysis shows that 45% of the lexical repairs were prosodically marked. The results further suggest that error repairs are marked more frequently than appropriateness repairs. Hence, 53% of the error repairs were marked prosodically, but only 19% of the appropriateness repairs.

Howell & Young (1991) argue that speakers mark repairs prosodically since they need to indicate that the forward flow of speech has stopped and make the alteration intelligible to the listener. It is further argued that both of these functions can be signalled prosodically. Their analyses show that speakers to appear pause just after an interruption. Furthermore, regardless of whether there was a retrace or not, there was an increased stress at the start of the alteration. This tendency was particularly prominent when there was retracing section. These results suggest that speakers pause when the flow of speech has been interrupted and mark the start of the alteration prosodically. Howell & Young (1991) also present experimental results on the comprehensibility of prosodically marked repairs by altering the prosody of a synthesis reproducing repairs. The results support that utterances with an increased stress in the beginning of the alteration and pauses at the interruption point were judged as more comprehensible than a corresponding utterance without these prosodic cues.

3.6. Hesitations

The editing phase in repairs typically contains some type of pause or behaviour that indicates uncertainty. These behaviours will be referred to as *hesitations* in this thesis. Frequently occurring hesitation phenomena include repetitions, fillers such as “um”, “uh”, “eh”, and silences in the middle of a speaker turns. Shriberg (1994) shows that the probability of a “fluent” sentence decreases with sentence length. This may not be surprising since an increased number of words increases the opportunity for “disfluency”. However, the probability of a sentence being initiated with “disfluent” speech also increases with

sentence length, which suggests that "disfluent" phenomena increase with increased cognitive load. Hence, phenomena such as silent pauses, fillers, and repetitions appear to be correlated with difficulties in the planning process.

One of the first to study such hesitation phenomena systematically from a linguistic perspective was Goldman-Eisler (c.f. Goldman–Eisler, 1961; Henderson et al., 1966; Goldman-Eisler, 1972). Goldman-Eisler focused on duration and organization of hesitations in relation to speech rate, and her findings suggest that hesitation phenomena follow regular patterns. These patterns are further argued to be intrinsic, determined by the underlying processes of speech production. According to Henderson et al. (1966): "*The psycholinguistic process appears, from this evidence, to be not a simple Markovian process, but one in which fairly regular periods of planning and internal organisation govern the final speech output for short periods ahead*". Furthermore, Butterworth & Goldman-Eisler (1979) suggest that speakers hesitate in order to plan future output, and speech production consists of two reoccurring phases, planning and execution. The planning phase contains higher frequencies of disfluent phenomena, whereas the execution phase is relatively effortless and fluent. This strict division of labour of planning exclusively during pauses has later been questioned. For example: the error segments immediately followed by repairs in Blackmer & Mitton (1991) data suggest that planning and articulation can occur simultaneously.

The frequency, type and position of hesitation phenomena are claimed to be determined by the amount of cognitive load required for planning future segments of speech. For example, findings by Goldman-Eisler (1958) and Lounsbury (1954) suggest that fillers and pauses within turns are related to the transitional probability of succeeding linguistic events, that is, the probability of a pause depends on the strength of the association between the surrounding lexical items. In support of this view, Tannenbaum et al. (1965) present results that suggest that the contextual probability of lexical items succeeding hesitations is lower than the probability of lexical

items in fluent context. However, Beattie & Butterworth (1979) argue that these studies do not control for whether the frequency of hesitations is related to transitional probability or word frequency. In order to address this issue, Beattie & Butterworth (1979) conducted a study investigating the contextual probability of fillers and pauses while controlling for transitional probability between lexical items and word frequency. The results showed that there was no difference in the distribution of hesitations depending on word frequency when contextual probability was held constant. However, when word frequency was held constant, words with less contextual probability were more likely to be proceeded by hesitations. In conclusion, contextual probability (but not word frequency) was correlated with hesitation probability.

A more recent series of experiments (c.f. Berthold & Jameson, 1999; Müller et al., 2001) explores the effects of an increased demand on the speaker's working memory by distracting visual or auditory stimuli. The results show that an increased cognitive load results in higher frequencies of fragmental utterances, repetitions, false starts, syntax errors, fillers and pauses. Though these findings show that an increased cognitive load results in higher hesitation rates, the distribution of these are not necessarily arbitrary. Oomen & Postma (2004) manipulated speech rate by adding time pressure to speakers describing visual objects connected by lines on a computer screen. Time pressure should increase cognitive load, but the speakers did not produce more fillers in the fast condition than during normal speech rate. However, there was a higher frequency of repetitions. Thus, speakers were more likely to repeat a word than to use a filler in the fast condition. In a later study, Schnadt & Corley (2006) used a similar speech production task, but instead of adding time pressure, the number of linked lines between objects was manipulated and some objects had been blurred. Analyses of the effects of these manipulations show that a higher number of possible lines from an object result in a higher frequency of disfluencies, primarily pauses and repairs. These finding suggest that a speaker produces more fillers

when there are more lexical items to choose from. These results are further supported by findings which show that filled pauses occur more frequently before long utterances (Shriberg, 1994).

Analysis of three different corpora including both face-to-face and telephone conversation in three different languages – Dutch, Swedish and Scottish English – show that pauses within turns are on average longer than silence between speakers (Heldner and Edlund, in press). These irregularities in pause – and between-speaker silence duration suggest that speakers can pause in the middle of turns for rather long periods of time without being interrupted. In an early study on hesitation phenomena in dialogue, Maclay & Osgood (1959) showed that fillers occur more frequently at phrase boundaries than (silent) pauses are more frequent within phrases. These findings further support that hesitation phenomena are not arbitrary and that fillers may be used to maintain the turn after a phrase final segments that would have otherwise been perceived as turn final.

Swerts (1998) presented another study on fillers and discourse structure. Analyses of Dutch monologues showed that phrases which follow major discourse boundaries more often contain fillers. Furthermore, fillers after major breaks often occur phrase initially, whereas fillers in phrases following minor breaks often occur in phrase-internal position. For an overview of psycholinguistic research on the intentionality of disfluencies, see Corley & Stewart (2008).

Other entities that occur frequently in conversation are so-called *cue phrases*. Whereas the behaviours discussed so far in this chapter often are viewed as errors or unintentional side effects of speech production, cue phrases are expressions with an indisputable linguistic status that have been studied in both spoken and written language.

3.7. Cue phrases

Cue phrases are expressions such as “oh”, “well”, “and”, “but”, “or”, “so”, “because”, “now”, “then”, “I mean”, and “y’know”. These expressions are often used to signal relations between different dis-

course segments. Cue phrases are also referred to as *discourse markers*, *pragmatic markers*, *discourse particles*, and *connectives*, to mention a few. Consider the following sentence (from Fraser, 1999 p. 383):

but when do you think he will really get there?

The bold words are cue phrases. If these are removed, the sentence will be structured as follows:

when do you think he will get there?

The example above illustrates that cue phrases are entities that can be removed without affecting the propositional content or the syntactic structure of the sentence. So what functions do these particles serve?

The entities and the relations that cue phrases hold are often ambiguous and it is difficult to give an exact definition of what cue phrases are. Much research within discourse analysis, communicative analysis, linguistics, and psycholinguistics has been concerned with these particles and what kind of relations they hold (see Schourup, 1999 for an overview). A rule of thumb is that cue phrases are words or chunks of words that have little lexical impact at the local semantic level but serve significant pragmatic function. Cue phrases come mainly from the syntactic classes of conjunctions, adverbs, and prepositional phrases (Fraser, 1999). Many cue phrases have different pragmatic functions in different contexts. Furthermore, a majority of these expressions can be used as a cue phrase as well as in a non-cue phrase position that is, in a sentential sense. For example, Fraser (1999) argues that the “and” in sentence A below is a cue phrase, but the “and” in B is not. This is since sentence A’s “and” can be removed without changing the syntax or the propositional content, whereas B’s “and” cannot.

A. John can't go. And Mary can't go either.
B. John and Mary can’t go.

3.7.1. Formal definitions of cue phrases

Early work by Cohen (1984) proposed that cue phrases (Cohen called them "clue words") used in arguments can help listeners process and reveal the argument's structure. While cue phrases have figured in earlier literature (c.f. Levinson, 1983), Deborah Schiffrin (1987) was one of the first to provide a comprehensive overview of these entities. According to Schiffrin (p.31): "*I operationally define markers as sequentially dependent elements which brackets units of talk*". Schiffrin uses units of talk to refer to all kinds of speech segments, including syntactic phrases, propositions, speech acts and tone units. By brackets Schiffrin suggests that cue phrases mark boundaries between units of talk in some way. Furthermore, since several of these markers appear to occur rather freely within a syntactic phrase, they are claimed to be independent of sentential structure. According to Schiffrin's perspective, cue phrases display the discourse relations, rather than create them.

Extensive work on cue phrases has also been presented by Fraser (c.f. Fraser, 1996; Fraser, 1999). Fraser refers to cue phrases as *pragmatic markers* and defines these as entities that signal relations between some part of a discourse segment (that they are part of) and some earlier discourse segment. In contrast to Schiffrin, Fraser proposes that cue phrases *signal* relations rather than illuminate other potential relationships.

Another definition of cue phrases is presented by Schourup (1999). Schourup describes cue phrases as entities that contain features such as *connectivity, optionality*, *weak clause association*, *initiality,* and *orality*. These features can be described as follows. Connectivity refers to cue phrases' ability to signal relations between discourse segments. By optionality Schourup suggests that cue phrases are syntactically optional. The term *weak clause association* is used to describe that cue phrases can occur outside of the syntactic structure or loosely attached to it. Initiality suggests that cue phrases often are positioned initiality in a discourse segment, whereas orality suggests that cue phrases are primarily used in speech. There are sev-

eral different cue phrases taxonomies, but only a few of them are based on dialogue. One example of such a dialogue taxonomy is presented by Louwerse & Mitchell (2003). Louwerse & Mitchell consider cue phrases as cohesive devices that indicate coherence relations between dialogue segments.

There are many different formal definitions of cue phrases and reviewing them all is beyond the scope of this overview. Instead, the rest of this chapter focuses on work that has studied the pragmatic functions and characteristics of these entities in more detail.

3.7.2. Cue phrases in use

In a series of experiments, Fox Tree & Schrock (1999) explored listeners' comprehension of the cue phrase "oh" in spontaneous speech. It was proposed that the high frequency of cue phrases in spontaneous speech results from the fact that when we produce speech online we have less time to structure discourse and therefore employ cue phrases as devices to signal discourse structure. The dialogue data used in the experiments were spontaneous narratives told by students in a face-to-face situation. The subjects pushed a button whenever they heard a target word. Some target words were preceded with an "oh" and some were not. The results show that the subjects' response times were faster when the target followed an "oh" than when it did not. Fox Tree & Schrock conclude that the "oh" function as a "change-of-state marker" which signals that extra attention should be paid to the upcoming speech.

Another study by Bestgen (1998) explores speakers' use of the connective "and" and sequential markers such as "then", "next", and "after" in narratives. The subjects were asked to produce descriptions of what someone was doing based on a list of activities. Each subject produced three narratives, twice with the list and once without. The thematic distance between two topics was judged by another group of subjects in advance. These judgements were used to create a scale of "break levels" that could be used as an independent variable of the thematic distance between two topics. The results show that sequen-

tial markers such as "then", "next", and "after" occurred more frequently between activities with major break levels than between two activities that were semantically related.

Bestgen further proposes that "and" can signal at least two discourse functions in dialogue; *signal* and *trace*. First, "and" can signal continuity, indicating that two phrases are highly related. "And" can also be used as a *trace* that is, to connect phrases that *lack* semantic coherence. In this context, "and" traces difficulties in production rather than signals continuity. The results show that the frequencies of "and" differed between break levels at session one and session three. These sessions were more "difficult" than session two since the subjects were either producing the narrative for the first time or producing it without the list as a reference. No such differences were found during the second session where the subjects had practised the narrative once as well as had access to the list. This observation supports the hypothesis that "and" can be used by speakers who have production difficulties.

In a series of experiments, Hirschberg and Litman explored the automatic classification of cue phrases based on a combination of prosodic and contextual features (c.f. Hirschberg & Litman, 1987; Litman & Hirschberg, 1990; Hirschberg & Litman, 1993). The classification task was to distinguish a word used as cue phrase (DISCOURSE) from the same word used in its sentential sense (SENTENTIAL). Two classification models were presented, one based on contextual features and one based on prosodic features. The analyses showed that intonational features play an important role, specifically pitch accent and prosodic phrasing.

Gravano (2009) explored acoustic and contextual features of *affirmative cue words* (ACW) in a corpus of non-face-to-face conversations in American English. ACW are words such as "alright", "okay", and "mm-hm". For example, it was shown that backchannels, short feedback utterances in the middle of another speaker's turn, typically end in a rising intonation. Turn-initial ACW were presented with a falling intonation and high intensity, whereas turn-final ACW were

produced with low intensity. Similar to Hirschberg and Litman, Gravano (2009) tries to identify contextual and acoustic features that can be used to discriminate between DISCOURSE and SENTENTIAL use of ACW. Furthermore, two additional classification tasks are proposed, to classify: (a) "whether such words convey acknowledgement or agreement", and (b) "whether they cue the beginning or the end of a discourse segment". SVM (Support Vector Machines) models are used to classify ACWs according to 11 different discourse functions associated with the classification tasks described above. The results show that the SVM models approach the error rate of trained human labellers. It is further shown that contextual features that capture the position of the ACW provide the most predictive power, while acoustic features are lower ranked predictors. Furthermore, phonetic features such as the identity and duration of the phones did appear to improve the classification accuracy.

One issue that has only been mentioned briefly hitherto is the segmentation of speech into different units. Both hesitations and cue phrases appear to affect *turn-taking* – that is, how speakers distribute the turn between each other during conversation. This topic will now be discussed in more detail.

3.8. Turn-taking in dialogue

Spontaneous speech is not produced in regular constant pace of vocalized segments, but in streams of fragments in varying sizes (Butterworth, 1975). Much work within linguistics, phonology, phonetics and speech processing has been devoted to the segmentation of spoken language into units of varying length. Written text is segmented by overt symbols such as blank spaces, full stops and commas, but speech has no such explicit cues to the underlying discourse structure. What types of phenomena are used to structure spoken language and how do we know when to speak?

Irregularities in pause duration and turn length suggest that interlocutors cannot use silence duration to discriminate pauses from ends

of turns. An early theory of turn-taking suggests that speakers identify appropriate places to speak by attending to various behavioural cues or signals in the message of the preceding speaker (c.f. Duncan, 1972, Duncan & Fiske, 1977). According to Duncan (1972 p.283): "The proposed turn-taking mechanism is mediated through signals composed of clear-cut behavioural cues, considered to be perceived as discrete". Duncan explored such turn-taking cues in a corpus of face-to-face dialogues in American English. Correlation analyses of these data show that the number of available turn-yielding signals is linearly correlated with listeners' turn taking attempts. When several signals are used in combination, there appears to be an additive effect. However, when speakers employed signals to suppress such attempts, the number of turn-taking attempts radically decreased, regardless of the number of turn-yielding signals.

Influential work by Sacks et al. (1974) describes human turn management as a set of principles motivated by the inclination to avoid gaps or overlaps. They describe a set of technique that speakers employ to allocate the turn and mechanisms for dealing with failures to do so. Furthermore, dialogue is composed by "turn-constructional units", syntactically complete dialogue units that have the property that interlocutors can predict the end of these units in advance. Hence, speakers have a mutual understanding of Transition Relevant Places (TRPs) (Ford & Thompson, 1996). A frequent assumption is that speakers can predict TRPs very precisely, and that a majority of speaker changes are directly adjoining without any overlap or silence. These theories are not compatible with turn-taking based on behavioural cues, since speakers appear need at least 200 milliseconds to verbally react to an auditory stimulus (Izdebski & Shipp, 1978). Instead, de Ruiter et al. (2006) suggests that humans predict upcoming turn-endings by lexico-syntactic content alone after showing that listeners' accuracy in predicting upcoming turn-endings in Dutch dialogues did not decrease when the intonational contour was removed (de Ruiter et al., 2006).

However, recent analysis of turn transitions in spontaneous face-to-face conversation in American English, German and Japanese have shown that pauses and overlaps are in fact normally (Gaussian) distributed (Weilhammer & Rabold, 2003), suggesting that perfectly adjoining transitions are rare. Moreover, analysis of three different corpora including both face-to-face and telephone conversation in three different languages – Dutch, Swedish and Scottish English – show that 41% to 45% of the speaker transitions are longer than 200 milliseconds (Heldner and Edlund, in press). The large number of speaker turns separated by gaps longer than 200 ms suggests that Duncan's theory of turn-taking based on behavioural cues near the end of previous turn is feasible.

The next section presents some of the behaviours that have been suggested as relevant for turn-taking.

3.8.1. Turn-taking cues

Duncan (1972) introduces a number of different turn-taking cues. Behaviours that have a turn-yielding effect include a rising or falling pitch contour, the termination of a hand gesture, a drop in loudness, and completion of grammatical pauses. Behaviours that suppress turn-taking attempts include an intermediate pitch level and socio-centric sequences (stereotyped lexical expressions). In a recent corpus analysis of non-face-to-face, spontaneous task-oriented dialogues in American English, a number of phenomena were found to take place at significantly higher frequencies before speaker changes than before speaker holds (Gravano (2009). These turn-yielding cues include falling or high-rising intonation, a reduced lengthening, a lower intensity level, a lower pitch level, points of textual completion, and longer inter-pausal unit duration. A flat or sustained pitch contour has been reported to have turn-holding functions (see for example Selting, 1996, Koiso et al., 1998). Cutler & Pearson (1986) present results that suggest that long segments of speech are more likely to be judged as turn final. In addition, turn-final speech segments have been shown to be significantly longer than turn-medial speech seg-

ments (Gravano, 2009). In line with Duncan's findings, Gravano's results show support for a linear relationship (positive correlation) between the number of simultaneously available turn-yielding cues and the number of turn-taking attempts.

There are contradictory findings regarding the effect of some turn-taking cues. For example, according to Duncan, a pitch level terminal-junction combination other than an intermediate pitch level in American English is associated with turn-yielding intentions. A more detailed analysis of a rising intonation suggests that a high-rise (H-H%) has turn-yielding effects and a plateau (H-L%) has turn-holding effects, whereas the effects of a low-rising contour (L-H%) are unclear (Gravano, 2009). Local et al. (1986), on the other hand, claim that a rising intonation has both turn-yielding and turn-holding functions in Tyneside English. Swedish has two basic intonation patterns, medial fall (H*L%) and fall-rise (H*LH%) (Bruce, 1977). Thus, in an analysis of the prosodic aspects of turn-taking in Swedish, Edlund & Heldner (2005) make a distinction between patterns with a final rise and a final fall. This analysis shows that a rising intonation was followed by an equal distribution of speaker changes and speaker holds (51% and 49% respectively), implying that the turn-taking effects of a rising intonation in Swedish are unclear.

Local et al. (1986) claim that increased phrase-final lengthening has turn-yielding functions in Tyneside English whereas Gravano (2009) presents results that show that increased phrase-final lengthening in American English have turn-holding effects. In line with Gravano, Ferrer et al. (2003) present results that suggest that the final rhyme of the phrase is lengthened in both cases, but that the lengthening before internal pauses is even longer than before end of turns. Furthermore, the duration of the lengthening is positively correlated with pause length.

3.9. Summary

This chapter has discussed research on human speech production. The research reviewed suggests that there are good reasons to believe that small variations in the way in which speech is delivered play a central role in communication. Many of these behaviours are typically referred to as "disfluencies", and considered as flaws or interruptions in the flow of "fluent" speech. However, research within psychology and psycholinguistics suggest that these behaviours affect how we perceive and respond to a dialogue contribution as well as influences our expectations of future contributions.

The view of repairs and hesitations as traces of underlying cognitive processes has been given a prominent role in this chapter (see for example Levelt, 1989). According to this perspective, spontaneous speech phenomena such as mid-word interruptions, fillers, silent pauses, and repetitions are side effects of problems in the planning process. However, researchers disagree as to what extent these behaviours are intentional. Do speakers strategically employ disfluencies as means of coordination or are these phenomena more or less automatized side effects of human speech processing? Our long experience of human interaction has likely influenced how and when we produce these phenomena and we may occasionally employ these behaviours strategically. Thus, the origin of these phenomena is likely located somewhere in a varying spectrum between intentional strategies and automatized "side effects". This thesis will not address this issue further. Instead, the following chapters focus on the central theme of the thesis, that is, how can interactional cues such as hesitations, repairs, cue phrases and turn-taking cues be employed in spoken dialogue systems to signal similar functions? Before addressing this issue, work on spoken language generation in dialogue systems will be reviewed.

4. Spoken language generation in dialogue systems

This chapter presents an overview of spoken language generation in dialogue systems and discusses some of the related issues.

4.1. Spoken dialogue systems

Spoken dialogue systems are computer applications that interact with users through spoken conversation. These applications deal with a number of challenging tasks that span several disciplines, including linguistics, psychology, and computer science. The processes of spoken dialogue systems are generally distributed over a number of modules ordered sequentially in a pipeline architecture (see Figure 5). The typical processing steps include *automatic speech recognition* (ASR), *natural language understanding* (NLU), *dialogue management*, *natural language generation* (NLG), and *text-to-speech* (TTS).

First, the system needs detect the user's speech and transform it into text. ASR is often error-prone since there is a lot of variation within the speech signal. The error-prone output from the ASR is a bottleneck that the rest of the system has to deal with.

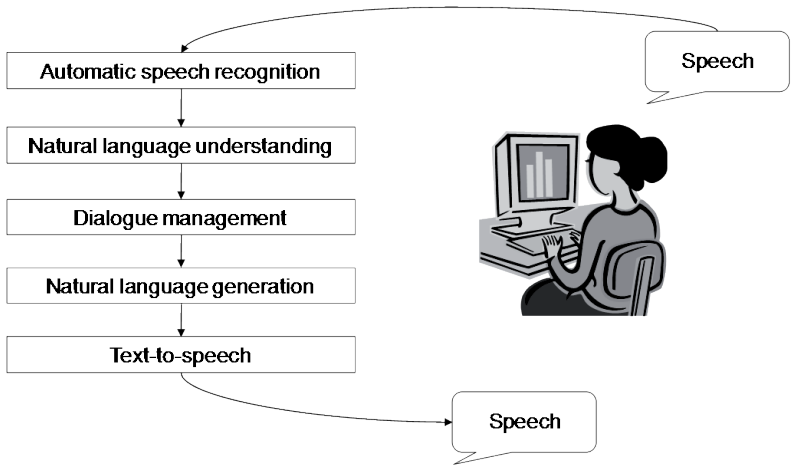


Figure 5. A dialogue system pipeline architecture

The ASR is also responsible for segmenting the continuous speech strings into manageable units. These units are typically "utterances", stretches of speech by one speaker surrounded by silence. The silence threshold used for segmentation differs between systems and is often tweaked to accommodate the requirements of a particular system. The next step, NLU, refers to processes involved when transforming the text into some abstract semantic representation. This involves syntactic and semantic analyses. Dialogue management is the processes responsible for controlling the flow of the dialogue, deciding what action to take based on previous input, and generating a response. The challenges of the dialogue manager include determining if the system has elicited adequate information from the user, contextual understanding, information retrieval and *content planning*. Content planning is the process of selecting what information to present back to the user. The response generated by the dialogue manager is normally on an abstract semantic level. Components responsible for NLG transform this abstract representation into a surface generation (often in some textual form) which can be passed on to the TTS engine and transformed into speech.

4.2. Natural language generation

Natural language generation is the process of deliberately constructing some kind of natural language output – speech or text – from an abstract semantic representation in order to meet some specified communicative goals. In some sense, NLG can be viewed as the inverse of NLU (Dale & Mellish, 1998). NLU transforms natural language input into abstract representations of meaning which can be processed by computers, while NLG transforms abstract representations of meaning into natural language (Figure 6).

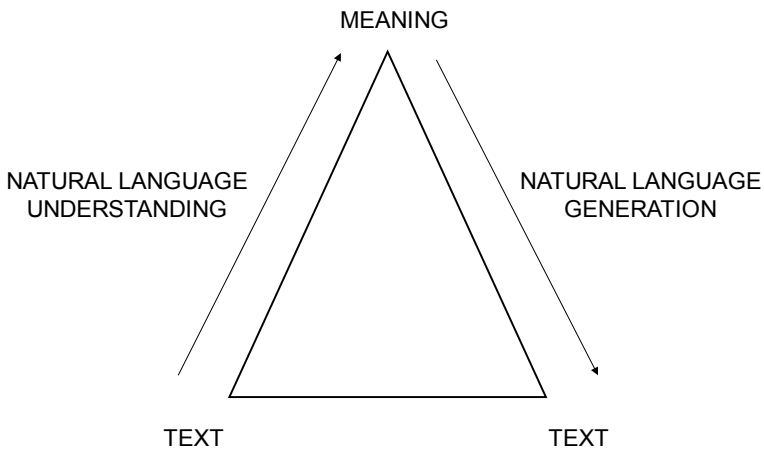


Figure 6. The processes of NLU and NLG

The focus within NLU is on hypothesis management, ruling out possible interpretations of natural language input and determining which the most appropriate one is. The challenges within this field are to deal with ambiguity, under-specification and ill-formed input. The focus of NLG is choice, i.e. choosing between different ways of realizing a message given a specific context.

4.3. Spoken language generation in dialogue systems

Computers were able to produce natural language output years before they were able to process natural language input. Because of the great challenges related to detecting and understanding spoken language, much effort has been put into the research areas of speech recognition and natural language understanding whereas natural language generation has gained less attention. Much research within NLG has been concerned with producing monologues as text or monologues to be synthesized as speech. However, due to the specific challenges of conversational speech, methods for monologue generation cannot be directly applied to spoken dialogue systems. A variety of different techniques has been used to generate spoken language in dialogue systems. Below is a classification of different approaches.

Speech interfaces that prioritize a human-sounding voice and predictability over flexibility typically use *canned speech*. Canned speech is recordings of human speech that are played back to the user in a timely fashion. The speech segments are typically fed in chunks that are scripted for that specific application without any intermediate levels of linguistic or semantic representations. The advantage of this method is that it requires little linguistic knowledge.

Template-based generation allows slightly more flexibility than canned speech. At its simplest form, template-based generation uses scripted segments of speech similar to canned speech but with slots to be filled. Speech can be generated using either larger segments of concatenated speech or text-to-speech (TTS) techniques. However, the abstract semantic representations of input are still mapped directly without any intermediate levels to a surface structure (Reiter & Dale, 1997). Canned speech and template-based generation are sometimes referred to as *shallow* generation since there is no theoretically based linguistic representation of output (Busemann & Horacek, 1998). The next section provides an overview of deep level processing approaches to spoken language generation.

4.4. NLG processes in spoken dialogue systems

For NLG methods that rely on in-depth linguistic analysis it is not obvious where the processes of NLG starts and what knowledge sources they rely on. Neither is there a universally accepted design, and most systems have individual architectural solutions. However, a rough overview of the NLG tasks is presented in Figure 7. The process are split over three different components, the dialogue manager which is responsible for determining "what to say" (content planning), the surface realizer responsible for producing a textual representation (surface realisation), and finally the TTS engine, which is responsible for the acoustic realization. In multimodal systems, processes for determining which modality to use need to be included in the model.

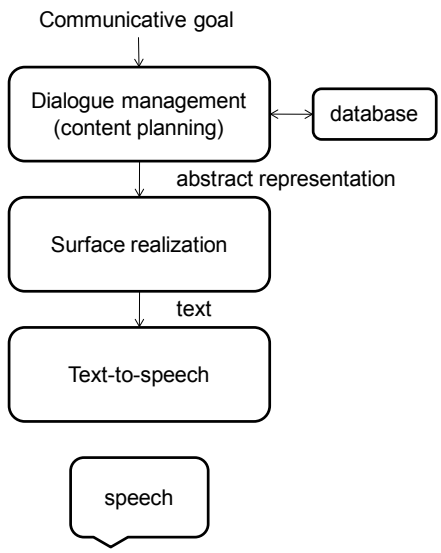


Figure 7. NLG in spoken dialogue systems

4.5. Input

There is no consensus on what the input to an NLG system should be. In the generation of text documents, the entire contents is available from start and there is a possibility to process the discourse as a whole as well as to go back and make changes. In dialogue systems, however, the system does not know how the dialogue will progress and it needs to rely on previous discourse. In order to keep track of the previous discourse, new dialogue contributions are stored in some type of *discourse model*.

4.5.1. Content planning

Content planning or content selection is the task of selecting the content or the communicative intent of the message that is about to be presented to the user. In dialogue systems, this task is generally performed by the dialogue manager (Theune, 2003). At this stage, the message is generated in a semantic abstract form, that is, some symbolic representation that the system uses internally for semantic and pragmatic information. Depending on the complexity of the domain, these representations range from shallow syntactic structures to deep semantics. Whereas there are off-the-shelf technologies available for ASR, TTS and parsing, there are few standard solutions to dialogue management and content planning. Dialogue systems operate in a variety of domains, and selecting what action to take based on a certain input is highly dependent on the domain. As a consequence, there are many individual approaches to content planning.

4.5.2. Surface realisation

In text generation, *surface realisation* is the process of constructing a grammatically correct sentence of the message. To choose between different surface realisations in dialogue systems, the natural language generation component needs access to information available in the discourse model. This information can include the user's lexical choices, whether this concept has been mentioned previously in the

dialogue, and confidence scores from ASR. In dialogue systems, surface realisation often refers to all processes that are not involved in content planning. The processes involved during surface realization include *lexicalisation*, *aggregation*, and *referring expression generation*.

4.5.2.1. Lexicalisation

Lexicalisation is the task of putting words to the concepts in the abstract message. A concept can be expressed in many different ways, and the task of lexicalisation is to choose the most appropriate word given a specific context. Since users tend to imitate the vocabulary of the system, a basic design principle is that the system should be able to recognize all the words that it can generate. Also, the system's choice of words may affect how it is perceived. For example, using a different lexical choice than the user may be perceived as a strategy to correct or refine the user's previous utterance. Thus, the system cannot maintain only an abstract semantic representation of previous utterances, but also needs to keep track of previous lexical choices.

4.5.2.2. Aggregation

Aggregation in text generation is the processes responsible for structuring the document linguistically into paragraphs and sentences. For example, the sentences "*Johan has a book*" and "*Mary has a book*" can be written as "*John and Mary have books*". Aggregation can also include determining how the information should be ordered. According to Appelt (1985), the task of aggregation in spoken dialogue systems is to make the utterances more concise, avoid repetitious language and make the system more intelligible. Since the system in general has no information about the utterances that will follow, aggregation has to be done incrementally on utterance level. Lemon et al. (2003) have implemented aggregation processes in an incremental fashion. Since future utterances are unknown the only way aggregation can be performed is by "retro-aggregating" new utterances with previous ones (see Example 2).

