Nja, ja and aa

Data collection and analysis of cue phrases with focus on generation of respsonses in dialogue systems

JENNY KLARENFJORD

KTH Computer Science and Communication

Master of Science Thesis
Stockholm, Sweden 2008
Nja, ja and aa

Data collection and analysis of cue phrases with focus on generation of responsives in dialogue systems

JENNY K L A R E N F J O R D

Master’s Thesis in Speech Communication (30 ECTS credits) at the School of Media Technology
Royal Institute of Technology year 2008
Supervisor at CSC was Anna Hjalmarsson
Examiner was Rolf Carlson

TRITA-CSC-E 2008:088
ISRN-KTH/CSC/E--08/088--SE
ISSN-1653-5715
Nja, ja and aa
-Data collection and analyses of cue phrases with focus on generation of responsives in dialogue systems

Abstract

This master’s thesis presents a data collection of human-human dialogue and analyses of cue phrases, for generation in DEAL, a game for conversation training in Swedish. The main attention has been on the responsive cue phrases, trying to distinguish what influences how we interpret different discourse-pragmatic functions of the word “ja” (Eng: yes). Two annotators labelled cue phrases in the corpus using an annotation scheme with ten different categories. The annotators reached high inter-annotator agreement (Kappa coefficient of 0.87) for the binary task, i.e. when classifying a word as a cue phrase or not. Inter-annotator agreement for the categorization of cue phrases was also high (Kappa of 0.82). The most frequent lexical item for the responsive cue phrases was “ja” (Eng: yes). A listening test with 21 subjects was performed to see which features, contextual or acoustic, are important for how different realizations of “ja” are interpreted. All stimuli were presented twice, isolated and in context. Over all inter-annotator agreements between the two conditions were fairly low. However, agreement between the subjects was higher for stimuli in context than for isolated stimuli, which suggests that contextual cues are important for how “ja” is interpreted. However, a few stimuli received high inter-annotator agreement between the two conditions, i.e. with and without context, and for a specific category. Acoustic analysis revealed a few features that might be characteristic for some of the different responsive categories. For example: long duration may be important for the Responsive Dispreference category and the rising shape of the f0-curve, containing two pitch peaks, may be characteristic for the Responsive New Information Category.
**Nja, ja och aa**

- Datainsamling och analys av diskursmarkörer med fokus på generering av responsiver i dialogsystem

**Sammanfattning**

Examensarbetet presenterar en datainsamling av spontan dialog samt analys av diskursmarkörer för generering i DEAL, ett spel för konversationsträning i svenska. Responsiva diskursmarkörer har varit i fokus, med mål att urskilja vad som påverkar hur vi tolkar olika diskurspragmatiska funktioner av responsiven ”ja”. Den inspelade dialogen märktes upp av två annoterare enligt ett annoteringsschema för diskursmarkörer innehållande tio olika kategorier. Överensstämmelsen mellan de båda annoterarna var hög (Kappa 0.87) för den binära uppgiften, dvs. att bedöma om ett ord är en diskursmarkör eller ej. Överensstämmelsen var också hög för kategoriseringen av diskursmarkörer (Kappa värde 0.82). Ordet ”ja” var den vanligast förekommande responsiven bland diskursmarkörerna. Ett lyssningstest med 21 deltagare utfördes för att se vilka egenskaper, kontextuella eller akustiska egenskaper hos ordet, som är viktiga för hur vi tolkar olika varianter av ”ja”. Alla stimuli av ”ja” presenterades två gånger, isolerade och i kontext. Överensstämmelsen vid val av kategori mellan dessa båda tillstånd var överlag relativt låg. Däremot var överensstämmelsen mellan testpersonerna högre för stimuli i kontext än för isolerade stimuli, vilket antyder att kontextuella egenskaper är viktiga för hur ”ja” tolkas. Några stimuli hade däremot hög överensstämmelse för de två olika tillstånden, dvs. med och utan kontext och även för en specifik kategori testpersonerna emellan. En akustisk analys visar några egenskaper som verkar vara karaktäristiska för ett par av de responsiva kategorierna. Lång varaktighet (duration) verkar vara utmärkande för Dispreferermarkeringande responsiver och en stigande form på f-kurvan, innehållande två frekvenstoppar, verkar utmärka de Nyhetsmottagande responsiverna.
# Table of contents

1 INTRODUCTION .................................................................................................................................1

1.2 PURPOSE AND METHOD ....................................................................................................................1

1.3 OVERVIEW OF THE PAPER ...............................................................................................................2

2 THEORETICAL BACKGROUND .............................................................................................................3

2.1 DIALOGUE SYSTEMS ..........................................................................................................................3
   2.1.1 Human metaphor ...........................................................................................................................3
   2.1.2 Perception of human-like language in dialogue systems ...............................................................4
   2.1.3 Entrainment in spoken dialogue systems ......................................................................................4

2.2 DEAL ..................................................................................................................................................6
   2.2.1 The game .......................................................................................................................................6
   2.2.2 The DEAL architecture .................................................................................................................7
     Speech recognizer .................................................................................................................................8
     Semantic parser ......................................................................................................................................8
     Dialogue management .........................................................................................................................8
     Speech output .........................................................................................................................................9

2.3 HUMAN-LIKE LANGUAGE GENERATION .......................................................................................10
   2.3.1 Human-human interaction .............................................................................................................10
     Mixed-initiative ..................................................................................................................................10
     Presence ..............................................................................................................................................10
     Common ground .................................................................................................................................11
     Referring expressions and fragmental utterances ...............................................................................11
     Error-handling ....................................................................................................................................11
     Turn-taking .........................................................................................................................................12
   2.3.2 Incremental generation ..................................................................................................................12
   2.3.3 Incremental speech in spoken dialogue systems ..........................................................................13

2.4 CUE PHRASES ..................................................................................................................................15
   2.4.1 Schiffrin .......................................................................................................................................15
   2.4.2 Fraser ..........................................................................................................................................16
   2.4.3 Further research on cue phrases ...................................................................................................17
   2.4.4 Cue phrases in dialogue systems ..................................................................................................18
     Using context and prosodic features for automatic identification ......................................................18
   2.4.5 Classification of cue phrases by Lindström ...............................................................................20
     Classification according to function .................................................................................................20
   2.4.6 Studies on responsive cue phrases .............................................................................................23
     Automatic identification of responsives .............................................................................................23
     A perception study ..............................................................................................................................24

3 METHOD ..............................................................................................................................................25

3.1 DATA COLLECTION OF SPOKEN DIALOGUES ............................................................................25
   3.1.1 The recording ...............................................................................................................................25
     Setting ...................................................................................................................................................25
1 Introduction

One of the big research areas within speech technology is the development of spoken dialogue systems, computer systems intended to interact with humans through speech. At the department Speech, music and hearing at KTH in Stockholm there is a long tradition of research related to spoken dialogue systems. One current research project is DEAL, a computer game developed for learning Swedish as your second language. It is a conversational dialogue system with speech as primary focus. The purpose of the game is to make the users talk as much as possible to practice their conversation skills. To do this the user needs to collaborate with the agent using speech. One research focus in DEAL is to design a system which generates speech in a human-like way. One reason for this is to make the game agent behave more believable, and another is that spontaneous conversational speech is a desired feature of a game agent when the aim of the game is to give second language learners conversation training in a second language. It is yet a complex task to make a dialogue system converse in a human-like way. There are many factors to analyse and implement in the system before reaching the goal of a human-like dialogue system. The goal is not however to create a system behaving and talking exactly like a human but human enough to maintain the user's suspension of disbelief, which means that the user ignore unimportant inconsistencies for him or her to get pleasure from the interaction with the system.

In spontaneous dialogue humans frequently use particles called cue phrases. These particles are often small words like the Swedish “ju, och, men, okej” (Eng: like, and, but, al right). They give information about how a new utterance relate to prior discourse. Moreover, humans produce speech incrementally. One way to try to make a dialogue system sound more human-like is to generate utterances in a similar fashion. This thesis will observe and analyse how humans express themselves in spoken dialogue with focus on cue phrases. The purpose is to find cue phrases which can be generated in DEAL.

1.2 Purpose and method

The focus of this master’s project was to collect and analyse cue phrases for generation in spoken dialogue systems. In order to do this in a human-like way, extended knowledge was needed on how these phrases are used in human-human dialogue. The main attention has been on the responsive cue phrases, trying to distinguish acoustic features which characterize different functions of the word “ja” (Eng: yes).

Is there a functional way of categorizing Swedish cue phrases? Which cue phrases are most frequently used within these categories? Which features, contextual or acoustic, are important for how different realizations of “ja” (Eng: yes) are interpreted?

By recording human-human dialogues in a role-play situation similar do the domain in DEAL, a suitable corpus was collected. Six people participated to record a total of eight different dialogues. These recorded dialogues were later transcribed and labelled by two annotators using annotation schemes for cue phrases, communicative acts, disfluencies and non-lexical features. The annotation was then used to study the cue phrases in more detail with focus on the responsive categories. A listening test was performed to see which features, contextual or acoustic, are important for how different realizations of “ja” (Eng: yes) are interpreted.
1.3 Overview of the paper

This master’s thesis begins with a theoretical part (part 2) describing features of human-human conversation, theories about cue phrases and previous experimental studies of them. DEAL is also presented. The methods are presented in part 3, describing the recording of dialogues, annotation and the implemented listening test on responsive cue phrases. The results are presented next (part 4), with data on cue phrase frequency and turn position for annotated cue phrase categories. The results from the listening test are also presented. Part 5 offers a discussion on used methods and results, and part 6 presents the conclusions. Part 7 offers suggestions of future research. This master’s thesis ends with acknowledgements (part 8), a reference list (part 9) and appendix (part 10).
2 Theoretical background

In the spoken dialogue system DEAL, a game for second language learners of Swedish to practice their conversation skills, the goal is to model more human-like language generation. The language generated in DEAL is crucial from a language learning perspective. To make the user talk to the agent as if talking to another human-being, the agent needs to behave in a human-like way. This section will describe what characterizes spontaneous speech and human-human interaction in speech with focus on desirable features for generation in DEAL.

2.1 Dialogue systems

Dialogue systems are systems which communicate with humans in natural language by using text or speech. The aim of the system is in most cases to collaborate with the user to solve a domain-specific task (buy a train ticket or solve a game task, for example). The system transforms the text- or speech input into a representation that can be interpreted and understood by the system and then produces a suitable response back to the user (McTear, 2002). The way that spoken dialogue systems generate utterances affects how users perceive the system and also how the user behaves when using the system (Wik et al., 2007a; Gustafson et al., 1997). One may think that adding more human-like features to a spoken dialogue system will make it easier to use since human-human conversation is very natural to us, but that depends on what kind of system one wants to create. So far machines are not as good as humans to manage all dimensions of human dialogue, thus it is sometimes more efficient to use more controlled and finite dialogues. In the next section (2.1.1) two types of metaphors which dialogue systems can be perceived within will be presented. Then the features of human-like language in spoken dialogue systems will be presented; how humans perceive dialogue systems that use a more human-like language and how users adjust their language according to the language used by the system. These following sections will motivate the importance of human-like speech in DEAL and similar dialogue systems.

2.1.1 Human metaphor

Users may interpret their interaction with spoken dialogue systems in more than one way (Edlund et al., 2006). Users may use different metaphors to make sense of human-computer interaction. Depending on which metaphor, the user may experience and interact with the system differently. (Edlund et al., in press). Different metaphors may change the expectations of what the system can or can not do. There are two main metaphors; the interface metaphor and the human metaphor. If the user interacts with the spoken dialogue system using an interface metaphor, the system is perceived as a machine interface, which is often a computer interface. Instead of using keyboard or a mouse to accomplish a task, speech is used as input. For example, instead of choosing a link in a web page using the mouse, the link may be entered by using speech as input to the system. If the user interacts with the spoken dialogue system using a human metaphor, the computer is perceived as an interlocutor with human-like conversational abilities. In these systems speech is not a substitute or one of many alternatives; it is the primary way for interacting and communicating with the system. If the user applies a human metaphor when interacting with a system, he/she may have expectations that the system will be capable of converse in a human-like way.

Developers of spoken dialogue systems may want that the interaction with their system should be coherent with a specific metaphor (Edlund et al., 2006). They then have to build a system that is internally coherent with the chosen metaphor. A system coherent with the human metaphor
requires different capabilities than a system coherent with the interface metaphor. When approaching an interface metaphor, human-machine experiments are useful to see how users interact with the interface. If the choice is to make the system coherent with a human metaphor, developers may instead initially study human-human interaction to find out how their system should be modelled. The system does not need to be an entire virtual human; nor does it need to behave like real humans (Edlund et al., in press). It should only act human enough to be perceived within a human metaphor, i.e. that the user speaks with the system as if it would respond to another human. Applications where the primary choice is displaying a human metaphor may for example be found in game characters and applications for entertainment purpose. DEAL is one of them.

2.1.2 Perception of human-like language in dialogue systems

To build dialogue systems consistent with a human metaphor the systems need to display human-like behaviour (Edlund et al., 2006). This does not mean that the system needs to behave exactly like a human for the user to perceive the system within a human metaphor. The system developers may rely on the users’ suspension of disbelief, which means keeping a persons ability to ignore unimportant inconsistencies for him or her to get pleasure from a work of fiction. As an example, this is what happens when people watch a movie or participate in a role-play. Still dialogue systems need to appear more natural for the user to almost believe he/she is interacting with another human.

Most spoken dialogue systems generate complete, inflexible and grammatically correct utterances, which may be perceived unnatural. Hjalmarssson & Edlund (in press) studied how adding human-like features in a simulated spoken dialogue system was perceived by non-participating listeners. Test subjects were asked to compare two versions of an artificial human-computer dialogue, one constrained and one unconstrained. The unconstrained version was a word-by-word replica of the human speaker. The other version was constrained lexically, removing filled pauses, repetitions, false starts and information not relevant for the task (for more information see Hjalmarssson & Edlund (in press)). The results show that subjects did not have trouble accepting and understanding more human-like dialogue systems; they rather perceive the version with human behaviour as more human-like, more intelligent and also more polite. There was no difference in perceived efficiency between the constrained and unconstrained versions. Since disfluencies often are perceived as flaws in the language, this result is interesting. For the development of DEAL these results suggest that adding more human-like features to the system might make the system more coherent with a human-metaphor.

2.1.3 Entrainment in spoken dialogue systems

Generating human-like speech can make the users speak more freely themselves. Research has shown that users of dialogue systems adapt the way the systems speak (Gustafson et al., 1997). This phenomenon is called lexical entrainment.

Gustafson et al. (1997) studied how the system vocabulary affects the lexical choices of the user. In the study several questions were asked from the system to the user using a wizard-of-oz experiment, simulating a dialogue system. They found that if the system asked the user questions, the user adopted the system’s lexical choices in his/her answers. When the system asked the users “Where do you want to go from?”, 60% of the users answered using an ellipse, such as “From Stockholm”. 57% of the users answered with complete sentences using the lexical items from the system question, such as “I want to go on Friday”. These results show that users adapt the language of the system both in terms of vocabulary and structure. However, it may also show that people do not expect the system to handle other than simple utterances since it uses simple utterances itself.
DEAL is an example of a game where the conversation itself is important, since the purpose of the game is to practice a language. With lexical entrainment in mind, a dialogue system that speaks like a human will encourage the users to do the same.
2.2 DEAL

At KTH there are one group of researchers working with a game called DEAL, a spoken dialogue system designed for conversation training for foreign students learning Swedish (Hjalmarsson et al., 2007). DEAL is a research platform combining elements from dialogue systems, language learning and computer games. Different from most other dialogue systems the main goal of the interaction is the talking itself (Brusk et al., 2007) and for the system to be fun combined with being educational (Hjalmarsson et al., 2007).