System.	I have cancelled flying to the base
System.	and the tower
System	and landing at the school.

Example 2. Dialogue excerpt taken from Lemon et al. (2003)

4.5.2.3. Referring expression generation

Referring expression generation is determining which expressions to use when referring to entities: definite descriptions or pronouns? A dialogue system that only uses definite descriptions is likely to be experienced as repetitive. Unjustified pronoun use, on the other hand, can cause misunderstandings that can be difficult to recover from. Thus, the choice of noun phrase should be made in order to provide the reader/listener with "sufficient" information to identify the intended referents. Much work on pronouns in computational linguistics has focused on anaphora resolution and the parsing of pronouns rather than how they are generated. According to the centering theory, some entities in an utterance are more central and this imposes constraints on the use of referring expressions (Grosz & Sidner, 1986). The leading assumption has been that a pronoun should be used whenever referring to an entity that is highly prominent in the local discourse. However, research that focuses on the generation of pronouns argues that the centering theory does not account well for patterns of pronouns in naturally occurring texts (c.f. Callaway & Lester, 2001; McCoy & Strube, 1999). Sentence boundaries, distance from last mention, discourse structure and ambiguity are identified as factors which influence pronominialization. In spoken dialogue, pronoun usage is also characterised by the relationship between the speaker and hearer ("you", "I"), how well entities have been established in the context, and vocal stress. The acoustic aspect of speech is an extra dimension that can be used to stress prominence in an utterance that is not revealed by its syntactic structure.

4.5.3. Acoustic realization

The output of spoken dialogue systems is realized using pre-recorded speech or a text-to-speech (TTS) engine. The acoustic realisation of a message plays an important role in language processing, yet few dialogue systems explore the full potential of signalling pragmatic and semantic functions through variations in acoustic and prosodic realization. Regardless whether the system uses pre-recorded or synthetic speech, the systems generally speak in complete sentences that are realized in a "neutral" manner without considering the context in which they are being realized. These realisations often sound monotonous and the intonational patterns are typically suited for monologues or read speech rather than conversational output. A dialogue system that generates more context-aware prosody can help users with lexical disambiguation as well as provide them with semantic and pragmatic information (Hirschberg, 2002). Yet, in order to generate the appropriate rhythm, fundamental frequency and vocal stress, the NLG components need to provide the acoustic realizer with more information than the simple text string. Part of speech tags can be used to resolve homographs, that is, words that have the same spelling but with different pronunciation, and make distinction between questions and non-questions (Jurafsky & Martin, 2000). Prosody is also crucial for the generation of non-lexical utterances such as "hmm", "um" and "aha" since the grounding functionality of these utterances is mainly conveyed through prosody (Ward, 2004). Another functional property of prosody is how intonation affects listeners' expectations of a speaker change. A flat intonation has turn-holding effects whereas a falling intonations increase the probability of a speaker change (Edlund & Heldner, 2005). Emotion is another dimension that can be expressed acoustically. For an overview of work on emotional speech synthesis see Schröder (2001).

4.5.3.1. Synthetic versus natural speech

In dialogue systems, there is a choice between using pre-recorded human speech and synthetic speech. Human speech has often shown

to be more intelligible than synthetic voices (for an overview see Winters & Pisoni, 2004). Human speakers can control their voices in order to produce fine-grained variations that can be used to signal different pragmatic and semantic functions as well as emotion. However, a synthesis is sometimes more practical since it can be manipulated automatically. Pre-recorded or "canned" speech requires new recordings for every new lexical entry and variability in intonation, whereas the vocabulary of synthesis can be extended and varied on-line. To take advantage of both methods, some systems mix human and synthetic speech. That is, the human recorded voice is used for fixed prompts while the synthesis is used for the dynamic content. However, it has been shown that task performance was higher in a speech interface for email and calendar access with synthesis only than when a mixed approach was used (Gong & Lai, 2001).

If a synthesis is used, especially if we aim for humanlikeness, an important issue is how to produce appropriate prosody. There is a choice between using *Formant synthesis* and *Diphone synthesis*. Formant synthesis, or rule-based synthesis, produces speech through rules on acoustic correlates of various speech sounds. No recordings of human speech are used at run-time. Formant synthesis often sounds "robot-like" but has a large range of parameters that can be manipulated, which allows for a high degree of control. These parameters are related to both the voice source and the vocal tract. Diphone synthesis, on the other hand, makes use of pre-recorded human speech and concatenated *diphones*, that is, adjacent pairs of phones typically cut in the middle to capture the transition between two phones. Some signal processing technique is used to manipulate prosodic features on-line. In general, the only parameters than can be manipulated in diphone synthesis are fundamental frequency, duration and intensity. Diphone synthesis is therefore less flexible than formant synthesis, but often considered to sound more humanlike.

Regardless whether we choose to use pre-recorded speech or synthesis, we need to know how to manipulate/record the appropriate prosody of these voices in order to build systems that can sound con-

versational. This is a difficult undertaking since there is no one-to-one mapping between specific prosodic/acoustic realization and a specific semantic or pragmatic function. Moreover, there is a large variety in prosodic realisations between speakers and even within a single speaker in different contexts.

We will now discuss how some of the features of human conversation that were discussed in the previous chapter are approached in spoken dialogue systems.

4.6. Humanlike spoken language generation

Unlike humans, most dialogue systems process speech in a strictly segmental order, utterance by utterance. The next processing step does not start before the previous one has generated a complete output. As a result, the output of dialogue systems is generated in a fashion that is more similar to written chats or text messaging than spoken dialogue. The system is fed with chunks of complete speaker turns and presents utterances in a similar fashion.

4.6.1. Incremental processing in dialogue systems

Kilger & Finkler (1995) argue that incremental processing is an efficient processing strategy in spoken dialogue systems. Compared to the typical modularized architecture described above, an incremental system processes speech in smaller segments than utterance by utterance. The system starts to process the user's speech before input is complete. Smaller segments of speech are fed through the system as the user speaks, and processed by different modules in parallel. Output is produced in a similar fashion.

Schlangen & Skantze (2009) present a model for incremental processing in dialogue systems based on the principles cited above. The model does not define the systems modules or the unit of information that is communicated. The different processing levels as well as how to segment the continuous speech string are decisions to be determined by the individual system designer. Instead, the model

provides an abstract structure of how *Incremental Units* (IUs) can be processed in a parallel fashion. In a strictly modularized system, each module waits until a previous module has produced a complete output before it starts processing. In the incremental model presented by Schlangen & Skantze, the modules pass on partial results and these results are processed by several modules in parallel. This process can be described as follows. Each module has a processor with a left buffer (LB) and a right buffer (RB) (see Figure 8). The LB buffer is fed with an information stream which is processed internally until enough information has been collected for the module to make an initial commitment and move the partial result into the RB. The RB of the module is also the LB buffer of another module. Thus, the model assumes that this is the very same IU and both modules can process this information in parallel. It is also assumed that the RB can be connected to several LB. The model was implemented in the domain of number dictation. A user study was conducted to evaluate the proposed incremental architecture. The results showed that naïve users preferred the incremental system to a non-incremental version, and perceived it as more humanlike (Skantze & Schlangen, 2009).

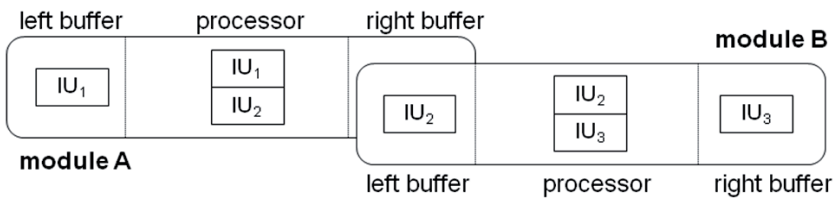


Figure 8. Two connected modules – from Skantze & Schlangen (2009)

Other approaches to incremental processing have mainly been concerned with syntactic processing (c.f. Kempen & Hoenkamp, 1987; De Smedt, 1990; Kilger & Finkler, 1995).

4.6.2. Repairs and hesitations in dialogue systems

Similar to human speakers, dialogue systems have to deal with the time constraints of producing utterances on-line in a conversation. In order to avoid long ambiguous delays, the system can benefit from processing spoken language incrementally. However, a dialogue system that starts to generate output before it has completed analyzing input has to deal with a higher level of uncertainty than dialogue systems that process utterances or syntactic phrases as whole. In order to produce utterances segment by segment, the system may need to rely on partial or incomplete interpretations of the users' utterances in order to predict the next dialogue move. While such predictions allow the system to rapidly initiate new responses, these predictions may also be premature. If so, the system needs to have strategies to repair these mistakes. Furthermore, a system that produces speech in an incremental fashion needs to deal with how to produce asynchronous output. If the system has already started to speak and underlying processes work asynchronously, the system risks running out of things to say. To avoid confusing silences where the user starts to speak before the system has provided a complete output, the system needs strategies to maintain the turn. In these situations, the system may employ hesitations or other types of interactional cues to signal turn-holding functions.

In a dialogue system context, repairs and hesitations have mainly been studied from a natural language understanding perspective, in order to identify and accommodate these phenomena in the users' speech. For example, Heeman & Allen (1999) present a statistical language model that incorporates part-of-speech (POS) tags in order to identify cue phrases, speech repairs and intonational phrases. The results show that these tasks are better addressed simultaneously than in isolation. It is further suggested that the benefit of identifying these phenomena early in the natural language understanding process is that these phenomena can be used by later processing steps to facilitate speech segmentation and understanding.

In another study, Goto et al. (1999) tried to identify filled pauses and word lengthening phenomena in spoken conversation. Similar to Heeman & Allen (1999), the aim is to identify these phenomena in order to explore their role in conversation and facilitate higher levels of processing. The method is based on acoustic analyses that try to identify small fundamental frequency variation and spectral envelope deformation. Experimental results show that the model can detect word lengthening and filled pauses in a system that process speech in real time with a recall rate of 84.9% and a precision rate of 91.5%.

To my knowledge, few dialogue systems generate repairs or hesitations. One system that produces fillers is presented in Pfeifer & Bickmore (2009). Initially, an empirical study of human-human interaction was conducted. The analyses of these data show that humans are likely to gaze away while producing fillers. Based on these findings, a user study with two versions of an embodied conversational agent, one with fillers and one without, was conducted. The results show that the subjects had mixed feelings towards computers that use conversational fillers. Some subjects indicated that they preferred the versions with fillers whereas some preferred the one without. In a study presented by Adell et al. (2007), however, it is shown that a synthesis that produces filled pauses is rated as more natural than a corresponding version without filled pauses.

Another study that explores listeners' perceptions of hesitation phenomena when produced by a synthesis is presented by Carlson et al. (2006). Two duration features – pause length and final lengthening – are manipulated to explore how listeners' perceive hesitations. The results indicate that the total duration increase is the valid cue rather than the contribution of either factor.

4.7. Cue phrases

The overview of cue phrases presented in the previous chapter suggests that it is difficult to accept one core definition, function or list of lexical entries that can be used to identify these markers. However,

some of the empirical findings that were reviewed suggest that cue phrases play an important role in conversation. These findings also illustrate important pragmatic functions that could be meaningful to signal in a spoken dialogue system. Below we list two important functions:

- Cue phrases provide listeners with information that signals how an upcoming speech segment relates to previous discourse. Cue phrases can also be used to signal conversational moves, for example, if a speech segment is a question or an assertion. Furthermore, cue phases such as "oh" can used to evoke the user's attention and highlight that the upcoming speech segment is important or contains new information.

- The research reviewed in the previous chapter suggests that cue phrases also play an important role at the interactional level of dialogue (c.f. Bestgen, 1998). Similar to the hesitations and other phenomena discussed earlier in this chapter, cue phrases can be used to maintain dialogue at the conversational level.

There are a few examples of studies that have tried to model cue phrases within the area of spoken dialogue systems. Heeman et al. (1998) present a machine learning approach to identify cue phrases in spoken dialogue systems. This model identifies cue phrases early in the processing stream by incorporating POS tagging during language modelling. It had been shown previously that there are correlations between specific cue phrases and specific conversational moves (Byron & Heeman, 1997). By employing this method, the system can identify cue phrases and use these to predict the speaker's conversational move.

There are also a few examples of systems that produce cue phrases (c.f. Graesser et al., 2001; Kim et al., 2000). AUTOTUTOR, presented by Graesser et al. (2001), is a conversational tutoring agent

that helps students learn about computer literacy. The system employs cue phrases to signal that an utterance is a question and uses markers such as "all right" and "let's go on" in order to signal a change in topic. The aim of these cue phrases is to indicate what actions are expected from the user and clarify the system's functions to the user.

4.7.1. Turn-taking in dialogue systems

One crucial aspect in dialogue systems is how to control the flow of dialogue contributions between the system and the user. Very few dialogue systems use sophisticated methods to manage turn-taking. These systems are generally poor both at detecting users' end of turns and at generating appropriate turn-taking behaviour to help users discriminate momentary pauses from ends of turns. A frequently used strategy is to interpret long silences as end of turns. While silence is an explicit, unambiguous indication that a speaker is momentarily not vocalizing, it is a crude detector of end of turns, as pause length *within* turns varies. For dialogue systems in English, the silence threshold for end of turn detection has been reported to range between 0.5 to 1 second (Ferrer et al., 2002). Yet, analyses of spontaneous dialogue in French show that silences within turns (*pauses*) may be longer than 1 second (c.f. Campione & Veronis, 2002). Moreover, Weilhammer & Rabold (2003) found that the mean duration for silences between turns (*gaps*) in spontaneous face-to-face conversation in American English was 380 milliseconds, which is shorter than 0.5 second. Consequently, if we use a silence threshold (0.5 to 1 second) to detect ends of turns, we end up with a system that has a longer mean response time than humans, but which still risks interrupting its users.

Apart from using silence for end of turn detection in spoken dialogue systems, one frequent strategy is to signal turn-taking artificially, as for example in push-to-talk systems, where the user takes and maintains the turn explicitly by pushing a button. However, while push-to-talk has shown to be an efficient strategy for improv-

ing task completion, the extra element of pushing a button appears to affect the way users interact with the system. For example, Fernández et al. (2007) found that, compared to free turn-taking, push-to-talk resulted in longer turns and less positive feedback. Allowing users to interact freely without artificial artefacts such as a button may not be a necessity to build successful spoken dialogue systems, but it is a crucial aspect if we want to build dialogue systems that interact with their users in a humanlike manner.

One approach to building dialogue systems with more flexible and robust turn-taking is presented by Raux (2008). In the proposed dialogue system architecture, dialogue management is distributed over two different modules, one high-level Dialogue Manager (DM) and one low-level Interaction Manager (IM). Whereas the DM operates on turn level, the IM operates on smaller speech segments controlling the system's reactive behaviour. Two different approaches as to how to improve the system's turn-taking behaviours are explored. One approach is to use an optimization method that dynamically sets the silence threshold used to detect the users' ends of turn. The model is based on a number of different dialogue features extracted automatically by the system. Compared to a fixed threshold, the dynamic turn-taking approach reduces the response latency by 22% while the detection rate is kept constant. The feature set with the best results is based on semantic features. The second approach is a flexible model to control the system's turn-taking behaviour. This model, the Finite-State Turn-Taking Machine (FSTTM), is based on turn-taking modelled as six different states. It is shown that this model can be used to reduce the system's response latency by up to 40% (compared to the base-line system).

4.8. Summary

This chapter has presented work on spoken language generation in dialogue systems. The level of complexity for different approaches varies from pre-recorded "canned" speech to more complex models

that incorporate intermediate processing levels of deep semantic and linguistic structures. Much work in this area is originally based on models for text or monologue generation. However, the task of conversational utterance generation has a number of additional issues. For example, the system needs to produce speech in real time in cooperation with another speaker and deal with the interactive aspects of this task. There is also a high level of uncertainty since speech recognition is error prone and future content of a dialogue is unknown. In order to deal with these issues, the system needs strategies to recover from potential misunderstandings and misrecognitions.

In order to address some of these issues, it is proposed that dialogue systems should employ behaviours from human conversation. An efficient processing strategy that is based on theories about human cognition is to process speech *incrementally*. This is an efficient strategy since processing of input is initiated before the user has stopped speaking. Another important characteristic of incremental processing is that processes on different levels work in parallel. However, the system needs strategies to deal with time constraints and premature interpretations of input. It is proposed that the system can employ *hesitations*, *repairs* and *turn-taking cues* as devices to deal with some of these issues. For example, by employing these behaviours, the system can signal refinements to an utterance or delays in the flow of speech. This will make dialogue systems appear as more humanlike and more intuitive to talk to.

Part II.
Human interaction as a model for spoken dialogue system behaviour

5. Spoken dialogue system behaviour: effects of variability

Part I served as a theoretical background of this thesis. Part II presents the data collections, data analyses and empirical studies that this thesis is concerned with.

Chapter 2 discussed the potentials of dialogue systems that behave as human conversational partners. Our aim is to build dialogue systems that behave similar to human speakers. Before we know how to control human conversational behaviours in a dialogue system, we want to explore how such a system is perceived. This chapter presents a listening test where subjects' attitudes towards a dialogue system with human behaviour were investigated.

5.1. Introduction

In a majority of today's commercial dialogue systems, speech is used in a way that is similar to a keyboard or mouse. In Chapter 2 it was argued that many speech interfaces are consistent with an *interface metaphor*, i.e. the users are inclined to perceive these systems as some type of machine interface. This assumption is motivated by the observation that many speech interfaces operate in the same domains and are based on interfaces that are manipulated with a keyboard or mouse. Furthermore, many dialogue systems are characterised by

design principles that restrict users' behaviours in order to address the technical issues of automatic speech recognition and the complexity of language understanding. Interactions with these systems and other types of machine interfaces may have influenced the public's opinions and expectations of speech interfaces in general. Systems consistent with an interface metaphor may have affected how we approach dialogue systems and encouraged us to adopt a machine-like style of interaction.

As there are no fully functional systems with human conversational capabilities, it is difficult to imagine how such systems will be perceived. A user study presented by Saini et al. (2005) suggests that it is difficult to let subjects judge systems that they have not experienced. In this study, two groups of subjects interacted with different version of a talking robot cat. After the test, each group was asked whether they would have preferred a system with the characteristics of the other, to them unseen version. Both groups responded that they preferred the version that they had interacted with. Other studies have shown that users have mixed feelings towards anthropomorphic interfaces. Pfeifer & Bickmore (2009) showed that about half of the subjects in a user study preferred a system that used conversational fillers whereas the other half preferred one that did not. Tomko & Rosenfeld (2004) showed that a majority of subjects preferred the Grafitti speech interface that incorporates a small subset vocabulary of standardized key words and phrases to the more conversational MOVILINE interface.

The aim of the study presented here was to further explore how dialogue systems with conversational behaviours are perceived. In order to consider the potentials of dialogue system with humanlike behaviour, a listening test that investigates listeners' perceptions of a conversational dialogue system was conducted. Since we are still far from being able to control such behaviour interactively, an *off-line human-human data manipulation* was used. This method makes use of human-human dialogue data to create an illusion of a dialogue system behaving very much like a human speaker, by replacing one

of the parties in a recording of human-human conversation with a synthetic voice. A corresponding version, manipulated to simulate a dialogue system with restricted human behaviour, was used as a baseline.

5.2. Dialogue data

The dialogue data used as a basis for this study was collected during the development of a dialogue system, namely the KTH Connector (Edlund & Hjalmarsson, 2005). The domain of KTH Connector is a "secretary" who helps its users by filtering streams of incoming telephone calls. The motivation is to assist people in a world where cell phones and laptops make us constantly available. These devices have many useful features, but the possibility of being contacted everywhere is sometimes a nuisance. When we are on a train, in a lecture or in another country we may not want to accept all kinds of telephone calls. Yet, we may not want to turn off the phone or computer to become completely isolated (and unreachable). The task of the KTH Connector is to establish intelligent communication links between people, connecting one or more persons in an appropriate manner at an appropriate time. The KTH Connector was part of the CHIL-project, an Integrated Project (IP 506909) under the European Commission's Sixth Framework Programme.

A data collection of human-human dialogues in the KTH Connector domain was conducted. The data collection was made with ten subjects, two posing as secretaries and eight as callers. The secretaries were provided with headset, a laptop with Skype VoIP software and a fictional employer's personal agenda. They were instructed to answer incoming Skype calls and to act as a personal secretary. The callers were instructed to make a phone call to book a meeting. Two different caller scenarios were used:

- Call a personal friend and try to book a private meeting
- Call a co-worker and try to book a business meeting.

The two scenarios were very similar apart from the personal- business dimension that was used to see how the type of task and personal relation affected their behaviour. The dialogues were in English and each dialogue was about 10 minutes long. All dialogues were transcribed orthographically including non-lexical behaviour such as repetitions, filler words and false starts. False starts were phrases that were interrupted and abandoned halfway. Table 1 presents examples of these behaviours.

Repetitions	I I am Mr Smith
Fillers	eh yeah I am not sure eh I can do it eh eh sometime after work
False start	that would that looks just fine on her schedule

Table 1. Example of repetitions, fillers and false starts from the KTH Connector dialogues

5.3. Stimuli preparation

Two dialogues, one from each scenario, were chosen as stimuli for the experiment. In each of them, the secretary's voice was replaced with a synthetic voice (the callers' voices were kept intact). A CONSTRAINED and an UNCONSTRAINED version were created from each dialogue. The UNCONSTRAINED version was a replica of what the human speaker's verbal behaviour. The CONSTRAINED version was the based on the same set of transcriptions as the unconstrained version but transformed in order to represent a restricted version of human behaviour. The transformations were done according to a set of *constraint rules*. The dialogues were created to differ in how information was presented while preserving its literal meaning. The constraint rules are based on the dialogue system design principles ar-

gued to restrict human behaviour in section 2.2.2. The following rules were applied to create the CONSTRAINED version:

- All fillers ("eh", "mm" and "ehm"), lexical repetitions and false starts were removed.
- All information ("small talk") which was not directly semantically relevant for the task was removed.
- Lexical variation was reduced.
- Short grounding actions such as "ok" were made more explicit (for example: "ok" was transformed to "ok a Chinese restaurant" to confirm the user's previous utterance).
- All pronouns (unless referring to an entity in the same utterance) were replaced with the nouns they were referring to.
- All incoming system (secretary) utterances that overlapped the callers speech were removed

Table 2 exemplifies the effects of applying the constraint rules i.e. creating the UNCONSTRAINED and the CONSTRAINED version. Jönsson & Dahlbäck (2000) refer to this type of dialogue transformation as *dialogue distillation*. Transcriptions of all stimuli dialogues presented in Appendix A.

unconstrained	mm she she has a dinner on Friday mm but she is available on Saturday and Sunday and on Thursday as well
constrained	Anna is available for dinner on Thursday Saturday and Sunday

Table 2. Example of constrained and unconstrained version

5.3.1. Synthetic speech production

All behaviours explored in this thesis are verbal behaviours, which all have an acoustic realization. Some prosodic analysis is provided, but our main focus is not on the details of the acoustic realization. The interaction explored in the present study has an acoustic nature and it is essential to present stimuli with spoken rather than written language. In the design of such experiments and in the design of spoken dialogue systems with increased humanlikeness, there is a choice between using a recorded human voice or a synthetic voice in dialogue system. While empirical evidence has shown that human speech is more expressive than synthetic voices (c.f. Winters & Pisoni, 2004), the behaviours explored in this thesis is produced with diphone synthesis. This choice is motivated by the possibilities of manipulating diphone synthesis on-line. In the present study, the secretary's voice was replaced with a synthetic voice primarily for three reasons:

- One aim of this study is to explore whether a synthesis can be used to produce the behaviours of interest. Since previous studies have shown that users have mixed feelings towards dialogue systems with human behaviour, the present study further investigates how these behaviours are perceived when produced in the context of a dialogue system with a synthesis. Furthermore, replacing the human voice with a synthesis creates an illusion of a human interacting with a machine.

- Many dialogue systems use a synthesis rather than a prerecorded human voice since synthetic voices can easily be updated and manipulated on-line (Reiter & Dale, 1997).

- The manipulations done to the original recordings in order to create the CONSTRAINED version would have resulted in audible discontinuities that might have affected the CONSTRAINED version negatively.

Both versions, UNCONSTRAINED and CONSTRAINED, were simulated using a diphone synthesis (Carlson & Granström, 2007). Apart from minor changes to transcriptions where the text-to-phone system had failed, no adjustments were done to the synthesis acoustically. The prosody of the original recordings was not reproduced since this would have possibly disadvantaged the CONSTRAINED version, which had no corresponding prosodic realization. Yet, this is expected to mainly disadvantage the UNCONSTRAINED version since this version contains more features such as filled pauses and short feedback utterances ("mhm") which are expected to rely more on prosodic realization (Gravano, 2009). The choice of synthesis was not principally important since both utterance generation strategies used the same voice and the focus of the study was on system behaviour and not on the utterances' acoustic characteristics.

5.3.2. Listening test

The experiment was made with 23 subjects (15 male and 8 female) between 23 and 65 years of age. None of the subjects had any professional experience of speech technology. The test was set up as a web-based form (see Appendix B). The subjects were led to believe that they were listening to recordings of users interacting with a fully functional dialogue system. The test contained the two stimuli dialogues divided into smaller units of about 2-3 utterances long. Each unit was presented with two sound clips: one UNCONSTRAINED and one CONSTRAINED version. The order of the clips was randomized and subjects were not aware of how the versions differed.

The task was to evaluate the system's behaviour by comparing the two different versions of system behaviour according to five dimensions. Four dimensions were chosen because they describe characteristics that are closely associated with human behaviour:

- Humanlikeness (the system behaves like a human would do in a similar situation)

- Politeness (the system acts polite towards the caller)
- Intelligence (in the context of this dialogue the system behaves intelligently)
- Display of understanding (the system behaves like it understands the caller well)

Efficiency was included as a fifth dimension since it is an often used metric in dialogue system evaluation (Walker et. al, 2000):

- Efficiency (the system tries to help the user in an efficient way)

The subjects were also requested to state which version they would prefer if they were to interact with a similar type of system. The test was forced choice, that is, the subjects chose which version was most prominent according to a particular dimension. If they considered both versions to be equally prominent they had the possibility to choose neither version (no difference).

After the test had been completed, the subjects were asked to rate how important they thought that the dimensions were if they were to interact with a similar system. This was done on a scale between 1, "not important at all", and 5, "very important".

5.3.3. Result analysis

In a first overall analysis, all judgements were analyzed together. This resulted in 1656 comparisons. The UNCONSTRAINED and CONSTRAINED versions were chosen in 41% and 28% respectively of these comparisons. In 31% of the comparisons, the subjects indicated that there was no difference between the two versions. Figure 9 presents how the judgments were distributed over the different dimensions. A McNemar test showed that differences in ratings between system versions were significant for three dimensions: *hu-*

manlikeness, *politeness* and *intelligence* (p<.05). The version based on human behaviour was rated as more *humanlike*, *polite* and *intelligent*. For the other two dimensions, *efficiency* and *display of understanding*, there were no significant differences in ratings between versions. Neither was there any preference for a particular version.

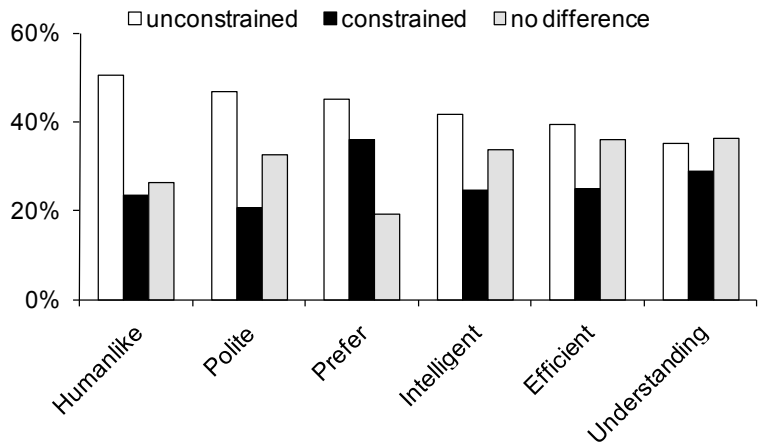


Figure 9. % judgments distributed over dimensions

In order to explore the relationship between the different dimensions, kappa coefficients were calculated pair-wise over the different categories. Thus, for each sound clip and subject, was there an agreement as to which version was selected for two different categories? For example: was the most efficient version also rated as most intelligent? Overall, the kappa values varied between fair (0.21-0.40) to moderate (0.41-0.60) agreement (see Table 3). The low Kappa values suggest that the subjects made individual judgments for each category. The highest kappa values (0.53) were between intelligence and display of understanding and efficiency and the preferred version. That is, the version rated as most intelligent was frequently also

rated as the version that best displayed understanding and the preferred version was frequently also rated as most efficient. Thus, the agreement between these dimensions is still only moderate.

	Human-like	Prefer	Polite	Intelligent	Display of understanding	Efficiency
Human-like						
Prefer	0.44					
Polite	0.26	0.39				
Intelligent	0.37	0.51	0.35			
Display of understanding	0.26	0.44	0.35	0.53		
Efficiency	0.24	0.53	0.22	0.46	0.35	

Table 3. Kappa coefficients calculated pair-wise over the 6 dimensions

5.3.4. Rating of efficiency

Efficiency in dialogue systems is traditionally measured using quantitative measures such as the number of words, syllables or utterances it takes to complete a task. To investigate the relationship between such objective metrics and the subjective judgment of efficiency, the number of syllables for each stimulus was calculated. The mean difference between the version judged as most efficient (regardless if CONSTRAINED or UNCONSTRAINED) and its corresponding version (M=0.90, SD=13.48, N=276), however, suggest that the version judged as most efficient contained *more* syllables. A paired t-test was

performed to investigate this further, but the difference was not significant and there is no support for such a relationship.

5.3.5. Ratings of the dimensions

After the listening test, the subjects were asked to rate how important the different dimensions were on a scale between 1 and 5 (Figure 10). *Intelligence* and *mutual understanding* was rated as most important whereas the two dimensions with the lowest ratings were *politeness* and *humanlikeness*. All dimensions had an average higher than 3, which can be interpreted as that none of the dimensions were considered as unimportant.

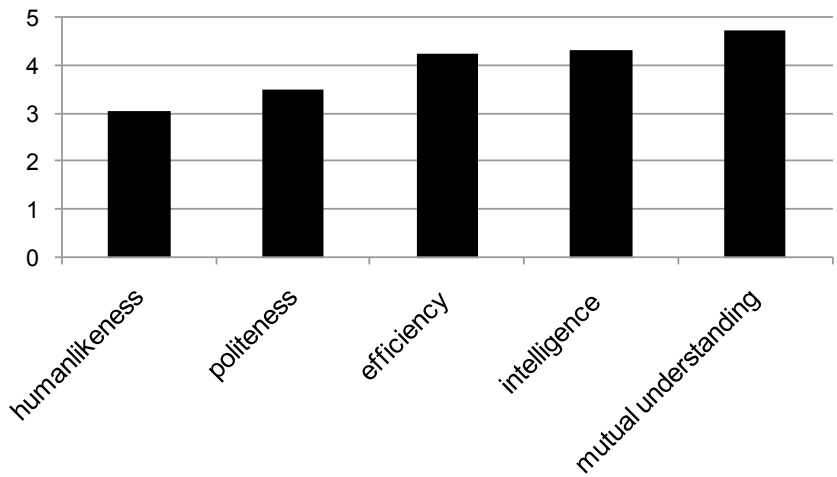


Figure 10. Average ratings of the dimensions' importance (1 – not important at all and 5 – very important)

5.3.6. Questionnaire comments

During and after the test the subjects were encouraged to submit comments about the reason for their decision and their subjective opinions about the different versions. The comments from different subjects were sometimes contradictory. This is in line with previous findings which have shown that different users have different preferences. Comments in favour of the UNCONSTRAINED version were enthusiastic about the use of pronouns and fragmental utterances. For example, "*It is nicer to talk to someone who doesn't always repeat what you are saying*" and "*version x* (the CONSTRAINED version) *repeats too much what has been said"*. In contrast, others indicated that they preferred the CONSTRAINED version's since it repeated what the caller had said. For example "*It's reassuring that the secretary repeats the names since they are not a common part of the English language and you feel more secure with the system*" and "*I like that nr 1* (the CONSTRAINED version) *checked that the whole number was right, but nr 2* (UNCONSTRAINED version) *acted more like a human*". Other comments pay tribute to the UNCONSTRAINED version "intelligence" and "naturalness" whereas others claim that it is "unnatural" when a machine is "*forced to make human mistakes*".

5.4. Discussion

Most dialogue systems available to the public are systems designed with an interface metaphor and it is reasonable to believe that humans engage in dialogues with machines based on experiences from these systems. The aim of the present study was to explore overall attitudes towards dialogue systems with a larger range of human behaviours compared to a more restricted version. Although there was no overall preference for the UNCONSTRAINED version, this study provides some encouraging results.

First, the acoustic realization is a potential bottleneck when building dialogue systems with humanlike behaviour. Humans have long experience in using their voices and it has a large range of different

acoustic characteristics that can be varied in order to express different semantic and pragmatic functions as well as emotions. It is difficult to postulate how the UNCONSTRAINED version would be perceived compared to a version with a human voice or with a synthesis with more conversational prosody. Still, the results presented here suggest that the UNCONSTRAINED version is perceived as more humanlike despite of its limited capabilities to express acoustic variation. These results diverge from findings reported by Nass et al. (2006) who claim that a synthesis is not human enough to use pronouns such as "I" or "me". The results presented here do not indicate that the CONSTRAINED version was preferred to the UNCONSTRAINED version despite its synthetic realization of a large range of behaviours that are rarely produced with synthetic voices including syntactic errors, slips of tongues and hesitations.

Another interesting result is the lack of correlation between length in syllables and perceived efficiency. Hence, long utterances are not necessarily associated with inefficiency. The preferred version, however, is relatively often rated as the most efficient version. What type of behaviour that is associated with perceived efficiency is an interesting area for future research. Some of the comments submitted in the post-experiment questionnaire indicate that several subjects preferred the UNCONSTRAINED version to the CONSTRAINED version because its "intelligent" confirmation strategies. The UNCONSTRAINED version was also described as more humanlike.

The results presented above are restricted in the sense that subjects listened to pre-recorded dialogues and were not able to interact with the two system versions. Thus, we were not able to explore how users adjust to different types of behaviour. Moreover, since each stimuli contained several turns, it was not possible to distinguish what type of behaviour that contributed to perceived humanlikeness. In the rest of this thesis, different types of human behaviour including *turn-taking cues*, *cue-phrases* and *disfluencies* will be explored in more detail. Next we will present DEAL, a dialogue system in a domain where conversational skills are strived after.

6. DEAL – a conversational spoken dialogue system

This chapter presents DEAL, a spoken dialogue system for second language learners of Swedish developed at KTH. DEAL is intended as a multidisciplinary research platform where challenges and potential benefits of combining elements from computer games, dialogue systems and language learning can be explored.

6.1. Introduction

There are many potential dialogue system domains where a human metaphor is beneficial. Examples of such domains include applications for entertainment and intelligent tutoring systems (ITS). In this thesis, the application used as a research platform tries to combine these two.

There is a growing trend among educational researchers to look at games and game design in order to make education more appealing and effective. A new and challenging domain for spoken dialogue systems is *serious games*, i.e. applications of interactive technology that have purposes other than solely to entertain, including training, advertising, simulation, or education (Iuppa & Borst, 2007). If successful, serious games will engage users who want to be entertained and/or educated. Encouraged by such motivations, users will be willing to talk to dialogue systems because it is fun, repeatedly and for

long periods of time. These applications may also encourage users to view dialogue systems in the light of new types of metaphors, and approach these systems in new ways. This is a tempting scenario.

From a dialogue research point of view, a serious game approach contributes with several novel and interesting objectives and challenges. These include how to design dialogue systems that are entertaining and intuitive to talk to. Furthermore, a dialogue system for language learning should use a language that suits the vocabulary and language complexity of language learning students on various levels. Since efficiency and information transfer in the traditional sense are no longer the main objectives, dialogue systems in a serious game context do not have to be predictable, rational or even co-operative. Instead, we need to focus on how to build systems that are fun, educational and addictive to talk to.

6.1.1. Acquiring conversational skills

Language learning can be modelled as a series of developmental steps going from declarative to procedural knowledge. First, an item is noticed in a meaningful contrastive situation, then it occurs repeatedly in meaningful input and is practised in communication until it is internalised, and finally automatized (Ellis, 2006). To automatize these processes when learning a second language, we need a meaningful situation where conversational skills can be practised repeatedly. Because of its complexity, learning a language requires substantial effort and the motivation varies both over time and between individuals. To practise conversational skills while playing a game may increase any existing motivation, and may even create a motive if there is none. For a more detailed discussion of speech applications and language learning see Wik & Hjalmarsson (2009)

6.2. Motivation

Game designers focus on finding ways to keep players engaged and motivated throughout a game. Nonetheless, dialogues in today's games have a restricted way of affecting the continuance of the game. The interaction is typically based on complex tree structures, where one action leads to a set of new choices. Choosing one line or topic has an immediate result, and the dialogue traverses a finite branching tree structure. With these types of dialogues, it is fairly trivial how to get the desired result, making it less interesting to engage in the interaction. Façade is an interactive drama project that introduces a drama manager to make the outcome of a dialogue less predictable (Mateas & Stern, 2003). In Façade, the story is divided into *beats*, an atomic unit of drama, where beats and transitions between beats can unfold in various ways depending on what type of input is provided by the user.

The NICE project is another example of a game dialogue system where dialogue is not just an add-on, but is used as the primary means for game progression (Gustafson et al., 2004). NICE is a fairytale system which allows children and adults to interact with animated characters in a 3D world. In order to move forward in the game and gain access to the goals and desires of the fairytale characters, the users have to interact with these personas through verbal and non-verbal communication.

Similar to NICE and Façade, the aim of DEAL is to use dialogue as the primary means for interaction and game progression. The practical motivation of DEAL is to build an application where conversational skills can be practised in a fun and meaningful context. The target audience is second language learners of Swedish who want to put recently learned vocabulary into practise. A similar approach is used in the tactical language training system (TLTS), a large-scale application that helps people acquire basic conversational skills in Levantine and Iraqi Arabic (Johnson et al., 2005). The choice of the initial domain for DEAL is the trade domain, but the system can be

extended to cover more domains in the future. In the trading domain scenario, DEAL sets the scene of a flea market where a talking animated agent is the owner of a shop where used objects are sold. The domain was chosen for several reasons:

- A trading situation is a restricted and universally well-known domain. It is something everyone is conceptually familiar with, regardless of cultural and linguistic background.

- A trading situation is from a language-learning point of view a very useful domain to master in the new language.

- The objects sold at a flea market can be a diverse set of items that can be tailored to suit the vocabulary mastered by a language-learning student.

- A flea market is a place where it is acceptable to negotiate about the price. Negotiation is a complex process that includes both rational and emotional non-rational elements. This opens up for interesting and complex dialogue.

These characteristics combined provide for a dialogue system situated in a well-known context but which also includes elements of surprise and challenge (i.e. getting a good price).

As discussed in chapter 3, human speakers tend to coordinate their linguistic behaviour. Research on linguistic entrainment in human-machine interaction has shown that users adopt the system's way of speaking (c.f. Brennan, 1996; Zoltan-Ford, 1991; Gustafson et al., 1997). Furthermore, research on second language acquisition (SLA) is diverse, with no single theory or model seen as the most appropriate. However, there seem to be a consensus about the value of conversational interactions. The more you talk the better it is. Thus, from a second language learning perspective, the language used in DEAL will be crucial. It is important that the system moti-

vates the users to talk a lot, and not only in short command-like utterances. This does not necessarily mean that the agent needs to be cooperative or polite. The seller can actually be rude and try to avoid the users' requests as long as this is done in a way that does not destroy the users' willingness to accept the shopkeeper in DEAL as a character with humanlike conversational capabilities.

6.3. Ville

DEAL is developed as a freestanding part of Ville, a framework for language learning developed at KTH (see Wik and Hjalmarsson, 2009). Ville is a virtual language tutor helping students to improve their listening and pronunciation skills in Swedish (see Figure 11).

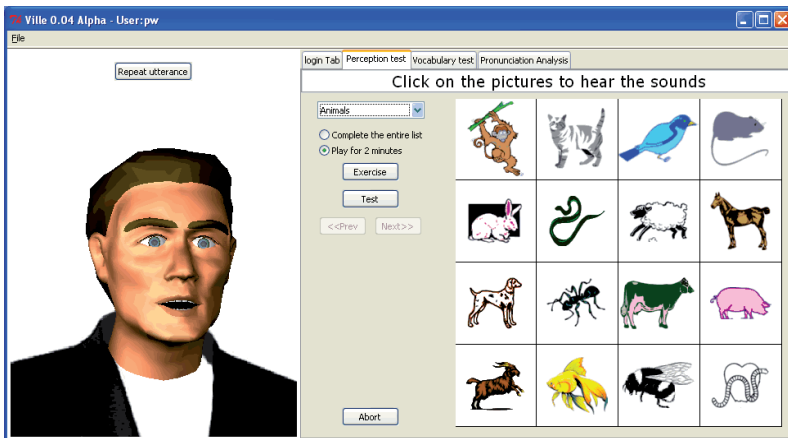


Figure 11. Ville user interface

Ville detects and gives feedback on pronunciation errors, and has challenging exercises that are used in order to teach new vocabulary, or to raise the students' awareness of particular perceptual differences between their first and second language. Ville has exercises on phone, syllable, word, and sentence level. DEAL adds the possibility to give conversation training. DEAL serves as an important complement to Ville; whereas Ville provides exercises on isolated speech segments,

i.e. phone, syllable, word, and sentence level, DEAL adds the possibility of practicing these segments in the context of a conversation. While Ville has the role of a teacher who gives you feedback and help when you encounter problems, DEAL has the role of a native speaker, for example, a person with a service occupation, whom you need to communicate with using your new language. Using DEAL as an integrated part of Ville, the system has knowledge about particular students' acquired vocabulary. This information can be used to tailor the language in DEAL as well as the goods being sold.

6.4. Dealing with DEAL

DEAL has two actors, a non-player-character (NPC) – the shopkeeper – and one human language student. The student is given a mission to buy items at a flea market getting the best possible price from the odd-looking shopkeeper. The shopkeeper can talk about objects and their properties and negotiate about the price of the objects. The most challenging part in DEAL, both from a "customer" (user) point of view and when designing the conversational agent, is negotiating about the price of objects. At first, dealing about price can seem like a fairly rational and straightforward procedure. However, negotiating is a complex multidisciplinary area of research that touches fields such as psychology, economics and political science. Negotiating about a price in a face-to-face situation involves a number of various parameters that are often affected by non-rational and emotional aspects. Second hand items may have rich interesting characteristics that make them interesting to talk about. For example the items can be defective, have a personal history or an affection value to the shopkeeper, all of which may have an impact on the negotiation process. The dialogue can unfold in different ways depending on what the user says (see Example 3).

Negotiation is implemented using a straightforward algorithm and a few heuristics. To introduce elements of game, a parameter that represents the agent's "willingness" has been integrated to reduce

the price of an item. The willingness parameter is the percentage share of the seller's original price that the ECA is willing to accept, after negotiating. The parameter has an initial value that may be affected depending on how the dialogue proceeds. To affect the outcome of the interaction, the player may try to influence the willingness of the shopkeeper to reduce the price.

U1. I'm interested in buying a toy.
S1. Oh, let me see. Here is a doll. (a doll is displayed)
U2. Do you have a teddy-bear?
S2. Oh, yeah. Here is a teddy bear. (a teddy bear is displayed, see Figure 12)
U3. How much is it?
S3. You can have it for 180 SEK
U4. I give you 1 SEK (willingness decrease)
S4. No way! That is less than what I paid for it.
U5. Ok how about 100?
S5. Can't you see how nice it is?
U6. But one ear is missing. (willingness increase)
S6. Ok, how about 150?
U7. 130?
S7. Ok, it is a deal!

Example 3. Dialogue example from DEAL

The outcome of the game is affected by what the user says. For example in utterance U4 the seller is offended by the user's low bid and his willingness to give the user a good price is reduced. However, when the user points out a flaw of the object the seller feels obligated to give the user a better price, i.e., his willingness increases.

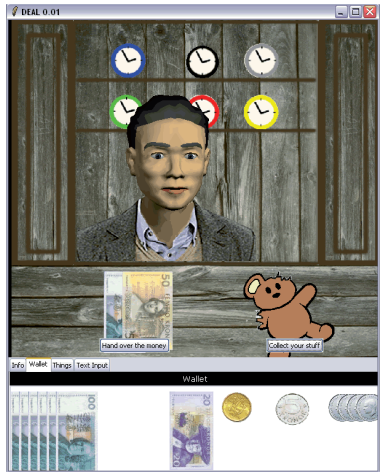


Figure 12. The DEAL GUI – a teddy bear with a missing ear

6.5. Implementation

The section describes the first implementation of DEAL. This implementation is based on a traditional modularized pipeline architecture that does not support incremental processing. A later incremental version of DEAL is presented in Chapter 9.

The first version of DEAL is implemented using components from the Higgins project (Skantze, 2005). This version also includes an off-the-shelf ASR engine, an action manager developed for DEAL purposes and a GUI with an ECA.

6.5.1. DEAL's Graphical User Interface

The graphical user interface (GUI) in DEAL is divided into six parts (see Figure 13). The top part contains the shopkeeper, an ECA, which is developed at KTH (Beskow, 2003).

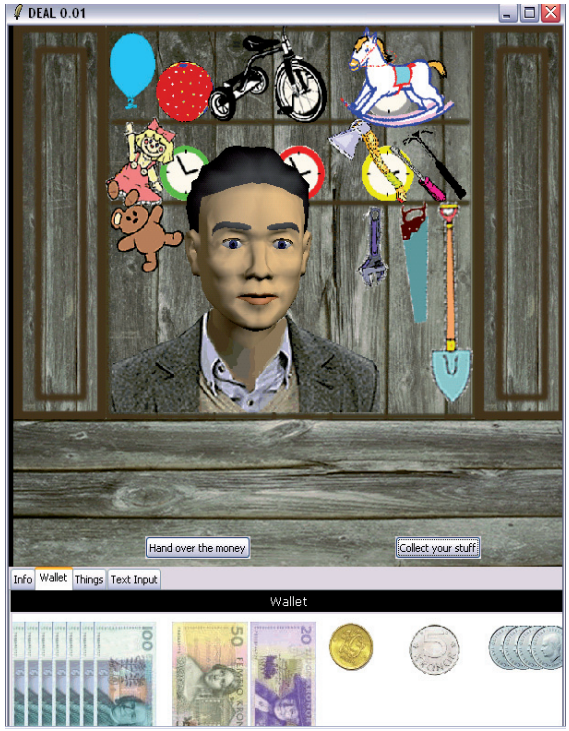


Figure 13. DEAL user interface

The ECA speaks with a synthetic voice, a diphone synthesis. The head produces lip-synchronized speech and is animated according to a set of randomized sequences of facial gestures and head movements. These movements are used to make the shopkeeper appear more humanlike. For example, the shopkeeper can tilt his head and raise or lower his eyebrows. Language is multimodal, and in second language learning, visual signals are an important source of information. Behind the ECA, some of the goods for sale are on display. These were included to give the users an initial idea of what kind of objects to talk about. The section just below the ECA is the shop-counter. Goods up for discussion are displayed here. Below the counter are four tabs. The info tab provides hints, for example if the

student has trouble remembering the vocabulary. The wallet tab displays the money the student has at his or her disposal. The things tab shows pictures of the object that the student has already bought. Finally, the text input tab provides the user with the option text filed where the user can submit typed input rather than speech as means of communication.

6.5.2. Architecture

Higgins includes modules for semantic interpretation and analysis. *Pickering* is a modified chart parser that supports continuous and incremental input from a probabilistic speech recognizer. Speech is unpredictable; chunking a string of words into utterances is difficult since pauses and hesitations are likely to be incorrectly interpreted as end of utterance markers. This will be even more evident for second language learners whose conversational skills are not yet automatized and whose language contains phenomena such as hesitations and false starts. Pickering uses context free grammars (CFG) and builds deep semantic tree structures. Grammar rules are automatically relaxed to handle unexpected, ungrammatical and misrecognized input robustly. The discourse modeller, *Galatea*, interprets utterances in context and keeps a list of the communicative acts (CA) in chronological order. Galatea resolves ellipses, anaphora and has a representation of grounding status which includes information about who added a concept, in which turn a concept was introduced and the concept's ASR confidence score.

6.5.3. Action management in DEAL

Action management in DEAL, that is deciding how to respond to the user's input, is done according to a set of simple rules. These are based on the *script*, or episodic knowledge structures that guide us when we interact with a shopkeeper in a shop in order to buy a product. Communicative acts used in DEAL include OBJECT-REQUEST, PROPERTY-REQUEST, PRICE-REQUEST, SUGGEST-PRICE, DEAL, and so on. The user can request objects, ask about object

properties, give price offers, and make deals. The haggling algorithm is a set of simple heuristics. These are based on the relation between the user's offer and the system's "retail price" which is stored in the system database. For example, if the user's offer is too low, the shopkeeper refuses to lower the price. However, if this offer approaches to the retail price, the shopkeeper willingly lowers the price. Some of the objects in the database have obvious visual defects (e.g. the missing ear in Figure 12) and if detected and pointed out by the student, the agent reduces the price.

The goal in DEAL is to build a reactive, mixed initiative dialogue system where the agent takes the initiative if the student fails to do so. As a first step in this development, we have implemented a few system initiatives and emotional reactions that illustrate the shopkeeper's attitudes towards what goes on in the dialogue. The shopkeeper's initiatives include suggestions of objects and prices if no user input is provided, trying to bring the dialogue to a close. What action is taken is based on the dialogue state; for example, if an object is in focus (on the table), the agent suggests a price for that object, and if no such object exists, a new object is presented. The shopkeeper also displays emotion, i.e. looks angry or happy, depending on how the dialogue progresses. Greetings from the user and closings of deals are responded to with a smile. However, after long sequences of haggling or price offers from the student that are too low (less than 10% of the agent's initial price suggestion), the agent looks angry. An important characteristic of the system is that the goals of the agent and the student partly differ. Both have the goal to complete a successful interaction; however, the agent wants to sell goods for as much as possible while the student wants to buy them for the lowest possible price. In terms of game-play, buying an object for a certain price must be challenging. To make the bargaining trickier, the agent is easily "fed up". After a fixed set of speaker turns haggling about the price of a certain object the agent claims to be bored and refuses to discuss that object anymore. Instead, he suggests a new one.

6.5.4. Generation in DEAL

The system's behaviour is crucial in DEAL. To encourage the user to talk to the system as if talking to another human being, the agent needs to be responsive and flexible. When aiming diverse and engaging dialogue, long response times and repetitive language using templates or pre-recorded speech are not acceptable.

The generation task in DEAL is distributed over four different modules; the communicative manager (CM), the action manager (AM), Ovidius (see Figure 14). In this un-incremental version of DEAL, the system processes output and input turn by turn. The discourse modeller Galatea forwards an abstract semantic interpretation of the user's input to the CM. The CM is responsible for managing the flow of dialogue contributions, making a shallow analysis of input before passing them on to the AM. The CM also modifies the new response generated by the AM before presenting it to the user in order to accommodate the current dialogue context. The first task of the CM is to act as a shallow error-handling filter, determining if the system needs to clarify some part of the incoming message. If the ASR confidence score for a particular entity is too low, below a certain threshold, the CM generates a clarification request without passing the message on to the AM. Hence, the user is requested to clarify the object with poor recognition before proceeding with any further analysis of input. If no clarifications are needed, the CM forwards the abstract interpretation of the user's message to the AM.

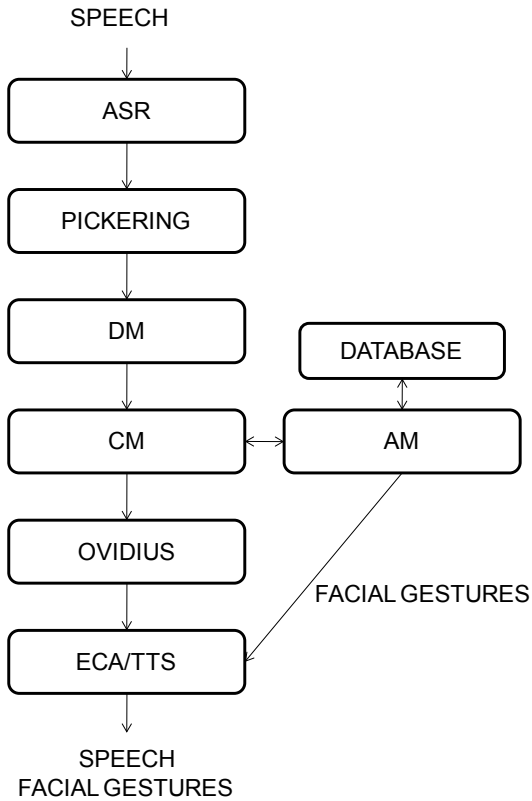


Figure 14. DEAL architecture

Immediately after this message has been passed on, the CM initiates a new utterance regardless if the AM has generated a response or not. In this way, the system can also utilize the time it takes to access objects in the external database to ground concepts in the user's incoming utterances (e.g. "ok a green watch"). This is done regardless of whether this object exists in the database or not (i.e. the object could already be sold or not exist in the database). If the object for some reason turns out to be unavailable, the system revises its previous grounding segment and suggests another object (see dialogue example 2, s1a and b).

U1. I want to buy a green watch.
S1a. Ok, a green watch…
S1b. … I'm sorry there is no green watch but I do have a red one.
U2. Do you have a yellow one?
S3a. mm a yellow watch…
S3b. … here is one

Example 4. DEAL Dialogue example – turn initial grounding and repair

If the user input contains no reference to a particular object, the CM generates neutral grounding fragment such as "mm" "ja" (Eng: "yes") or a filler word such as "eh" or "ehm".

While the CM initiates a new utterance, the AM is responsible for deciding which action to take based on the new input from the user, or, if no input is detected, initiate a new action based on the previous dialogue state. When the AM has generated a response, it is passed back to the CM which is responsible for modifying the response based on the dialogue context. For example, the CM decides how entities should be referred to, e.g. determines whether to use referring expressions or full noun phrases, as well as turning full propositions into elliptical constructions. The decisions are based on how well the entities are grounded in the dialogue, based on the confidence scores from the ASR and if, how and when these entities have been previously mentioned. Furthermore, the system's own previous speech segments are also used as a basis for these modifications. Hence, Galatea, the discourse modeller, not only keeps track of the user's conversational acts (CA) but also its own previous CAs. How a concept was referred to in the first part of an utterance is used to determine how the shopkeeper should refer to this object in the second part. As exemplified above (2, S3a and b), if the first part refers to the object with a full noun phrase (e.g. "a yellow watch"), the system

uses a pronoun in the second part (e.g. one).The message modified by the CM is forwarded to Ovidius, still in its abstract semantic form.

Ovidius is a part the Higgins project and the module responsible for realising the textual representation. Ovidius takes an abstract representation of a system CA as input and generates a text that is subsequently realised acoustically by a speech synthesiser. Ovidius uses a set of template rules, working much like inverted Pickering grammar rules – they match the semantic tree structures and produce text strings. The acoustic realisation in the current version of DEAL is a combined set of pre-synthesized prompts, and on-line text-to-speech generation. Feedback and other cue phrases as well as filled pauses are pre-synthesized prompts while the rest of the dialogue is synthesized speech generated online. The pre-synthesised elements have prosodic features, including F0 contour, speaker rate and energy, automatically extracted from the DEAL corpus which is described in the next chapter.

6.6. Summary

This chapter has presented DEAL, a spoken dialogue system that provides second language learners of Swedish conversation training in a fun and challenging context. DEAL is a free-standing part of Ville, a framework for language learning developed at KTH. Whereas Ville has exercises on phone, syllable, word, and sentence level, DEAL adds the possibility to give conversation training. The DEAL domain is a *serious game* approach, where system designers try to combine elements from computer games and intelligent tutoring systems (ITS) in order to make education more appealing and effective. This approach contributes with several novel and interesting challenges. Rather than building systems that are co-operative, prompt, and accurate, we need to focus on how to build systems that are fun, educational and addictive to talk to.

The DEAL domain is a trading domain. DEAL has a non-player-character (NLP), a shopkeeper who is in charge of a flea market. The

user's task is to engage in a dialogue with the shopkeeper in order to buy goods, getting the best possible price. The system architecture is based on modules developed within the Higgins project, two modules responsible for dialogue management developed for DEAL purposes, and an embodied animated character (ECA) developed at KTH. The ECA is capable of producing lip-synchronized speech and showing happiness or anger depending on how the dialogue progresses. Dialogue management is distributed over three different modules, Galatea (a discourse modeller), a conversational manager and an action manager. Galatea interprets utterances in context and keeps a list of the communicative acts in chronological order. The conversational manager is responsible for managing the flow of dialogue contributions. The CM also perform grounding functions, for example it makes clarification requests, generate feedback utterances and choose between full noun phrases and pronouns depending on how well grounded a concept is in the previous context. The action manager is responsible for deciding which action to take based on the user's input, or, if no input is detected, initiating a new action based on the previous dialogue state.

The next chapter describes a data collection of human-human dialogues in the DEAL domain.

7. DEAL data collection

This chapter presents a data collection of human-human dialogue. The aim of this effort is to collect instances of the behaviours that we aim to model in the DEAL dialogue system. The behaviours analyzed in this chapter are so-called *cue phrases*, expressions that are used to signal discourse structure and often occur in a turn-initial position (Heeman et al., 1998). These elements are relevant for turn-taking, and a dialogue system that masters these elements could be more intuitive to talk to. The data collection was performed in an early stage of system development to serve as a basis for DEAL.

This chapter is structured as follows. First, the details of the data collection – the domain, the general set up, the recordings and the task that was given to the participating speakers – are described. Some basic descriptive statistics of the dialogue data are also presented. This is followed by a section on the manual annotation of cue phrases and lexical and prosodic analyses of these elements.

7.1. Introduction

A natural first step in our striving for humanlike dialogue systems is to attain an appropriate corpus of dialogue data that contains the dialogue behaviours we aim to model. As discussed in Chapter 2, speakers tend to behave differently when they believe that they are talking to a machine. Their previous experiences of speech interfaces or other types of machine interfaces appear to affect their behaviour.

In order to obtain examples of conversational behaviours that are not coloured by previous experiences of speech interfaces, a corpus of human-human dialogues were collected. The motivations for collecting this corpus were: (1) To gather data that is representative of the conversations to be held with the DEAL system. These data can be used to build language models for different system components including ASR, NLU and NLG. (2) The DEAL domain has some specific characteristics that are of special interest:

- The dialogues are concerned with negotiation. This is a complex process that contains both rational and non-rational elements. Characteristic to this task is that the interlocutors have different goals. The customer's task is to get a price that is as low as possible, whereas the seller's task is to get a price that is as high as possible. Still, both parties have a mutual goal, i.e. to engage in a successful conversation.

- Negotiation is about making and responding to offers. Such conversational acts are likely to contain referring elements that signal how new dialogue contributions relate to previous offers or counter arguments. The dialogues are therefore expected to contain a lot of cue phrases, elements that are explored in more detail at the end of this chapter.

- Dialogues in the DEAL domain are also expected to be characterized by high planning demands. Both the sellers' and customer's task required arithmetic calculations. The sellers were required to calculate how much the price could be lowered based on a fixed retail price. The buyers, on the other hand, needed to calculate how much money they could spend on certain goods in order to have enough money left to buy all the things that the task required. These calculations were hypothesized to aggravate planning and cause speaker hesitations such as fillers and pauses.

7.2. The DEAL data collection

The DEAL data collection was made with six participants, four male and two female. The subjects were recruited at the Department of Speech, Music and Hearing. All were native speakers of Swedish, between the ages of 20 and 45. Two of the speakers were recruited as shopkeepers and four as customers. Since we had no access to professional shopkeepers, the speakers playing the role of the shopkeeper participated in four dialogues with two different buyers each in order to allow these speakers to get familiar with the situation. Eight dialogues were collected all together. Two of the four subject pairs knew each other.

The dialogues recorded were spontaneous, human-human, face-to-face conversation. In order to collect data in a similar setting as the final application, the task and the recording environment were set up to mimic the DEAL domain and role-play.

7.2.1. Experimental setup

The recordings were held in the speech lab at the Department of Speech, Music and Hearing at KTH. Before each dialogue session, the speakers were instructed separately in different rooms. The subjects recruited as customers were given a mission: To buy a set of goods at the best possible price from the shopkeeper. Before each dialogue, they were given a specific scenario. For example, to buy three tools to repair a house or to buy toys for their niece's birthday. The customers were given a certain amount of toy money, however not enough to buy what they were instructed to buy without bargaining. The subjects requited as sellers were instructed to sell things to the customers, getting as large profit as possible (English translations of the sellers' and customers' instructions are provided in Appendix C and Appendix D respectively). Before the first dialogue, the subjects filled out a questionnaire about their previous price negotiation experiences. A similar questionnaire was filled out after each dialogue to explore whether the speakers employed any particular

negotiation strategies during that specific dialogue session. The questions in the questionnaire are presented in Appendix E.