It started with Ville, a virtual language tutor developed at KTH that helps students to improve their language skills in a new language (Wik, 2004). While Ville focuses on listening skills and pronunciation skills, DEAL is a complement to Ville, with focus on improving the conversation skills of the foreign language students (Hjalmarsson et al., 2007). DEAL is a platform for practice the vocabulary you learn with Ville, but in a context. In DEAL, spoken dialogue is not just an add-on, it is the primary way for the user and the system to interact in order to progress forward in the game (Wik et al., 2007b). Since the dialogue is in focus, to succeed with this game the dialogues need to be fun, natural, and motivating. The aim is to motivate users to talk a lot, and to not only use short utterances when interacting with the system.

2.2.1 The game

The main task of the game is for the language learning student to take part in a role-play in a trade domain. The trade domain used for this game is a flea market, a place where people can sell used objects, often with a reduced price. The motivation behind choosing the trade domain is because it is universally known, useful to manage in a foreign language and a way to learn useful vocabulary. The user interface contains an animated talking agent (ECA) developed at KTH (Beskow, 2003). The ECA can produce lip-synchronized speech, express extra linguistic signs such as nodding and frowning, and eyebrow movements. Using an ECA in DEAL is important since language learners often incorporate visual information to a greater extent than native speakers (Granström & House, 2005). The ECA is the owner of the flea market and the user takes the role of the customer. The game player can look at the objects, talk about their properties with the ECA and negotiate about the price. The main task for the game player is to buy a set of items from the shopkeeper with a certain amount of money and try to get the items for best possible price. The money may not be enough so negotiation has to take place for the player to be able to succeed with the mission. At the same time the shopkeeper negotiates to get as much money as possible for his objects. Some of the items have defects and the user can point out these flaws to negotiate towards a better price. By doing this the shopkeeper’s willingness increases and he feels obligated to suggest a lower price. In Ville the ECA acts like a teacher, but in DEAL the ECA acts like a conversation partner keeping a motivating dialogue between him and the student (Wik et al., 2007a).

When playing the game there are a few steps that the system and the student need to go through in order to reach a deal (Brusk et al., 2007). The first state is the initiation of the dialogue. Both the ECA and the student can initiate the dialogue, for example by a greeting act. The shopkeeper can also initiate the dialogue by a request to the customer, asking what he/she wants to buy. The next state, the main state, is the trading state, which consists of several steps. First the shopkeeper needs to find an object that the customer is willing to purchase. When that task is finished the next step is negotiation, which is a complex process. Resolve exchange is the state when the parties have reached an agreement and want to conclude the deal. The end state is when there has been a deal or the trade has been cancelled.
A traditional information-seeking spoken dialogue system (for example systems used to by train tickets) and a spoken dialog system made for conversational role-play (for example DEAL) will be judged differently in their efficiency completing their specific tasks. For the first type of system, the information exchange between the system and the user is essential. If the user interacts with the system to buy a ticket to Stockholm, he/she will not be pleased if the ticket is to Gothenburg instead. Using speech to interact with the system may also be one of several alternatives. Depending on the situation, different ways to interact with the system may be more preferable than others. The focus is not on the interaction itself, it is on the ticket. Buying the right ticket in a convenient way, using a small amount of speaker turns and words, are preferable features for this system in order to be perceived as efficient. However, in DEAL the conversation itself is important, and not exactly what information is exchanged between the user and the system. The goal with DEAL is for the user to talk a lot to practice his/her conversation skills in a new language (Wik et al., 2007a).

The ECA does not necessarily need to behave as in traditional dialogue systems, where it presumably gives the information and does what the user asks for. The ECA in DEAL can be rude or even act in other unpleasant ways (Wik et al., 2007a). What is important is for the user to maintain a suspension of disbelief. In DEAL, it is preferable with an ECA who behaves in a human-like way, so that the user can act as if he/she is talking to a human.

### 2.2.2 The DEAL architecture

The dialogue system architecture implemented in DEAL is based on another architecture called Higgins, which is developed by Skantze (2007) at KTH, an off-the-shelf Automatic Speech Recognition (ASR) system, a dialogue manager developed for DEAL purposes and also a graphical user interface (GUI) with a conversational agent. Below is a figure showing the proposed architecture of DEAL.
Speech recognizer
The task of the speech recognizer is to transform the speech input, a continuous-time signal, into a string of discrete units such as phonemes or words (McTear, 2002). This is a difficult task since the speech signal contains a high degree of variability; variability of linguistic features such as that the same phoneme can have different acoustic realizations in different contexts, but also speaker variability; differences between speakers but also within speakers. Age, gender, and regional origin are all examples of factors that are different between different speakers. Factors within speakers might be physical, such as tiredness and change of mood. Recognition of spontaneous continuous speech is also difficult because it is hard to distinguish isolated words from the signal.

Semantic parser
The sequence of words extracted by the recognizer is then analysed by the semantic parser. The task of this component is to produce a meaning representation of the utterance (McTear, 2002). This includes syntactic analysis to determine how the words are grouped together, and also semantic analysis to determine the meaning. The parser used in DEAL is called Pickering (Skantze, 2007). Pickering analyses continuous and incremental input from the speech recognizer. This processing can then be done while the user is still talking. Since speech is unpredictable with hesitations and pauses (see chapter 2.3) it is difficult to chunk the string of words into utterances. Pickering builds semantic tree structures using context free grammar (CFG) that can be relaxed to robustly handle unpredictable, ungrammatical and misrecognized input.

Dialogue management
The main task for the dialogue management component is to control the flow of the dialogue (McTear, 2002). This includes determine if adequate information has been collected from the user in order to precede the communication with external applications such as databases. From the databases information that the user asks for can be retrieved. The external applications contain domain specific information, in DEAL this might be objects which can be find in the game, their prices etc. In the end the dialogue management will produce an answer back to the user. The dialogue management is maybe the most complex component in the dialogue system. It controls the interaction between the system and the user, and it also cooperates with several
other components of the system. In DEAL, the tasks of the dialogue manager are distributed over several modules, a *discourse modeller* and a set of *action managers*. The discourse modeller is called Galatea (Skantze, 2007). Galatea makes an interpretation of utterances in context and also keeps a chronological list of the communicative acts (CA). Galatea also keeps a list over mentioned entities and information about them in the dialogue. The action managers initiate system actions, such as answering the users’ questions, promote certain objects or suggest prices. The discourse modeller and action managers are separated because the discourse modelling is more general while action selection is more specific for the system domain. The DEAL action manager (DAM) cooperates with a *communication manager* (CM) to decide on what to say on an abstract semantic level.

One requirement in DEAL is to keep a constant and not too long response time. Too long delays in the responses from the dialogue system might confuse the user who might think that the system did not hear what was said and try again. The dialogue system therefore needs to hold the turn while producing the rest of the output message, and then give the turn back to the user. This is done by the CM. While CM holds the turn, the DAM constructs a response for the CM to modify to fit the context and then pass on.

**Speech output**

Ovidius, the *surface generator*, operates on the communicative act sent by the CM and generates a text. It works similar to Pickering but inverted, using grammar rules to map semantic structures into text strings. The *speech synthesizer* realizes this text acoustically. First it analysis the text and decides upon different parameters how to generate the message, and then it generates speech (McTear, 2002).
2.3 Human-like language generation

In virtual games it is often desirable if the game agent possesses a believable behaviour, and as a tool for learning languages in DEAL, a system speaking in a human-like way is to prefer. Researchers are increasing their efforts trying to model conversational language. However, the output components have not received the same amount of attention as the input components of spoken dialogue systems. One possible reason is that some characteristics of spontaneous speech are regarded as flaws of human language, and therefore seen as irrelevant for task completion. A good landmark is to create a system which is able to produce what it is able to understand. In the DEAL project, the development of generating human-like utterances is in process. This part describes spontaneous speech, with focus on features important for generation of conversational output in spoken dialogue systems.

2.3.1 Human-human interaction

What distinguishes a dialogue from a monologue is that it consists of at least two participants contributing to discourse (Clark, 1996). Conversations can exist face-to-face or over telephone. We use the spoken language mostly to interact with others (Brown & Yule, 2004). While written language often is used to transfer information, we use speech to establish and continue relationships. Humans easily express complex meaning, engage in social relationships and solve problems by using spoken language. Spoken language is constructed by units of speech, which often are called utterances. An utterance is often defined as a sequence of speech made by one person, and before and after this sequence of speech is silence (Schiffrin, 1987). An utterance can therefore vary in size from a single word or lexical item to longer sequences of words. When communicating through speech, other signals than words are also present (Clark, 1996). Gestures are often used, using your body to signal extra cues in addition to the discourse. Examples of regular gestures are ‘head nod’ signaling yes, ‘head shake’ signaling no and ‘thumb up’ to signal approval. Below are more characteristics of human-human interaction, characteristics that need to be considered when developing spoken dialogue systems within a human metaphor.

Mixed-initiative

A typical characteristic of conversation is the two-way flow of information (Nickerson, 1976). If this feature does not exist, it is not a dialogue, but a monologue. However, the flow of information can be more or less controlled by one of the participants. In many dialogue systems the system is in charge of the flow, which may not feel like a conversation. A dialogue system can be controlled by the user, by the system or by mixed-initiative (McTear, 2002). If the system is mixed-initiative, each participant can take control of the conversation by for example asking the other participant questions and not only answer them. The control will change during the session. A spoken dialogue system offering mixed-initiative must have information about the interpersonal roles of the participants and also have control over dialogue history, task and goal. A conversational dialogue system needs to be mixed-initiative in order for the user and the system to collaborate, discuss and explore a specific task (Allen et al., 2001).

Presence

In a conversation, a speaker is often aware of if and to what degree the other participants are giving him/her attention or not (Nickerson, 1976). The participants give signals confirming attention in various ways, by uttering responses or by using facial expressions. Acknowledging response or other responses are important in dialogue systems. Slow responses may make the user unsure of if the system is about to generate an utterance or not. When time passes the user may wonder if his/her input was understood correctly, or if the system did not hear the message. Silence is an issue. It is therefore important for a spoken dialogue system to generate responses while processing the input information and incrementally producing the answer. Being capable of
giving responses, such as verbal feedback, while the user speaks, is also a feature which may strengthen the human metaphor (Edlund et al., 2006).

**Common ground**

In conversation there are different sorts of knowledge. *Common knowledge* is the knowledge about the world around us, and without it conversations as we know them could not occur (Nickerson, 1976). *Shared knowledge* is knowledge brought to the conversation that contains information about each other and other special information shared by the participants. However, common knowledge is also created during conversations. *Common ground*, the things we know about each other and that we know that the other one knows, needs to be established in order for the people participating in a conversation to understand each other. Grounding is the processes responsible for trying to establish a mutual understanding (Clark, 1996). When a message is understood by the hearer, the hearer often gives signals to ensure that he/she heard and understood the message. A response (*yes, mm, uh huh*), nod or a smile may signal understanding and acceptance of what has been uttered by the speaker. Signal of understanding may also be done by the hearer when initiating the next turn, by repeating some part of or the entire message. Responsive cue phrases may be used both to show presence and to signal understanding. This will be described in the section called *Cue phrases*.

**Referring expressions and fragmental utterances**

A message in dialogue can be generated through full sentences, elliptical constructions or by using different anaphoric expression (Schlangen, 2003). However, in dialogue complete sentences are seldom used. The speaker may use pronouns when referring to entities instead of using full names of people and objects. However errors can occur if the listener do not identify or do misunderstand which entity is referred to. Utterances are often fragmental, as when using ellipsis. Using fragmental utterances is efficient instead of always answering questions with whole sentences.

- Who bought the car?
- Erik

**Error-handling**

Spoken conversation contains several sources of uncertainty. Yet humans have a significant capability of adapting from difficulties, prevent, detect and recover from problems occurred while speaking (Skantzze, 2007). Speakers may produce ambiguous referring expressions and fragmentary utterances which may be hard to understand. There is always a risk that the listener will not understand or misinterpret the message. Humans can easily repair and recover when miscommunication has occurred. If the listener does not understand or accept the message from the speaker he/she may give negative evidence of understanding by asking the speaker to repeat the message. If the hearer feels unsure that he/she has understood the message, a clarification request may be used so that the speaker can repeat a specific piece of the utterance. See example from Skantzze (2007 p.39);

- I have a red building on my left.
- Did you say the building was red?

The speaker may also correct the hearer if noticed that the hearer has misunderstood the delivered message.
Turn-taking

A speaker turn can contain several utterances (Yule, 1996). A turn may be described as one or several utterances in a sequence produced after and before a shift in speaker within the conversation. The order of which they speak is not predetermined but usually it is clear for the participants who are in control. People are often good at recognizing when to speak. The participants often wait until the one talking shows that he or she has finished before taking the turn. There are several ways for a participant to signal that he or she has finished a turn. Strategies to signal a completion point is for example to end a turn with prosodic features such as silent pauses or using specific intonation patterns (rises and falls in the intonation) (Edlund & Heldner, 2005). Another way is to formulate the utterance so it is perceived as a question. This may be done using cue phrases such as or, then or well, attached at the end of the turn. These cue phrases will be described more deeply in the section Cue phrases (2.4). There are also ways to indicate for a participant that he or she wants to take the speaking turn. Speakers can use body shifts or facial expressions to signal that they have something to say, and they can also start to make short repeated sounds while the speaker is talking. In most cultures, there tends to be an avoidance of silence between speaking turns. Usually one of the participants stops talking when two people start to talk at the same time. Below is an example by Yule (1996 p.128):

A: Didn’t you *know why-
B: *But he must’ve been there by two
A: Yes but you knew where he as going

Here person A stops to speak when B start speaking. [*] is showing where speak is overlapping.

The way participants in a conversation take the turn may affect how they are perceived by the other participants (Yule, 1996). A person that often cuts in on other speakers might be seen as ‘rude’, and a person that mostly keeps waiting for a chance to take a turn that never seems to occur might be seen as having a ‘shy’ personality. This knowledge of turn-taking might be used as strategies when participating in a conversation. There are also other strategies, avoiding normal completion points, when trying to hold a turn. Examples are prosodic or phonetic features such as “held” articulation by lengthening the vowels (Edlund & Heldner, 2005). We often use this strategy when we need to work out what to say while actually saying it. Since we expect a completion point to be a pause or a marker that shows the end of a turn, the strategy to keep the turn is then to avoid these markers and unfilled pauses (Yule, 1996). Instead connectors might be used, such as and, and then, so, but to show for the other participants that there is something more to say, that the turn is not yet finished. Connectors will be described more deeply in the section Cue phrases (2.4). Another way to avoid completion points, is to place the pauses at points where it is semantically obvious that the message still is incomplete; don’t pause at the end of a sentence. To make it more clear, the one keeping the turn might ‘fill’ the pauses with hesitation markers such as er, umm, uh, ah etc.

These types of strategies may be seen as undesirable in language, since they are not grammatically correct. But they are often present in spontaneous speech and used by the most people, and they can be seen as part of speech, what makes conversation work, since we use them to control the interaction (Yule, 1996). The strategies are part of spoken social interaction. Other ways of signal turn-taking are by facial expressions, eye gaze, syntax and intonation (Clark, 1996).