During the recordings, the speakers were alone in the recording room, placed sitting face-to-face in front of each other. The shopkeeper sat behind a table, a "counter", with images of objects pinned to the wall behind him (see Figure 15). Some of the goods had minor defects. These flaws were used to open up for interesting negotiation. Each customer interacted with the same shopkeeper twice. There was no time limit. The speakers were instructed to go on for as long as they liked or until they had completed the task.

Figure 15. The DEAL data collection

The recordings were made with a computer at 16 kHz using the Wavesurfer[4] software (Sjölander & Beskow, 2000) and two close-talk microphones. One microphone was attached to each speaker and the dialogues were recorded as a stereo recording with one speaker on each channel. The dialogues were also video recorded and the sellers' head movements were tracked using motion capture. The research presented in this thesis focus on verbal language behaviours and none of the visual data is presented here.

7.3. Dialogue data

Each dialogue lasted about 12.5 minutes, making for about 1 hour and 40 minutes of speech in total in the corpus. In the questionnaires, the participants reported to have used a wide variety of strategies and arguments to convince the other party of a reasonable price. Here are some of the strategies provided by the shopkeepers (the comments were extracted from the questionnaires and translated into English):

- I tried to play hard to get and not lower my price too easily
- I tried to make friends with the customer and argued that the goods were of good quality and that the broken goods could be easily repaired
- I told him that he is my friend and that I could give him a special price that gives me no profit
- I tried to figure out how much money she had and gave her as few things as possible for her money

Here are some of the strategies provided by the customers (translated into English from the questionnaires):

[4] www.speech.kth.se/wavesurfer

- Point out flaws of the objects and say that I didn't have enough money even though I did.
- I gave the shopkeeper compliments
- I lied about how much money I had.
- I tried to get a package deal
- I threatened him by saying that I wasn't going to buy anything.

Many of the dialogues were initiated and ended with small talk such as greetings and polite questions about the speakers' fictional businesses, friends and family.

7.4. Dialogue segmentation, transcription and alignment

The dialogues were first transcribed orthographically using an annotation tool that was developed by Gabriel Skantze at the department of Speech, Music and Hearing as a part of the Higgins project (http://www.speech.kth.se/higgins/). The dialogues were transcribed by two different transcribers from the department. Except for lexical transcripts, the dialogues were also manually annotated with verbal behaviours such as laughter, lip-smacks, hemming, audible inhalations and exhalations. In the rest of this thesis, *words* refers to all lexically transcribed tokens, include fillers such as "mm", "ehm" and "eh", even though these tokens that are not traditionally represented in a lexicon. Table 4 list the ten most frequent words in the DEAL corpus. The words are listed in order of frequency with the most frequent word first.

Order of frequency in corpus	Lexical Transcription (English translation)	quantity
1	det (it *neuter*)	720
2	ja (yes)	629
3	ju (of course/actually)	515
4	jag (I)	505
5	den (it non-neuter)	494
6	är (is)	480
7	du (you)	387
8	men (but)	310
9	för (for)	279
10	så (so)	276

Table 4. The 10 most frequent words in the DEAL corpus

The transcriptions were subsequently time-aligned with the speech signal. This was done using forced alignment with N-align, the CTT aligner tool (Sjölander, 2003). In order to obtain reliable timings for all tokens including phenomena such as lip-smacks and filler words that were not a part of N-align's lexicon, the timings from the forced alignments were manually verified.

The next step was to segment the time-aligned dialogues into operationalizable units.

7.4.1. Dialogue segmentation

Spoken conversation contains two or more speech signals that are continuous in nature. In order to perform more detailed analyses, these signals need to be segmented into smaller units. Frequently employed units in the literature include "utterances" (Clark, 1996), "speaker turns" (Sacks et al., 1974), and "dialogue acts" (c.f. Core & Allen, 1997). The units are based on different information, including syntactic, semantic/pragmatic or prosodic. The definitions, however, are sometimes vague and differ between researchers. Many require some kind of manual annotation. To avoid manual labour, the

DEAL corpus was segmented into operationalizable *talkspurts* as defined by Norwine & Murphy (1938 p. 282):

> *"A talkspurt is speech by one party, including his pauses, which is preceded and followed, with or without intervening pauses, by speech from the other party perceptible to the one producing the talkspurt. Obvious exceptions to this definition are the initial and final talkspurts in a conversation. There may be simultaneous talkspurts by the two talkers; if one party is speaking and at the same time hears speech from the other double talking is said to occur."*

As a first step, the dialogues were automatically segmented into interpausal units (IPUs), a sequence of words surrounded by silence longer than 200 milliseconds (ms). 200 ms was used as a segment criterion since Swedish has long plosives. If we include silences shorter than 200 ms, we risk extracting plosive stops where we aim to extract pauses between speech segments. These IPUs were subsequently segmented into talkspurts. An illustration of the segmentation of talkspurts is presented in Figure 16. Here, a talkspurt is a maximal sequence of IPUs so that between two adjacent IPUs there is no speech from another speaker.

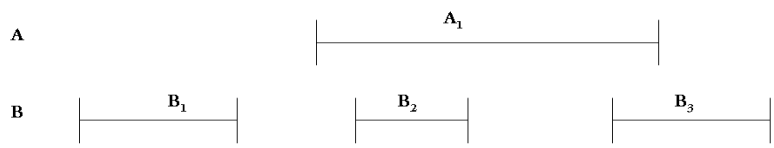


Figure 16. Illustration of the talkspurt segmentations

There were 2036 talkspurts in the dialogues. Table 5 presents some basic statistics for the eight dialogues. The sellers' talkspurts were in general longer than the customers' were.

Dialogue	Knew partner?	Sex	Length in seconds	Number of talkspurts		Average talk-spurt length in seconds		Words per talkspurt	
				total		total		total	
				both speakers		both speakers		both speakers	
				cus-tomer	seller	cus-tomer	seller	cus-tomer	seller
1	y	m/f	864	251		4.90		8.73	
				118	133	2.66	3.58	9.07	8.42
2	y	m/f	618	221		1.75		5.63	
				105	116	1.54	1.91	5.63	5.77
3	n	m/f	842	307		2.48		6.92	
				147	160	1.10	3.73	3.66	9.70
4	n	m/f	1032	364		2.68		8.17	
				169	195	1.67	3.54	5.80	10.23
5	n	m/m	797	226		2.86		9.75	
				110	116	2.55	3.10	8.56	10.87
6	n	m/m	798	276		2.49		8.57	
				137	139	2.28	2.71	7.55	9.57
7	y	m/m	492	174		2.32		7.99	
				81	93	2.11	2.45	6.94	8.90
8	y	m/m	603	217		2.58		8.77	
				102	115	1.79	3.16	6.57	10.73
Average			756	254.5		2.54		8.05	
				121.1	133.38	1.93	3.11	6.61	9.36

Table 5. The DEAL dialogue data – basic statistics

A talkspurt duration histogram is presented in Figure 17. This histogram illustrates that there are many talkspurts shorter than one second. A large part of these very short talkspurts is likely instances of what is typically referred to as "back channels" (Ward & Tsukahara,

2000) or "continuers" (Goodwin, 1986), that is, brief feedback uttered while attending to another interlocutor's speech. Instead of manually annotating these speech segments, Edlund et al. (2010) introduces an auxiliary unit*, very short utterance* (VSU), which can be detected automatically. Edlund et al. (2010) further shows that manually annotated feedback utterances in the Columbia Games Corpus can be identified with good precision based on duration alone. 71% of all VSUs (talkspurts shorter than 1 second) in this corpus were annotated as different kind of feedback utterances (*back-channels* or *affirmative cue words*). If we adopt this definition, 44% (901) of the talkspurts in the DEAL corpus are VSUs.

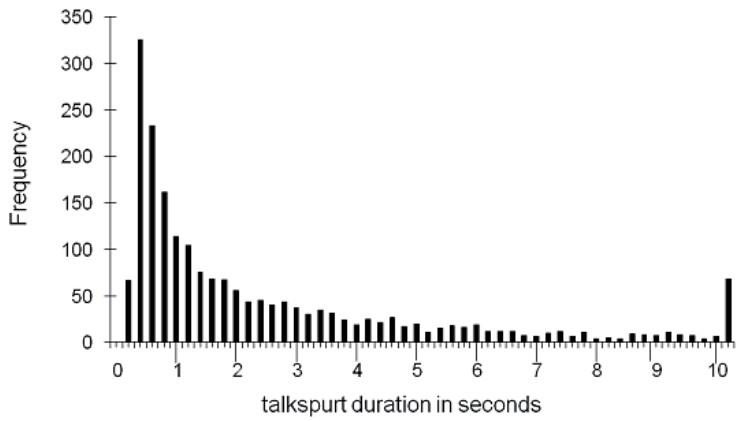


Figure 17. Talkspurt duration histogram (each bin spans 200 ms)

The different dialogue units that can be derived from the segmentation described above are illustrated in Figure 18. The units in this illustration can be briefly described as follows, *talkspurts* are stretches of speech surrounded by another interlocutor's speech whereas *IPUs*

are stretches of speech surrounded by silences longer than 200 ms. Silences within talkspurts are referred to as *pauses* and silences between talkspurts are referred to as *gaps*. *Very short utterances* (VSUs) are talkspurts shorter than one second.

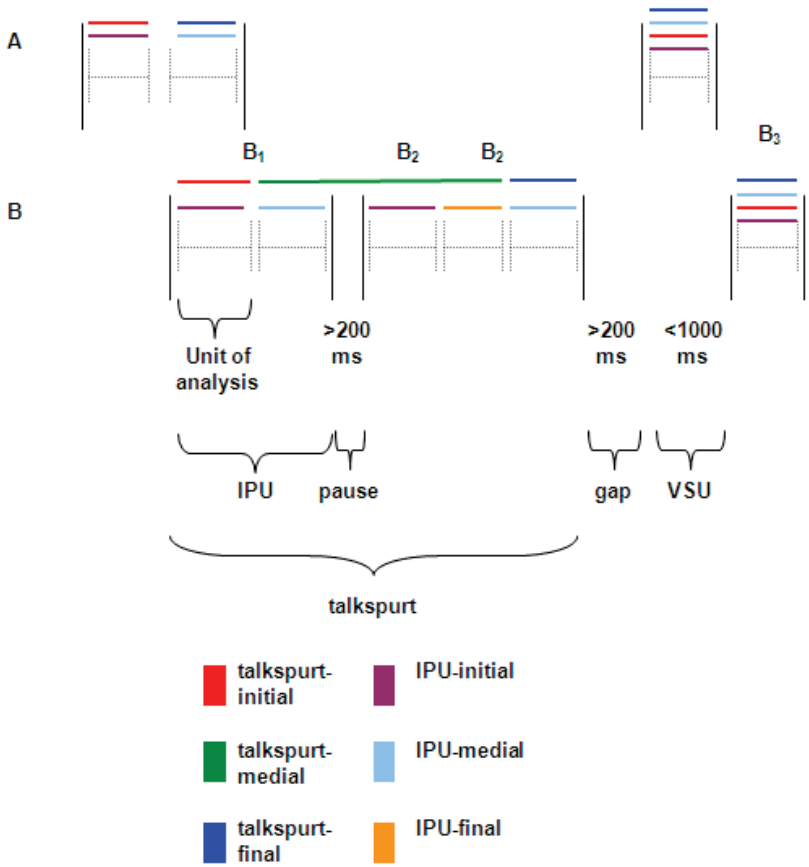


Figure 18. An illustration of different dialogue units in the DEAL corpus

The different dialogue units in Figure 18 can further be used to describe different positions of the behavioural phenomena that we aim to analyze. For example, a word or a sequence of words can be described as positioned in a talkspurt-initial or an IPU-initial position. Using this scheme, it is possible to make even more fine-grained descriptions of where certain phenomena occur based on combinations of these units.

Based on the data preparations described above, the manual annotation of cue phrases in the DEAL corpus will now be presented.

7.5. Cue phrases in the DEAL corpus

The definition of cue phrases used in this thesis is broad. All types of lexical entities that the speakers use to hold the dialogue together at different communicative levels are considered. Furthermore, cue phrases can be a single word or sequences of words, occupy various positions, belong to different syntactic classes, and be realized with different prosodic contours.

7.5.1. Annotation

The classification scheme used is an adaptation of a number of cue phrase categories suggested by Lindström (2008). Lindström defines cue phrases as words or expressions that regulate conversation or modify utterances. In line with many other cue phrase definitions (see Section 3.7 for an overview), Lindström argues that cue phrases are optional elements since they have relatively little propositional content. An example of this property is presented in Example 5.

Seller: det är ett bra pris för den röda klockan
English translation: that is a good price for the red clock
Customer: *men* jag har ingen nytta av två analoga klockor
English translation: *but* I have no use of two analogue clocks

Example 5. Dialogue excerpt from the DEAL corpus

The cue phrase "men" (Eng: "but") in the customer's utterance can be removed without changing the propositional content at the level of the isolated dialogue segment. Yet cue phrases are essential on a higher level of discourse since they affect how utterances are perceived in relation to previous discourse. The "men" above facilitates the interpretation of this utterance as a response to a previous price suggestion from the seller.

The cue phrase classes adapted from Lindström are based on functional categorizations. The different categories describe cue phrases' main functionality in a given context. Such categorization relies on the interpretation of the interlocutors' intentions. Underlying intentions cannot be attained, only interpreted by outside assessors or the speaker in hindsight. However, since cue phrases' role in conversation is difficult to define in terms of syntactic or lexical properties, the approach to manually annotate cue phrases in terms of their function is an attempt to explore how these elements are perceived in context. Once this is done, it is possible to explore specific cue phrases categories in more detail. The cue phrase classification scheme in this thesis includes:

- 3 classes of connectives: ADDITIVE CONNECTIVES, CONTRASTIVE CONNECTIVES and ALTERNATIVE CONNECTIVES. Typical examples of these classes are "och" (Eng: "and"), "men" (Eng: "but"), and "eller" (Eng: "or"), respectively. The connectives indicate how segments of speech connect to previous dialogue contributions.

- 3 classes of responsives: RESPONSIVE, RESPONSIVE NEW INFORMATION, and RESPONSIVE DISPREFERENCE. The responsives serve important grounding functions, acknowledging the previous dialogue contribution. The responsive cue phrases include feedback expressions such as "ja", "mm", and "a" (Eng: yes). The three different categories describe the acknowledging speaker's attitude. RESPONSIVE is used

when the speaker does not display any particular attitude. RESPONSIVE NEW INFORMATION indicates that that the information perceived was new or somehow unexpected. RESPONSIVE DISPREFERENCE indicates that the previous speaker's contribution was perceived, but that the acknowledging speaker does not agree. Isolated responsives uttered when listening to another interlocutor's talk correspond to what is typically referred to as back-channels or continuers.

- RESPONSE ELICITING are expressions used to elicit information from a conversation partner. Typical examples are "eller hur" (Eng: "right?"), "då" (Eng: "then" as in "shall I go then?"), and "väl" (Eng: "surely" as in "surely you are not angry?").

- REPAIR CORRECTION are editing expression such as "jag menar" (Eng: "I mean"), and "eller" (Eng: "or").

- MODIFYING cue phrases are expressions that modify attitude or value to the speaker's statement. Examples are "liksom" (Eng: "so to speak"), "ju", (Eng: "of course"), "åtminstone" (Eng: "at least") and "faktiskt" (Eng: "as a matter of fact").

In order to evaluate the applicability of the annotation scheme described above, two annotators labelled one of the dialogues according to these nine classes. It was noted that speakers sometimes used expressions to refer to a previous segment of speech. Therefore, a tenth class was added:

- REFERRING: a typical example of referring is "som vi sa" (Eng: "like we said").

Subsequently hesitations were also added as a cue phrase category:

- FILLERS refer to filler words such as "eh", "ehm".

Filler words are not typically considered cue phrases, however, these tokens have similar functionalities as other types of cue phrases. Filler words provide important pragmatic information, i.e. indicating that the speaker is hesitating. It has further been shown that these expressions create certain expectations from the interlocutor's and that they attend and adjust their behaviour according to these phenomena. For example, Brennan & Schober (2001) show that fillers help listeners compensate for disruptions and delays in spontaneous speech, and a corpus study of Dutch fillers showed that these tokens can highlight discourse structure (Swertz, 1998).

An overview of all cue phrases classes is presented in Table 6. For each cue phrase category, the first row presents an example of this category in context from the DEAL corpus, the word(s) in bold is the annotated cue phrase, and the second row presents the three most frequently used for that particular class. The most frequently used instance is the first word; the secondly most used instance is the second word and so on.

Additive Connectives (CAD)
och grönt är ju fint [**and** green is nice]
så, också, då [so, also, then]
Contrastive Connectives (CC)
men den är ganska antik [**but** it is pretty antique]
men, alltså, fast [but, thus, although]
Alternative Connectives (CAL)
som jag kan titta på **istället** [which I can look at **instead**]
eller, istället [or, instead]
Responsive (R)
ja jag tycker **ju** det [yeah I **actually** think so]
ja, mm, a [yes, mm, yeah]
Responsive New Information (RNI)
jaha har du några sådana [**right** do you have any of those]
ja, ok, mm [yes, ok, mm]

Responsive Dispreference (RD)
ja men det är klart dom funkar [**yeah** but of course they work]
ja, nej, nä [yes, no, no]
Response Eliciting (RE)
vad ska du ha för den **då** [how much do you want for that one **then**]
då, va, eller hur [then, what, right]
Repair Correction (RC)
nej nu sa jag fel [**no** now I said wrong]
nej, jag menar [no, I mean]
Modifying (MOD)
ja jag tycker **ju** det [yeah I **actually** think so]
ju, faktiskt, ja [actually, as a matter of fact, yeah]
Referring (REF)
fyra hundra kronor **sa vi** [four hundred crowns **we said**]
som sagt, nu igen, vad sa vi [as said, now again, what did we say]
Fillers (FILL)
hon är väl en **eh** fem sex år [she is about **eh** fix six years old]
eh, ehm, hm

Table 6. The DEAL cue phrase classification scheme

The fillers had already been annotated during the initial orthographic transcriptions. All other cue phrase categories were manually annotated in a subsequent annotation step. This annotation included a two-fold task, to decide if a word was a cue phrase or not – a binary task – but also to classify what functional class it belongs to according to the annotation scheme. The annotators could both see the transcriptions and listen to the recordings while labelling. Two of the eight dialogues were annotated by both annotators. In order to assess inter-annotator agreement, a kappa coefficient was calculated on word level. The kappa coefficient for the binary task, to classify if a word was a cue phrase or not, was 0.87 ($p<.05$). The kappa coefficient for the classification task was 0.82 ($p<.05$). The agreement in percentage distributed over the classes is presented in Figure 19.

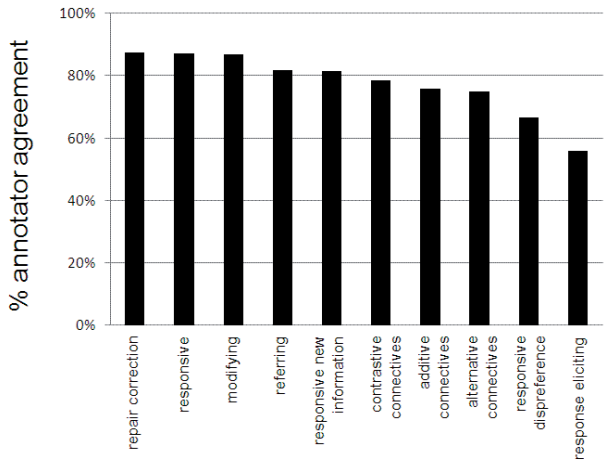


Figure 19. % Cue phrase annotator agreement (2 labellers)

7.5.2. Data analyses

The high inter-annotator agreement for the cue phrase classification task suggests that the categorisations were reliable, but what kind of features characterize these elements? The rest of this chapter is concerned with general descriptive statistics and some exploratory data analyses of the cue phrases in the DEAL corpus.

Analyses of the cue phrase distribution shows that 76% of all talkspurts contained at least one cue phrase, and 23% of all words were labelled as cue phrases. In line with previous observations (c.f. Heeman et al., 1998), it was observed that cue phrases often occur in a talkspurt-initial position. 51% of the talkspurts in the corpus were initiated with some type of cue phrase. This can be compared to 68.2% in the TRAINS corpus (Heeman et al., 1998). However, it is difficult to compare the frequency of cue phrase between these corpora since the dialogue units and cue phrases classifications differ.

Figure 20 shows how the cue phrases were distributed over the different cue phrase categories.

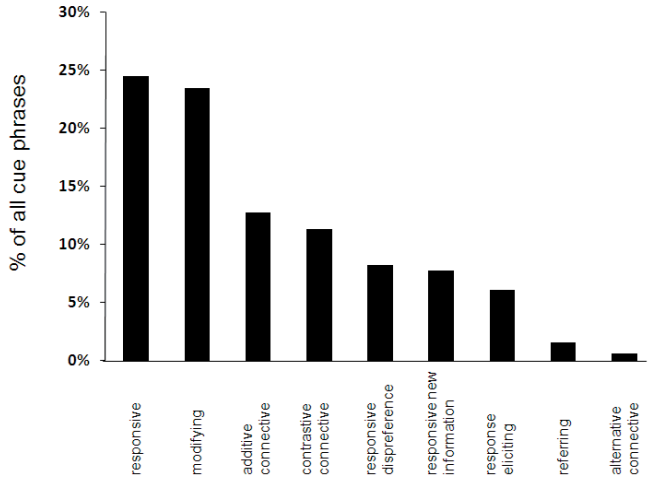


Figure 20. Cue phrase distribution over the different categories

CONNECTIVE ALTERNATIVE, REPAIR CORRECTION and REFERRING were excluded from further analyses since these had a frequency of less than 100 instances.

91% of the cue phrases are single words, 7% are two words and the remaining 2% are three words or longer. Table 6 suggests that many words annotated as cue phrases are not exclusive to one single cue phrase category, but appear in several classes. Many of these words are high frequency words. 53% of the instances of the ten most frequently occurring words in the corpus (see Table 4) are annotated as cue phrases. Of all words labelled as cue phrases, the 8 most frequent tokens, all with a total frequency larger than 100, are listed in falling order of frequency in Table 7. Column 2 displays the most frequent cue phrase classification for that word and column 3 displays the second most frequent category. For the five most frequent words annotated as cue phrases – "ja", "ju", "men", "m", and "eh" – more than 60% of these instances in the corpus are annotated with a single cue phrase category, RESPONSIVE, MODIFYING, CONTRASTIVE CONNECTIVES, RESPONSIVE, and FILLERS respectively.

Order of frequency in corpus	Transcription	Most frequent classification		Second most frequent classification	
		Classification	%	Classification	%
1	ja	Responsives	60%	No cue phrase	10%
2	ju	Modifying	97%	No cue phrase	2%
3	men	Contrastive connectives	86%	Additive connectives	6%
4	m	Responsives	77%	Responsive Dispreference	9%
5	eh	Fillers	100%	-	-
6	då	No cue phrase	42%	Response Eliciting	29%
7	a	Responsives	71%	Responsive New Info	12%
8	så	No cue phrase	51%	Additive connectives	38%

Table 7. Cue phrase classification of the 8 most frequent words annotated as cue phrases in the DEAL corpus

Figure 21 presents the positions of the different cue phrase categories over talkspurts. Words not annotated as cue phrases are represented as "other". The chart illustrates that the three responsives – RESPONSIVE, RESPONSIVE DISPREFERENCE and RESPONSIVE NEW INFORMATION – mainly occur in a talkspurt-initial position or are a talkspurt all on their own. All of the responsives that were complete talkspurts were very short utterances (VSUs, shorter than 1 second). As expected, RESPONSE ELICITING occurs mainly in a talkspurt-final position. The rest of the cue phrase categories – MODIFYING, CONTRASTIVE CONNECTIVES, ADDITIVE CONNECTIVES, FILLERS, and words not annotated as cue phrases – occur mainly in a talkspurt-medial position.

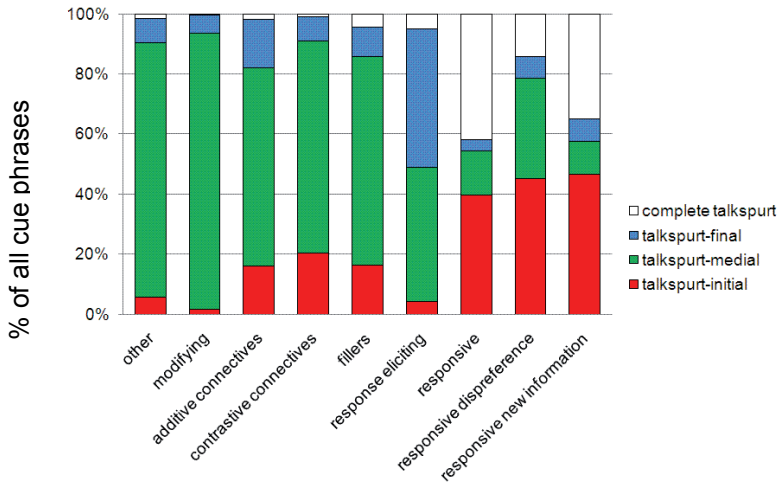


Figure 21. Cue phrase distribution over talkspurts

The distribution of cue phrases was also analyzed over IPUs. These data are presented in Figure 22.

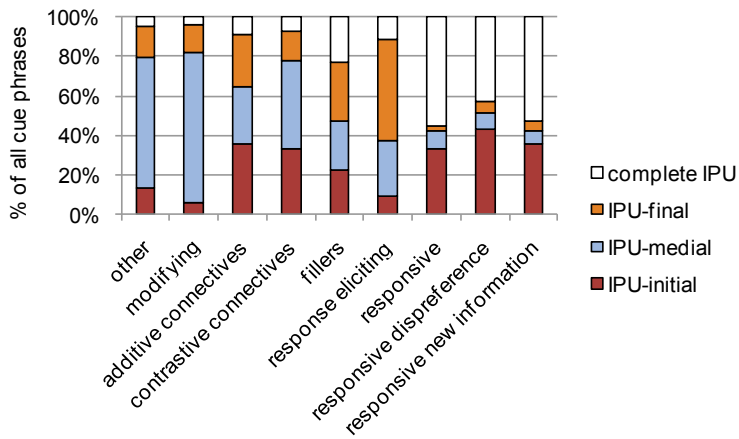


Figure 22. Cue phrase distribution over IPUs

The IPU distribution shows similar trends as the talkspurt distribution. However, the IPU distribution shows that a relatively small part of the cue phrases occur in a IPU-medial position. This suggests that cue phrases often are adjacent to a pause or a speaker change.

Figure 23 presents the proportion of instances in IPU-medial position in relation to the instances that are adjacent to a pause or/and speaker change, that is, regardless if a complete IPU or in a IPU-initial or IPU-final position.

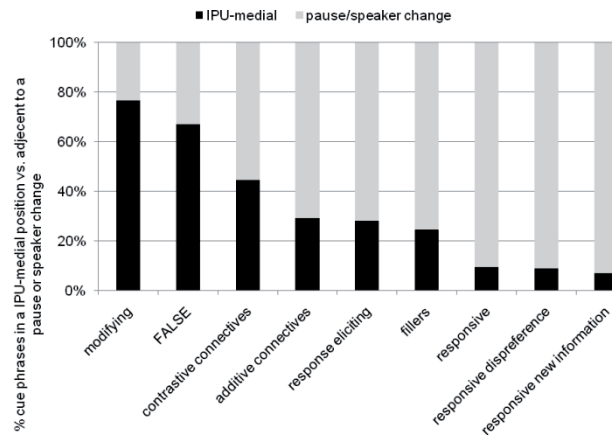


Figure 23. % cue phrases in an IPU-medial position vs. cue phrases adjacent to a pause or speaker change

With the exception of the MODIFYING cue phrases, a majority of all cue phrases categories are adjacent to a pause or a speaker change. This can be compared to tokens not annotated as cue phrases (other), which occur in an IPU-medial position more than 60% of the time. A Chi-square test of independence was employed to further explore this difference. The results from this test suggest that the cue phrases occur more frequently adjacent to a pause or a speaker change than words not annotated as cue phrases, $C^2(1, N= 16185) = 1059$, $p<.00$.

7.5.3. Responsive cue phrases

Lexical analyses suggest that the transcription alone cannot be used to classify cue phrases. The responsive cue phrases – RESPONSIVE, RESPONSIVE DISPREFERENCE, and RESPONSIVE NEW INFORMATION – occur frequently in the corpus (1374 instances). 42% of these cue phrases occur in talkspurt-initial position and 36% are VSUs. The most frequent lexical items for all categories is "mm", "ja", and "a" (Eng: variations of "yes"). As there is a lot of lexical overlap between these categories, the responsive cue phrases were analyzed acoustically in order to explore if these can be distinguished prosodically. First, the duration, pitch and intensity for "mm", "ja" and "a" were extracted. The average F0 (in Hz) per cue phrase were automatically extracted using Snack (www.speech.kth.se/snack/). The fundamental frequency was transformed into a semitone scale using 261.63 Hz (middle C) as a reference value. The semitone scale was normalized per speaker (each value was shifted by the mean value per speaker and dialogue). Figure 24 shows average F0 values in (normalized) semitones for the different responsive cue phrase categories. More than 50% of the responsive tokens were uttered simultaneously as the other speaker. These overlaps had to be excluded from the data analysis since the recordings were not completely channel-separated and crosstalk could conceivably interfere with the results. This resulted in a relatively low number of data points that could be used for analyses (315 RESPONSIVE, 78 RESPONSIVE and 52 RESPONSIVE NEW INFORMATION).

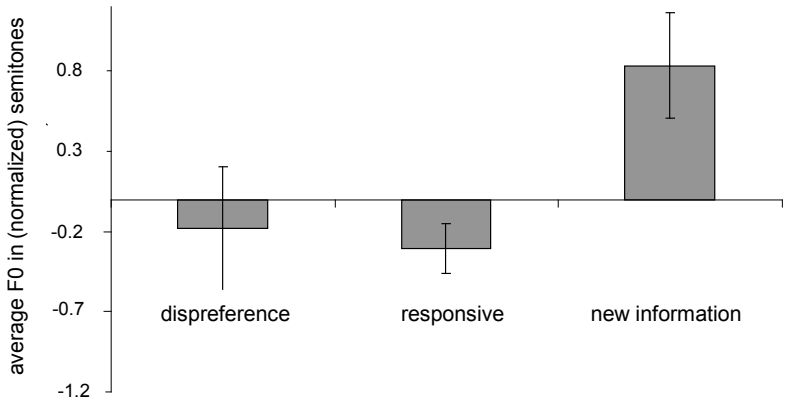


Figure 24. Average F0 in (normalized) semitones. Error bars represents the standard error.

Figure 24 suggests that RESPONSIVE NEW INFORMATION is about one semitone higher than the two other RESPONSIVE categories. A one-way ANOVA with the different cue phrase categories RESPONSIVE DISPREFERENCE, RESPONSIVE, and RESPONSIVE NEW INFORMATION, as a between-subject factor was used to test for differences in F0. No significant differences in pitch between RESPONSIVE cue phrase categories were found. It should be noted that RESPONSIVE NEW INFORMATION has only 52 non-overlapping tokens that could be used for statistical analyses.

The average intensity (dB) per cue phrase was also automatically extracted using Snack. The intensity was normalized over speaker and dialogue. Data plots suggest that all the cue phrase categories are similar in intensity, thus, the mean difference is smaller than one dB. It is questionable if such a small difference is noticeable in dialogue.

Finally, a one-way ANOVA with the different cue phrase categories as a between-subject factor was conducted to test for differences in duration. The three most frequent one-syllable words “ja”, “a” and “mm” were explored in this analysis. Significant differences in dura-

tion between the categories were found. F (4, 1050) = 21.13, p < .00. Tukey's post hoc comparisons of the three groups indicate that RESPONSIVE DISPREFERENCE is longer in duration than all other the other RESPONSIVE cue phrases (see Table 8).

Cue phrase annotation		Difference in mean duration (milliseconds) i – j	Standard error	p-value
i	j			
Responsive Dispreference	No cue phrase	157	0.02	0.000
	Responsive	170	0.02	0.000
	Responsive New Information	125	0.03	0.000

Table 8. Differences in average duration between the cue phrase categories RESPONSIVE, RESPONSIVE NEW INFORMATION, and RESPONSIVE DISPREFERENCE (Tukey's p<.05, df=4).

7.5.3.1. ROC-curve analyses

The results presented above show that there are differences in duration between RESPONSIVE DISPREFERENCE and the other RESPONSIVE cue phrase categories. To explore duration as a discriminative feature, ROC (relative or receiver operating characteristic) curves were used (c.f. Metz, 1978). ROC-curves are mainly used to study the accuracy of a diagnostic test in terms of how well it discriminates diseased cases from normal cases. More specifically, ROC-curves illustrate the relationship between true positive rate (TPR) and false positive rate (FPR) as a discrimination threshold is varied. The shape of an ROC-curve illustrates the overall accuracy of a test in terms of the sensitivity, the probability that a test result will be positive when the target condition is present, versus the specificity, the probability that a test is negative when the target condition is not present. Each

point in an ROC-curve represents the sensitivity versus the specificity for a particular cut-off value. The closer the curve is to the upper left corner of the graph, the higher is the accuracy of the test. A useless test that is no better than chance at identifying true positives has an area of 0.5 (illustrated by a no discrimination line in Figure 25). A test with perfect discrimination has an area of 1.00.

True Positive Rate (TPR) is the percentage of RESPONSIVE cue phrases that was correctly classified as RESPONSIVE DISPREFERENCE based on duration as the threshold for these values are varied. False Positive Rate (FPR) is the percentage of tokens incorrectly classified as negative as the threshold values are varied. The ROC-curve for classifying RESPONSIVE DISPREFERENCE based on duration is plotted in Figure 25. It illustrates how well tokens annotated with a RESPONSIVE DISPREFERENCE can be separated from the same tokens that are not annotated with that cue phrase category using duration.

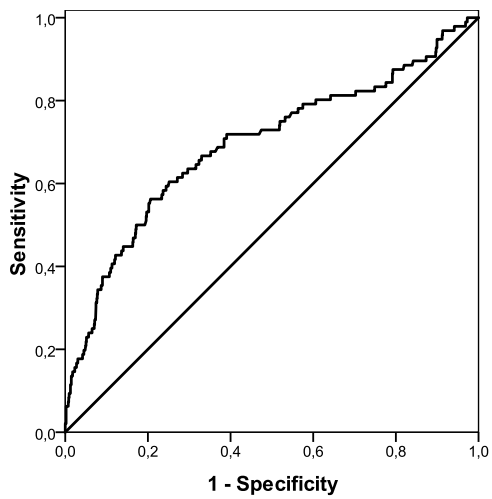


Figure 25. Receiver Operating Characteristic Curve for classifying RESPONSIVE DISPREFERENCE based on duration (in ms)

The area under the curve for these ROC-curves is 0.69. This suggests that the duration of the cue phrase carrie some discriminative power, but the accuracy is rather weak. Although duration alone may not be sufficient to discriminate RESPONSIVE DISPREFERENCE from other RESPONSIVE cue phrases, the difference in duration can be used as guidelines for speech production in spoken dialogue systems.

7.5.1. Talkspurt initial intensity

Previous research suggests that an increase in intensity has turn-claiming functions. For example, Ström & Seneff (2000) increas intensity in order to signal that user barge-ins are disallowed in particular dialogue states. Theoretical support for such manipulations is provided by an early line of research on interruptions in dialogue (Meltzer et al., 1971), who suggest that the outcome of speech overlaps is affected by prosodic characteristics and show that the greater the increase in amplitude, the greater the likelihood of "interruption success". Moreover, they show that the success of interruptions, that is who retains the floor, is based on how much higher the intensity of the interruption is compared to the previous speaker's intensity or compared to the speaker's own intensity at the end of that speaker's previous turn.

51% of all talkspurts in the DEAL corpus were initiated with cue phrases. The majority of the talkspurt-initial cue phrases in the DEAL corpus were high frequency monosyllabic words. Short, high-frequency function words are typically not associated with stress, although on listening, they give the impression of being louder than other talkspurt-initial vocalizations. To verify this observation, the first word in each talkspurt was extracted and analyzed. The talkspurt-initial cue phrases were annotated with different cue phrase categories: 587 (28%) talkspurt-initial words were annotated as either RESPONSIVE, RESPONSIVE DISPREFERENCE or RESPONSIVE NEW INFORMATION. 189 (9%) of all talkspurt-initial words were annotated as CONNECTIVES. The third most frequent talkspurt-initial cue phrase category was FILLERS (57, 3%).

The intensity in decibel of the first word of each talk spurt was extracted using Snack. All talkspurts following a one word only talkspurt from the other speaker were excluded as an approximation to avoid speech following backchannel responses. 300 (33%) of the speaker changes contained overlapping speech. These overlaps were excluded from the data analysis since the recordings were not completely channel-separated and crosstalk could conceivably interfere with the results.

Since the distance between the lips and the microphone was not controlled for during the recordings, the values were first normalized per speaker and dialogue (each value was shifted by the mean value per speaker and dialogue). Figure 26 presents the average normalized intensity for talkspurts initiated with cue phrases and other words.

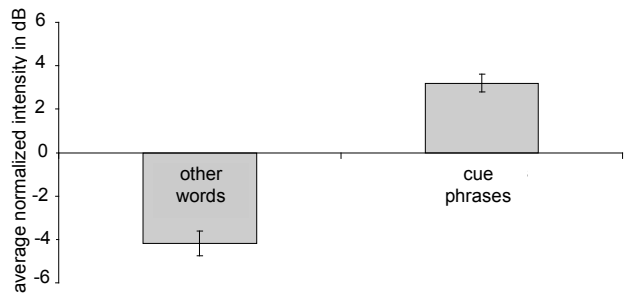


Figure 26. Average normalized vocal intensity in dB for talk-spurt initial words. Error bars represents the standard error.

An independent samples t-test was conducted between the intensity of talkspurts initiated with cue phrases and other talkspurt-initial words. There was a significant difference in intensity between talkspurts initiated with cue phrases (M=3.20 dB, SD=6.99) and talkspurts initiated with other words (M=-4.20 dB, SD=9.98), $t(597)=10.55$, $p<.00$. This shows that, on average, talkspurts initiated with cue phrases were significantly louder (on average 6 dB) than talkspurts initiated with other words.

Next, an additional approach to measuring talkspurt-initial intensity was explored. This was done to investigate whether the voice intensity in the interlocutor's immediately preceding speech can be used as a reference point in order to measure intensity as an inter-speaker relation over the course of a dialogue. This approach is motivated by research which suggests that speakers adjust their vocal intensity online in order to accommodate the surrounding acoustic context. For example, speakers tend to raise their voice unintentionally when background noise increases to enhance their audibility; this is the so-called Lombard effect (Pick et al., 1989). Speakers also adjust intensity based on their dialogue partners (Natale, 1975) and the distance to their listeners (Healey et al., 1997).

In order to explore the vocal intensity as an inter-speaker relation continuously over the dialogue, the average intensity of the last word of all talk spurts was extracted. In order to avoid the need for global analysis over speakers and dialogues, only the (un-normalized) difference in intensity between the last word of the immediately preceding talkspurt and the first word of a new talkspurt was calculated. The inter-speaker differences in intensity for talkspurt-initial cue phrases and other words are presented in Figure 27.

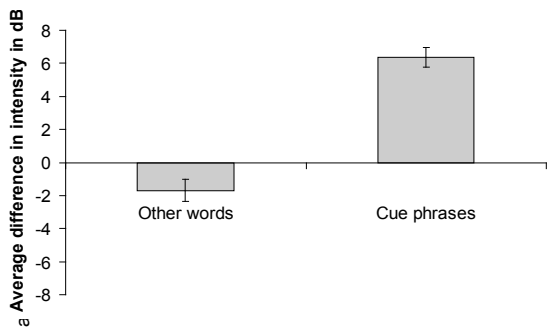


Figure 27. Average difference in intensity (in dB) for talkspurt-initial words. Error bars represents the standard error.

An independent samples t-test was conducted to explore the difference in voice intensity as an inter-speaker relation. There was a significant difference in intensity between talkspurts initiated with cue phrases (M=6.14 dB, SD=11.86) and talkspurts initiated with other words (M=-1.52 dB, SD=13.07); t(595)=7.48, p<.00. This suggests that the increase in intensity was significantly larger for talkspurts initiated with cue phrases (about 7 dB) than for talkspurts initiated with other words.

7.6. Discussion – implications for dialogue systems

The analyses of cue phrases presented in this chapter focus on cue phrases uttered adjacent to speaker changes and pauses within talkspurts. Words uttered adjacent to these events are important to master in DEAL in order manage turn-taking and to provide the user with information on the system's continued plan of interaction as well as on how the new dialogue contribution relates to previous dialogue contributions.

Cue phrases that occur frequently at a talkspurt-initial position are different type of feedback expressions that were annotated as responsives in the DEAL corpus. The annotation was based on the interpretation of the speakers' attitudes, expressing either neutral feedback (RESPONSIVE), non-agreement (RESPONSIVE DISPREFERENCE) or surprise (RESPONSIVE NEW INFORMATION). A dialogue system with a repertoire of such talkspurt-initial feedback expressions can rapidly initiate new talkspurts after the user has stopped speaking. In dialogue systems capable of incremental processing, the system can employ responsives to initiate talkspurts without response delays and simultaneously continue planning the rest of the message.

In order to explore the characteristics of the three responsive categories further, these elements were analysed lexically and prosodically. The results show that the three RESPONSIVE categories are very similar and it is difficult to find intrinsic characteristics that can be

used to automatically distinguish between these categories. However, as our principal motivation is to generate responsives rather than recognize them, two minor differences in prosodic characteristics can be noted. First, although not statistically significant, plotting the data (see Figure 24) suggests that RESPONSIVE NEW INFORMATION is produced with a slightly higher F0 than the other responsives. Furthermore, RESPONSIVE DISPREFERENCE is slightly longer in duration than the other two categories. These results can be compared to Freese & Maynard (1998) who claim that a high pitch is associated with good news while a decrease in speech rate is argued to be associated with bad news. Gardner (2001) further suggests that the prosody of short feedback expression also regulate turn-taking and that a fall-rise contour characterize continuers while a falling pitch marks topic completion. However, to further explore the pragmatic functions of prosody for these tokens here, more data is needed.

This chapter also explored the intensity of talkspurt-initial cue phrases. The majority of the talkspurt-initial cue phrases were high frequency monosyllabic words such as “yes”, “mm” and “ok”. The most frequent talkspurt-initial words that were non-cue phrases were “den” (Eng: “it”), “vad” (Eng: “what”), and “jag” (Eng: “I”). Thus, similar to talkspurt-initial cue phrases, these tokens were high-frequency one-syllable words, items that are not typically associated with prosodic stress. Yet, the results show that talkspurt-initial cue phrases are produced with higher intensity than other talkspurt-initial words are. In the light of previous research, which suggests that increased intensity has talkspurt-claiming functions, one can speculate that speakers produce talkspurt-initial cue phrases with increased intensity in order to convincingly claim the floor before having formulated a complete utterance. It is also possible that the increase in intensity is used to indicate that the feedback expression is the initiation of a talkspurt rather than a backchannel.

Finally, it is proposed that intensity may be better modelled relative to the intensity of the immediately preceding speech rather than in absolute speaker normalized terms. Speakers adjust their intensity

to the current acoustical environment and such a dynamic inter-speaker relative model may accommodate the current acoustic context over the course of a dialogue. In support of this proposition, results presented in this chapter showed that the increase in intensity can be calculated dynamically over the dialogue using the end of the previous speaker's talkspurt as a reference point. Inter-speaker relative measures are also motivated practically. Extracting objective measures of intensity is problematic since contextual factors such as the distance between the microphone and the lips are difficult to control between dialogues and speakers, but the effects are mitigated by dynamic and relative measures. This is not to say that measuring intensity over the course of a single dialogue is trivial. Variation due to for example unforeseen alterations of the distance between the lips and the microphone during the dialogue are still problematic, but it is less of a problem within a session than between different sessions.

7.7. Summary

This chapter has presented a data collection of human-human dialogues in the DEAL domain. The dialogues were annotated for cue phrases with high inter-annotator agreement. By manually annotating cue phrases, a set of tokens that can be used to signal different pragmatic functions in the DEAL domain has been identified. The descriptive statistics and explorative data analyses presented can be used as guidelines for when and how to produce cue phrases in dialogue systems.

Many of the cue phrases explored in this chapter are related to how the conversational floor is passed from speaker to speaker. The next chapter explores such turn-taking behaviours in more detail.

8. The additive effect of turn-taking cues in human and synthetic voice

A crucial aspect in dialogue systems is to control the flow of dialogue contributions between the system and the user. This chapter presents a study that explores how different behaviours affect listeners' expectations of a turn change. The existence of turn-taking cues is based on the previous findings which suggest that listeners attend to the small variations in which speech is delivered and that these variations have pragmatic implications that influence who speaks when in a conversation (Duncan, 1972). The experiment was set up like a game where the participants listened to recorded dialogues, and when the recording stopped, they were to guess who would be the next speaker. One motivation of this experiment was to explore whether verbal turn-taking behaviours can be realized with a synthetic voice. For this reason, the stimuli contained both human-human dialogues and dialogues where one of the speakers had been replaced with a synthetic voice.

8.1. Introduction

One challenge is to build dialogue systems that produce dialogue contributions in a timely fashion. This task partly relies on the system's abilities to interpret the users' interactive behaviour in order to

know when it is appropriate to speak. The other part is to provide the user with similar information. In order to do this, the system needs to maintain an updated model of its future dialogue plans. If speech segments are produced incrementally, in short segments, the system needs to indicate whether such a segment is talkspurt-medial or talkspurt-final. This will help users discriminate between pauses and gaps and make the interaction with the system more intuitive. Whether a speech segment is syntactically or semantically complete or not has shown to play an important role when interlocutors determine if it is an appropriate place to speak (de Ruiter et al., 2006). However, a syntactically complete phrase is not necessarily talkspurt-final and vice versa. Previous research has shown that additional to lexical cues, speakers rely on a number of prosodic and visual behaviours (c.f. Duncan, 1972; Novick et al., 1996; Gravano, 2009).

This chapter presents an experimental study that investigates how behavioural turn-taking cues form a complex signal and affect listeners' interpretations of turn-taking behaviour in dialogue. The motivation is to investigate the possibilities of generating turn-management cues with a synthetic voice with the future aim to employ such cues in spoken dialogue systems in order to make turn-taking in these systems more intuitive.

8.2. Methods for exploring turn-taking

Many of the behavioural cues that have been suggested as relevant for turn-taking are subtle, and speakers probably employ and react to these behaviours automatically. These characteristics make it difficult to identify and explore the effects of these behaviours.

Theories of turn-taking have been strongly influenced by work done within conversation analysis (CA) (c.f. Sacks et al., 1974). Studies following a CA tradition have experts analyzing a few isolated instances of the relevant phenomena. Other research has covered larger sets of data by studying correlates of turn-switches in corpora of varying sizes (c.f. Duncan, 1972; Gravano, 2009). However, as

pointed out by de Ruiter et al. (2006), observations of correlations between certain behavioural phenomena and turn-endings do not necessarily imply causality.

There are few examples of studies that have investigated turn-taking experimentally. For example, Schaffer (1983) and Oliveira & Freitas (2008) studied the role of prosody in turn-taking by analyzing the judgments of non-participating listeners in perceptual experiments. In order to isolate the prosodic realization from the semantic influence, the stimuli in Schaffer's experiment were band-pass filtered to render the utterances intelligible. The results show a great variability in the listeners' use of intonation and do not support a clear-cut effect of prosody alone. Work by de Ruiter et al. (2006) questions the role of prosody in turn-taking entirely and suggests that humans predict upcoming turn-endings by lexico-syntactic content alone after showing that listeners' accuracy in predicting upcoming turn-endings did not decrease when the intonational contour was removed. However, manipulating dialogues off-line and analyzing these out of context can be problematic since this may result in stimuli that never would occur in a real dialogue setting. To tackle this problem, the present study use recordings of un-manipulated dialogues as stimuli. Furthermore, similar to how dialogue is perceived in its original context, the subjects were allowed to follow longer dialogues segments chronologically. The motivation of this approach was to let the subjects get familiar with the speakers and the content of the dialogues.

8.3. Turn-taking cues

Before introducing the present study, the concept *turn-taking cue* needs to be discussed. Duncan uses the term "signal" which implies that speakers employ these behaviours deliberately. Yet, these phenomena are likely more or less automatized. For instance, there are acoustic phenomena, e.g. drop in energy or inhalations that guide interlocutors in their turn-taking. The likely origin of these "signals"

is the anatomy of our speech organs. If we plan to continue speaking, we keep the speech organs prepared and if we plan to finish, we release them (Local & Kelly, 1986). Whether conscious or not, these non-verbal phenomena appear to affect the addressees' interpretation of the message. However, since the behaviours are not necessarily deliberate, the present study employs the term *cue* rather than signal. Here, *turn-taking cues* refer to all perceivable phenomena relevant for turn-taking, regardless of whether they are conscious or not.

Previous research has mainly focused on turn-yielding cues, cues that indicate that a talkspurt is about to be completed. In this study, however, cues to maintain the turn, so-called turn-holding cues, are also considered. This decision is motivated by the need for devices to maintain the turn when producing speech incrementally. If the system has initiated a talkspurt, but the rest of the message is delayed, the system needs to indicate that the upcoming pause is not an appropriate place for the user to speak.

8.4. Method

The aim of this study is to explore the possibilities of using turn-taking cues to generate appropriate turn-taking behaviour in spoken dialogue systems. Thus, in addition to stimuli with human-human dialogue, the experiment included stimuli where one of the human interlocutors was replaced with a synthetic voice. The motivation to use a synthesis rather than a pre-recorded human voice in a dialogue system is that synthetic voices are easier to update and manipulate on-line (Reiter and Dale, 1997). For example, no new recordings are needed to manipulate prosody or to extend the system's vocabulary.

8.4.1. Stimuli preparation

The DEAL recordings were used as a source of stimuli in the experiments. First, suitable dialogue segments to use as stimuli needed to be extracted. The initial effort was to identify different types of turn-taking cues in these data. Since annotating the entire data set with

turn-takings was too time-consuming, only the final segments of IPUs, i.e. stretches of speech just prior to a silence longer than 200 ms, were annotated. Findings presented by Izdebski & Shipp (1978), suggest that speakers need at least 200 milliseconds to react verbally to an auditory stimulus. Hence, the present study is concerned with non-overlapping turn-changes. This decision is based on the methodological issues of creating stimuli for overlapping speaker changes. The subjects in the present experiment were presented with continuous dialogue segments and whenever the recording was paused, their task was to predict if there was going to be a speaker change or not. If the entire IPU was used as a stimuli, overlapping speaker changes would reveal who the next speaker was. On the other hand, interrupting the recording immediately before an overlap, in the middle of an IPU, would have made the experimental task more complex. Thus, rather than simply asking "who will speak next?", we would need to ask "is there going to be a speaker change soon?" or "is the current speaker about to finish?". To avoid these issues, only non-overlapping speaker changes were considered. In order to obtain examples of turn-yielding as well as turn-holding cues, IPUs in different talkspurt positions (prior to both gaps and pauses) were extracted and annotated for turn-taking cues.

According to the definition of turn-taking cues presented in the previous section, the work presented here is concerned with perceivable phenomena that are relevant for turn-taking. The approach used to identify such behaviours was to manually annotate a number of phenomena that have been argued as relevant for turn-taking in the previous literature. To extract behaviours that were perceivable to human listeners rather than automatically extracting features, two human annotators were used.

While, the original dialogues were face-to-face interactions, the subjects in the experiment were presented with audio only. The motivation of this decision was to focus on the lexical and acoustic cues that can potentially be reproduced in a synthetic voice. It has been hypothesized that speakers use fewer prosodic cues in face-to-face

conversation since the visual channel provides an additional set of turn-taking devices (c.f. Schaffer, 1983). However, the results presented by Schaffer show no support for this assumption.

Duncan (1972) has been criticized for not reporting inter-annotator agreement or formal description of his "signals" (Beattie et al., 1982). An important part of this work is therefore to provide a detailed description of how the turn-taking cues were identified.

8.4.2. Annotation of turn-taking cues

The present study explores six different categories of turn-taking cues. The cues were chosen to represent a fair distribution of different turn-taking phenomena. Except for semantic completeness, the cues explored are discrete in nature. The advantage of discrete cues is that these can be produced in dialogue system without necessarily modelling these phenomena continuously over the course of the dialogue. The cue categories explored were *intonation*, lexical completion (*semantic completeness*), *phrase-final lengthening*, other speech production phenomena such as perceivable breathing and lip-smacks, and some frequently occurring cue phrases (see Table 9).

Category	Turn-yielding cues	Turn-holding cues
Intonation	fall	flat
Phrase-final lengthening	no phrase-final lengthening	long phrase-final lengthening
Speech production phenomena	audible expirations	audible inhalations, lip-smacks
Disfluencies	-	abrupt halts and repetitions
Cue phrases and filled pauses	response eliciting	connectives filled pauses
Semantic completeness	complete	incomplete

Table 9. Turn-taking cue categories

Two annotators were used for labelling. The annotators were researchers at the Department of Speech, Music and Hearing at KTH with good knowledge of linguistics and phonetics. In order to avoid influences from other cues, the cues were annotated one at a time and the labellers' task was to identify the target phenomena only, without knowing who the next speaker was and without considering any turn-taking issues.

8.4.2.1. Intonation and phrase-final lengthening

When identifying prosodic cues automatically, Gravano (2009) analyzed the last 200 ms of IPUs. In the present study, it was noted that 200 ms of speech was too brief to annotate manually. Instead, for the annotation of intonation and phrase-final lengthening, the last 500 milliseconds of IPUs were analyzed manually. During annotation, the stimuli were presented to the annotators in isolation and in random order in order to reduce influences of the prosodic realization of adjacent speech and the lexical context. For intonation, the target labels were *flat*, *rising* or *falling* pitch contour whereas the target labels for phrase-final lengthening were *long*, *short* and *no* phrase-final lengthening. The inter-annotator agreement for both tasks were 69% overall agreement or kappa 0.37. To address the poor inter-annotator agreement, two precautions were taken. First, only stimuli where both annotators agreed were considered to contain cues. Secondly, the reliability of the manual annotations was further explored in terms of how well these correspond to automatically extracted measures of fundamental frequency (F0) and speaker rate.

As an automatic measure of intonation, the change in F0 during the last 200 ms of the IPU was automatically extracted using Snack and z-score-normalized over speaker and dialogue. As a measure of phrase-final lengthening, speaking rate was calculated over IPUs as the number of syllables per second. Negative durational data is impossible and the distribution of syllable durations will therefore be skewed to the left. This was confirmed by histograms of the distribution of speaker syllable rate per second. Since it has been suggested

that the log-normal law is a better fit to duration data (c,f. Campione & Veronis, 2002), speaking rate was calculated per second and transformed into a logarithmic scale (base 10). The syllable rate was also z-score-normalized over speaker, dialogue and phoneme.

8.4.2.2. ROC-curve analyses of intonation and phrase-final lengthening

To explore the relationship between the automatic measures and the manual annotations, ROC (relative or receiver operating characteristic) curves were used (see 7.5.3.1 for a more detailed discussion). Here, True Positive Rate (TPR) is the percentage of IPUs with a specific prosodic cue (labelled by both annotators) that was correctly classified as positive based on automatically extracted values of F0 and syllables rate as the threshold for these values are varied. False Positive Rate (FPR) is the percentage of IPUs incorrectly classified as negative as the threshold values are varied. The ROC-curves for intonation and phrase-final lengthening are plotted in Figure 28 and Figure 29 respectively. The aim is to illustrate how well IPUs annotated with a specific prosodic cue can be separated from IPUs that are not annotated with that cue using automatically extracted values of F0 and syllable length. The shapes of the curves suggest that threshold values for automatically extracted F0 and syllable rate can be selected to identify the manually annotated prosodic cues with high accuracy, that is, well above chance. The accuracy for flat intonation (area under the curve 0.84) is higher than for falling intonation (area under the curve 0.72). Area under the curve for long phrase-final lengthening is 0.72 and 0.77 for no lengthening. The discriminative power of these tests, that is, the possibility to identify these manually annotated cues using automatically extracted prosodic features, suggests that the annotators indeed were labelling something dependable, despite the low kappa values. The ROC-curve for rising intonation, however, suggests that the accuracy for this test is poor. For this reason, this cue was excluded from further analyses.

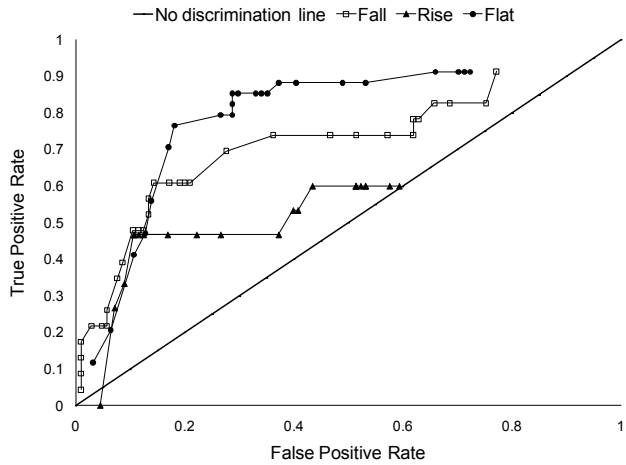


Figure 28. ROC curve for falling, rising and flat intonation

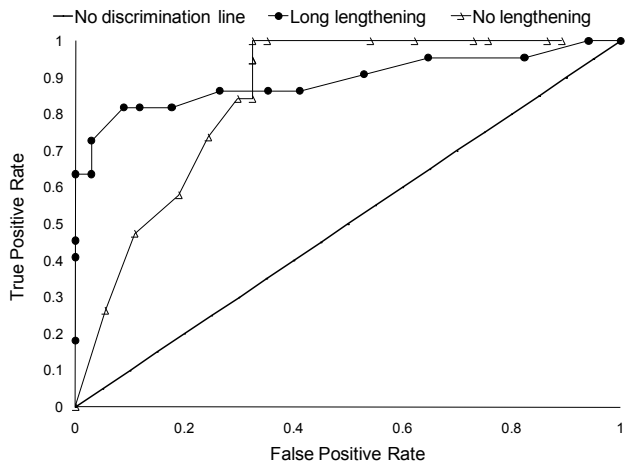


Figure 29. ROC curve for phrase-final lengthening

8.4.3. Semantic completeness

The *semantic completeness* cue represents the lexical content in the dialogues. This cue corresponds to what is often referred to as lexico-syntactic, lexical or syntactic completion points in a dialogue. The manual annotation of semantic completeness was performed as follows: Transcriptions of IPUs were presented incrementally to the annotators, and for each segment, they were asked to label whether the current IPU "was a complete response to the previous turn". The two annotators were provided with the previous lexical dialogue context, but the tool used for annotation only displayed the dialogue up to the target IPU. After each judgment, the dialogue segment up to the next target IPU was provided incrementally. The annotators had access only to the orthographic transcriptions of the dialogues and did not listen to the recordings. Non-lexical elements such as lip-smacks and breathing were removed from the transcripts, since they are considered to represent acoustic information, information that is already represented in other cues. Inter-annotator agreement for semantic completeness was high (Kappa 0.73).

8.4.4. Cue phrases

An observation in the previous chapter was that many cue phrases are closely associated with the initiation, continuation or termination of talkspurts.The five turn-taking categories explored as turn-taking cues were the three connectives, ADDITIVE CONNECTIVES, CONTRASTIVE CONNECTIVES and ALTERNATIVE CONNECTIVES (for example "and", "but", and "or" respectively). The connectives were considered to have turn-holding functions. The fourth category was FILLERS, which were also considered to have turn-holding functions. The fifth cue phase category investigated was RESPONSE ELICITING, i.e. lexical expressions used to elicit information from listener(s) (for example "eller hur?" Eng:"right?"). RESPONSE ELICITING cue phrases were typically placed at the end of turns and considered to have turn-yielding functions.

8.4.5. Speech production phenomena

There are a number of speech production behaviours that are not typically associated with content of speech, but which are highly correlated with either the termination or continuation of a talkspurt. Examples of such behaviours are breathing and lip-smacks. The speaker may not be aware of these behaviours, but they may help listeners identify appropriate places to speak. These behaviours were also explored as potential turn-taking cues. Exhalations are associated with the completion of talkspurts and therefore hypothesized to have turn-yielding effects. Inhalations and lip-smacks were considered to indicate an intention to continue speaking and therefore hypothesized as turn-holding. Annotation of these phenomena was already available in the original transcriptions of the DEAL corpus.