2.3.2 Incremental generation

The fact that humans can make use of languages has been very important for human development. Without using a language we could not have created music or even performed warfare (Yule, 1996). Human language is unique and most of the time we use it in speech. For the body to produce speech, air passes out through our mouths. We move our mouths and depending on how the mouth is moved, we produce various speech sounds, such as vowels and consonants. These speech sounds, produced after each other in sequence, become utterances.
Written language always has the possibility of being changed and corrected subsequently. This is one major difference between spontaneous speech and written text. Clark (1996) argues that speaking and listening are participatory actions, and that these actions are not independent of each other. In a dialogue speech is produced incrementally; we start to produce speech before a complete utterance has been planned, using information from different sources in parallel (Brennan, 2000). When we start talking we do not have pre-planned utterances ready to be generated. We plan what to say and how to say it while talking. This may lead to generation of disfluencies such as unfilled pauses (silence), filled pauses (eh, hm), truncated words, restarts, mispronunciations and corrections (Clark & Wasow, 1998). Such disfluencies are frequently used in spontaneous speech and are often considered “flaws” or “mistakes” in the message to be delivered. However, Brennan (2000) has studied the comprehension of disfluent speech and showed that disfluencies may bear valuable information, helping the hearer to understand the message in a certain context. Disfluencies may also help to improve, correct and change what we say as we speak. They are used more frequently in longer utterances, since these are more difficult to plan in advance. Spontaneous speech also contains non-lexical features such as laughs, hesitations, hawks and sighs.

2.3.3 Incremental speech in spoken dialogue systems

Unlike humans, dialogue systems often speak in grammatically “correct” utterances. Since DEAL is made for conversation training for foreign students learning Swedish, it is important that the dialogue system can generate human-like speech. It might encourage the students to speak more, and to speak more freely. When trying to learn a language, the more you talk the better it is. As could be read about entrainment (part 2.1.3), research has shown that users of spoken dialogue systems often adopt the way the system speaks (Gustafson et al., 1997). In DEAL this might be crucial since it is important for the system to speak human-like and not only in short command-like utterances if the language learning student will be capable to adopt a good Swedish language (Hjalmarsson et al., 2007).

If the system should be able to produce human-like language, the generation of speech needs to be incremental, or at least be perceived by its users as producing speech online as the dialogue progresses. At the moment there is research at KTH about how to redesign the modules and architecture of the system in DEAL to process information in an incremental way. In most spoken dialogue systems the process works like this: The speech recognizer segment the input into utterances. This segmentation is based on a silence threshold. The interpretation of the utterance from the process is the base from which the dialogue system produces an output. First the system produces a communicative act (CA). The CA is then changed into text and then synthesised to give the output to the user. As mentioned earlier, this is not how humans process speech. Humans start planning what to say next while the other person is still speaking. When a person starts to speak he/she often has an idea of what to say but not exactly how to say it. Our input and output process works in parallel, and we can plan what to say while realizing new contributions.

DEAL needs to manage dialogue in an incrementally way since it should be capable of producing human-like speech. The dialogue system in DEAL needs to start processing the input before the user has finished his/her turn and the dialogue system also needs to start to produce output before it has decided exactly what to say and how to say it. To be capable of generating speech incrementally, the system needs to be capable of returning initial fragments of utterances immediately, before it has processed the entire utterance (Callaway, 2003). The dialogue system needs to give a response to what has been said by the user, before a too long pause passes by. Thus, the dialogue system needs to generate a response to the user before it has a complete plan of what to say. One strategy is to mimic human disfluencies in spontaneous speech. Silent pauses, filled pauses, corrections and repetitions of what has been said can be produced to buy time while speaking. It signals that we will soon say something or that we need time to think of what
to say or how to continue what has been started and meanwhile the system has time to plan the rest of the utterance. The system also needs to allow realistic turn-taking and give feedback in order to generate speech incrementally. Another demand for a system generating speech incrementally is to be capable of keep track of what has been said earlier in the dialogue to be able to correct itself and to handle how input and output speech relate to previous discourse (Clark & Wasow, 1998). The system also needs to handle interruptions (Skantze, 2005). If the user barges in, in the middle of the systems utterance, the system needs to remember what it planned to say and what it managed to say before the interruption occurred. Using cue phrases may be a good way of beginning to speak before a whole utterance is planned. Cue phrases may also be used to handle turn-taking and to relate speech to previous discourse. Cue phrases are important features of human-human interaction and may also be useful in spoken dialogue systems. Next, cue phrases will be described in more detail.
2.4 Cue phrases

What are cue phrases? First, many names are used to describe them, for example they are also called discourse markers, pragmatic markers, connectives, and discourse particles. In this report these markers are referred to as cue phrases. They are linguistic devices used to signal relations between different segments of speech.

Consider this sentence (Lindström, 2008 p.56);

“**men** det är väl inte så konstigt **eller**”
(Eng: “**but I suppose** it is not so strange **or**”)

The bold words are examples of cue phrases. If removed, the utterance will look like this;

“det är inte så konstigt”
(Eng: “it is not so strange”)

According to Lindström’s definition of cue phrases, the proposition of the utterance above is not changed, and nor is the grammatical structure of the utterance when removing the cue phrases. The removed words, however, modify the utterance in different ways, and also regulate the conversation in different directions. Thus they are important in spontaneous speech.

Cue phrases are commonly used in spontaneous speech. They are also used in written text, but appear more frequently in spoken language (Louwense & Mitchell, 2003). When reading about cue phrases there are different opinions in how these phrases should be defined, which lexical entities are included and how they can be categorized depending on their origin, part of speech or function (Schiffrin, 1987; Fraser, 1996; Lindström, 2008). The features of cue phrases and which words are included may differ between languages, but there may be a general correspondence between them. Even if these features vary between languages, there are also differences in opinion between researchers within one language. However, the underlying importance of cue phrases in language is an opinion which unites researchers.

Louwense & Mitchell (2003) try in their work, called *Taxonomy of discourse markers in dialog*, to reveal why spoken discourse contains so much more cue phrases than written discourse. Based on a survey, they identified that spoken discourse contained ten times more cue phrases than written. Moreover, spoken dialog had almost twice as many cue phrases as monologs, and informal dialogue had twice as many as formal dialogue. Since spontaneous conversation is created by two or more participants the dialogue might take unexpected directions. Accordingly, the participants cannot plan their messages in advance and such a complex structure might require the use of cohesion cues. Cohesion cues enable the participants to comprehend the information being communicated within the conversation. As an example, the participants need to give feedback and acknowledgement, fill in words and interrupt each other, to show that they are still following the conversation and participate in the interaction. Also Louwearse and Mitchell argue, as does Schiffrin (1987), that cues may also be communicated through non-verbal items such as nods and gestures or variations in intonation. Spoken dialogue is not as structured and organized as written text, therefore we need to compensate with a higher use of cue phrases. Theories of two important researchers in this area, Schiffrin and Fraser, will be described next.

2.4.1 Schiffrin

Early work in the area of cue phrases was carried out by Schiffrin (1987). By recording sociolinguistic group interviews and carefully analyzing these utterance by utterance she was one of the first researchers to do thorough empirical studies of these entities which she refers to as discourse markers. Schiffrin had two different definitions of cue phrases; operational definition and
experimental definition. Her operational definition is that cue phrases are sequentially dependent elements which bracket units of talk. Schiffrin uses units of talk in her definitions because it includes sentences, speech acts, propositions, speaker turns and other forms that an utterance may take. A speaker turn can consist of one or several utterances. Since cue phrases are bound to all these forms of utterances, Schiffrin uses units of talk when defining the use of cue phrases. By brackets Schiffrin suggests that cue phrases mark the boundaries of units of talk in some way. These cue phrases can mark a unit as long as a conversation or as short as a single word. They can both initiate and terminate a unit, and Schiffrin suggests that they do different kind of work depending on their position. A cue phrase positioned initially may signal how the new information relates to previous discourse and define what kind of message will follow. By sequential dependence Schiffrin suggests that cue phrases are not dependent on smaller units of talk which compose the discourse, but work on discourse level. Schiffrin also suggests that these cue phrases relate syntactic units of language and make them fit into a context, whether it is a textual context or a context of spoken discourse. The most common cue phrases used for sequentially dependent reason are and and but.

The items Schiffrin (1987) group together as cue phrases are oh, well, and, but, or, so, because, now, then, I mean and y’know. However Schiffrin mentions that it is difficult to make them members of a single word class called cue phrases, since they are member of other word classes such as conjunctions (and, but, or, because), adverbs (now and then). However her second definition, the theoretical definition is that cue phrases are members of a functional class of verbal (and nonverbal) devices. These devices provide contextual coordinates for ongoing talk. Schiffrin suggests that members of this functional class are members because they possess certain characteristics which make them fit into the first definition, i.e. as cue phrases that sequentially bracket units of talk.

According to Schiffrin (1987) the structural relations and meaning relations in spoken discourse, or in texts, are still there even if no cue phrase is used. However, she suggests that once a cue phrase is used, a specific relation has been chosen to the exclusion of another relation. This will explain Schiffrin’s opinion, that cue phrases do not create structure relations in discourse, they only display these relations.

2.4.2 Fraser

Fraser (1999) uses the term pragmatic markers or pragmatic formatives and defines them as lexical expressions which do not contribute to the propositional content of an utterance but that signal different types of messages. Fraser also describes a third type of pragmatic formatives which he calls ‘commentary pragmatic markers’ that includes what he calls discourse markers. He further defines these markers as sharing one common property: that they force a connection between the discourse segments that they are part of and previous discourse segments. If a pragmatic marker is positioned in an utterance, it relates this segment to prior discourse. Fraser yet excludes some expressions to be called pragmatic markers, for example expressions such as frankly and obviously, which do not signal connections between discourse segments. Instead they signal a separate message, a comment, to the following segment. These are the markers he calls ‘commentary pragmatic markers’. Among other specific expressions, he also excludes particles such as even, only, just and markers such as hmm, well and oh to be pragmatic markers.

Fraser (1999) characterizes cue phrases semantically as relating two discourse segments, while not contributing to either of their propositional meaning. Cue phrases signal the connection of the basic message to previous discourse (Fraser, 1996). They are distinct from the propositional content of an utterance, but give clues signalling the possible communicative intentions of the speaker. This characteristic is similar to the one suggested by Schiffrin. Another characteristic is that the meaning of a cue phrase is procedural and not conceptual (Fraser, 1999). This means that they only provide an instruction on how to interpret the utterance to which the cue phrase is attached to (Fraser, 1996). A third characteristic about cue phrases are that every individual cue phrase has a particular core meaning. When using the marker so, it signals that the following
utterance is to be interpreted as a conclusion which follows from earlier discourse. Cue phrases often have an initial position in an utterance.

Cue phrases contribute to how we interpret an utterance, and not to the propositional content of the utterance. The greatest difference between Schiffrin’s and Fraser’s opinions about the items called cue phrases, is that Schiffrin also includes non-verbal items in the definitions of cue phrases because she suggests they work in the same way as verbal cue phrases. Fraser does not include non-verbal items in his definition, only verbal, lexical items.

2.4.3 Further research on cue phrases

Lindström (2008) has studied cue phrases in Swedish and Swedish-Finnish language. He uses the term discourse markers and defines them as a description of words or expressions that regulate the conversation or works as signals modifying an utterance. Lindström’s opinion about these markers is also that they may be excluded without changing the main utterance, concerning both its proposition and its grammatical structure. Even though these markers do not affect the main proposition of the message, they still affect how the message is interpreted by its receiver. These markers signal how a message is relevant in discourse and how the speaker relates to his/her message and what sort of response he/she expects to achieve.

Lindström’s definition is similar in many ways to the traditional descriptions of cue phrases by Schiffrin (1987) and Fraser (1996). However, Lindström has a broader view on cue phrases. While Fraser (1999) only includes expressions which signal connections between discourse segments, Lindström (2008) also include cue phrases which can signal a modification of the present utterance. Lindström also sees some disfluencies as cue phrases, including filled pauses such as eb, ab and bm. He also includes repairing phrases, used when the speaker needs to correct what has been said.

Schourup (1999) characterizes cue phrases as containing features such as connectivity, optionality, weak clause association, initiality and orality. The connectivity feature reminds of how the researchers mentioned earlier define cue phrases; i.e. as sequentially dependent (Schiffrin, 1987) and relating two or more utterances (Fraser, 1996). A cue phrase may then be used to relate expressed propositional content in a current utterance that may not have been expressed in a previous utterance. By optionality Schourup (1999) suggests that cue phrases are almost “universally regarded as syntactically optional”, which means that a cue phrase can be removed without affecting the syntax of the main utterance. And even if removed, the relationships between the units are still there, but no longer clearly cued to the hearer. With weak clause relation Schourup implies that cue phrases occur either outside the syntactic structure of the utterance or loosely attached to it. Another characteristic is initiality, which suggests that cue phrases are positioned initially in segments of discourse. Orality suggests that most of these cue phrases occur primarily in speech.

Louwarse & Mitchell (2003) describe cue phrases as verbal and non-verbal devices that mark transition points in communication. They suggest that using cue phrases facilitates the construction of a mental representation of what is described in the message. Cue phrases are used to instruct the participant in the conversation how to consider an upcoming utterance. Cue phrases are used to bring the interactions between the participants together by drawing the attention of a participant, serve as turn-taking signals or to signal agreement or other responses. Nods and gestures are usually used for similar purpose, and might therefore also be considered as cue phrases.

More recent research of cue phrases tends to include more words, non-verbal phrases and facial gestures. The focus is on the function and purpose they might have in different contexts. Even if some characteristics are similar or resemble each other in the different definitions of cue phrases,
they are still different. This fact means that an item may be described as a cue phrase in one definition but not in another, which may create problems when trying to analyse them.

2.4.4 Cue phrases in dialogue systems

*Flaws* (hesitations, restarts etc) and cue phrases in spoken language can make the user perceive the system as more human-like, polite and intelligent and also be used to hide short-comings of the system (Edlund et al., 2006; Hjalmarsson & Edlund, in press). If the goal is generating incremental speech in spoken dialogue systems, the utterances will not be planned completely before generated. As mentioned earlier, human spoken interaction involves features such as turn-taking, disfluencies and grounding. Using cue phrases in spoken dialogue systems may be an effective or even necessary strategy to cope with some of these features.

It is not easy to find research on cue phrases within spoken dialogue systems. Nor is it easy to find research on how cue phrases can be generated in spoken dialogue system. More attention has been given to developing sufficient models for automatically identifying cue phrases in spoken dialogue (Heeman et al., 1998). Heeman et al. (1998) suggest that such models might help the speech recognizer in the spoken dialogue system to enhance recognition of spontaneous speech. Grote & Steede (1998) argue that the selection of cue phrases is important when there is a wish to produce cohesive text in automatic text generation. They suggest a specialized lexicon for cue phrases. This lexicon will list relevant constraints and preferences associated with each cue phrase. The lexicon will then be used in the generation process, choosing ways of expressing relations in the context among alternative cue phrases. A similar lexicon may be created for generation in spoken dialogue systems.

**Using context and prosodic features for automatic identification**

When developing models for automatic identification of cue phrases in spoken discourse, the models use known features and functions of the cue phrases in order to identify them (Heeman et al., 1998). Researches are therefore enthusiastic to find and model these features for each cue phrase with the possibility to develop an identification model. These features may also be used when adding cue phrases in generation components of dialogue systems. The context in which they occur may be used to indicate where and in which utterances cue phrases should be generated and their acoustic realizations may be used to indicate how cue phrases should be generated in order to approach human-likeness. Analyzing cue phrases is an important step to better understand the content and structure of the dialogue (Zufferey & Popescu-Belis, 2004).

Studies of detecting valuable features of cue phrases (or other features of the spoken language) for automatic identification in spoken dialogue systems often begin with analysis of human-human dialogues. Analysis of human-human conversation is valuable to see how human use language to interact in real dialogues (Jönsson & Dahlbäck, 2000).