Another set of behaviours explored as potential as turn-taking cues in the present study were lexical repetitions and interruptions. The DEAL transcriptions included annotations of repeated words and phrases. Such repetitions are often considered as signs of difficulties to plan or produce upcoming utterances (Shriberg, 1994). This makes them potential turn-holding cues.

The transcriptions also included annotation of speaker interruptions; these were annotations of abrupt stops in the middle of the speech flow. According to Levelt's main interruption rule, speakers stop the flow of speech immediately when a problem is detected (Levelt, 1989). Hence, speaker interruptions suggest that the speaker has detected a problem in previous speech segment and that this is about to be altered. This makes them potential turn-holding cues.

8.4.6. Stimuli selection

The stimuli used in the experiment were dialogue segments played to the subjects in chronological order. Just subsequent to a particular set of IPUs, the dialogue playback stopped and the subjects guessed who was the next speaker. The target IPUs were selected to get a fair distribution of IPUs prior to pauses and gaps and a variety of different turn-taking cue categories. However, it was difficult to find segments

in the data that fulfilled all requirements, and a perfect weighted range was impossible to obtain because some combinations did not occur in the data. In the end, 125 IPUs were selected as stimuli. The number of cues over IPUs is presented in Table 10.

Turn-holding cues	Turn-yielding cues			
	0	**1**	**2**	**3**
0	6	21	13	3
1	24	11	3	
2	30	7		
3	6			
4	1			

Table 10. Number of turn-yielding and turn-holding cues over stimuli IPUs

8.4.7. Re-synthesis of dialogues

One motivation of this work was to investigate whether cues could be reproduced in a synthetic voice and perceived as having similar functions. In order to create the synthesized stimuli, a corresponding reproduction of the male party in the dialogues was created by replacing his voice with a diphone synthesis. This was done using Expros, a tool for experimentation with prosody in diphone voices (Gustafson & Edlund, 2008). Expros automatically extracts fundamental frequency and intensity from the human voice and creates a synthetic version using these parameters. Some manual alterations were made to the phonetic transcriptions in order to correct mispronunciations. Since breathing and lip-smacks could not be re-synthesized, the original human realizations were kept and concatenated with the synthetic voice using the manually verified timings. The synthetic version, thus, has timings, intonation, intensity, as well as concatenated lip-smacks and breathing that correspond to the original recordings.

8.5. Experimental setup

The experiment included four dialogue segments from four different dialogues. The segments were between 116 to 166 seconds long. The dialogues were two party dialogues with three different speakers, one male and two female. The male speaker (S1) participated in all four dialogues and the two female speakers (S2 and S3) in two dialogues each. In the experiment, the recording stops playing just subsequent to a target IPU, allowing the subjects to make a judgement. Each subject listened to two human-human dialogues and two dialogues where one party was replaced with the diphone synthesis. The re-synthesized dialogues differed between subjects. Stimuli presented with a synthesis to one subject were presented with human voice to another subject and vice versa.

The experimental setup was designed as a game where the subject received points based on whether they could guess who would be the next speaker. Two movie tickets were awarded to the "best" player. The GUI of the test (see Figure 30) included two buttons with "pacmans" and a button allowing the subjects to pause the test. The speakers in the dialogues were recorded on different channels and the movements of the face with the left position on the screen corresponded to the sound in the subject's left ear, and vice versa. The pacman buttons represented the speakers in the dialogues, and when the corresponding interlocutor spoke, the pacman opened and closed its mouth repeatedly. The subjects' task was to listen to the dialogues and guess who the next speaker would be by pressing a corresponding button. To make the subjects aware that the playback had stopped, the faces changed colour. Each time the playback stopped, the mouse pointer was reset to its original position, in the middle of the pause button, and the subjects had to move the mouse pointer from the pause button to one of the pacman buttons in order to make their judgement. This was done to control the conditions before each judgment, enabling comparisons of reaction times.

To elicit judgements based on intuition rather than afterthought, speed was rewarded. The faster subjects responded, the fewer minus points they incurred when they were wrong and the more bonus points they received if they were right. Whether they made the right choice or not was actually unimportant, but it was used as an objective rewarding system to motivate the user to respond immediately and make the experiment more fun. Whether the subject was right or wrong was based on which interlocutor vocalized first.

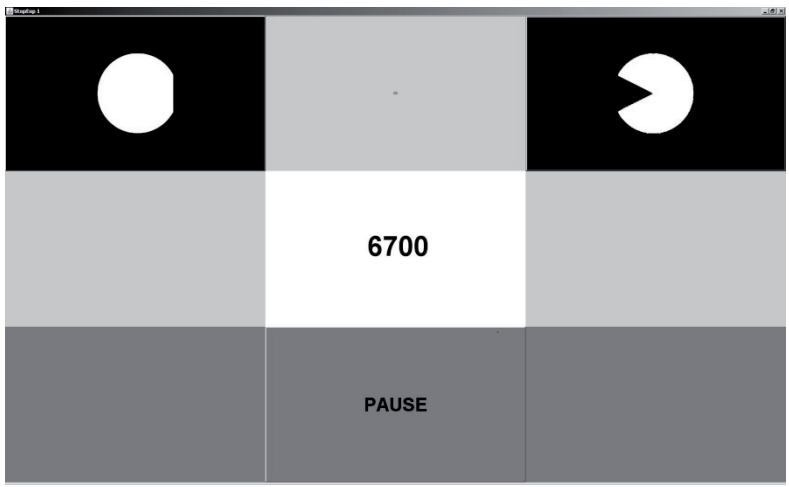


Figure 30. Experiment GUI

8.5.1. Pilot experiment

A pilot experiment was conducted to test the experimental setup and features of the GUI. The pilot experiment included 10 subjects, 5 male and 5 female, between the ages of 31 and 58. Based on the results from this experiment and comments from the subjects, a few changes were made to the experimental design before the final experiment. Training effects were controlled by changing the order of the dialogues. There was also a 210 second long training session to allow the subjects to become familiar with the task.

8.5.2. Experiment

The experiment included 16 subjects, 9 male and 7 female, between the ages of 27 and 49. All were native Swedish speakers except for two who had been in Sweden for more than 20 years. Five of the subjects were working at the department of Speech Music and Hearing, but the majority had no experience in speech processing or speech technology.

8.6. Results

This section analyzes the effects of both individual and combined sets of turn-taking cues. Initially, we present results on the individual cues. Our motive is to investigate whether the annotated behaviours affect the subjects' judgements as hypothesised.

8.6.1. The effect of individual turn-management cues

To explore the effect of individual turn-taking cues, namely, whether a turn-holding cue increased the expectations of a hold, and turn-yielding cues increased the expectations of change, the judgements for all stimuli with a particular cue were compared to the overall distribution of change and hold. The cues investigated were all the cues presented in Table 9. Intonation contour and speaker rate were based on automatic extractions of these features as described in Section 8.4.2.1. The thresholds were extracted from the ROC-curves (see Figure 28 and Figure 29) where minimizing the false positive rate (FPR) was prioritized over high true positive rate (TPR) to get discrete categories.

Figure 31 presents the percentage of judgements for a speaker change versus hold over the different cue categories. Increased phrase final lengthening is listed separately since we did not have any clear hypothesis about the effect of this cue.

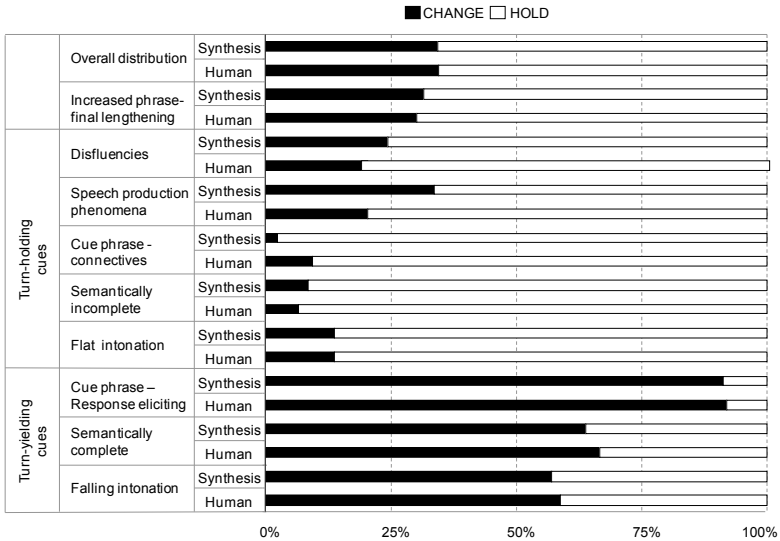


Figure 31. % judgments for change and hold over the different cue categories. Results include both synthetic and natural voice (*difference in distribution between groups is NOT significant)

Chi-square tests of independence were employed to investigate whether the judgement distribution between change and hold for all stimuli containing a particular cue type differed from judgement distribution of change and hold when this cue was absent. The results from these tests are presented in the top row of Table 11. The distribution of change and hold for all cues except phrase-final lengthening differs significantly (each cue was tested individually) from the overall distribution of change and hold for ($p<.05$, by chi-square test of independence with 2 x 2 contingency tables). These results were also checked for the direction, that is, whether turn-holding cues resulted in a higher number of judgments for hold than the overall distribution and vice versa. The results support the conclusion that the turn-takings cues were perceived as hypothesised.
In order to examine the potentials of realizing turn-taking cues with a synthetic voice, Chi-square tests of independence were also calcu-

lated comparing the judgement distributions of when a particular cue was present and not for the human and synthetic voice independently (Table 11 row 2 and 3). These results show ($p<.05$) that when split over natural and synthetic voices, the same results hold. Chi-square tests of independence were also employed to explore the impact of the different cues types on each individual subject. By doing this, any bias for a particular outcome in the subject's overall judgement distribution is considered. The number of subjects for whom the distribution of change and hold differ when a particular cue is present is presented in Table 11 row 4. It should be noted that some of the cues were less frequent than others. The total number of stimuli that contain a particular cue is presented in parenthesis after the cue category label (Table 11).

Previous literature frequently mentions phrase-final lengthening as a turn-taking cue (c.f. Gravano, 2009, Local et al., 1986 and Ferrer, 2003). Since we did not find any turn-taking effects of phrase-final lengthening in the experiment, this phenomenon was explored in more detail by analysing the turn-taking decisions made by the original speakers in the dialogues.

First, speaker rate was calculated as syllables per second and computed for each of the five last syllables of the IPU. Speaker rate was also calculated as vowels per second (z-normalized over speaker and dialogue) and computed for each of the five last vowels. Speaking rate for the last two vowels and syllables are displayed in Figure 32. First, an independent-samples t-test was conducted to explore overall phrase-final lengthening. Hence, regardless of whether the IPU was followed by a speaker change or not, there was a significant difference between the last ($M=0.67$, $SD=0.85$) and the last but one vowel ($M=0.12$, $SD=0.71$); $t(1440)=13.35$, $p=0.00$. However, there were no significant differences in phrase-final lengthening between IPUs prior to pauses and IPUs prior to gaps. Neither was there any significant difference in speaker rate between IPUs prior to pauses and IPUs prior to gaps over the preceding four syllables (compared pair-wise from the end).

		Turn-yielding cues			Turn-holding cues					Phrase-final lengthening
		Falling intonation (27)	Semantically complete (49)	Cue phrase response-eliciting (6)	Flat Intonation (40)	Semantically incomplete (49)	Cue phrase connectives (22)	Disfluencies (8)	Non-lexical speech phenomena (18)	Long phrase-final lengthening (23)
All data df=1, N=1993	X^2	133.0	539.6	137.6	173.1	407.1	6.38	6.58	11.0	0.19
	p	.00	.00	.00	.00	.00	.01	.01	.00	.98
Human voice df=1, N=1421	X^2	75.8	306.8	81.2	92.14	238.1	75.9	5.8	18.6	.0
	p	.00	.00	.00	.00	.00	.00	.02	.00	.16
Synthesis df=1, N=572	X^2	58.2	233.1	56.6	81.3	170.0	62.9	-	-	0.8
	p	.00	.00	.00	.00	.00	.00	-	-	.37
n subjects with difference in judgment distribution p<.05		12/16	16/16	15/16	16/16	16/16	0/16	0/16	0/16	0/16

Table 11. Differences between the judgment distribution (change and hold) per cue compared to the overall judgment distribution (Chi-square test of independence. Some comparisons could not be made because these configurations did not contain enough data points (cells with a frequency less than 5).

In order to further investigate phrase-final lengthening, lexical stress were derived from the transcriptions and an independent-samples t-test was conducted between IPUs prior to a pause and IPUs prior to a gap for lexically stressed and unstressed syllables separately. Still, no differences in phrase-final lengthening between talkspurt-final and talkspurt-medial IPUs were found.

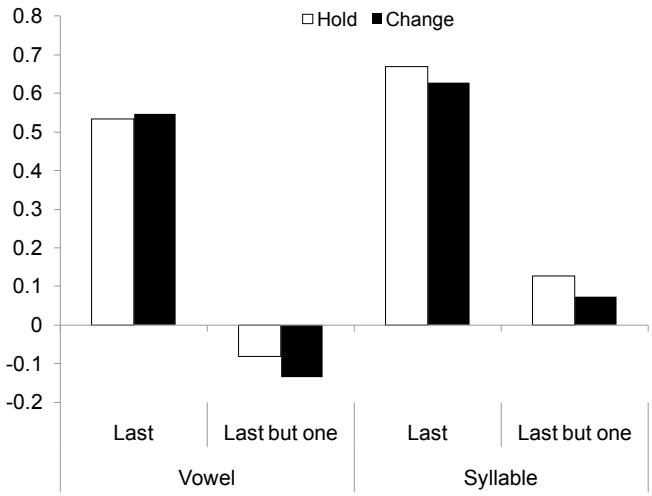


Figure 32. Vowels and syllables per second z-normalized over speaker and dialogue for the two last vowels and syllables of the IPU for HOLD versus CHANGE.

8.6.2. The additive effect of turn-management cues

This section presents results from analyzing the combined effect of the turn-taking cues. All turn-taking cues in Table 9 except phrase-final lengthening were explored. Phrase-final lengthening was excluded since the judgement distribution for this cue did not differ significantly from the overall judgement. For simplicity, all cues were given equal weight (1) and the relative contribution of the different cues was not considered.

8.6.3. Reaction times and judgement agreement

Because of properties of durational data (see discussion 8.4.2.1) the reaction times were transformed into a logarithmic scale (base 10). The average reaction times differed considerably between subjects (from 933 ms to 1510 ms) and were therefore z-normalized over each subject. A one-way ANOVA with % judgement agreement (75%, 85%, 95% and 100%) as a between-subject factor was used to test for differences in reaction times over judgement agreement. Stimuli with high agreement, regardless of the number of cues, were judged significantly faster than stimuli with low agreement, $F(3, 1989) = 34.55$, $p = .00$. The average reaction time for stimuli with 75%, 85%, 95%, and 100% judgement agreement are presented in Figure 33. For completeness, each point is labelled with its average $\log^{10}$ value (un-normalized) in milliseconds. All differences, except between 75% and 85% agreement, are significant (Tukey's test, $p<.05$, see Table 12). Analyses were done with four outliers, the two longest and the two shortest reaction times, excluded.

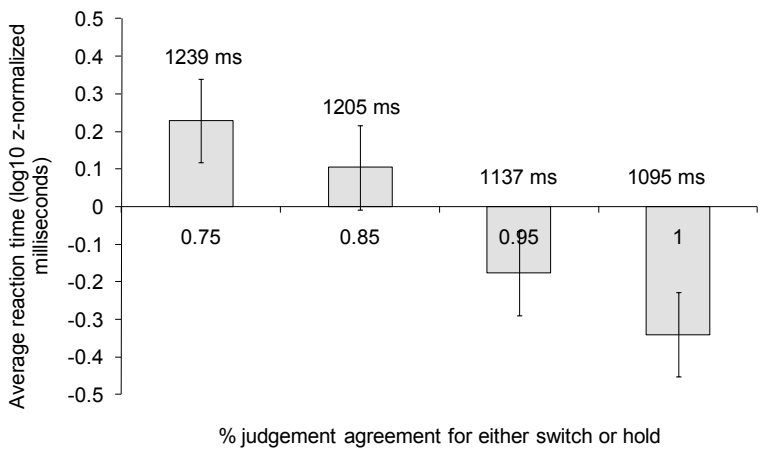


Figure 33. Average reaction time log10 z-normalized milliseconds over IPUs with % agreement. Error bars represents the standard error.

Judgement agreement		Difference in mean response time i – j		Standard error	p-value
i	j	$\log^{10}$ z-value	$\log^{10}$ in ms		
75%	85%	0.112	31	0.06	0.285
	95%	0.398	100	0.07	0.000
	100%	0.561	142	0.06	0.000
85%	95%	0.286	70	0.06	0.000
	100%	0.449	112	0.06	0.000
95%	100%	0.163	42	0.05	0.046

Table 12. Differences in average response time between 75%-85%, 75%-95%, 75%-100%, 85%-95%, 85%-100% and 95%-100% judgement agreement (Tukey's p<.05, df=3). Significant differences in bold.

To study the additive effect of the turn-management cues, the distribution of judgements for change and hold was compared over stimuli with different numbers of cues. Thus, stimuli with one turn-holding cue were compared to stimuli with two turn-holding cues and so on. The results of these comparisons are presented using a bubble chart (see Figure 34). Some cue combinations were rare (Table 10) and since small variances in the data will affect the results for these cues, cue combinations represented in fewer than five IPUs were excluded. The bubble chart is used to enable comparisons of all cue combinations, that is, including stimuli annotated to occupy both turn-holding and turn-yielding cues. The number of turn-yielding cues is displayed on the x-axis and turn-holding cues on the y-axis. The diameters in the bubble charts represent the percentage of judgments for change versus hold. Each bubble is labelled with the percentage values for change and hold.

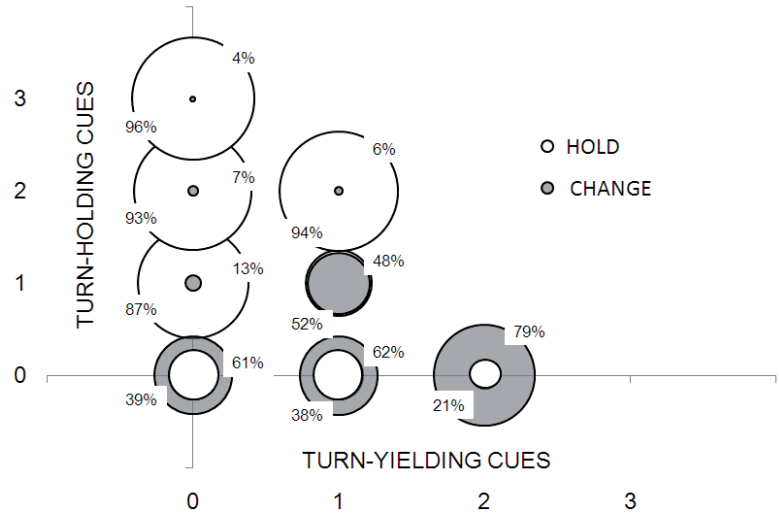


Figure 34. The distribution of judgments for change versus hold. Each bottom left bubble is labelled with the %hold and each top right bubble is labelled with %change.

Chi-square tests of independence were employed to explore the impact of the different number of cues on judgement distribution between change and hold. Thus, the distribution of change and hold was compared between 1 and 2 cues, 1 and 3 cues and so on (see Table 13). Turn-holding cues and turn-yielding cues were compared separately. For the overall data set ("All"), all steps differ significantly except between 2-3 turn-holding cues (Chi-square test of independence $p<.05$). The impact of different number of turn-management cues was also compared over the different speakers and for the synthesis separately. There is a significant relationship between the number of turn-management cues and the judgement distribution over all speakers as well as for the synthetic voice (Chi-square test of independence $p<.05$).

Chi-square comparison				Speaker				
				S1	S2	S3	Synthesis	All
Turn holding cues	0	1	X^2	90.73	53.19	29.82	38.62	235.66
			N	511	368	256	311	1297
			P	.00	.00	.00	.00	.00
	1	2	X^2	18.94	22.87	5.26	14.58	81.87
			N	667	197	95	384	1196
			P	.03	.00	.03	.00	.00
	2	3	X^2	-	-	-	-	235.66
			N	-	-	-	-	687
			P	-	-	-	-	.26
Turn yielding cues	0	1	X^2	114.72	96.68	51.59	46.19	238.29
			N	203	400	217	419	1689
			P	.00	.00	.00	.00	.00
	1	2	X^2	55.72	44.51	41.31	20.56	57.92
			N	350	272	193	196	881
			P	.00	.00	.00	.00	.00
	2	3	X^2	-	14.25	-	-	11.77
			N	-	64	-	-	304
			P	-	.00	-	-	.00

Table 13. Differences in the distribution of change and hold judgments between 1-2,1-3,1-4, 2-3,4-4 and 3-4 turn management cues. Turn-holding and turn-yielding cues were compared separately. Significant differences in bold (Tukey's $p<.05$, $df=1$). Some comparisons could not be made because these configurations did not contain enough data points (cells with a frequency fewer 5 than 5).

The reaction times for IPUs with different number of cues were analyzed using one-way ANOVA with the number of turn-management cues, regardless whether turn-holding or turn-yielding, as a factor (the statistics are calculated on IPUs without contradictory cues). There was a significant effect of number of cues on reaction times; $F(4, 1988) = 4.01$, $p=.00$. Post hoc comparisons using the Tukey HSD test were used to explore these differences in more detail. These results are presented in Table 14. Although not all steps differ significantly, there is a strong trend: the more turn-management cues, the faster reaction time. An independent-samples t-test was conducted to explore differences in reaction time between IPUs with a majority of turn-holding cues from IPUs with a majority of turn-yielding cues. IPUs with a majority of turn-holding cues (M=-0.23, SD=0.95) were judged significantly faster than IPUs with a majority of turn-yielding cues (M=0.03, SD=1.03); $t(1229)=-4.3$, $p=0.00$.

Turn-management cues		Difference in mean response time i – j		Standard error	p-value
i	j	$\log^{10}$ z-value	$\log^{10}$ in ms		
1	2	0.372	93	0.11	0.00
	3	0.440	96	0.11	0.00
	4	0.695	117	0.13	0.00
2	3	0.067	3	0.07	0.843
	4	0.323	23	0.10	0.00
3	4	0.255	23	0.10	0.058

Table 14. Differences in average response time between 1-2, 1-3, 1-4, 2-3, 2-4 and 3-4 turn-management cues (Tukey's $p<.05$, $df=3$). Significant differences in bold.

8.6.4. Differences between synthetic and human voice

An independent-samples t-test was conducted to explore differences in reaction time for human and synthetic voice. For this comparison, only stimuli based on the male speaker (S1) were included. Stimuli based on the other speakers were excluded since their voices did not have a corresponding synthesized version. No significant differences in reaction times between synthetic and human voice were found.

8.6.5. Differences between speakers

To explore the reaction times for the different speakers, a one-way ANOVA was conducted with speaker as factor. The results show that there is a significant effect of speaker, $F(2, 1452)=16.06$, $p=.00$. Post hoc comparisons using the Tukey's HSD test indicated that the mean reaction time for speaker S2 ($M=0.14$, $SD=1.04$) and S3 ($M=0.03$, $SD=1.02$) differed significantly from speaker S1 ($M=-0.19$, $SD=0.99$), $p=.00$ for S1*S2 and $p=.00$ for S1*S3. However, no differences were found between speaker S2 and S3. To explore if this difference was an effect of differences in cue frequency over IPU and speaker, a Kruskal-Wallis test was conducted for average number of cues per IPU over speakers, but no significant differences were found. However, the differences in judgement distribution between change and hold for different number of cues over the different speakers suggest that there was an additive effect of the turn-taking cues regardless of speaker (see Table 13).

8.7. Discussion

Duncan (1972) has previously shown that a number of behaviours affect turn taking in dialogue. If used in combination, the number of turn-taking cues was linearly correlated with listeners' turn-taking attempts. The present study further explores these findings by examining the effect of such turn-taking cues experimentally. The objective of the present study was to investigate the possibilities of gener-

ating turn-taking cues with a synthetic voice. In order to explore this objective, the experiment included dialogues realized with a human voice as well as dialogue where one of the speakers was replaced with a synthesis. Analyses of the reaction times show that stimuli with high judgement agreement, regardless of the number of turn-taking cues, were judged significantly faster than stimuli with low agreement. The judgment agreement and reaction times were further used as measures to analyze the effects of the different turn-taking cues.

First, the effect of individual turn-taking cues was explored. For each cue, the judgement distribution between change and hold was analyzed. The results show that all except one of the turn-taking cues explored in the present study affected the judgements as hypothesized. The exception was phrase-final lengthening which did not have a significant effect on the listener's judgements. The judgement distribution for different cues further suggests that some cues had a major impact, affecting a large majority of the judgements, whereas some cues were less influential. These differences between cues suggest that some cues are more central than others are and that the additive effect of turn-taking cues is not necessarily linear.

8.7.1. Implications for dialogue systems

The aim of the present study is to explore the potentials of using turn-taking cues in spoken dialogue systems. Primarily, dialogue system designers should consider cues that affect a majority of judgements and users accordingly. Such cues include semantic completeness, turn-yielding cue phrases and a falling and flat intonation. It should be noted that some cues are more straightforward to employ in dialogue systems than others are. For example, cue phrases and falling and flat intonation are all discrete behaviours that can be produced locally just prior to a pause or turn-ending, without the need for syntactic or semantic analyses. Semantic completeness is an influential cue, but in order to employ semantic completeness as a cue, the system needs keep track of whether a dialogue segment is complete or not. If the state differs from the system's continued plan of

generation, it needs to generate a phrase that changes semantic completeness in a way that is consistent with the current dialogue context. This is a fairly complex task.

The present study further explores the additive effect of turn-taking cues. The results show that the more cues with the same pragmatic function, the faster the reaction time and the higher the agreement on the expected outcome. Thus, as hypothesized, the higher the number of turn-yielding cues, the higher the expectations of a turn-change and the higher number of turn-holding cues, the higher the expectations of a speaker continuation. This is in line with Duncan's findings.

The objective of the present study was to identify turn-taking strategies that can be produced with a synthetic voice in order to communicate appropriate places for dialogue system users to take the turn. The results show that turn-taking cues presented with a synthesis have a similar effect as cues presented with a human voice. As for cues presented with a human voice, an increased number of simultaneous turn-holding cues increased the expectations of a hold, and an increased number of turn-yielding cues increased the expectations of a change. No differences in reaction times were found between the two conditions. Furthermore, analyses of the judgement distribution indicate that the effects of the individual cues as well as the additive effect of the cues are very similar for the synthetic and the human voice (see Table 13).

8.7.2. Differences between speakers

The experiment was designed to allow the subjects to follow the dialogues in chronological order and get familiar with the speakers and the dialogues in a way that is similar to how dialogue is perceived in a real conversation. However, this restricted the number of dialogues and speakers used as stimuli. The analyses of the reaction times suggest that one speaker was judged more easily than the other speakers were. A possible explanation is that speaker S1 occurred more frequently and the subjects got familiar with this speaker's particular

turn-taking strategies. Still, the additive effect of the cues was similar for all speakers. Differences between speakers and how speakers adjust their turn-taking strategies to their dialogue partners are interesting areas for future research.

8.7.3. Turn-holding vs. turn-yielding cues

Finally, turn-holding cues were judged significantly faster than turn-yielding cues, and the judgement distribution shows that 87% of the listeners expected stimuli annotated with one turn-holding cue to be followed by a hold, whereas only 62% of the listeners expected stimuli annotated with one turn-yielding to be followed by a change. This may not be surprising since the overall distribution between internal pauses and silences between speakers (gaps) were 85% hold and 15% change in the DEAL corpus. Hence, the likelihood that the current speaker will continue is much higher than the likelihood of a speaker change. However, it is likely that the outcome of turn-holding cues also is more predictable than turn-yielding cues for other reasons. While turn-holding cues indicate a continued plan of interaction, turn-yielding cues can simply indicate completion and signal that the floor is open for anyone to speak. It is therefore possible that many of the internal pauses in the original dialogues in fact were transition relevant places but that the interlocutor did not take the opportunity to speak.

For stimuli with contradictory cues, i.e. stimuli with both turn-yielding and turn-holding cues, the judgements were almost equally distributed between hold and change.

8.8. Summary

This chapter has presented a perception experiment where subjects listened to dyadic dialogues and judged whether a speech segment was going to be followed by a speaker change or not. The experiment included both stimuli realized with a human voice and stimuli where one of the speakers was replaced with a synthesis. In line with

Duncan (1972), it was shown that more turn-taking cues with a particular pragmatic function – turn-yielding or turn-holding – the faster the reaction time to make the judgement and the higher agreement among subjects on the expected outcome. Furthermore, the synthesis affects listeners' expectations of a turn change in a way that is similar to a human voice.

9. A user experiment with an incremental version of DEAL

The dialogue behaviours analyzed in this thesis – cue phrases, hesitations, and turn-taking cues – have been suggested as useful behaviours to model in humanlike spoken dialogue systems. These more or less intentional behaviours have been proposed as devices to initiate new talkspurts, buy more time for cognitive processing (hesitate) and indicate to the users when it is an appropriate time to speak. It has further been proposed that a system whose behaviour is similar to a human dialogue partner can make the interaction with such systems more intuitive.

This chapter presents a study that explores some of these behaviours online in a dialogue system setting using a Wizard-of-Oz setup.

9.1. Introduction

The aim of the present study is to explore if different types of cue phrases can be used to incrementally initiate talkspurts in order to provide the users with a fast response. As discussed in 4.7.1, a bottleneck that often delays the system's response is the silence threshold used for ends of turn detection. However, even if dialogue systems used sophisticated techniques to detect ends of turns, the system may need to perform some time-consuming process, and is therefore not able to produce an immediate response. For example, it is typically assumed that automatic speech recognition (ASR) needs to process speech in real-time in order to produce fast responses. However,

given additional time, the ASR component can perform more deep level processing in order to produce results that are more robust. Another example is speech interfaces that before giving a complete response need to extract information from some external resource such as a database or the internet. If the search is time-consuming, devices that can be used to buy extra time for processing are valuable.

The scenario explored in the present study is a Wizard-of-Oz setting. A common problem in WoZ user studies is the time it takes the Wizard to process the user's incoming utterances, and therefore for the system to respond. Therefore, this is an interesting test case for the incremental response generation model presented here.

The approach taken to reduce system response times in this chapter is to initiate talkspurts using different types of cue phrases. These expressions are initiated incrementally as soon as the end of the previous user's talkspurt has been detected, before the processing of input has been completed. Except for providing the user with a rapid response, these cue phrases provide the listener with important pragmatic information, e.g. acknowledging the user's input, and at the same time indicating that the system claims the floor and is about to generate a complete response.

9.2. Experimental method

9.2.1. Experimental setting

An incremental version of DEAL was implemented in Jindigo. Jindigo is a Java-based open source framework for implementing and experimenting with incremental dialogue systems developed by Gabriel Skantze at the Department of Speech, Music and Hearing at KTH. Jindigo based on the abstract framework of incremental speech processing in dialogue system presented by Schlangen & Skantze (2009) (for details see 4.6.1). The modules in the Jindigo framework process sequences of *Incremental Units* (IUs), chunks of "information" that trigger connected modules into action. Each

module has a left buffer and a right buffer. The left buffer of each module receives new IUs, processes it, and forwards it to the right buffer. From the right buffer the IU is passed on to the next modules left buffer and so forth.

9.2.1.1. System architecture

The system architecture of the Jindigo implementation of DEAL is presented in Figure 35[5].

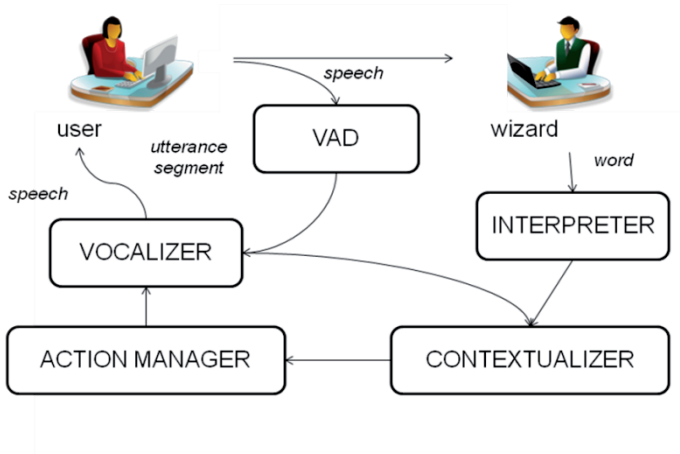


Figure 35. The system architecture used in the Wizard-of-oz experiment.

Instead of using ASR, the users' speech is transcribed on-line by a Wizard. A Voice Activity Detector (VAD) is used to detect when the user turns silent. As soon as the VAD has detected that the user has stopped speaking, the system initiates a response based on what the Wizard has transcribed so far. The Interpreter tries to find an optimal sequence of top phrases and their semantic representations (similar to Pickering, see Skantze & Edlund, 2004). The Contextualizer packages these phrases into communicative acts (CAs) and maintains a list of CAs that serves as a discourse model similar to Galatea (see

[5] The incremental version of DEAL was implemented by Gabriel Skantze

Skantze, 2008). Based on the current dialogue context, as represented by the Contextualizer, the action manager generates a SpeechPlan, which is passed on to the vocalizer, which transforms the system's response into speech. Since there were no mature models for the Interpreter, the Wizard was allowed to adapt the transcription to match the models, while preserving the semantic content.

For comparison, a non-incremental version of the same system was configured. In this version, the user's utterances were not forwarded to the rest of the system before the Wizard had completed the transcription and committed by pressing the return key.

9.2.1.2. Incremental units

The DEAL system has no sophisticated method to detect ends of turns. Instead, a VAD was used to segment the user's speech into UtteranceUnits using a silence threshold of 500 ms. The UtteranceUnits are segmented into smaller IUs, which are forwarded incrementally through the system. The IUs are segmented by a short silence threshold of 50ms. The SpeechPlan generated by the system is also segmented into smaller incremental units, so-called SpeechSegments. The SpeechSegments correspond to a pre-synthesized audio file, or a word or phrase synthesized on-line by the system.

An important aspect in DEAL is to produce a varied and flexible output. To relieve the action manager from this burden, the vocalizer keeps track of which audio files and lexical expressions that the system has recently used and tries to produce a varied output. Furthermore, a speech segment can be *optional*, indicating that it can be skipped if the rest of the SpeechPlan is already ready to be produced by the vocalizer. Optional elements include fillers and other stalling phrases such as "wait a minute".

9.2.1.3. Revisions

As soon as the VAD detects a silence, the system initiates a response using a cue phrase (see description below) based on what the Wizard has transcribed so far. For example, if the Wizard has transcribed the beginning of a price request (e.g. “vad kostar...” Eng: “how much...”), the system initiates a price presentation (e.g. “den kostar...” Eng: “it costs...”). If the user’s CA cannot be predicted, a less specific talkspurt initiation is used (e.g. “eh”, “vänta lite”, Eng: “eh”, “just a minute”). Occasionally a talkspurt initiation is based on a premature hypothesis. For example, if the user change his mind as in “How much is... sorry I mean do you have a blue doll?”. To deal with this, the system needs to handle *revisions*. In Jindigo, the modules react to three different situations, namely: IUs are *added* to a buffer, which triggers processing, IUs are *revoked* when an earlier module has made a premature hypothesis, which may trigger a revisions of the module’s own output, and finally, modules *commit* to an IU, which means that this IU cannot be revoked any longer. The Vocalizer keeps track of how much of the SpeechPlan has been realized and when a user CA is revoked, the system generates a new SpeechPlan and compares this plan to the part of the previous plan that has already been realized. If these differ, a self-repair is generated, a so-called overt repair. If not, the concerned IUs are altered before being articulated, a so-called covert repair (Levelt, 1983). An illustration of covert and overt repairs is presented in Figure 36.

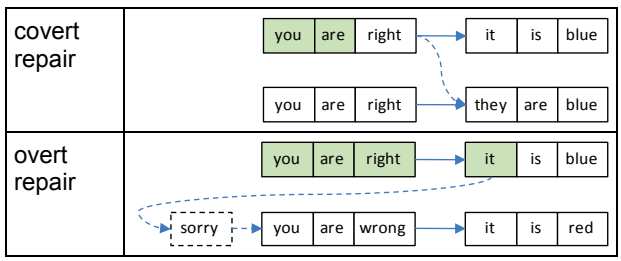


Figure 36. Different types of repairs. The shaded boxes show which units have been realised, or are about to be realised, at the point of revision.

SpeechSegments have the property *committing*, which indicate whether a SpeechSegment needs to be repaired or not. For example, fillers are non-committing and therefore do not need to be repaired.

9.2.1.4. Talkspurt-initial cue phrases

As a basis for the turn-initial cue phrases used in this experiment, the acoustic and lexical realizations of cue phrases in the DEAL corpus were used. The motivation to use speech segments derived from human-human dialogue recordings was to make the system sound human and convincing in terms of both lexical choice and intonation. With a repertoire of different realizations of fillers and responsives such as “ja”, “eh” and “mm”, the system can avoid sounding monotone and repetitive.

The cue phrases extracted from the DEAL corpus included a number of different responsives, such as “ja”, “a”, “mm” (Eng: “yes”), and a number of different fillers such as “eh”, “ehm”. Also, a number of talkspurt-initial phrases that initiate different conversational acts were extracted. These phrases were extracted to initiate system specific CAs such as to present a price, e.g. “it costs..”, “this one is...”, or present an object, e.g. “here is a..”, “how about this...”. Finally, a set of stalling phrases such as “let me see” or “just a minute” were extracted. The transcripts were used to identify the phrases and their corresponding sound files were extracted automatically based on their timings. The sound files were then re-synthesized using Expros (for details see section 8.4.7 or Gustafson & Edlund, 2008). In order to confirm that the re-synthesized phrases did not sound out of place when combined with the rest of the system’s response, the re-synthesised were matched with the different system CAs and manually verified.

During runtime, the pre-synthesized audio files were concatenated with SpeechSegments that were synthesized and generated online (for example references to objects, prices, etc).

9.2.1.5. System domain

The incremental version of DEAL talks about the properties of goods for sale and negotiates about the price. The price can be reduced if the user points out a flaw of an object, argues that something is too expensive, or offers lower bids. However, if the user is too persistent haggling, the agent gets frustrated and closes the shop. Then the user has failed to complete the task. Figure 37 shows the GUI that was shown to the user. The object on the table is the one currently in focus. Example objects are shown on the shelf. Current game score, money and bought objects are shown on the right.

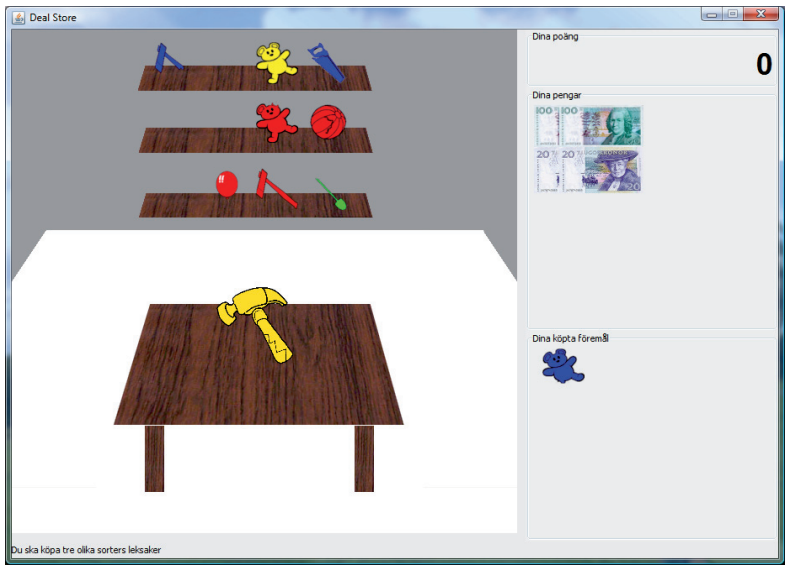


Figure 37. The user interface in DEAL.

An example interaction with the incremental version of the system is shown in Example 6. Utterance S.11 exemplifies a self-correction, where the system prepares to present another bid, but then realizes that the user's bid is too low even to consider.

S.1 [welcome] [how may I help you]
U.2 I want to buy a doll
S.3 [eh] [here is] [a doll]
U.4 how much is it?
S.5 [eh] [it costs] [120 crowns]
U.6 that is too expensive
how much is the teddy bear?
S.7 [eh] [you can have it for] [let's see] [40 crowns]
U.8 I can give you 30 crowns
S.9 [you could have it for] [37 crowns]
U.10 I can give you 10 crowns
S.11 [let's say] [or, I mean] [that is way too little]

Example 6. An example DEAL dialogue (translated from Swedish). Speech segments are marked in brackets.

The Wizard in the experiment starts to type as soon as the user starts to speak and may alter whatever he has typed until the return key is pressed. The word buffer is updated in exactly the same manner as if it had been the output of an ASR.

9.2.2. Experimental design

In order to compare the incremental and non-incremental versions of the system, we conducted an experiment with ten participants, four male and six female. The participants were given a mission: to buy three items (with certain characteristics) in DEAL at the best possible price from the shopkeeper. The participants were further instructed to evaluate two different versions of the system, the incremental version and the non-incremental version. However, they were not informed how the versions differed. The participants were led to believe that they were interacting with a fully working dialogue system and were not aware of the Wizard-of-Oz set up. Each participant interacted with the system four times, first two times with each ver-

sion of the system, after which a questionnaire was completed (a translated version of the questionnaire is presented in Appendix G). Then they interacted with the two versions again, after which they filled out a second questionnaire with the same questions. The order of the versions was balanced between subjects.

The mid-experiment questionnaire was used to collect the participants' first opinions of the two versions and to make them aware of what type of characteristics they should consider when interacting with the system second time. When filling out the second questionnaire, the participants were asked to base their ratings on their overall experience with the two system versions. In the questionnaires, they were requested to rate which one of the two versions was most prominent according to eight different dimensions: which version they *preferred*; which was more *humanlike*, *polite*, *efficient*, and *intelligent*; which gave a *faster response* and *better feedback*; and with which version it was *easier to know when to speak*. All ratings were done on a continuous horizontal line with one system version on each end of the line. The centre of the line was labelled with "no difference".

The participants were recorded during their interaction with the system, and all system messages were logged.

9.3. Results

The average response time for the incremental and non-incremental system version was calculated. Figure 38 presents the difference between the two versions. As expected, the incremental version started to speak more quickly (M=0.58s, SD=1.20) than the non-incremental version (M=2.84s, SD=1.17). Furthermore, the talkspurt-initial cue phrases produced by the incremental version resulted in longer utterances. It was harder to anticipate whether it would take more or less time for the system to complete the talkspurt in the incremental version. Both versions received the final input at the same time. On the one hand, the incremental version initiates utterances with speech segments that contain little or no semantic

information. Thus, if the system is in the middle of such a segment when receiving the complete input from the Wizard, the system needs to complete this segment before producing the rest of the utterance. Moreover, if a response is initiated and the Wizard alters the input, the incremental version needs to make a repair which takes additional time. On the other hand, it may also start to produce speech segments that are semantically relevant (e.g. "it costs..."), based on the incremental input, which allows it to finish the utterance more quickly. As the figure shows, the average response completion time for the incremental version (M=5.02s, SD=1.54) is about 600 ms faster than the average for non-incremental version (M=5.66s, SD=1.50), ($t(704)=5.56$, $p<.0001$).

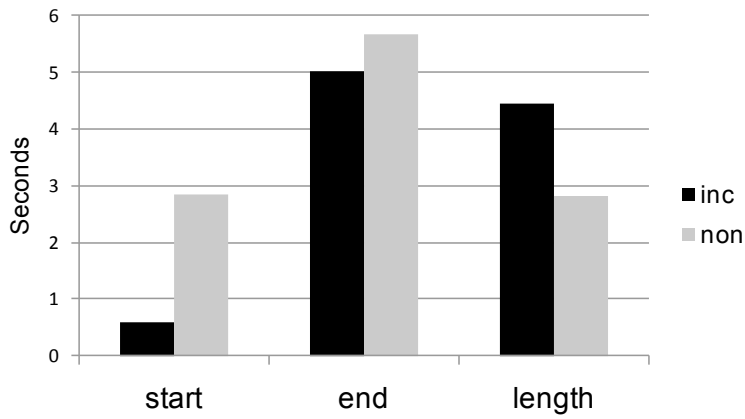


Figure 38. The first two column pairs show the average time from the end of the user's utterance to the start of the system's response, and from the end of the user's utterance to the end of the system's response. The third column pair shows the average total system utterance length (end minus start).

In general, subjects reported that both system versions worked very well. After the first interaction with the two versions, the subjects found it hard to point out any difference, as they were focused on solving the task. However, after the second interaction, most of them had a more clear opinion. The results presented here are based on the second questionnaire.

The marks on the horizontal continuous lines were measured with a ruler based on their distance from the midpoint (labelled with "no difference") and normalized to a scale from -1 to 1, each extreme representing one system version. A Wilcoxon Signed Ranks Test was carried out, using these rankings as differences. The results are shown in Table 15. As the table shows, the two versions differed significantly in three dimensions, all in favour of the incremental version. Hence, the incremental version was rated as more polite, more effective, and better at indicating when to speak.

Dimension	diff	z-value	p-value
preferred	0.23	-1.24	0.214
humanlike	0.15	-0.76	0.445
polite	0.40	-2.19	0.028*
efficient	0.29	-2.08	0.038*
intelligent	0.11	-0.70	0.484
faster response	0.26	-1.66	0.097
feedback	0.08	-0.84	0.400
when to speak	0.35	-2.38	0.017*

Table 15. The results from the second questionnaire. All differences are positive, meaning that they are in favour of the incremental version.

In order to explore whether the users entrained to the different system version, the user utterance length and user response time were analyzed. However, no significant differences between the interactions with the two versions were found. The cumulative user response time presented in Figure 39 suggests that the response time was very long for both versions (a majority is longer than one sec-

ond). It is possible that the complexity of the task affected the users' response time and this might explain why the users did not entrain to the incremental version's rapid response times. An interesting area for future research is to study user response time in a dialogue system context that allows for more rapid turn-taking.

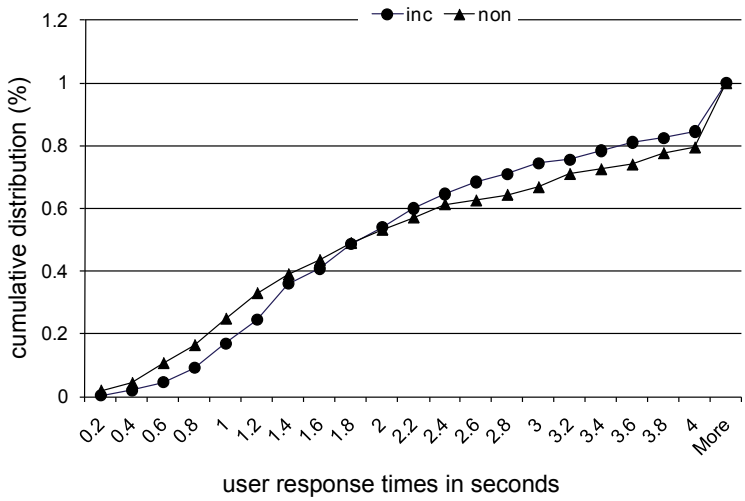


Figure 39. Cumulative distribution of user response time measured from the end of the system's utterance to the start of the user's utterance

9.4. Discussion

There are several ways to improve the model. First, in this version, the only cue phrase categories used were fillers and responsives. However, the system could use additional cue phrase categories including different types of connectives and employ these based on the previous discourse history. Secondly, when the user has finished speaking, it should (in some cases) be possible to anticipate how long it will take until the processing is completed and thereby choose a more optimal path (by taking the length of the SpeechSegments into consideration). Third, a lot of work could be done on the dynamic

generation of SpeechSegments, considering syntactic and pragmatic constraints, although this would require a speech synthesizer that was better at convincingly produce conversational speech.

9.5. Summary

The present study has presented a first step towards incremental speech production in dialogue systems. The results are promising: when there are delays in the processing of the dialogue, it is possible to produce talkspurt-initial cue phrases incrementally in order to make the interaction more efficient and pleasant for the users.

The experiment also shows that it is possible to achieve fast turn-taking and convincing responses in a Wizard-of-Oz setting. This opens up new possibilities for the Wizard-of-Oz paradigm, and thereby for practical development of dialogue systems in general.

10. Conclusions and future work

This thesis has explored human conversational behaviour as a model for spoken dialogue systems. While the long-term and high-reaching objective, to build an artificial dialogue partner, has far from been accomplished, this thesis presents empirical findings that serve as a step in this direction. The work presented has implications for both the design of spoken dialogue systems and for how to approach this kind of research methodologically.

The listening test presented in chapter Part II explored listeners' perceptions of a humanlike spoken dialogue system. This study introduced a method where human-human dialogue data was manipulated in order to simulate two different versions of a spoken dialogue system. The UNCONSTRAINED version was a replica of a human speaker. The CONSTRAINED version was based on the same set of transcriptions as the UNCONSTRAINED version, but transformed to represent a restricted version of human behaviour. A methodological concern in this study was how to explore the effects of small variations in conversational behaviour. Rather than asking the subjects to assess the target behaviours explicitly, their task was to rate the different system versions according to a number of dimensions, some of which are not typically associated with machines.

The results suggest that behaviours such as fillers, revisions and fragmental utterances, are not necessarily perceived negatively in the

context of a dialogue system. The UNCONSTRAINED version was rated as more *humanlike*, *polite* and *intelligent* than the CONSTRAINED version. This was despite the fact that both versions used a machine-like synthetic voice that did not have conversational prosody. Furthermore, the results show that the number of syllables per utterance was not correlated with subjective ratings of efficiency. This finding suggests that turn length does not necessarily affect the users' perception of system efficiency.

There were, however, a number of limitations with this study. The experimental setup did not allow the subjects to interact with the two system versions on-line, and the interactive effects of the two system versions could not be explored. Another limitation was that the effects of the individual behaviours could not be isolated.

One motivation to generate conversational behaviour in dialogue systems is to create a dialogue system that is perceived as more humanlike per se. Yet, one important aim of this thesis is to explore how interactional cues can be employed to cope with irregularities in the system's flow of speech caused by the underlying processes of spoken language generation. The idea is to use human behaviour as conceptual metaphors of dialogue system processes. For example, planning utterances that require high cognitive load for humans is not necessary correlated with increased processing demands for the dialogue system and vice versa. However, if interactional cues can be employed in a way that affect listeners in the same way as if produced by a human speaker, such phenomena can be used to signal irregularities in the system's flow of speech.

The DEAL dialogue system and data collection have played significant roles in this thesis. While there is still much work to do before DEAL behaves like a conversational partner, the system has served as a platform where human conversational behaviour could be modelled and experimented with in a dialogue system context. The DEAL domain was further used as a basis for a data collection. The aim of this data collection effort was to obtain examples of the behaviours of interest when produced in a conversational context. The

corpus was manually annotated with cue phrases with high inter-annotator agreement. Some of these cue phrases were later extracted and used in an incremental version of the DEAL dialogue system. Some acoustic and lexical analyses of cue phrases were also presented. These analyses suggest that talkspurt-initial cue phrases are produced with high vocal intensity. The duration and fundamental frequency of turn-initial feedback expressions were also explored in order to discriminate between three different types of responsives. However, only relatively small differences between these categories could be identified, and it is likely that the interpretations of these expressions rely principally on context.

Chapter 8 presented an experimental study that compared turn-taking cues presented with a human and a synthetic voice. The aim of this study was to identify turn-taking cues that can be employed in dialogue systems in order to help users to identify appropriate places to speak. The results from this study show that turn-taking cues realized with a synthetic voice affect the expectations of a turn change just as in the corresponding human version. Furthermore, the results show that the more turn-taking cues with the same pragmatic function (turn-yielding or turn-holding) the higher the agreement among subjects on the expected outcome. The analysis of the individual turn-taking cues suggests that some cues are more central than others are, and that the additive effect of turn-taking cues is not necessarily linear. Cues with major impact include *semantic completeness*, *turn-yielding cue phrases* and a *falling* and *flat* intonation.

The final study presented in this thesis was a user study exploring the incremental production of turn-initial cue phrases using a Wizard-of-Oz setup. The DEAL version used in this study was an incremental version implemented in Jindigo, a Java-based open-source framework for implementing and experimenting with incremental dialogue systems. This version initiates talkspurts incrementally using feedback expressions, fillers and phrases associated with specific communicative acts in order to provide the user with rapid response. The results show that the incremental version had shorter response

times and was rated as more efficient, more polite and better at indicating when to speak than a non-incremental implementation of the same system.

Much effort in this thesis has been devoted to collecting, identifying and evaluating the effects of the behaviours of interest. This has been a difficult and time consuming undertaking, since there is no one-to-one mapping between specific lexical or prosodic realization and a specific semantic or pragmatic function. Moreover, there is a large variety in prosodic realisations between speakers and even within a single speaker in different contexts.

The turn-taking experiment presented in chapter 8 is an example of how to explore the effects of conversational behaviour online. While the subjects did not participate in the actual dialogues, the experimental setup made it possible to collect naïve subjects' reactions to turn-taking behaviour in terms of reaction times.

The first listening test and the turn-taking experiment both studied off-line manipulations of behaviours extracted from human-human dialogues. The final Wizard-of-Oz setup, however, explored the effects of incremental speech production in an interactive setting.

The motivation of employing human conversational behaviour is neither efficiency nor accuracy and therefore new evaluation criteria are needed, since traditional evaluation measures such as task success and dialogue length are not necessarily sufficient. This thesis has argued that human conversational behaviour can be used to make the interaction with dialogue systems more *intuitive*. A major concern is how to measure intuitiveness. One suggestion proposed in the introductory chapter is to explore humanlikeness in terms of whether the system encourages users to speak in a way that is similar to how they speak to a human partner. This approach has, however, not been thoroughly explored here. An interesting area for future research is to explore the effects of conversational behaviour in terms of dialogue symmetry. Furthermore, in line with the approach suggested by Moore (2007), dialogue systems somehow need to verify online that the interactional cues that they produce are perceived accordingly.

Cue phrases and turn-taking cues are not dichotomous variables, and the system can never be certain of how its contributions will be perceived. Thus, a more appropriate model for a conversational dialogue system is a perceptual control system with a perceptual feedback loop that controls and compares the outcome to the system's original intention. This also addresses another important issue that has not been focussed on in this thesis, namely, dialogue entrainment.

Users are well aware of the difficulties that pauses, fillers and mispronunciations can cause when speaking to a dialogue system. In order to avoid speech recognition errors and misunderstandings, speakers appear to plan their utterances ahead to avoid such behaviour. A design principle is to generate only entities that the system can understand. As follows, dialogue systems that produce hesitations and revisions are likely to receive similar behaviour from their users. Many of the results presented in this thesis also have implications for the understanding and identification of these phenomena. However, in order to deal with the increased complexity that comes with these kinds of behaviours, it is important to carefully consider and what kinds of user expectations come with increased humanlikeness.

Finally, to build a conversational partner is a far-fetched and high reaching goal and the work presented in this thesis is only one step in this direction. Anthropomorphic interfaces are hardly without controversy, but humanlike conversational interfaces open up for new and interesting research challenges. Rather than trying to build dialogue systems that are prompt and accurate, researchers and dialogue system designers in this area will be encouraged to explore and profit from the features that are so exceptional to spoken language. Spoken conversation will be used as means of communication in new and inventive domains, and last but not least, this research will increase our understanding of the cognitive processes responsible for human speech processing.

References

Adell, J., Bonafonte, A., & Escudero, D. (2007). Filled pauses in speech synthesis: towards conversational speech. In *Proceedings of 10th International Conference on Text, Speech and Dialogue (LNAI 07)* (pp. 358–365).

Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., & Stent, A. (2001). Towards conversational human-computer interaction. *AI Magazine, 22*, 27-37.

Allwood, J., & Haglund, B. (1992). *Communicative Activity Analysis of a Wizard of Oz Experiment*. Technical Report, Göteborg University, Gothenburg, Sweden.

Appelt, D. E. (1985). Planning English referring expressions. *Artificial Intelligence, 26*(1), 1-33.

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research, 32*, 25-36.

Bard, E. G., & Lickley, R. J. (1997). On not Remembering Disfluencies. In *Eurospeech 97*. Rhodes, Greece.

Beattie, G., & Butterworth, B. (1979). Contextual Probability and Word Frequencies as determinants of Pauses and Errors in Spontaneous Speech. *Language and Speech, 22*(3).

Beattie, G. W., Cutler, A., & Pearson, M. (1982). Why is Mrs. Thatcher interrupted so often?. *Nature, 300*(23), 744-747.

Berthold, A., & Jameson, A. (1999). Interpreting Symptoms of Cognitive Load in Speech Input. In Kay, J. (Ed.), *User modeling: Proceedings of the Seventh International Conference (UM99)* (pp. 235-244). Vienna: Springer-Wien.

Beskow, J. (2003). *Talking heads - Models and applications for multimodal speech synthesis*. Doctoral dissertation, KTH, Department of Speech, Music and Hearing, KTH, Stockholm.

Bestgen, Y. (1998). Segmentation markers as trace and signal of discourse structure. *Journal of Pragmatics, 29*(6), 753-763.

Blackmer, E. R., & Mitton, J. L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition, 39*(3), 173-194.

Boyce, S., & Gorin, A. (1996). User Interface Issues for Natural Spoken Dialog Systems. In *Internat. Symp. on Spoken Dialogue, ISSD* (pp. 65-68).

Brennan, S., & Schober, M. (2001). How Listeners Compensate for Disfluencies in Spontaneous Speech. *Journal of Memory and Language, 44*(2), 274-296.

Brennan, S., & Williams, M. (1995). The Feeling of Another's knowing: Prosody and Filled Pauses as Cues to Listeners about the Metacognitive States of Speakers. *Journal of Memory and Language, 34*, 383-398.

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

Brysbaert, M., Fias, W., & Noël, M-P. (1998). The Whorfian hypothesis and numerical cognition: is "twenty-four" processed in the same way as "four-and-twenty"?. *Cognition, 66*, 51-77.

Burke, K. (1969). *A grammar of motives.* Los Angeles, CA, USA: University of California Press.

Busemann, S., & Horacek, H. (1998). A Flexible Shallow Approach to Text Generation. In *In Proc. 9th International Workshop on Natural Language Generation*. Canada.

Butterworth, B., & Goldman-Eisler, F. (1979). Recent Studies on Cognitive Rythm. In Siegman, A., & Feldstein, S. (Eds.), *Of speech and time: temporal speech patterns in interpersonal contexts* (pp. 211-223). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Butterworth, B. (1975). Hesitation and semantic planning in speech. *Journal of Psycholinguistic Research, Volume 4*(Number 1).

Byron, D. K., & Heeman, P. A. (1997). Discourse marker use in task-oriented spoken dialog. In *Proceedings of Eurospeech 97*. Rhodes, Greece.

Callaway, C., & Lester, J. (2001). Pronominalization in generated discourse and dialogue. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Philadelphia, Pennsylvania.

Campione, E., & Veronis, J. (2002). A large-scale multilingual study of silent pause duration. In *ESCA-workshop on speech prosody* (pp. 199-202). Aix-en-Provence.

Caporeal, L., & Heyes, C. (1997). Why anthropomorphise? Folk psychology and other stories. In Mitchell, R. W., Thompson, N. S., & Miles, H. L. (Eds.), *Anthropomorphism, anecdotes, & animals* (pp. 59-73). Albany: State University of New York Press..

Caramazza, A. (1997). How Many Levels of Processing Are There in Lexical Access?. *Cognitive Neuropshychology, 14*(1), 177-208.

Carlson, R., & Granström, B. (2007). Rule-based Speech Synthesis. In Benesty, J., Sondhi, M. M., & Huang, Y. (Eds.), *Springer Handbook of Speech Processing* (pp. 429-436). Springer Berlin Heidelberg.

Cassell, J. (2007). Body language: lessons from the near-human. In Riskin, J. (Ed.), *Genesis Redux: Essays on the history and philosophy of artificial life* (pp. 346-374). University of Chicago Press.

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., & Yan, H. (1999). Embodiment in Conversational Interfaces: Rea. In *CHI'99* (pp. 520-527). Pittsburgh, PA, US.

Chomsky, N. (1965). *Aspects of the theory of syntax.* The MIT Press, Cambridge, MA.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition, 84*(1), 73-111.

Clark, H. H. (1996). *Using language.* Cambridge, UK: Cambridge University Press.

Cohen, R. (1984). A computational theory of the function of clue words in argument understanding. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics* (pp. 251-258).

Core, M., & Allen, J. (1997). Coding Dialogs with the DAMSL Annotation Scheme. In *In Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines* (pp. 28-35). Camebridge.

Corley, M., & Stewart, O. (2008). Hesitation Disfluencies in Spontaneous Speech: The Meaning of um. *Language and Linguistics Compass, 2*(4), 589-602.

Corley, M., MacGregor, L., & Donaldson, D. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition, 105*(3), 658-668.