**Methods for automatic identification of cue phrases**

Heeman et al. (1998) present a method for identifying cue phrases in spontaneous speech. The method is based on machine learning and incorporates part-of-speech (POS) tagging. Heeman et al. suggest that adding POS tagging in the language model will be an improvement since these can be used to identify cue phrases. For the automatic identification of cue phrases, they used a decision tree learning algorithm. This decision tree includes binary questions about the words and POS tags. They suggest that each POS tag should have a word classification tree so it will be capable of dealing with words with several functions. Their annotating strategy for cue phrases was to use four POS tags; one for detecting single words of acknowledgements (*okay, right, yes, no*), one for detecting interjections with discourse purpose (*oh, well, hm, mm, like*), one for detecting coordinating conjunctions used as cue phrases (*and, but, so*), and the last one to detect adverbials used as cue phrases (*then, now, anyway*) (Heeman et al., 1998).
Contextual and prosodic features
Zufferey & Popescu-Belis (2004) focused on the word *like*. They tried to classify the word into its discourse- and sentential function, and compare the classification results for human annotators and automatic classification. As an example, *like* operates as a cue phrase, but also has other functions such as being a verb, a preposition, etc. As a cue phrase *like* introduces an approximation of what is being said (*It took* *like* five minutes). So what is important is being able to separate *like* as a cue phrase from other functions of *like* in a dialogue. Zufferey and Popescu-Belis first performed two experiments to evaluate the performance of human annotators and their ability to separate *like* as a cue phrase from its sentential meaning. The classifications were made on a transcription from a natural dialogue and on a transcription of a movie dialogue. In the first experiment the human annotators only had the written transcription, while in the second experiment they were also able to listen to recordings of the utterances while annotating. Results showed that in the first experiment Kappa coefficients are quite low, $\kappa = 0.40$ for the natural corpus and $\kappa = 0.65$ for the movie corpus. The difference in inter-annotator agreement between the dialogue corpus and the movie corpus shows that it is easier to annotate pre-planned dialogues, because they are less ambiguous. When adding the acoustic information in experiment two, inter-annotator agreement increased to $\kappa = 0.74$. This indicates that there is acoustic information which helped the human annotators in this task.

For the automatic classification Zufferey & Popescu-Belis (2004) studied the method of using collocation to erase instances where *like* cannot be a cue phrase, for example as in *I like, seems like*, *feel like, just like* etc. The results show that using collocation filters is an efficient method. Zufferey and Popescy-Belis also tried adding prosodic cues such as duration of the spoken word *like* and the duration of the pause before and after the word. This shows, using collocation-based filters is more efficient than using only time-based features. However, they suggest that this method may be more efficient for other cue phrases. Results from their study show that human annotators perform better when using prosodic information while annotating. For the automatic classification of *like* using mainly collocation filters was the most effective feature. Overall, the performance of the automatic identification was close to the human annotators. Kappa value was 0.63 for the automatic identification (compared to $\kappa = 0.74$ by the human annotators). One interesting result from this experiment is that human annotators performed considerably better when they were able to use acoustic cues, listening to the sound files while annotating. However in the automatic identification and classification of *like*, the prosodic cues used were not improving the results. This suggests that maybe other prosodic features than duration is needed.

Heeman et al. (1998) examined the role of cue phrases in the beginning of utterances, excluding acknowledgements. In the Train Corpus, 68.2% of the utterances were initiated by a cue phrase. When excluding the acknowledgements and turn-initial filled pauses, and also turns that only consist of cue phrases or filled pauses, 44.1% of the turns began with cue phrases. In earlier studies of cue phrases in task-oriented human-human dialogues Byron & Heeman (1997) investigated the relation between cue phrases and speech acts (function form of the utterance, such as *Request, Respond, and Confirm* etc. (Clark, 1996)) (Heeman et al., 1998). No strong relations between cue phrases and particular speech acts were found. Hence, they also studied relations between cue phrases and prior speaker turn. Results show that cue phrases are less used when the previous turn is the initiation of an adjacency pair. Cue phrases are more frequently used when the previous speech act functions to acknowledge, respond or inform, or when it concludes an adjacency pair. Then the next utterance may use a cue phrase to relate to prior discourse. This suggests that there is a higher probability that a cue phrase is used when there are no strong expectations about the upcoming utterance.

Speech repair and intonational phrases
Heeman & Allen (2000) studied how identifying speech repair and intonational phrases in spoken language can help to identify cue phrases. They suggested that a better result would be obtained for each task if solving them simultaneously. Cue phrases have been found to often be used in the editing term while speaking. Using cue phrases helps to signal that a speech repair has occurred. Speech repairs occur frequently in spoken language. As an example, the cue phrase *ah* is
often attached to or within speech repairs. Identifying cue phrases will help detecting the speech repairs, but also the other way around. Cue phrases are also often used at utterance boundaries, and consequently interact with intonational phrasing. Phrase boundaries and speech repairs correlate with acoustic cues, such as in pauses between words. These cues have been treated as noise, but could instead give evidence on how and when these events occur. This might improve speech recognition and also help later processing, such as syntactic and semantic analyses. The results show improvement in detecting cue phrases when using speech repairs and intonational phrases as features. Heeman and Allen suggest that adding more acoustic cues, rather than only using pauses of silence, may improve the results even more.

In the identification studies, the goal is to automatically identify all cue phrases in an utterance. However, the generation component in DEAL does not need to be capable of generating all types of cue phrases. The goal is to find non-ambiguous cue phrases with different functions in the language. To find such cue phrases for use in DEAL, dialogues in a similar domain need to be analysed. The cue phrases need to be categorized by function and then each category needs further analyses to extract what characterize the cue phrases in the different classes and in which context they are used. A proposed way to categorize cue phrases according to function is described in the next section.

2.4.5 Classification of cue phrases by Lindström

Lindström (2008) divides cue phrases into two main categories; discourse particles, markers which primarily or only function is to regulate the discourse, and secondary discourse markers, words from other word classes which also function as cue phrases. Discourse particles offer responses, by giving feedback to the person speaking. Discourse particles are also used to mark transitions between utterances and within utterances. They also give focus to an utterance, or to a single item within an utterance. They may also convey how the speaker feels about his/her utterance. These particles may have several functions depending on their position within the utterance and depending on the context of the utterance. Example of particles are ja, jo, mm, nej, jaha, jasså, okej, vellå, förstår du, eh, då, va, liksom, typ, ju, väl and nog (Eng: yes, no, okay, alright, you know, like, as, suppose). Secondary discourse markers consist of words or expressions from other word classes, which may be used in a way which resembles the function of discourse particles. The boundary between discourse particles and secondary discourse markers is not definite. Some of the secondary discourse markers are used so frequently to regulate the discourse that they are on their way to function primarily as discourse particles. Example of markers are hallå, och, men, eller, alltså, försthen, faktiskt, eivärk, titta, tycker jag and sån här (Eng: hello, and, but, or, therefore, by the way, actually, exactly, look, I think, like this)

Classification according to function

Lindström (2008) also presents a classification scheme based on the function of the cue phrases. This way of classifying cue phrases is presented and described below with examples of the most frequently used Swedish words within each category. The classification scheme will be described one main category at a time.

Responsives

This category of cue phrases can stand alone but is more frequently used attached to an utterance at initial or terminal position, or within an utterance (Lindström, 2008). The responsives are among the smallest units in Swedish language that may be used to interact and participate in a conversation. They may have different sub-functions. For example answers to yes/no questions have a responsive function, and so do feedback and acknowledgement, which signal to the speaker that the hearer keeps up with what is being said. The listener agrees to what the speaker talks about or at least confirms that he/she is listening (responsive). The use of these cue phrases may also be used to show uncertainty and doubtful feelings of what the speaker is saying, or simply rejecting what has been said (dispreference). Another function of these cue phrases is to show the speaker that what is being said is new to the listener (new information).
### Appealing

This class of expressive cue phrases are used to call for attention. (Lindström, 2008). They are used to get attention from the hearer when changing the direction of the discourse, e.g. “Wait, there is something I have to tell you”. The appealing cue phrases signal that something surprising or some new information will be communicated. This class is often used when the speaker wants to introduce a new topic into the conversation, or change the topic in the discourse and he/she needs attention from the hearer.

<table>
<thead>
<tr>
<th>Function</th>
<th>CUE PHRASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appellerande/Appealing</td>
<td>dude, hörd, hallá, hej</td>
</tr>
<tr>
<td>Allmänt appellerande</td>
<td>you, listen, hello, hi</td>
</tr>
<tr>
<td>General appealing</td>
<td></td>
</tr>
<tr>
<td>Point marking</td>
<td>větu, serdu, förstådu, se, tank</td>
</tr>
</tbody>
</table>

*Figure 2.4: Appealing cue phrases*

### Response Eliciting

These particles are often attached to a mother utterance, initially, within the utterance or at a terminal position (Lindström, 2008). They seldom stand as an utterance alone. Some of these particles function as appealing the hearer or searching for a response; trying to draw attention from a hearer and get his/her understanding in what has been said, trying to reach common knowledge. One of these particles is the Swedish word *va* (Eng: huh, what) which is positioned at the end of an utterance. By attaching *va* to an utterance the speaker demand for a response from the hearer. The Swedish word *då* (Eng: then) has a similar function. When used as a cue phrase it is attached at the end of the utterance and is used to clarify the utterance as a question or assumption.

<table>
<thead>
<tr>
<th>Function</th>
<th>CUE PHRASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fragande/Asking</td>
<td>va</td>
</tr>
<tr>
<td>Kontrollerande</td>
<td>What</td>
</tr>
<tr>
<td>Elicitierande</td>
<td>va, väl, eller, då, förstådu</td>
</tr>
<tr>
<td>Response Eliciting</td>
<td>isn’t it, or, then</td>
</tr>
</tbody>
</table>

*Figure 2.5: Asking functions of cue phrases*
Connectives
These cue phrases are conjunctions which mark relations in discourse (Lindström, 2008). These may be relations of attachment (att, och) (Eng: that, to, and) which indicates that the following utterance is an addition to the previous message. These cue phrases may also be relations of contrast (fast, men) (Eng: though, but) between utterances, signalling that the speaker says something from another perspective than what was previously said. These cue phrases are mostly positioned initially in an utterance; but are sometimes also attached at the end of the utterance. These connective cue phrases can connect whole turns to each other. The connective word eller (Eng: or) introduces alternatives to what was previously said.

<table>
<thead>
<tr>
<th>FUNCTION</th>
<th>CUE PHRASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Konnektiv/Connectives</td>
<td>att, och, så, alltså, förresten, jamenar, då and, so, thus, by the way, then, I mean</td>
</tr>
<tr>
<td>Additiv</td>
<td></td>
</tr>
<tr>
<td>Additive</td>
<td></td>
</tr>
<tr>
<td>Kontrastiv</td>
<td>fast, men, i alla fall, i och för sig</td>
</tr>
<tr>
<td>Contrastive</td>
<td>though, but, anyway, anyhow</td>
</tr>
<tr>
<td>Alternativ</td>
<td>eller, jamenar</td>
</tr>
<tr>
<td>Alternative</td>
<td>or, I mean</td>
</tr>
</tbody>
</table>

*Figure 2.6: Connective functions of cue phrases*

Modifying
These particles are mostly positioned within the utterances but also have initial and terminal positions (Lindström, 2008). They are used much more frequently in spoken language than in written text. The modifying cue phrases have become more regularly used as cue phrases. These particles modify the utterance according to the truthfulness of a message. This class of cue phrases colour utterances, and invite the hearer to create meaning together with the speaker. The most frequently used modifying cue phrase in the Swedish language is ju (Eng: of course). The function of the cue phrase ju is to indicate that the knowledge about a situation is available, either through the speaker’s own experience, or through his/her and the hearer’s shared experience. The word väl (Eng: I suppose, surely) signals an assumption or a suggestion and an appeal to the hearer to cooperate on this assumption. By using väl the speaker assumes an approving response by the hearer. By using the word nog (Eng: probably) the speaker signals that he/she has enough reason to believe what he/she says is true or a least probably true.

The Swedish word liksom (Eng: so to speak, like) is used by the speaker to weigh the appropriateness of the utterance or the whole message. The word typ (Eng: like) has a similar function, but works more as an exemplifying word. The cue phrase bara, origin from the word bara (Eng: only, just), is mostly used by young people. This group contains cue phrases whose function is to give comments on the action of the message, but also cue phrases which act as evaluating, such as faktiskt (Eng: as a matter of fact) and tyvärr (Eng: unfortunately). The cue phrase faktiskt signals that the speaker is being frank about what was just said.

<table>
<thead>
<tr>
<th>FUNCTION</th>
<th>CUE PHRASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nyanserande/Modifying</td>
<td></td>
</tr>
<tr>
<td>Typifierande</td>
<td>ba, liksom, typ, sån här/där, så här/där so to speak, like</td>
</tr>
<tr>
<td>Typifying</td>
<td></td>
</tr>
<tr>
<td>Garderande</td>
<td>i alla fall, åtminstone, i och för sig, vetja, nu anyway, at least, now</td>
</tr>
<tr>
<td>Guarding</td>
<td></td>
</tr>
<tr>
<td>Evaluatorande</td>
<td>allvarligt (talat), faktiskt</td>
</tr>
<tr>
<td>Evaluating</td>
<td>seriously, as a matter of fact</td>
</tr>
<tr>
<td>Epistemisk</td>
<td>ju, väl, nog</td>
</tr>
<tr>
<td>Epistemic</td>
<td>of course, well, probably</td>
</tr>
</tbody>
</table>

*Figure 2.7: Modifying functions of cue phrases*
Repairing
The repairing cue phrases are often attached to a mother utterance, initially, within the utterance or at a terminal position (Lindström, 2008). They seldom stand alone. The Swedish cue phrase *ja menar* (Eng: I mean) acts as a correction, and is used to specify in more details what is said, or to reformulate what was previously uttered. This category also include a special type of cue phrases, cue phrases such as *eh, åh, ôh and hm*. These are used when the speaker has been interrupted, or when searching for what to say and how to say it. The repairing cue phrases are mostly at an initial position or within an utterance, but can also terminate a segment of speech when there is a wish to keep the turn.

<table>
<thead>
<tr>
<th>FUNCTION</th>
<th>CUE PHRASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reparandë/Repairing</td>
<td></td>
</tr>
<tr>
<td>Ordsökande</td>
<td><em>eh, åh, ôh, hm, dedär, liksom, vända eh, hm, like, wait</em></td>
</tr>
<tr>
<td>Korrigande</td>
<td><em>nej, ja menar, eller, alltså</em></td>
</tr>
<tr>
<td>Correcting</td>
<td>no, I mean, or, in other words</td>
</tr>
</tbody>
</table>

Figure 2.8: Repair functions of cue phrases

The categories and the words are translated into English; however this does not necessary mean that the words have the same discourse function in English. In this classification scheme the cue phrases’ syntactic function is not considered. Lindström (2008) mentions that this way of classifying cue phrases (into discourse-pragmatic function categories) might seem too strict. However, he believes that the advantage with this type of classification is that it gives a good overview on different discourse functions. However, this classification may more easily be used for other purposes, since it reminds of a dictionary.

2.4.6 Studies on responsive cue phrases
Responsive cue phrases are important in positive and negative feedback and for acknowledging and grounding purposes. The responsives function may depend on both their context and acoustical features. If the purpose is to generate responsive cue phrases in a dialogue system, it is important to study if the same lexical responsive might be interpreted differently depending on acoustical realization. Studies on the importance of context and acoustic features in identification and classification have been made for English responsive cue phrases.

Automatic identification of responsives
Gravano et al. (2007a) present a machine learning experiment where the task is to classify the discourse function of English single responsive words in a task-oriented spoken dialogue corpus. Examples of such responsives are *alright, okay* and *mm-hm*. Gravano et al. found that it was not enough to classify these responsives into 2 categories: cue phrases and sentential meaning. So in addition they added two tasks. One was to detect the cue phrases acknowledgement and agreement from all the rest of the words. The second task was to detect the boundary functions of discourse segments, i.e. to classify them according to their position; if they signal the beginning or end of a discourse segment. Three annotators classified the single responsive words in conversations of spontaneous task-oriented speech. The annotators could choose from 11 categories when classifying them. When annotating they had access to both transcript and speech.

Inspired by the work of Heeman et al. (1998) and the use of POS tagging for identifying cue phrases they designed the machine learning experiment. A number of features were extracted, for example contextual, acoustic and prosodic features from the affirmative words but also from the surrounding context. The features included text-based features, timing features, word acoustic/prosodic features and also acoustic/prosodic features from previous turn. The results show that models based on contextual features is close to the error rate of the human labellers.
However, using models with acoustic features only presented little improvement. According to the result by Gravano et al. (2007a) contextual features and text-only outperform the acoustic features.

A perception study
Gravano et al. (2007b) performed a perception study to find out the importance of contextual and acoustic features in the disambiguation of the three discourse-pragmatic functions of the English word okay. The results of the study showed that contextual features were more helpful in deciding the discourse function of okay, than the acoustic features from the word alone.

In their perception test subjects had to listen to a number of spoken okay from a recorded corpus. The perception test was designed to study how the subjects perceived different realizations of okay. Two versions of each stimulus were prepared, one with the surrounding context included, and one without. The stimuli with context contained two speaker turns for each stimulus, the turn including the token okay and also the previous turn. The stimuli in isolation consisted of only the word okay. The task for the subjects was to listen to each of them and then chose a suitable meaning out of three. They could choose from the meanings of acknowledgement/agreement, backchannel, and cue of an initial discourse segment. The test was divided into two parts (Gravano et al., 2007b). In the first part the subject task was to listen to the randomly presented isolated stimuli and categorize them. In the second part the stimuli were also presented randomly, but with its context. To help identify the tokens of okay the test subjects were shown the transcription while listening.

The word okay can have many functions in the English language. It can be used to convey an acceptable evaluation (the movie was okay) (Gravano et al., 2007b). It can also be used in dialogues to indicate that a hearer is listening to what the speaker says or signal agreement or acknowledgement or in its ‘cue phrase’ use, to start or finish a discourse segment. This study was done in order to find out how speakers may recognize and interpret these variations in meaning of the word okay. Gravano et al. wanted to find out the role of the discourse context when classifying cue phrases, or if they can be classified simply from the word alone. The inter-annotator agreement in labelling was higher in the condition of context than of isolated stimuli ($\kappa =0.120$ for isolated stimuli and $\kappa =0.293$ for stimuli in context). The results suggest that context, when available, plays a central role in the disambiguation of okay. Acoustic, prosodic and phonetic features of these classifications were examined trying to detect how listeners interpret the tokens. The only acoustic cue that improved the results (without contextual cues) was the feature of pitch digression at the right edge of okay.
3 Method

This section describes the methods used for this master’s thesis. First a data collection of spoken human-human dialogue is described. The dialogues were transcribed and labelled with cue phrases, communicative acts, disfluencies and non lexical speech sounds. The last part describes a listening test, carried out to find characteristic features of responsive cue phrases.

3.1 Data collection of spoken dialogues

If the user interacts with the spoken dialogue system using a human metaphor, the computer is perceived as an interlocutor with human-like conversational abilities (see chapter 2.1). The user may have expectations for the system to be able to carry a more human-like conversation. DEAL is a game where it is desirable to display a human metaphor. Since the choice is to make the system coherent with a human metaphor, it is preferable to initially study human-human interaction to find out how the system should be modelled. Analyzing human-human dialogues is a good way to study how humans interact with each other through speech. Recording such dialogues in a domain reminding of the system’s may also reveal which functions are needed for the system to meet the user’s general goals. Consequently the dialogues have a high validity (Jönsson & Dahlbäck, 2000). However, using human-human dialogues may also have drawbacks. The users may not have the same expectations when using a computer system as when they interact with another human being. The language used when talking to a person may also be different from talking to a computer. However, the system in development, DEAL, wants to approach spontaneous human-human dialogue. Thus, analyzing spontaneous dialogue in a setting reminding of DEAL is then a good start to reach such a goal.

3.1.1 The recording

Setting

The recordings took place in a studio. Trying to resemble the domain in DEAL, the setting was a simulated flea market. A little shop was staged, using a table in front of a wall. The shop had objects for sale in three different categories; tools, toys and clocks. The objects were pieces of paper in different sizes and colours. Some objects were in good condition and some were defective in some way. Toy money was used for the buying process. Two subjects participated in each dialogue. One subject played the role of a shopkeeper and the other played a customer. Some of the objects were pinned to the wall behind the shopkeeper and some were placed on the table. In some recordings, depending on the scenario, there was a photograph of a child and one of a couple, placed on the table. The shopkeeper was provided with a list with suggested prices of the objects. The role-play was recorded with two close-talking microphones, one microphone for each subject, and a digital video camera. A program called wavesurfer was used to record the dialogues.

Subjects

Six subjects participated in the data collection. There were four men and two women between 24 and 45 years of age. Two of these subjects work in the area of speech technology. Both subjects playing the role of the shopkeeper have experience in trading as a customer. One of them also has experience trading as a seller. Among the subjects playing the customer, two had earlier experience in trading; both of them as a customer, and one of them as a seller. Experiences in trading were mostly from experience of other countries than Sweden. The role of the shopkeeper
might take some time to feel comfortable with, so the two subjects who played this role participated in four scenarios each, while the subjects playing the role of the customer participated in two scenarios each.

Procedure
First the two subjects (one shopkeeper, and one customer) answered a questionnaire separately, answering questions on their experience in trading (see Appendix 10.1). Subsequently, the subjects were instructed separately (see Appendix 10.2). The instruction included information about the recordings, the scenario, and the subjects were also given a certain amount of money to shop for. Participating in this recording could almost be seen as participating in a game. Both subjects were given separate and conflicting goals to achieve. The subject playing the role of the shopkeeper worked hard not to reduce the prices of his objects, trying to earn as much money as possible. For the customer, the goal was to solve the task in the scenario, for example to buy some particular objects with a limited amount of money. These two goals were conflicting since the customer could not buy the objects without bargaining with the shopkeeper. The two subjects were not informed about each others goals. The two subjects were alone in the studio while performing the role-play. After each scenario the subjects separately answered another questionnaire about the role-play they just performed (see Appendix 10.3). Then they received new instructions about the next scenario and a new recording began. 3 different scenarios were used (see Appendix 10.2). The different scenarios targeted different type of objects and were used to try out possible scenarios that can be used in DEAL. Below is an example (translated from Swedish) of a scenario that was used:

Customer
You have recently bought an object for renovation. A nice, old house where both the building and the garden require proper renovations. However, you do not have any tools and before you go and buy new tools in the expensive hardware store, you have 350 Swedish crowns to spend at the flea market. Try to get at least 3 different objects to repair your new house. Examine the goods carefully, so you do not get swindled to buy something that does not work.

Shopkeeper
You are in a great mood and feel reasonable. If the customer finds flaws on your objects, you are willing to reduce the price.

A total of eight scenarios were recorded, each between 8 to 18 minutes. The overall audio recorded was 1 hour and 45 minutes.
3.1.2 Annotation of recorded dialogues

This part will describe the transcription and annotation of the dialogues in more detail. The transcription and annotation of the dialogues were made using the Higgins Annotation Tool, developed in Td at TMH by Gabriel Skantz. The dialogues were transcribed by hand orthographically, which means that the spoken utterances were written in a form that to a great extent makes use of the words in the written language. Non-lexical entities, such as laughter, hawks, hesitations and sighs were also transcribed.

Figure 3.1: The Higgins Annotation Tool

Annotation scheme for cue phrases

In this part the development of the annotation scheme for cue phrases will be described. Cue phrases were labelled according to their main function in given context. As described in part 2.4 cue phrases can be categorized in several ways. The first annotation scheme that was used was a categorization developed by Lindström (2008) based on cue phrase function. The scheme also included two cue phrase categories by Fraser; Inferential cue phrases and Elaborative cue phrases. They were added because the categorization by Lindström missed categorization of refinements, and conclusion of what had been previously said.

Two annotators independently annotated one of the transcribed dialogues according to the initial annotation scheme. The annotators had access to both the transcript and the sound recording of the dialogues while annotating. The annotated data showed that a lot of words were labelled as cue phrases; however there were some difficulties in using this annotation scheme for cue phrases. The annotation scheme was sometimes indistinct. It also had too many and too narrow categories. This led to insecurity as for which category to use and the annotation agreement
between the labellers were fairly low for the different categories. Some broad categories were easy
to annotate, such as the responsive cue phrases. However, it was considered difficult to identify
which specific sub-category of the responsives it was.

The annotation scheme was further developed and tested by the annotators before there was a
consensus about a final annotation scheme for the cue phrases. The categories from Fraser were
rarely used and therefore excluded. Narrow and ambiguous categories that were often confused
were merged into one. Each category was described in detail (see Appendix 10.4), trying to avoid
different opinions in how to use the different categories when labelling the cue phrases. During
the development of the annotation scheme the annotators had the choice of labelling a word as
cue phrase without choosing a specific category to see if there were any missing categories in the
annotation scheme. This resulted in a new category called Referring, used to label cue phrases
(often phrases with more than one word) explicitly referring back to prior discourse, e.g. “som
sagt”, “som jag sa tidigare”, “vad var det hon hette nu igen”, “vad var det vi sa” (Eng: “as said”,
“as I said before, “what was her name again’”). In the end, hand labelling a dialogue is a
subjective task. Below are the categories in the annotation scheme presented. Three of the
dialogues were labelled with two annotators, and the five dialogues remaining were labelled by
one annotator.

**Categories;** Connective Additive (CAD), Connective Contrastive (CC), Connective Alternative
(CAL), Responsive (R), Responsive New Information (RNI), Responsive Dispreference (RD),
Response Eliciting (RE), Repair Correction (RC), Modifying (MOD) and Referring (REF).

Annotated examples from each category are presented below (fig 3.2).

<table>
<thead>
<tr>
<th>Cue phrase</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Connective Additive</td>
<td>ja och grönt är ju fint</td>
</tr>
<tr>
<td>CAD</td>
<td>[yes and green is nice]</td>
</tr>
<tr>
<td>Connective Contrastive</td>
<td>ja men det är helt omöjligt förstå du</td>
</tr>
<tr>
<td>CC</td>
<td>[yes but it is totally impossible you know]</td>
</tr>
<tr>
<td>Connective Alternative</td>
<td>ja eller din systerdotter</td>
</tr>
<tr>
<td>CAL</td>
<td>[yes or your niece]</td>
</tr>
<tr>
<td>Responsive</td>
<td>ja men det låter ju väldigt överkommligt</td>
</tr>
<tr>
<td>R</td>
<td>[yes that sounds reasonable]</td>
</tr>
<tr>
<td>Responsive New Information</td>
<td>jaha din systerdotter tycker om gröna saker</td>
</tr>
<tr>
<td>RNI</td>
<td>[oh your niece likes green things]</td>
</tr>
<tr>
<td>Responsive Dispreference</td>
<td>ja men jag vet inte</td>
</tr>
<tr>
<td>RD</td>
<td>[yes but I don’t know]</td>
</tr>
<tr>
<td>Response Eliciting</td>
<td>vad är hon intresserad av då</td>
</tr>
<tr>
<td>RE</td>
<td>[what is she interested of then]</td>
</tr>
<tr>
<td>Repair Correction</td>
<td>klockan nej tre hundra femtio menade jag</td>
</tr>
<tr>
<td>RC</td>
<td>[the clock no I meant three hundred fifty]</td>
</tr>
<tr>
<td>Modifying</td>
<td>ja hon gillar ju dockor</td>
</tr>
<tr>
<td>MOD</td>
<td>[yes she actually likes dolls]</td>
</tr>
<tr>
<td>Referring</td>
<td>den är ju som sagt hel och fin</td>
</tr>
<tr>
<td>REF</td>
<td>[as said it is whole and nice]</td>
</tr>
</tbody>
</table>

*Figure 3.2: Examples of labelled cue phrases from the annotations*

The dialogues were also labelled with communicative acts and disfluencies (see Appendix 10.5)
Inter-annotator agreement

When having two or more annotators labelling the same dialogue there is an opportunity of testing the inter-annotator agreement. Inter-annotator agreement may be explained with how well the annotations of the dialogue agree with each other.

The Kappa coefficient

Inter-annotator agreement can be measured using the Kappa coefficient (Carletta, 1996). There is always a possibility that some agreement between the two annotators happens by chance. To correct for expected chance in agreement, Kappa statistics \( \kappa \) are used.

The Kappa formula:

\[
\kappa = \frac{(P(A) - P(E))}{(1 - P(E))}
\]

\( P(A) \) is the relative observed agreement among the raters,

\[
P(A) = \frac{A}{N}
\]

\( N \) are all the items the raters have to classify, and \( A \) are the items rated identically by the two raters. \( P(E) \) is the probability that agreement is due to chance.

When there is no agreement between the two raters, except the expected agreement by chance, \( \kappa = 0 \). When there is a perfect agreement, \( \kappa = 1 \). A Kappa value of 0.8 and above is considered as representing almost perfect agreement and reliability. A Kappa value between 0.67 and 0.8 are a substantial agreement. Inter-annotator agreement was calculated on the annotated DEAL corpus (see chapter 4).

Raw agreement

Inter-annotator agreement may also be measured in other ways. Raw agreement \( (P(A)) \) does not consider the agreement that may happen by chance. Using raw agreement may be useful descriptive statistics. Raw agreement may be less complex than Kappa and can also be very useful. Suppose that there are \( N \) items presented, and that \( A \) of these are perceived identically by the two raters. Then the agreement rate \( P(A) \) may be measured with the formula:

\[
P(A) = \frac{A}{N}
\]

This formula is called proportion of overall agreement and it is the same formula which is used to calculate \( P(A) \) in the Kappa formula. Raw agreement was used to analyse results from the listening test (see chapter 4).
3.2 Listening test

There were three different types of responsive in the annotation scheme: Responsive, New Information Responsive and Dispreference Responsive. The word “ja” (Eng: yes) were the lexical item most frequently used within all these categories. For the dialogue system to be capable of generating appropriate responses to the user there is a need to find what influence the interpretation of these responsive has. To do this a listening test was performed. Moreover, the listening test was used to see which features, contextual or acoustic, are important for how “ja” is interpreted. This part will describe the test in more detail.

3.2.1 Method

The stimuli in the test were a number of sound files, all with responsive from the different categories as annotated in the data collection. All stimuli were different realizations of the Swedish word “ja” (Eng: yes). “Ja” in its sentential form was also used as stimuli (“yes” as answer to yes/no questions). 23 stimuli were used in the test. Similar to Gravano et al. (2007b) each stimulus was presented twice; one time in isolation and one time in context, i.e. together with prior speaker turn. These stimuli were all taken from the recordings (presented with the real speaker’s voice). 7 of the stimuli were also presented with synthesized voice. 21 subjects participated in the test. Many of them were from the Stockholm area, but also subjects from other parts of Sweden participated. The subjects listened to the stimuli and for each stimulus they chose an appropriate category. In addition to the three responsive categories for cue phrases, the subjects could also choose the sentential meaning of “ja”. The test was forced choice, meaning that the subjects were forced to choose one of the categories for each presented stimulus. The stimuli were presented randomly.

Previously presented studies show that it is easier to categorize cue phrases if sound is added to the transcription and if the cue phrases are presented in context (Zufferey & Popescu-Belis, 2004; Gravano et al., 2007b). In this test, all stimuli were the same lexical item “ja”. Presenting the stimuli both in isolation and with context may give information on whether the subjects perceive the responsive differently within a context, or if the acoustical features are dominant when choosing a category.