Cutler, A., & Pearson, M. (1986). On the analysis of prosodic turn-taking cues. In Johns-Lewis, C. (Ed.), *Intonation and discourse* (pp. 139–155). London: Croom Helm.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies – why and how. In *Proceedings from the 1993 International Workshop on Intelligent User Interfaces* (pp. 193-200).

Dale, R., & Mellish, C. (1998). Towards Evaluation of Natural Language Generation. In *Proceedings of the First International Confer-*

ence on Language Resources and Evaluation (pp. 555-562). Granada, Spain.

Damian, M., & Dumay, N. (2007). Time pressure and phonological advance planning in spoken production. *Journal of Memory and Language, 57*(2), 195-209.

de Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: a cognitive cornerstone of conversation. *Language, 82*(3), 515-535.

De Smedt, K. (1990). IPF: An incremental parallel formulator. In Dale, R., Mellish, C., & Zock, M. (Eds.), *Current research in natural language generation* (pp. 167-192). London: Academic Press.

Dell, G. (1986). A spreading activation theory in sentence production. *Psychological review, 93*, 283-321.

Dijksterhuis, A., & Bargh, J. (2001). The perception-behavior expressway: Automatic effects of social perception on social behavior. *The perception-behavior expressway: Automatic effects of social perception on social behavior, 33*, 1-40.

Don, A., Brennan, S., Laurel, B., & Shneiderman, B. (1992). Anthropomorphism: from Eliza to Terminator 2. In *In Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 67-70).

Duncan, S., & Fiske, D. (1977). *Face-to-face interaction: Research, methods and theory.* Hillsdale, New Jersey, US: Lawrence Erlbaum Associates.

Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology, 23*(2), 283-292.

Dybkjær, H., Bernsen, N. O., & Dybkjær, L. (1993). Wizard-of-Oz and the trade-off between naturalness and recognizer constraints.

In *Proceedings of the 3rd European Conference on Speech Communication and Technology, Berlin*.

Edlund, J., & Beskow, J. (2009). MushyPeek - a framework for online investigation of audiovisual dialogue phenomena. *Language and Speech, 52*(2-3), 351-367.

Edlund, J., & Heldner, M. (2005). Exploring prosody in interaction control. *Phonetica, 62*(2-4), 215-226.

Edlund, J., & Hjalmarsson, A. (2005). Applications of distributed dialogue systems: the KTH Connector. In *Proceedings of ISCA Tutorial and Research Workshop on Applied Spoken Language Interaction in Distributed Environments (ASIDE 2005)*. Aalborg, Denmark.

Edlund, J., House, D., & Skantze, G. (2005). The effects of prosodic features on the interpretation of clarification ellipses. In *Proceedings of Interspeech 2005* (pp. 2389-2392). Lisbon, Portugal.

Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Communication, 50*(8-9), 630-645.

Edlund, J., Heldner, M., Al Moubayed, S., Gravano, A., & Hirschberg, J. (2010). Very short utterances in conversation. In *Proc. of Fonetik 2010*. Lund, Sweden.

Ellis, N. (2006). Selective Attention and Transfer Phenomena in L2 Acquisition: Contingency, Cue Competition, Salience, Interference, Overshadowing, Blocking, and Perceptual Learning. *Applied Linguistics, 27*, 164-194.

Fernández, R., Schlangen, D., & Lucht, T. (2007). Push-to-talk ain't always bad! Comparing Different Interactivity Settings in Task-oriented Dialogue. In *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue* (pp. 25-31). Trento, Italy.

Ferreira, F., & Swets, B. (2002). How Incremental Is Language Production? Evidence from the Production of Utterances Requiring the Computation of Arithmetic Sums. *Journal of Memory and Language, 46*, 57-84.

Ferrer, L., Shriberg, E., & Stolcke, A. (2002). Is the speaker done yet? Faster and more accurate end-of utterance detection using prosody. In *Proceedings of ICSLP* (pp. 2061-2064).

Ferrer, L., Shriberg, E., & Stolcke, A. (2003). A prosody-based approach to end-of-utterance detection that does not require speech recognition. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*. Hong Kong.

Ford, C., & Thompson, S. (1996). Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In Ochs, E., Schegloff, E., & Thompson, A. (Eds.), *Interaction and grammar* (pp. 134-184). Cambridge: Cambridge University Press.

Fox Tree, S., & Schrock, J. (1999). Discourse Markers in Spontaneous Speech: Oh What a Difference an Oh Makes. *Journal of Memory and Language, 40*, 280-295.

Fox Tree, J. E. (1995). Effects of false starts and repetitions on th eprocessing of subsequent words in spontaneous speech. *Journal of Memory and Language, 34*(6), 709-738.

Fraser, N., & Gilbert, N. (1991a). Effects of System Voice Quality on User Utterances in Speech Dialogue Systems. In *EUROSPEECH-1991* (pp. 57-60).

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

Fraser, B. (1996). Pragmatic markers. *Pragmatics, 6*(2), 167-190.

Fraser, B. (1999). What are discourse markers?. *Journal of Pragmatics, 31*(7), 931-952.

Fromkin, V. (1973). *Speech errors as linguistic evidence.* The Hague: Mouton.

Garrett, M. F. (1975). The analysis of sentence production. In Bower, G. (Ed.), *Psychology of learning and motivation: Vol. 9* (pp. 133-177). New York: Academic Press.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy?. *Trends in Cognitive Sciences, 8*(1), 8-11.

Goldman-Eisler, F. (1958). The Predictability of Words in Context and the Length of Pauses in Speech. *Language and Speech, 1*(3), 226-231.

Goldman-Eisler, F. (1972). Pauses, clauses, sentences. *Language and Speech, 15*, 103-113.

Goldman-Eisler, F. (1961). A comparative study of two hesitation phenomena. *Language and Speech, 4*, 1826.

Gong, L., & Lai, J. (2001). Shall we mix synthetic speech and human speech?: impact on users' performance, perception, and attitude. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 158-165). Seattle, USA.

Goodwin, C. (1986). Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies, 9*(2-3), 205-217.

Goto, M., Itou, K., & Hayamizu, S. (1999). A Real-Time Filled Pause Detection System for Spontaneous Speech Recognition. In *Eurospeech 99*. Budapest, Hungary.

Graesser, A. C., VanLehn, K., Rosé, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine, 22*(4), 39-51.

Gravano, A. (2009). *Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue*. Doctoral dissertation, Columbia University.

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics, 12*(3), 175-204.

Gustafson, J., & Edlund, J. (2008). expros: a toolkit for exploratory experimentation with prosody in customized diphone voices. In *Proceedings of Perception and Interactive Technologies for Speech-Based Systems (PIT 2008)* (pp. 293-296). Berlin/Heidelberg: Springer.

Gustafson, J., Larsson, A., Carlson, R., & Hellman, K. (1997). How do system questions influence lexical choices in user answers?. In *Proc of Eurospeech '97, 5th European Conference on Speech Communication and Technology* (pp. 2275-2278). Rhodes, Greece.

Gustafson, J., Bell, L., Boye, J., Lindström, A., & Wirén, M. (2004). The NICE fairy-tale game system. In *Proceedings of SIGdial*. Boston.

Gustafson, J., Heldner, M., & Edlund, J. (2008). Potential benefits of human-like dialogue behaviour in the call routing domain. In *Proceedings of Perception and Interactive Technologies for Speech-Based Systems (PIT 2008)* (pp. 240-251). Berlin/Heidelberg: Springer.

Hauptmann, A., & Rudnicky, A. (1988). Talking to computers: An empirical investigation. *International Journal of Man-Machine Studies, 28*(6), 583-604.

Hayes-Roth, B. (2004). What Makes Characters Seem Life-Like?. In Prendinger, H., & Ishizuka, M. (Eds.), *Life-Like Characters: Tools, Affective Functions, and Applications* (pp. 447-462). Germany: Springer.

Heeman, P., & Allen, J. (1999). Speech repairs, intonational phrases, and discourse markers. *Computational Linguistics, 25*(4), 527-571.

Heeman, P., Byron, D., & Allen, J. (1998). Identifying Discourse Markers in Spoken Dialogue. In *Proceedings of AAAI Spring Symposium on Applying Machine Learning and Discourse Processing* (pp. 44-51). Stanford, CA, USA.

Heldner, M., & Edlund, J. (in press). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*.

Henderson, A., Goldman-Eisler, F., & Skarbek, A. (1966). Sequential temporal patterns in spontaneous speech. *Language and Speech, 9*(4).

Hirschberg, J., & Litman, D. (1987). Now let's tal about now: identifying cue phrases intonationally. In *Proceedings of ACL 87*.

Hirschberg, J., & Litman, D. (1993). Empirical studies on the disambiguation of cue phrases. *Computational Linguistics, 19*(3), 501-530.

Hirschberg, J. (2002). Communication and prosody: Functional aspects of prosody. *Speech Communication, 36*(1-2).

Howell, P., & Young, K. (1991). The Use of Prosody in Highlighting Alterations in Repairs from Unrestricted Speech. *The Quarterly Journal of Experimental Psychology, 43a*(3), 733-758.

Iuppa, N., & Borst, T. (2007). *Story and simulations for serious games : tales from the trenches.* Focal Press.

Izdebski, K., & Shipp, T. (1978). Minimal reaction times for phonatory initiation. *Journal of Speech and Hearing Research, 21*, 638-651.

Jefferson, G. (1986). Notes on 'latency' in overlap onset. *Human Studies, 9*(2/3), 153-183.

Johnson, W., Vilhjalmsson, H., & Marsella, S. (2005). Serious games for language learning: How much game, how much AI?. In *12: th International Conference on Artificial Intelligence in Education*. Amsterdam.

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing. An Introd to Natural Language Proc, Computational Ling, and Speech Recogn.* Prentice Hall, Inc..

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

Kamm, C., Walker, M., & Rabiner, L. (1997). The role of Speech Processing in Human-Computer Intelligent Communication. *Speech Communication, 23*(4), 263-278.

Karsenty, L. (2002). Shifting the design philosophy of spoken natural language dialogue: From invisible to transparent systems. *International Journal of Speech Technology, 5*(2), 147-157.

Kempen, G., & Hoenkamp, E. (1982). Incremental sentence generation: implications for the structure of a syntactic processor. In *Proceedings of the 9th conference on Computational linguistics* (pp. 151-156). Prague, Czechoslovakia.

Kempen, G., & Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science, 11*(2), 201-258.

Kilger, A., & Finkler, W. (1995). *Incremental Generation for Real-Time Applications*. Technical Report RR-95-11, German Research Center for Artificial Intelligence.

Kim, J. H., Glass, M. S., Freedman, R., & Evens, M. W. (2000). Learning the use of discourse markers in tutorial dialogue for an intelligent tutoring system. In Gleitman, L. R., & Joshi, A. K.

(Eds.), *Proceedings of the cognitive science 2000* (pp. 262–267). Mahwah, NJ: Lawrence Erlbaum Associates.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech, 41*, 295-321.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By.* Chicago: University of Chicago Press.

Lemon, O., Gruenstein, A., Gullett, R., Battle, A., & Peters, S. (2003). Generation of collaborative spoken dialogue contributions in dynamic task environments. In *In Proceedings of the AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*.

Levelt, W., & Cutler, A. (1983). Prosodic Marking in Speech Repair. *Journal of Semantics, 2*(2), 205-218.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14*(1), 41-104.

Levelt, W. J. M. (1989a). *Speaking: From Intention to Articulation.* Cambridge, Mass., USA: MIT Press.

Levelt, W. (1989b). *Speaking-From Intention to Articulation.* The MIT Press.

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

Lickley, R. J., & Bard, E. G. (1996). On not Recognizing Disfluencies in Dialogue. In *ICSLP*. Philadelphia, USA.

Lindström, J. (2008). Diskursmarkörer. In *Tur och ordning; introduktion till svensk samtalsgrammatik* (pp. 56-104). Stockholm, Sweden: Norstedts Akademiska Förlag.

Litman, D., & Hirschberg, J. (1990). Disambiguating Cue Phrases in Text and Speech. In *Proceeding of COLING 90*.

Local, J., & Kelly, J. (1986). Projection and "silences": Notes on phonetic and conversational structure. *Human studies, 9*(2-3), 185-204.

Lounsbury, F. G. (1954). Transitional probability, linguistic structure and systems of habit family hierarchies. In Osgood, C. E., & Sebeok, T. A. (Eds.), *Psycholinguistics: A survey of theory and research problems* (pp. 93-101). Baltimore: Waverly press.

Louwerse, M. M., & Mitchell, H. H. (2003). Towards a taxonomy of a set of discourse markers in dialog: a theoretical and computational linguistic account. *Discourse Processes, 35*, 199-239.

Maclay, H., & Osgood, C. (1959). Hesitation phenomena in spontaneous English speech. *Word, 15*(1), 19-44.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* San Francisco, USA: W. H. Freeman.

Marslen-Wilson, W., & Tyler, L. (1980). The temporal structure of spoken language understanding. *Cognition, 8*, 1-71.

Mateas, M., & Stern, A. (2003). Façade: An experiment in building a fully-realized interactive drama. In *Game Developer's Conference: Game Design Track*. San Jose, California, US.

McCoy, K., & Strube, M. (1999). Generating anaphoric expressions: Pronoun or definite description?. In *Proceedings of the Workshop on the Relation of Discourse*. Maryland, USA.

Meltzer, L., Hayes, D., & Morris, M. (1971). Interruption Outcomes and Vocal Amplitude: Explorations in Social Psychophysics. *Journal of Personality and Social Psychology, 18*(3), 392-402.

Meyer, A. (1996). Lexical access in phrase and sentence production: results from picture–word interference experiments. *Journal of Memory and Language, 35*(4), 477-496.

Moore, K. (2007). Spoken language processing: Piecing together the puzzle. *Speech communication, 49*, 418-435.

Mori, M. (1970). The Uncanny Valley. *Energy, 7*(4), 33-35.

Müller, C., Großmann-hutter, B., Jameson, A., Rummer, R., & Wittig, F. (2001). Recognizing Time Pressure and Cognitive Load on the Basis of Speech: An Experimental Study. In Baumer, M., Gmytrasiewicz, P., & Vassileva, J. (Eds.), *User Modeling: Proceedings of the Eighth International Conference (UM2001)* (pp. 24-33). Berlin-Springer.

Nass, C., Brave, S., & Takayama, L. (2006). Socializing Consistency From Technical Homogeneity to Human Epitome. In Zhang, P., & Galletta, D. (Eds.), *Human-computer interaction in management information systems: Foundations* (pp. 373-390). Armonk, NY: M.E. Sharpe.

Naumann, J. D., & Jenkins, M. A. (1982). Prototyping: The New Paradigm for Systems Development. *MIS Quarterly, 6*(3), 29-44.

Nespor, M. (1986). The phonological word in Greek and Italian. In Andersen, H. (Ed.), *Sandhi phenomena in the languages of Europe*. Mouton du Gruyter.

Norman, D. A. (1998). *The design of everyday things.* MIT Press.

Norwine, A. C., & Murphy, O. J. (1938). Characteristic time intervals in telephone conversation. *The Bell System Technical Journal, 17*, 281-291.

Novick, D., Hansen, B., & Ward, K. (1996). Coordinating Turn-taking with Gaze. In *4th International Conference on Spoken Language Processing* (pp. 1888-1891). Philadelphia, USA.

O'Connell, D. (1992). Some intentions regarding speaking. *Journal of Psycholinguistics Research, 21*, 59-65.

Oliveira, M., & Freitas, T. (2008). Intonation as a cue to turn management in telephone and face-to-face interactions. In *Speech Prosody 2008* (pp. 485). Campinas, Brazil.

Oomen, C., & Postma, A. (2004). Effects of Time Pressure on Mechanisms of Speech Production and Self-Monitoring. *Journal of Psycholinguistic Research, 30*(2), 163-184.

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

O'Connell, D., & Kowal, S. (2005). Uh and Um revisited: Are they interjections for signaling delay?. *Journal of Psycholinguistic Research, Volume 34*(Number 6).

Pfeifer, L., & Bickmore, T. (2009). Should Agents Speak Like, um, Humans? The Use of Conversational Fillers by Virtual Agents. In *Intelligent Virtual Agents* (pp. 460-466). Springer Berlin / Heidelberg.

Pickering, M. J., & Garrod, S. (2006). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*, 169-226.

Postma, A. (2000). Detection of errors during speech production: a review of speech monitoring models. *Cognition, 77*(2), 97-131.

Raux, A. (2008). *Flexible Turn-Taking for Spoken Dialog Systems*. Doctoral dissertation, School of Computer Science Carnegie Mellon University.

Reeves, B., & Nass, C. (1996). *The Media Equation.* Stanford, CA, US: CSLI Publications.

Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering, 3*(1), 57-87.

Riccardi, G., & Gorin, A. L. (2000). Stochastic Language Adaptation over Time and State in Natural Spoken Dialogue Systems. *IEEE Trans. Speech Audio Proc., 8*, 3-10.

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

Saffer, D. (2005). *The Role of Metaphor in Interaction Design*. Master's thesis, Carnegie Mellon University.

Saini, P., de Ruyter, B., Markopoulos, P., & van Breemen, A. (2005). Benefits of Social Intelligence in Home Dialogue Systems. In *Proceedings of Interact 2005 - Communicating Naturally through Computers, Rome, Italy, 2005.*.

Schaffer, D. (1983). The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics, 11*, 243-257.

Schiffrin, D. (1987). *Discourse markers.* New York, USA: Cambridge University Press.

Schlangen, D., & Skantze, G. (2009). A general, abstract model of incremental dialogue processing. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*. Athens, Greece.

Schlangen, D. (2009). What we can learn from Dialogue Systems that don't work: On Dialogue Systems as Cognitive Models. In *Proceedings of DiaHolmia (Semdial 2009)* (pp. 51-58).

Schnadt, M., & Corley, M. (2006). The influence of lexical, conceptual and planning based factors on disfluency production. In *twenty-eighth meeting of the Cognitive Science Society*. Vancouver, Canada.

Schourup, L. (1999). Discourse markers. *Lingua, 107*(3-4), 227-265.

Schröder, M. (2001). *Emotional Speech Synthesis: A Review.* Aalborg, Denmark.

Selting, M. (1996). On the interplay of syntax and prosody in the constitution of turnconstructional units and turns in conversation. *Pragmatics, 6*, 357-388.

Shneiderman, B., & Maes, P. (1997). Direct manipulation vs. interface agents: Excerpts from debates at IUI'97 and CHI'97. *Interactions, 4*, 42-61.

Shneiderman, B. (1995). Looking for the bright side of user interface agents. *Interactions, 2*(1), 13-15.

Shriberg, E. (1994). *Preliminaries to a theory of speech disfluencies*. Doctoral dissertation, University of California.

Sjölander, K., & Beskow, J. (2000). WaveSurfer - an open source speech tool. In Yuan, B., Huang, T., & Tang, X. (Eds.), *Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing* (pp. 464-467). Beijing.

Sjölander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech. In *Proc of Fonetik 2003, Umeå University, Dept of Philosophy and Linguistics PHONUM 9* (pp. 93-96).

Skantze, G., & Edlund, J. (2004). Robust interpretation in the Higgins spoken dialogue system. In *Proceedings of ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*. Norwich, UK.

Skantze, G., & Schlangen, D. (2009). Incremental dialogue processing in a micro-domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*. Athens, Greece.

Skantze, G. (2008). Galatea: A discourse modeller supporting concept-level error handling in spoken dialogue systems. In Dybkjær,

L., & Minker, W. (Eds.), *Recent Trends in Discourse and Dialogue*. Springer.

Skantze, G., House, D., & Edlund, J. (2006). User responses to prosodic variation in fragmentary grounding utterances in dialogue. In *Proceedings of Interspeech 2006 - ICSLP* (pp. 2002-2005). Pittsburgh PA, USA.

Ström, N., & Seneff, S. (2000). Intelligent barge-in in conversational systems. In *Procedings of ICSLP-00*.

Swerts, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics, 30*(4), 485-496.

Swertz, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics, 30*(4), 485-496.

Tannenbaum, P., Williams, F., & Hillier, C. (1965). Word predictability in the environments of hesitations. *Journal of verbal learning and verbal behaviour, 4*, 134-140.

Theune, M. (2003). *Parlevink Language Engineering Group*. Technical Report, University of Twente, Parlevink Language Engineering Group.

Tomko, S., & Rosenfeld, R. (2004). Speech graffiti vs. natural language: assessing the user experience. In *in Proceedings of HLT-NAACL 2004*.

Van Wijk, C., & Kempen, G. (1987). A Dual System for Producing Self-Repairs in Spontaneous Speech: Evidence from Experimentally Elicited Corrections. *Cognitive Psychology, 19*, 403-440.

Ward, N., & Tsukahara, W. (2000). Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics, 32*(8), 1177-1207.

Ward, N. (2004). Pragmatic functions of prosodic features in non-lexical utterances. In *Proceedings of Speech Prosody* (pp. 325-328).

Watanabe, M., Hirose, K., Den, Y., & Minematsu, N. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech communication, 50*(2), 81-94.

Weilhammer, K., & Rabold, S. (2003). Durational aspects in turn taking. In *ICPhS 2003*. Barcelona, Spain.

Weizenbaum, J. (1966). ELIZA - A computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery, 9*, 36-45.

Wheeldon, L., & Lahiri, A. (1997). Prosodic Units in Speech Production. *Journal of Memory and Language, 37*(3), 356-381.

Wik, P., & Hjalmarsson, A. (2009). Embodied conversational agents in computer assisted language learning. *Speech communication, 51*(10), 1024-1037.

Winters, S., & Pisoni, B. (2004). Perception and Comprehension of Synthetic Speech. *Progress Report Research on Spoken Language Processing, 26*.

Wooffitt, R., Fraser, N. M., Gilber, N., & McGlashan, S. (1997). *Humans, computers and wizards.* London and New York: Routledge.

Zoltan-Ford, E. (1991). How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies, 34*(4), 527-547.

Appendix

Appendix A. Dialogue excerpts used as stimuli in the listening test

Dialogue segment		unconstrained	constrained
1:1	U	aha ok I see this is Peter and I have agreed to arrange a dinner	
	S	ok	-
	U	eh with her I guess	
	S	ok yes mmm when would you like to meet her	ok you want to have dinner with Anna when would you like to meet her
	U	eh yeah I am not sure eh I can do it eh eh on sometime after work eh end of this week	
	S	mm ok	ok you want to meet Anna for dinner in the evening at Thursday, Friday, Saturday or Sunday
	U	it is Wednesday today	
1:2	S	mmmm she she has a dinner on Friday mm but she is available on Saturday and Sunday and on Thursday as well	Anna is available for dinner on Thursday, Saturday and Sunday
	U	on Thursday that is tomorrow	
	S	yeah	yes, Thursday is tomorrow
	U	mm tomorrow let's see yeah why not tomorrow	
1:3	S	ok	ok
	U	ok	
	S	mmm what time would suit you?	what time would you like to meet Anna for dinner on Thursday?
	U	for me it would be fine at eh six thirty	
	S	six thirty that would that looks just fine on her schedule	ok six thirty

1:4	S	where would you like to meet her?	where would you like to meet Anna?
	U	do you know what kind of restaurant she would like to go to?	
	S	mmm I don't would you like me to talk to her and get back to you?	no I don't know which restaurant Anna would like to go to should I call and ask Anna?
	U	yeah why not	
1:5	S	ok fine could I just mmm take your number	ok can you please give me your telephone number
	U	my number is 123-4567	
	S	4567 and that was Peter?	ok your telephone number is 123-4567 and your name is Peter?
	U	yeah that is right	
1:6	S	ok and you have no particular suggestions yourself?	do you have any restaurant suggestions?
	U	well I could go to a Chinese restaurant	
	S	ok well I hear what she is interested in doing.	ok a Chinese restaurant I will call Anna and ask her about which type of restaurant she prefers
1:7	S	ok I will be back to you in a little bit	ok I will call you back shortly
	U	yeah ok	
	S	thank you	thank you
	U	ok shall I wait?	
	S	sure eh no I call you back	no I will call you back
	U	ok mm	
	S	ok	ok
	U	fine	
	S	thanks	thank you
	U	bye	

	S	bye	bye
2:1	**S**	hello this is jane anna's secretary	hello, this is anna's secretary
	U	ok ok I should set up eh meeting with Y	
	S	ok where would you like to meet her?	ok where would you like to meet her?
	U	when?	
	S	yes when would you?	when would you like to meet her?
	U	tomorrow	
	S	ok	ok
	U	after lunch	
	S	that looks fine her sched-	ok
2:2	**S**	when would you like to	when would you like to come
	U	eh at eh one thirty	
	S	one thirty ok that's fine for about how long do you need to meet her	one thirty ok for how long will the meeting go on
	U	eh I eh guess for eh one or two hours	
	S	ok	ok
3:1	**S**	I'm sorry I didn't get you name	what is your name
	U	I I am john smith	
	S	john	ok
	U	John yes J O H N	
	S	J O H N	J O H N
	U	yeah S M I T H	
	S	ok got that	S M I T H
4:1	**S**	and you said one to two hours so I should book her in til til about about three three thirty?	I'll book the meeting until three thirty?

	U	one thirty	
	S	yeah but starting at one thirty but until about three thirty?	ok starting at one thirty until three thirty?
	U	yeah until three thirty yeah that is true	
	S	yeah ok we will do that	ok
5:1	**U**	it will be some more people too	
	S	ok what are their names	ok shall I take their names
	U	eh Carl Anderson	
	S	ok Carl Andersonr	yeah ok and
	U	Peter Pan	
	S	Peter Pan	-

Appendix B. Listening test GUI

Appendix C. Seller Instructions

You are about to participate in a role-play that is situated at a flea market. It is a bit like acting, but try to use your normal voice and dialect. The text below presents a scenario that takes place at a flea market. Read and consider these background facts and try to make use of them when you participate in the dialogues.
You are the owner of a flea market that sells a diverse set of goods. You have bought the goods that you are selling from other flea markets and auctions. Since the flea market is your only source of income, it is important that you make a profit from the things you sell and try to make to customers pay as much as possible.
You will participate in 4 recording.

A few voluntary recommendations before you start:

- Use the double retail price as your initial offer.
- Try to make a profit of at least 10%.
- If the customer's offer is extremely low or this person use sneaky strategies, do not lower the price. Instead, try to point out how excellent the goods you are trying to sell are.
- Lower your price only after the customer has started to approach the price you had in mind.
- Consider lowering the price if the customer has reasonable arguments for doing so.

Appendix D. Customer instructions

You are about to participate in a role-play that is situated at a flea market. It is a bit like acting, but try to use your normal voice and dialect. The text below presents a scenario that takes place at a flea market. Read and consider these background facts and try to make use of them when you participate in the dialogues.

[One of the two following scenarios was presented to the customer:]

Scenario 1: 375 sek

You are trying to find three presents for your niece's birthday, but you only have 375 sek. You walk in to a small flea market and find an odd-looking shopkeeper.

Scenario 2: 250 sek

You have recently bought an old and beautiful house but the building as well as the garden needs a restoration. However, you have no tools and you only have 250 sek. Find out if the flea market has any tools before you go over to the expensive hardware store. Try to find at least three tools. Behind the counter inside the flea market is an odd-looking shopkeeper.

Appendix E. DEAL data collection questionnaire

1. Gender: Male **( )** Female **( )**

2. Age: ___

3a. [Before the test] Du you have any experience of price negotiation?

Alternatively:

3a. [After each dialogue] Did you use any particular negation strategy in the dialogue?

No ( )

Continue at the next page!

Yes ( )

3a. Did you negotiate as a seller?

( ) yes ( ) no

If you used a particular strategy, describe this strategy briefly below:

_____

_____

3b. Did you negotiate as a customer?

( ) yes ( ) no

If you used a particular strategy, describe this strategy briefly below:

_____

_____

Appendix F.

Token	nonCP	CP	Total
ja	89	542	631
ju	23	487	510
men	20	286	306
m	21	207	228
a	12	152	164
nej	24	96	120
nä	6	67	73
också	7	65	72
alltså	1	67	68
eller	16	42	58
faktiskt	5	42	47
nog	15	24	39
hm	12	25	37
precis	8	25	33
okej	1	32	33
mm	6	24	30
jo	9	19	28
asså		28	28
väl	2	23	25
mhm	1	19	20
jaha	3	16	19
fast	3	14	17
menar	5	10	15
ändå	5	10	15
förstår	4	11	15
heller	1	13	14
liksom	1	12	13

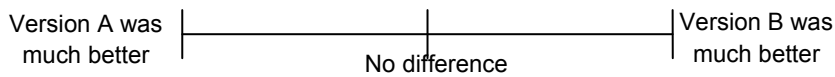
Appendix G.
An incremental version of DEAL: user questionnaire

Page 1
Sex: Age:

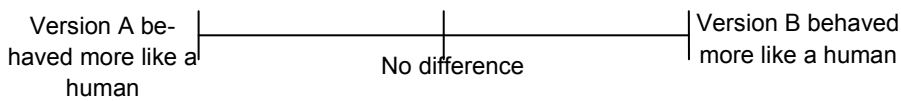
Did you experience any differences between the two system version of DEAL, and if so, what kind of differences?

Page 2
You are about to answer a number of questions where your task is to compare the two versions of DEAL. Try to not respond based on how well you succeeded with tasks, but on how you experienced the different versions. Put marks on the line.

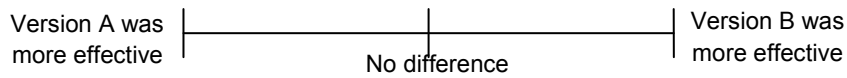
1. **Which version did you prefer?**



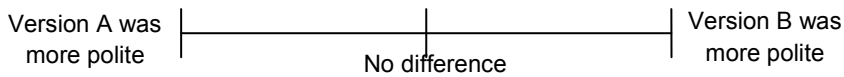
2. **Which version behaved more like a human?**



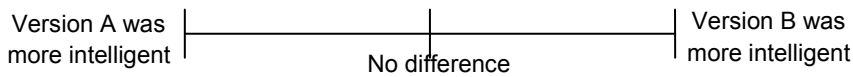
3. **Which version was more effective in its communication?**



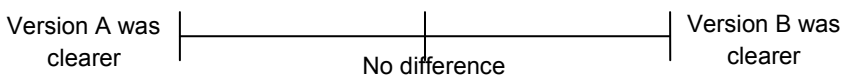
4. Which version was more polite?



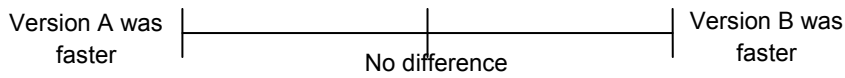
5. Which version behaved more intelligent?



6. Which version was better at confirming what you had said?



7. Which version gave the fastest response?



8. With which version was it clearer when to speak?