3.2.2 Selection of stimuli

The stimuli were chosen from dialogues labelled by two annotators where the two annotators had labelled them with the same category. There was not the same number of stimuli from each category in the listening test. Five stimuli were chosen labelled as Responsive New Information, six as Responsive Dispreference, six as Responsive and six stimuli with “ja” (Eng: yes) in its sentential meaning. There were only a few instances of “ja” in the Responsive New Information category to choose from. Therefore, stimuli from this category were also chosen from dialogues which were only annotated by one. However, there was still not enough “ja” to choose from in these dialogues, since the chosen stimuli needed to be positioned initially in speaker turns and where there was no overlap between the speakers’ utterances. All stimuli were different realizations of the Swedish word “ja” (Eng: yes). This included variations such as “aa”, “nja” or “aej”.

The synthesised stimuli were based on the original realizations in the recorded dialogues. The f0-curve and duration was automatically extracted from the original recording to create the corresponding synthesised version. The pitch was lowered for stimuli with female speaker voices, to avoid the male synthesised voice sounding whining.
3.2.3 Experimental setting/Platform

The test was set up as a webpage. The subjects needed a computer, an internet connection and headphones. The webpage, which was the platform for the test, was programmed in the computer scripting language php and in the markup language html. Below is shown how the test part was presented. There was a sound player that presented the stimuli automatically. Below the sound player were the four categories that the subject could choose from. At the bottom of the page a description of the categories were presented. A number displayed how many stimuli were left.

![Test interface](image)

*Figure 3.4: Snapshot from the test*

3.2.4 Pilot tests

Before the real test was conducted, pilot tests were done. Five subjects participated in the pilot test, one of them working in the area of speech technology. The procedure was similar to the one described below, but only using stimuli from the original dialogues and not presented in randomized order. The subjects were asked to give comments after completed the test. These comments were valuable and helped improve the test. The introduction was described better, the definition and the names of the categories were changed to better be understood and the sound player was changed to start automatically for each stimulus. The results from the pilot test were compared with the original annotation, and a few stimuli from each category were categorized the same by almost all subjects, indicating that there might be characteristic stimuli from each category. These were chosen to be presented also in synthesised voice in the test, trying the synthesiser used in DEAL.
3.2.5 Subjects
15 men and 6 women participated in the test. Different dialects and native language may affect how the stimuli are perceived. 17 of the subjects reported having Swedish as native language. The other reported having native language such as English, Russian, Turkish and Norwegian. However all of them have lived in Sweden for at least 20 years and one of them has lived his/her whole life in Sweden. 4 of the subjects reported having a dialect from Stockholm. Since it is a perception test, studying general perception of Swedish responsive cue phrases, the only demand of the subjects is to manage the Swedish language.

3.2.6 Procedure
Entering the test application, the first page contained a form. The form asked for personal information about the subjects such as name, email address, gender, native language and dialect. The subjects were told that the personal information was handled anonymously. Mowing forward in the test, the subjects were given a presentation of the test and an introduction to the test (see Appendix 10.7). The subjects were asked to use headphones and to carefully read the definition of the different categories.

The order in which the stimuli were presented was randomized for every subject. The same stimulus (isolated and with context) did not appear next to each other. The subjects were presented with one stimulus at a time, and asked to choose the most appropriate one of the four categories. The subjects were able to listen to the sound files as many times as they needed. When all the stimuli were categorized the subject could give comments about the test.
4 Result analysis

This section presents results from the data collection and the listening test on responsive cue phrases. Agreement in categorization within and between test subjects will be presented as well as acoustic features of the stimuli where the categories received high annotator agreement. First general data on cue phrases collected from the DEAL corpus will be presented.

4.1 Result analysis of cue phrases in the DEAL corpus

Eight dialogues were recorded. Altogether they had a recording time of 1 hour and 45 minutes. The length of each dialogue varied from 8 to 18 minutes. The total number of words collected was 16357. On average the shopkeeper used 13.4 words per speaker turn, and the customer used an average of 8.5 words per speaker turn. 81% of the speaker turns contained at least one cue phrase. 21% of all words in the dialogues were annotated as cue phrases. 86% of the speaker turns contained at least one cue phrase or disfluency.

4.1.1 Inter-annotator agreement

Three of the dialogues were annotated by two different annotators and Kappa was calculated to check for agreement between the annotators. Two different Kappa coefficients were calculated. One on word level for the binary task if a word was classified as a cue phrase or not and one on the classifications of the cue phrases into categories. The Kappa for the binary task, when classifying a word as a cue phrase or not, was 0.87 (p=0.05). The Kappa for the second task, classifying the cue phrases according to their category was also high, 0.82 (p=0.05). Below is a chart of percentage agreement in the different categories.

![Chart 4.1: % agreement for the categories](chart)

Chart 4.1 presents the annotation agreement in percentage for the different categories. The annotation agreement is high. Both the Modifying (MOD) category and the Responsive (R) category have an annotator agreement of 87%. However, three of the classes; Referring (REF), Connective Alternative (CAL) and Repair Correction (RC) had very few instances.
4.1.2 Frequency

Some categories of cue phrases were more frequent than others. Chart 4.2 presents the distribution of cue phrases over the different categories.

![Chart 4.2: Cue phrase distribution over the different categories](image)

Chart 4.2 shows that the modifying cue phrases were most frequent in DEAL corpus (27%). Responsive cue phrases were also common (20%). Together with Responsive Dispreference (9%) and Responsive New Information (8%), the three responsive categories had 37% of all cue phrases in the DEAL corpus.

Table 4.1 presents the most frequent cue phrases in all the categories. The instances of cue phrases in the categories of Repair Correction and Referring were mostly phrases containing more than one word. The Response Eliciting category also contained longer phrases. Table 4.1 gives examples of such phrases.

<table>
<thead>
<tr>
<th>Responsive (R)</th>
<th>Connective Additive (CAD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ja (60%), mm (25%) [yes]</td>
<td>så (24%), och (22%) [so, and]</td>
</tr>
<tr>
<td>Responsive New Information (RNI)</td>
<td>Contrastive Connective (CC)</td>
</tr>
<tr>
<td>ja (33%), okej (7%) [yes, okay]</td>
<td>men (69%), alltså (15%) [but, therefore]</td>
</tr>
<tr>
<td>Responsive Dispreference (RD)</td>
<td>Connective Alternative (CAL)</td>
</tr>
<tr>
<td>ja (32%), nej (29%)</td>
<td>eller (62%), istället (24%) [or, instead]</td>
</tr>
<tr>
<td>Modifying (MOD)</td>
<td>Referring (REF)</td>
</tr>
<tr>
<td>ju (60%), faktiskt (6%) [actually, of course, as a matter of fact]</td>
<td>&quot;som jag sa tidigare&quot;, &quot;vad var det vi sa&quot; [&quot;as I said before&quot;, &quot;what did we say&quot;]</td>
</tr>
<tr>
<td>Response Eliciting (RE)</td>
<td>Repair Correction (RC)</td>
</tr>
<tr>
<td>då (38%), &quot;tycker du inte det?&quot;, &quot;vad säger du om det då?&quot; [then, &quot;don't you think so&quot;, &quot;what do you say about that&quot;]</td>
<td>nej (19%), &quot;jag menar&quot;, &quot;nej nu sa jag fel igen&quot; [no,&quot;I mean&quot;, &quot;no, now I said the wrong thing again&quot;]</td>
</tr>
</tbody>
</table>

*Table 4.1: Frequently used cue phrases in the different categories*
For more information of used cue phrases in the different categories, see Appendix (10.6). More information on the use of Responsive cue phrases in DEAL corpus is presented in part 4.2 about the listening test.

### 4.1.3 Turn position distribution

The turn distributions of all cue phrases in the DEAL corpus are shown in Chart 4.3.

![Chart 4.3: Turn position distribution for each category](chart)

Chart 4.3 shows that cue phrases are most frequently used within a speaker turn (62%). Next comes cue phrases positioned initially (27%). Cue phrases at terminal position are rare (11%).
4.2 The outcome of the listening test

This part presents results from the listening test. Annotator agreement is made on whether the same category is chosen for the isolated stimulus as for the same stimulus in context. Agreement in category decisions between subjects will also be revealed as well as acoustic features for stimuli with high annotator agreement. At the end results from the synthesised stimuli will be presented. But first, some frequency data on responsive cue phrases used in DEAL corpus will be presented.

4.2.1 Responsives in DEAL corpus

Following charts (Chart 4.4, Chart 4.5 and Chart 4.6) show the frequency of different used responsive cue phrases in the DEAL corpus.

![Chart 4.4: Use of Responsive cue phrases](image)

Chart 4.4 presents the frequency of different Responsive cue phrases. 60% of the cue phrases in the Responsive category were variations of “ja” (Eng: yes), also including similar phonetic realizations such as “aa” and “ah”. “mm” is the second most used cue phrase in this category, with 25%.

![Chart 4.5: Use of Responsive Dispreference cue phrases](image)

Chart 4.5 presents the frequency of different used cue phrases in the Responsive Dispreference category. Also for this category the most common cue phrase was “ja” (Eng: yes) including similar cue phrases such as “aa” and “ah”. These were 32% of all cue phrases in this category. “nej” and “nä” (Eng: no) are almost used as frequently as “ja”, with 29%. The bar “other”
contains seldom used cue phrases. In this chart, that bar is high compared to the other responsives, having more than one fifth of the cue phrases in this category. Examples of such other cue phrases are “förstås” (Eng: of course) and “jo” (Eng: well, sure).

![Chart 4.6: Use of Responsive New Information cue phrases](image)

Chart 4.6 presents the frequency of different used cue phrases in the category of Responsive New Information. Also in this category, “ja” (Eng: yes) was most frequently used, with 33%. Next was “okej” (Eng: okay) with 7%. From the above charts (4.4, 4.5 and 4.6) it is obvious that phonetic variations of “ja” (Eng: yes) is the most common responsive cue phrase in DEAL corpus. It is used 48% of the times were there is a responsive cue phrase.

### 4.2.2 Categorizations of the responsives in the listening test

Chart 4.7 presents how frequently the different categories in the test were chosen. The Responsive category is the most frequently chosen category in the listening test with 34%. Least used is Responsive New information with 14%.

![Chart 4.7: Chosen categories by the test subjects](image)
Chart 4.8: Chosen categories for isolated stimuli and stimuli in context

Chart 4.8 presents the category distribution over the different categories for the isolated stimuli and stimuli in context separately. The bars in the chart show that the categories are almost as frequently used for the stimuli with context as for the stimuli without context.

Chart 4.9: Chosen categories by the test subjects compared to annotation categories

Chart 4.9 compares how the annotators categorized the cue phrases with how the subjects in the listening test perceived them. The cue phrases annotated as Responsive New Information and Responsive Dispreference (bar RNI and RD) were often perceived the same by the annotators and the listening test subjects. The cue phrases annotated as Responsive are often perceived differently by the test subjects.

4.2.3 Categorization agreement

This part presents raw agreement in categorization within and between the test subjects. In this test agreement within subjects refers to the proportion of subjects that agreed on a category for a stimulus with and without context. Agreement between refers to the subjects above that also agreed on the same category. In the test, 23 different stimuli were presented, each presented both with and without context. Inter-annotator agreement was calculated using Fleiss’s Kappa for agreement on isolated stimuli and stimuli with context separately. $K=0.286$ for isolated stimuli and $k=0.367$ for stimuli in context. Chart 4.10 shows the proportion of times subjects agreed that the isolated stimuli and corresponding stimuli in context belonged to the same category. However, a high value does not say that there is an agreement between the subjects on a single category.
Chart 4.10: Agreement in categorising stimuli with and without context for all subjects

Chart 4.10 shows that 7 of the stimuli have a category agreement over 70% with and without context.

Chart 4.11 presents the proportion of times subjects agreed on a specific category for a particular stimulus. The stimuli presented in the chart are those with high agreement (70% or more) from the chart above (Chart 4.10).

Chart 4.11: Category agreement between subjects for the stimuli with high agreement within the subjects

Chart 4.11 shows that the category agreement was highest for stimuli 7, 8, 9 and 10. A large majority of the subjects categorized these stimuli as Responsive Dispreference cue phrases. Stimuli 2 and 23 also had relatively high agreement on a specific category: Responsive New Information cue phrases. Stimulus 17 was categorized both as a Responsive cue phrase and as an answer to a yes/no question, the later being the most frequent.
4.2.4 Acoustic features of stimuli

This section presents acoustic features of the stimuli with high agreement. The features studied for the stimuli were duration in milliseconds, the shape of the pitch curve and also the waveform to see if there is any pattern in the intensity of the stimuli. The frequency in the pitch curves are considered to further analyse the shape of the pitch curve, to see how much the curve rises or falls in pitch difference. The data of the features were studied and gathered from a program called Wavesurfer. Before the stimuli and their gathered features are described, the mean durations for the responsives are presented.

Duration

Duration is one acoustic feature studied for these stimuli.

![Chart 4.12: Mean duration for all stimuli](image)

Chart 4.12 presents the mean duration for all stimuli in the listening test. The categories chosen for each stimulus, is the category with highest agreement on the isolated stimulus. The Responsive Dispreference cue phrases had much longer duration than the other, while responses to yes/no questions had the shortest mean duration. The mean duration for Responsive Dispreference cue phrases was 737 milliseconds compared to the Yes/No words with 261 milliseconds. Responsive cue phrases had a mean duration of 347 milliseconds and Responsive New Information had a mean duration of 339 milliseconds. The mean duration for all annotated responsive cue phrases is 446 milliseconds.
Responsive Dispreference

This category has very high category agreement for the stimuli with and without context, and also high agreement on a specific category. For stimulus 7, 95% of the test subjects chose the same category for the isolated stimulus and the stimulus in context. All of them categorized the stimulus as Responsive Dispreference. For stimulus 8, 86% of the test subjects chose the same category for the isolated stimulus and the stimulus in context. All of them perceived the stimulus as Responsive Dispreference. The isolated stimulus was perceived as Responsive Dispreference by 100% of the test subjects. For stimulus 9, 76% of the test subjects agreed on the isolated stimulus and the stimulus in context. All of them perceived the stimulus as Responsive Dispreference. For stimulus 10, 81% of the test subjects agreed on the isolated stimulus and the stimulus in context. All of them perceived the stimulus as Responsive Dispreference.

Stimulus 7  
Speaker: Woman

<table>
<thead>
<tr>
<th>Hz</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Fig 4.7: $V_o$-curve for stimulus 7

Fig 4.8: Waveform of stimulus 7

Duration: 689 milliseconds  
Pitch curve: Relatively straight between 216 Hz and 232 Hz.  
Intensity: The waveform shows small intensity in the beginning of the stimulus which then increases and falls in the end.

Stimulus 8  
Speaker: Woman

<table>
<thead>
<tr>
<th>Hz</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
</tr>
<tr>
<td>500</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Fig 4.9: $V_o$-curve for stimulus 8

Fig 4.10: Waveform of stimulus 8

Duration: 800 milliseconds  
Pitch curve: Relatively straight (except for change in pitch at one place).  
Intensity: The waveform shows small intensity in the beginning of the stimulus which then increases and falls in the end.
Stimulus 9  Speaker: Woman

![Fig 4.11: F0-curve for stimulus 9](image1)

**Fig 4.12: Waveform of stimulus 9**

Duration: 1011 milliseconds
Pitch curve: Falling down with a pitch difference of 50 Hz. One part on the second half of the curve is rising with 30 Hz.

Stimulus 10  Speaker: Man

![Fig 4.13: F0-curve for stimulus 10](image2)

**Fig 4.14: Waveform of stimulus 10**

Duration: 1064 milliseconds
Pitch curve: Rising slowly with a pitch difference of 49 Hz.
Intensity: The waveform shows small intensity in the beginning, which then increases.

Studying these four stimuli, there is one feature that is relatively similar, the duration. The duration between the stimuli varies from 689 to 1064 milliseconds. However they are all longer than mean duration of all stimuli (446 ms). Except for one of the stimulus, they are also longer than the mean duration of the Responsive Disprefer cue phrases (737 ms). The pitch curves for two of the stimuli (7 and 8) have a relatively straight shape. These stimuli also had the highest category agreement between the test subjects. The other stimuli (9 and 10) also have a relatively straight pitch curve compared to stimuli perceived as other categories.
Responsive New Information

Presented here are acoustic features of the stimuli with highest category agreement as belonging to the Responsive New Information category. For stimulus 2, 71% of the test subjects chose the same category for the isolated stimulus and the stimulus in context. 62% of them categorized the stimulus as Responsive New Information. For stimulus 23, 73% of the test subjects chose the same category for the isolated stimulus and the stimulus in context. 64% of them perceived the stimulus as Responsive New Information.

<table>
<thead>
<tr>
<th>Stimulus 2</th>
<th>Speaker: Woman</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hz</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

*Fig 4.3: F<sub>0</sub>-curve for stimulus 2*

<table>
<thead>
<tr>
<th>451</th>
</tr>
</thead>
</table>

*Fig 4.4: Waveform of stimulus 2*

Duration: 365 milliseconds
Pitch curve: Rising from 119 Hz to 245 Hz. Pitch difference of the rise is 126 Hz. Two pitch peaks, one in the middle and one at the end of the curve.
Intensity: The waveform shows that the intensity is greatest the first half of the stimulus.

<table>
<thead>
<tr>
<th>Stimulus 23</th>
<th>Speaker: Man</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hz</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

*Fig 4.5: F<sub>0</sub>-curve for stimulus 23*

<table>
<thead>
<tr>
<th>675</th>
</tr>
</thead>
</table>

*Fig 4.6: Waveform of stimulus 23*

Duration: 356 milliseconds
Pitch curve: Rising from 98 Hz to 236 Hz. Pitch difference of the rise is 138 Hz. Two pitch peaks, one in the middle and one at the end of the curve.
Intensity: The waveform shows that the intensity is greatest the first half of the stimulus, and then decreases.

Studying both these stimuli, there are some features similar between them. Both the pitch curves are rising. And the pitch difference is higher than that of Responsive stimuli. It appears to rise steeper than the Responsive stimuli. Both the pitch curves have two pitch peaks, one near the
middle of the curve, and one close to the end of the curve. After the last peak, which is the maximum peak, the curves go down again. The duration of these stimuli are also similar (365 ms and 356 ms), and compared to the chart with mean duration (Chart 4.12) these stimuli are a little bit longer than the mean duration of the cue phrases in this category. Compared to the mean duration of all responsive cue phrases (446 ms), these stimuli are slightly shorter. Another stimulus where most (but still only 43%) of the subjects perceived the isolated stimulus as Responsive New Information, the pitch curve has a similar shape as the two above.

Yes/No answer
The first presented stimulus (17) had high category agreement, 76%, for stimulus with and without context, but not on a specific category. 48% agreed on the Y/N answer category and 28% of the test subjects agreed on the Responsive category. For stimulus 18, 62% of the test subjects agreed on the isolated stimulus and the stimulus in context. 52% of them perceived the stimulus as Y/N answer and the rest (10%) perceived them as Responsive. For stimulus 19, 67% of the test subjects agreed on the isolated stimulus and the stimulus in context. 57% of them perceived the stimulus as Y/N answer and the rest (10%) perceived them as Responsive.

### Stimulus 17
**Speaker:** Woman

![Fig 4.15: Fo-curve for stimulus 17](image)

Duration: 269 milliseconds
Intensity: The waveform shows greatest intensity in the beginning, which then decreases.

### Stimulus 18
**Speaker:** Woman

![Fig 4.17: Fo-curve for stimulus 18](image)

Duration: 234 milliseconds
Intensity: The waveform shows two intensity peaks, one in the beginning and one close to the end.
Stimulus 19  Speaker: Woman

Fig 4.19: F0-curve for stimulus 19

Fig 4.20: Waveform of stimulus 19

Duration: 300 milliseconds
Intensity: The waveform shows greatest intensity in the beginning, which then decreases.

There is one specific feature that is similar to these stimuli perceived as answers to Yes/No questions. This feature is not found in the stimuli perceived as other categories. The shape of the pitch curve for the three stimuli above (17, 18 and 19) first falls, and in the middle it rises again. In the two first stimuli (17 and 18) this feature is more significant.

Another yes/no answer with very low agreement (43%) and also low agreement on a specific category (14%) has a pitch curve similar to the ones above in this category. However, the duration is 345 milliseconds which is higher than mean duration for this category (261 ms). 345 milliseconds is duration that lies between the mean duration for the Responsive cue phrases (346 ms) and the Responsive New Information cue phrases (339 ms). This one has a pitch difference of 50 Hz and duration of 345 milliseconds. This stimulus is however more perceived as belonging to the other categories.
Responsive

Presented here are acoustic features of a stimulus which received fairly high inter-annotator agreement as belonging to the Responsive category. 62% of the test subjects chose the same category for the isolated stimulus and the stimulus in context. 57% of them perceived the stimulus as Responsive. At the annotation process, both annotators labelled this stimulus as Responsive New Information.

Stimulus 3  Speaker: Man

![Fig 4.1: F0-curve for stimulus 3](image)

**Fig 4.1: F0-curve for stimulus 3**

![Fig 4.2: Waveform of stimulus 3](image)

**Fig 4.2: Waveform of stimulus 3**

**Duration:** 271 milliseconds

**Pitch curve:** Rising from 120 Hz to 177 Hz. Pitch difference of the rise is 55 Hz.

**Intensity:** The waveform shows that the intensity is greatest the first half of the stimulus.

The pitch curve (fig 4.1) is rising, but not as steep as the stimuli perceived as Responsive New Information (fig 4.3 and fig 4.4). Other stimuli where the test subjects categorized the “ja” more frequently as Responsive category for the isolated stimulus, also show a rising pitch curve with differences from 20 Hz to 100 Hz. These stimuli also have greater intensity in the first half of the waveform. The same stimuli are also frequently categorized as Responsive New Information or as Yes/No answer. One of the stimuli initially annotated as a Responsive cue phrase was more frequently categorized as Responsive Dispreference in the listening test. Pitch difference for that stimulus was 90 Hz and duration 427 milliseconds, which is over mean duration for the Responsive category. The intensity is greatest in the middle of the waveform.
4.2.5 Agreement on synthesised stimuli

Chart 4.13 presents the agreement when categorizing isolated stimuli and corresponding synthesised stimuli.

![Chart 4.13: % category agreement between isolated stimuli and corresponding synthesised stimuli](image)

Chart 4.13 shows that stimuli S2 and S3 have much higher category agreement than the others, both between the isolated stimuli and the correspondent synthesised stimuli (81% and 90%), but also on a specific category (76% and 90%). Both are perceived as Responsive Dispreference. S1 corresponds to stimulus number 1 annotated as Responsive New Information. In the listening test 47% of all test subjects categorized the isolated stimulus as Responsive. Responsive is also the most frequently chosen category for the synthesised stimulus. S2 and S3 correspond to stimuli nr 9 and nr 10 annotated as Responsive Dispreference. In the listening test 95% along with 90% of the subjects perceived them as Responsive Dispreference in the isolated case, and the agreement for the synthesised stimulus are also high for this category. S4 and S5 correspond to stimuli nr 11 and nr 14 annotated as Responsive cue phrases. In the listening test 52% and 48% of the subjects perceived them as Responsive. However, in their synthesised version, over 60% perceived them as Y/N answer. S6 and S7 correspond to stimuli nr 17 and nr 19 which are answers to yes/no questions in the DEAL corpus. In the listening test 57% and 71% of the subjects perceived them as Y/N answer, but in their synthesised version these stimuli were mostly perceived as Responsive cue phrases.

Looking at the pitch curves, intensity and duration they are the same as their corresponding stimulus. However voice quality may also be important for how we perceive what is said, and the synthesised voice sounds different from real voice. Still, S2 and S3 had a high agreement on Responsive Dispreference, which also their corresponding stimuli with real voice had. Next, the results in this chapter will be discussed further.
5 Discussion

In this study dialogue data was collected and used to analyse cue phrases, with DEAL in focus. Recording and analyzing cue phrases in human-human dialogue from a setting similar to the domain in DEAL provide extended knowledge on the use of cue phrases; which ones are most frequently used, and in which context they are used. The responsive cue phrase “ja” (Eng: yes) is frequently used in the DEAL corpus with different discourse-pragmatic functions. One purpose of this master’s project was to study what kinds of features influence its interpretation. The method and results will now be discussed.

5.1 Reflections on data collection

As described in the theoretical part of this master’s project, human-like speech is preferable in a spoken dialogue game like DEAL. In this report, human-like refers to also including features which are typically not generated in spoken dialogue systems, such as cue phrases and disfluencies. The data collection was performed to study how humans produce spontaneous speech in human-human conversations in the DEAL domain. The data collection set-up was a situation reminding of the trade domain in DEAL. A role-play was established where two subjects negotiated about objects and their prices trying to reach a deal. To succeed both subjects needed to engage in the negotiation; the customer, struggling to reduce the price, and the shopkeeper trying to earn as much money as possible for his objects. The participants seemed engaged in the role-play, and they were alone in the studio while recording to make them feel comfortable and avoid external disturbances. According to the study by Louverse and Mitchell (see part 2.4) spoken discourse contained ten more times cue phrases than written discourse, and informal dialogue had two times more cue phrases than formal dialogue. The dialogues collected in this project were not pre-planned, the participants got specific tasks to get started, but the speech was still spontaneous. The dialogues collected are rich of cue phrases and disfluencies: 86% of the speaker turns contained at least one cue phrase or disfluency, which supports that they were produced spontaneously. The dialogues are a good foundation to start analysing how cue phrases are used in a role-play trading situation, which can be suitable to reproduce in DEAL.

5.2 Reflections on annotation

Choosing a suitable annotation scheme for cue phrases is difficult. Different studies on cue phrases suggest different ways of classifying them. One way of developing a suitable annotation scheme for a specific situation and kind of dialogue is to try different classifications. Depending on the goal of the annotation, different schemes may be used. Trying to define discourse-pragmatic functions for cue phrases in DEAL led to the cue phrase classification developed by Lindström (2008). Lindström suggests a classification based on different functions cue phrases might have and provides examples of words and phrases in Swedish for the different functions. However, the classification by Lindström had too many and too narrow categories, and some of them were also ambiguous. To find a reliable annotation scheme for the DEAL corpus Lindström’s annotation scheme was reduced, some categories were merged and one category was added.

In the last version of the annotation scheme for cue phrases, it has gradually felt like one category is missing. Labelling words and phrases with the connective categories was sometimes difficult because it felt like none of the categories were suitable. The missing category was a conclusion/cause category, “Connective Conclusion”. However, “Connective Conclusion” and Connective Additive may use the same cue phrases, since conclusion and causal utterances are
adding something to previous discourse. Thus conclusion and also causal functions in DEAL can use phrases from Connective Additive. However, this is a general problem that categories are “missing” in the annotation scheme. If categories are added every time, the annotation scheme may grow to become too big and then too complex to use. It is very hard to find a perfect and comprehensive annotation scheme. To develop a suitable annotation scheme for a corpus and then have more than one annotator to use this scheme to classify cue phrases in a corpus is not an easy task. Most important is to find an annotation scheme that works well enough and suits the purpose. Bearing in mind the high Kappa value (0.87 for the binary task and 0.82 on the classification task) for the annotation on DEAL corpus, the annotation scheme described in this master’s thesis may well be one of those well defined for its purpose. Both annotators were well understood with the annotation scheme and reached a common ground about the annotation. However, a new annotator might need some practice to get used to the annotation scheme.

As mentioned above, the Kappa values for the annotation of cue phrases in DEAL corpus reached very high agreement. A kappa of 0.87 for the binary task and 0.82 on the classification task was calculated. When there is no agreement between the two raters, except the expected agreement by chance, $\kappa = 0$. When there is a perfect agreement, $\kappa = 1$. A Kappa value of 0.8 and above is considered as representing almost perfect agreement and reliability. This verifies again that the annotation scheme was well defined for its purpose.

The annotator agreement for Responsive and Modifying cue phrases are high (each 87%), and these cue phrases are also most common among the cue phrases in the DEAL corpus (see Chart 4.2). A well described definition of them in the annotation scheme may be the answer. Since they appear so frequently, maybe the annotators learn to look for them and disagreement on a few words here and there will not affect the agreement that much. Referring and Repair Correction cue phrases also had relatively high agreement (82% and 88%), but they are seldom annotated in the corpus. This could make it harder to reach a high agreement, since they first have to be classified as cue phrases, and then they should be classified with the same discourse-pragmatic function. Since few in these categories were annotated (see Chart 4.2), only a few disagreements would affect the value of agreement very much. One explanation of the high agreement is that Referring cue phrases and Repair Correction cue phrases are quite different from the cue phrases in other categories which make them easier to distinguish from the others.

5.3 Reflections on cue phrases in DEAL corpus

Conclusion from the general data analyses on cue phrases in DEAL corpus support that modifiers and responsives are the most frequently used cue phrases in Swedish trading conversations (see Chart 4.2). “Ja” (Eng: yes) is the most common responsive (48%), which signals different discourse-pragmatic functions depending on context and acoustic features. Among the connectives, depending on category, different cue phrases are most common. “så” (Eng: so) and “och” (Eng: and) are the most frequent words in the Connective Additive category, (24% and 22%), “men” (Eng: but) are most common in the Connective Contrastive category (69%), and “eller” (Eng: or) the most frequent word in the Connective Alternative category (62%). These results are not surprising, since they are used to define the different categories by several researchers.

One interesting result, however, is the high amount Modifying cue phrases in DEAL corpus (27%) (Chart 4.2). Studies trying to define cue phrases and their functions often focus more on connective cue phrases and other cue phrases connecting the upcoming utterance with previous discourse. Modifiers often colour the present message (in middle position and end position) and the upcoming message when used initially. They give cues to the hearer how to perceive the present message or even the messenger. Lindström views them as important in his definitions, while other researches seem to ignore their importance or do not include them as cue phrases.
However, since they are used so frequently, they seem to have a really important role in the conversations.

Cue phrases from the responsive categories were mainly used initially in speaker turns (Chart 4.3). In the theory of cue phrases by Fraser (1999), Schiffrin (1987) and Schourup (1999) connectives are argued to often be positioned initially in utterances to connect an utterance to previous discourse. However, in the DEAL corpus the connectives were more frequently used within the speaker turns. This difference can likely be explained by the fact that cue phrases in DEAL were often used in combination signalling functions at different discourse levels. Many speaker turns begin with a response such as “yes” and “mm”, signalling that the message is perceived, and is then followed by a connective cue phrase such as “and”, “so”, “but” and “or”, signalling how the new information relates to previous discourse. Consequently, removing the responsive from the initial positions, the position distribution for the connectives may have looked differently.

Only 11% of the cue phrases are positioned at a terminal position of the speaker turns in DEAL corpus (Chart 4.3). 26% of them are from the category Response Eliciting followed by the Modifying and Connective Additive cue phrases. 47% of the cue phrases in the Response Eliciting category are positioned terminally. In Lindström’s definition of Response Eliciting (see 2.4.5) some of the cue phrases in this category, such as “va” (Eng: huh), “eller” (Eng: or) and “då” (Eng: then) are attached to the end of the utterance when the speaker wants to have a response from the hearer. This is also what Yule (1996) suggests, that these cue phrases can be used to signal a completion point of a speaker turn, giving away the turn to the hearer (see 2.3.1). To further compare the position distribution results in the DEAL corpus with other categories by Lindström is difficult. He analyses the position distribution of cue phrases on an utterance level. In DEAL corpus, speaker turns are used when analyzing position distribution of cue phrases. If Lindström suggests that a cue phrase is attached initially to an utterance, these cue phrases can also be positioned within the speech segments of the DEAL corpus, since one speaker turn may consist of one or several utterances. This is also true when trying to compare the position distribution results with Schiffrin, Fraser or Schourup. According to Schiffrin, cue phrases are attached to the boundaries of units of speech. Since Schiffrin defines a unit of speech as including utterances, speaker turns and other speech segments, it is difficult to compare. Fraser and Schourup also refer to utterances when suggesting positions of cue phrases. Because of this, the distribution of cue phrases in DEAL corpus, at initial, middle and terminal position, will be different from their description of positions for the cue phrases.

Among the responsive cue phrases, Responsive was the category with highest annotator agreement (87%). Then comes Response New Information (81%) and last is Response Dispreference with lowest annotator agreement (67%) (Chart 4.1). The lexical item “ja” (Eng: yes) is most common in all these categories (60% of the Responsive, 32% of the Response Dispreference, and 33% of the Responsive New Information). There is still relatively high agreement on the first two mentioned categories, and one may wonder which features affect how these items are perceived. Using a dialogue system, there may be situations where the system only responds with a single responsive, without needing a whole sentence. Then the system needs to generate the “right” responsive if different acoustic realizations are interpreted differently. Consequently, it was interesting to do a listening test, trying to reveal which features influence how we interpret “ja”.

### 5.4 Reflections on the listening test

As stated in the theoretical part (2) of this report, using responsive cue phrases are important for signalling attention, understanding, agreement or also disagreement to the message. These responsive are also used to control turn-taking and perform error handling etc. It is clear that these entities are important in spoken conversation.
Stimuli selection for the listening test was based on the original cue phrase annotation. Using the annotation as a base for selecting stimuli for the test was good to get a mixed group of stimuli, with different acoustic realizations. “ja” in its sentential meaning was also included as a category since this is a frequently used function of “ja”. Choosing the word “ja” (Eng: yes) for the test was based on its frequent use (48%) in the responsive categories (see charts 4.4, 4.5 and 4.6). Another way could have been to use different words in the test, but then the word itself will affect how it is perceived and not only its acoustic realization. However, the different realizations of the lexical item “ja” in the test were somehow different from each other phonetically. In the test, there were variations of the “ja” such as “aa”, “eja” and “nja”. This might have affected how the stimuli were perceived.

Comments on the test revealed that the subjects sometimes had difficulties in choosing a category for each presented stimulus. Some felt that categories were missing, and other felt that the existing categories sometimes were too much alike making it hard to distinguish them from each other. Responsive and Yes/No answer were the categories most difficult to distinguish from each other. This might be one reason why there was no high category agreement on the stimuli with and without context (see Chart 4.10) and also for a specific category in these categories (see Chart 4.11). Comments also revealed that it was hard to find candidates for the Responsive New Information category. It was also less chosen in the listening test. One explanation is that the test subjects had a clear vision on how they think a “ja” (Eng: yes) in this category should sound like, and there were few of these in the test. Another explanation is that the definition of the category used in the test was not described well enough. Even if “ja” is the lexical item most common in the Responsive New Information category, it is only used 33% of the times in DEAL corpus for this category (see Chart 4.6). Using other words or phrases, such as “okej” (Eng: okay), “just det” (Eng: exactly) and “jahå” (Eng: oh) may be perceived as more representative to this category.

5.4.1 Results

High subject agreement between isolated stimuli and the same stimuli in context supports the idea that the subjects at least partially rely on acoustic features when choosing a category for that stimulus. When removing context, the contextual features which can potentially influence the interpretation of stimuli are removed, and left is the acoustic realization.

To calculate results, raw agreement was used. Raw agreement may not be seen as a good way of generating results, since it does not consider agreement by chance as in Kappa coefficient. However, raw agreement is useful because it reveals descriptive statistic that might be easier to understand than complex statistic measures. The Result chapter (4) presents features for all stimuli where agreement was over 70%. Also some stimuli between 60% and 70% agreement were described. Some of them were perceived as different categories between the subjects, however still having high agreement (a similar category distribution) between the isolated stimulus and stimulus in context. High category agreement for stimuli with and without context, but low agreement on a specific category between the subjects might reveal that the test subjects use the acoustic features to perceive function, but that the stimulus still is ambiguous since different test subjects perceive them differently. A low proportion of category agreement for a stimulus with and without context may show that the contextual features are used to interpret the function of the stimulus when the stimulus is presented in context. Also here, this may be a result of an ambiguous stimulus, which makes contextual features more important to interpret this stimulus.

The Kappa values for the inter-annotator agreement were 0.286 for stimuli without context and 0.367 for stimuli with context (see part 4.2.3). These are low kappa values, but still slightly higher than the values in the perception study by Gravano et al. (2007b), where inter-annotator agreement were $\kappa=0.210$ of isolated stimuli and $\kappa=0.293$ for stimuli in context of the word “okay”. Both in their test and in the test of this master’s project, inter-annotator agreement was
higher for the stimuli in context. This result might suggest that contextual features are more important for how we interpret the discourse function of both “okay” and “ja” (Eng: yes). In the annotation of DEAL corpus, inter-annotator agreements for the responsive categories were much higher in the listening test. In the listening test, the test subjects did not have a chance to practice; nor did they have as much contextual information as the annotators who had access to the complete dialogues. This might have affected the low agreement values in the listening test.

Using raw agreement, there was high proportion category agreement for stimuli with and without context for seven stimuli in the listening test (Chart 4.10). They all had an agreement over 70%. Six of these also had a high agreement on a specific category between the subjects. Four of these were agreed on as Responsive Dispreference. One of these stimuli had a category agreement of 95% with and without context, but also on a specific category (see part 4.2.4). For the other three stimuli the agreements were 76% to 86%, both for category agreement with and without context, and on this specific category. The other two stimuli with high agreement were perceived as Responsive New Information. One of these had a category agreement of 73% for stimulus with and without context, and 64% agreement on a specific category. The other stimulus had similar results, 71% category agreement for stimulus with and without context and 62% agreement on a specific category. This might suggest that Responsive Dispreference and Responsive New Information had the stimuli with most clear acoustic features, which may be characteristic features for these categories.

The responsives in the listening test were most frequently categorized as Responsive and least frequently as Responsive New Information. Test subjects gave comments on the difficulties in using the Responsive New Information category. Still, there is higher agreement on stimuli perceived as Responsive New Information than on stimuli perceived as Responsive. The Responsive category might have been used as the default interpretation of “ja”, used when there was uncertainty about which category to choose, but also Responsive cue phrases are more common among the responsives in spoken language (48%, see Chart 4.2), which might have affected the high use of the category in the test. The other categories might be perceived as more narrow.

For the stimuli with synthesised voice the agreement on category between the synthesised stimulus and stimulus in isolation with real speaker’s voice were low. There was also low agreement on a specific category for the synthesised stimuli. However, this was not true for two of the synthesised stimuli (see Chart 4.11). They showed high agreement on category when compared to their correspondent stimuli with real voice (81% and 90%). They also had a high agreement on a specific category; Responsive Dispreference (76% and 90%). Characteristic features in the Dispreference category then might be revealed even in the synthesised stimuli. One suggestion is that the duration is the most important feature for how these stimuli are interpreted. For the other stimuli it was harder to agree on a category, but their correspondent stimuli also showed relatively low agreement for a specific category. Stimuli perceived as Responsive was perceived as Yes/No answer among the synthesised, and vice versa. Watching the pitch curve and waveform did however not reveal an explanation to these results. This suggests that other acoustic features, such as voice quality, might be important for how these categories are perceived, features that might not have been presented in the synthesised stimulus or changed. To be able to generate responsive cue phrases, where the user perceive the intended intention of the system response, adding the right features to the responsives are important. Features may not be perceived the same with the synthesised voice, or they might need to be exaggerated for the user to perceive them in a correct way.

Overall, the agreement varied a lot between the different stimuli. This may show that context is an important feature for how we perceive the function of responsives. Gravano et al. (2007b) suggest that the only acoustic feature improving the results (without contextual features) was the feature of pitch digression at the right edge of “okay”. Still, they had studied the minimum, mean and maximum pitch, and also the intensity and duration for each stimuli of “okay”.
Acoustic features

When studying the acoustic features of the stimuli, pitch, intensity and duration were in focus (see part 4.2.4). Pitch for the different stimuli are difficult to compare since the stimuli are spoken by different persons. Especially between men and women there are big differences in fundamental frequency. Men often speak with a lower pitch than women. Nevertheless, pitch difference between the pitch maximum and pitch minimum might be a valuable feature, so also the shape of the pitch curve, even if this also varies between speakers. Intensity features are even more different to decide on. The intensity depends much on how the person is located according to the microphone while recording. If located closer to the microphone, naturally the intensity will be higher. However, as with the pitch, difference in intensity within stimuli might be studied, to see if there are any similar results for the stimuli in the same category. Yet it is important to have in mind that even if the microphones were attached to the body, and therefore considered as situated on a constant distance to the mouth, when moving the head up or down or in a sideways direction, there will still be a difference in distance. Duration measures might however be more trustworthy than the other acoustic features, since it can not be affected of the equipment, the setting or the use of different speakers with different voice quality and pitch. However, also duration may vary between speakers. If someone talks very fast, a slightly longer duration of a specific word or phrase may seem comparatively long. For this reason duration values needs to be normalized to each subject’s speaker rate to make comparisons more reliable.

All the Responsive Dispreference stimuli, with high category agreement, had longer duration than mean duration for all stimuli (446 ms). They varied from 689 ms to 1064 ms which may show that duration features might be important features for how we perceive responsiveness. Long duration responsive might be seen as hesitating or being doubtful about what has been said. The pitch curves with low pitch difference between maximum and minimum had higher category agreement in this category. The other had a slowly increasing pitch curve. This might indicate that responsives with a constant pitch or at least a pitch increasing slowly will be perceived as more Response Dispreference than others. The intensity increases. It is relatively low in the beginning of the stimuli to reach a peak somewhere between the middle and the end of the stimuli. After the peak the amplitude is falling again. Yet, long duration was the feature most noteworthy for the Responsive Dispreference stimuli.

When comparing the pitch curves of the stimuli with high agreement on Responsive New Information, the pitch was increasing more than in the other stimuli. The pitch differences were 126 and 138 Hz. The other stimuli with high agreement on a category showed lower difference in pitch, around 16 to 61 Hz. This might suggest that for a responsive to be perceived as a Responsive New Information, an increasing pitch curve is needed, with a high difference in pitch. The two pitch curves also contained to peaks, one in the middle of the stimuli, and the highest peak close to the end of the stimuli, and then the pitch falls again. These peaks were seldom seen in the other stimuli, and not in one of the stimuli with agreement on another category. This might suggest that, generating an increasing pitch with two pitch peaks, first one low and then the maximum peak, might be useful to generate responsives to be perceived as Responsive New Information.

There was not high agreement for any stimuli as Responsive category. However, studying the stimulus most agreed on as Responsive presents a pitch curve rising with a difference on 55 Hz. Stimuli agreed on as Yes/No answer show similar tendencies, with a pitch difference around 50 Hz. Stimuli agreed on as Yes/No answers had pitch curves that were somewhat different from the other pitch curves. They start by falling, and halfway they rise again. Even if this feature is not very clear, these tendencies are not seen in the other stimuli. Intensity is higher in the beginning of the stimuli, similar to the Responsive stimuli. Duration is short for both, but mean duration is shortest for the Yes/No stimuli (261 ms compared to Responsive mean duration which is 347 ms). This might explain why these have lower category agreement, since they might easily be confused.
6 Conclusion

The focus of this master’s project was to collect and analyse cue phrases for generation in spoken dialogue systems. In order to do this in a human-like way, extended knowledge was needed on how these phrases are used in human-human dialogue. By recording human-human dialogues in a role-play situation similar do the domain in DEAL, a suitable corpus was collected. One purpose of this master’s project was to find an annotation scheme for cue phrases suitable for DEAL corpus. The annotation scheme used in this project was based on the discourse-pragmatic functions of cue phrases suggested by Lindström (2008). Using this annotation scheme the inter-annotator agreement reached a Kappa value of 0.87 for the binary task, i.e. when classifying a word as a cue phrase or not. Inter-annotator agreement for the categorization of cue phrases had a Kappa value of 0.82. These values are very high, a kappa over 0.80 is considered to be a good reliability (Carletta, 1996). This indicates that the annotation scheme suited its purpose very well and that both annotators were well aware how to use the annotation scheme.

Another purpose of this master’s project was to analyse the cue phrases in DEAL corpus to see which cue phrases are used for different discourse-pragmatic functions and to study the frequency use for the different categories. Results show that Modifying and responsive cue phrases occur most frequently in DEAL corpus. This result is interesting since studies trying to define cue phrases and their functions often focus more on connective cue phrases and other cue phrases connecting the upcoming utterance with previous discourse. Modifiers often colour the present message (in middle position and end position) and the upcoming message when used initially. They give features to the hearer how to perceive the present message or even the messenger. Since they are used so frequently, they seem to have a really important role in the conversations. “ju” (Eng: actually, of course) are regularly used among the modifying cue phrases. Studying the use of responsive cue phrases shows that “ja” (Eng: yes) is the responsive most frequently used.

The main attention has been on the responsive cue phrases, trying to distinguish acoustic features which characterize different functions of the word “ja” (Eng: yes). A listening test was performed to see which features, contextual or acoustic, are important for how different realizations of “ja” are interpreted. 23 stimuli of “ja” were presented twice, isolated and in context. 21 subjects participated in the test. The Kappa values for the inter-annotator agreement were 0.286 for stimuli without context and 0.367 for stimuli with context. These are low Kappa values. However, a higher value for stimuli in context supports that contextual features are more important for the interpretation of “ja”. Yet, a few acoustic features of “ja” were found that may be important for the interpretation of some categories. Duration may be important for how we interpret responsive in the Responsive Dispreference category, and the shape of the f0-curve may affect how we interpret responsive in the Responsive New Information category.

Through this master’s project, the importance of cue phrases in spontaneous human-human conversation has been argued. Adding cue phrases may help to approach a human-like interaction between the user and the animated talking agent in DEAL. Cue phrases may be used to accomplish several features of human conversations, such as showing presence and acknowledge the speaker, grounding, and turn-taking among others. Adding cue phrases in right context with right acoustic features are important for how they are interpreted by the user. Thus it is important to analyse these contextual and acoustic features for each category of cue phrases, and also for specific cue phrases. Analysing the responsive “ja” (Eng: yes) was a first attempt to study the importance of contextual and acoustic features for how this word is interpreted.
7 Future Research

Much can be done to carry on the research of cue phrases for generation in spoken dialogue systems. The data collection is a good foundation for more research, and other experiments may also be done in order to collect more information about how cue phrases are used, how they can be generated, and also how they are perceived by the users if generated by the system.

7.1 Future research on data collection

1 hour and 45 minutes of dialogue were recorded. It was then transcribed orthographically and annotated with three annotation schemes; for cue phrases, communicative acts and disfluencies. This is a great collection of data for further studies. To enable generation of cue phrases in a human-like way, it is important to study patterns on how they are used in DEAL corpus. One way is to study if different cue phrases occur in combination with certain words or phrases. Machine learning or other statistical methods can be used to train the system to automatically identify when to generate cue phrases using features from the surrounding context. Disfluencies may also be a cue for how cue phrases are generated in spoken language. Studies on how cue phrases are used in combination with disfluencies such as pauses, filled pauses and corrections can also be interesting and of use.

In DEAL corpus, every speaker turn was annotated as different communicative acts (CA) (see Appendix 10.5). To further study the patterns on how cue phrases are used, communicative acts and cue phrases can be studied together. One way is to see for every interesting cue phrase, how frequently they are used in different CA’s. Also, one cue phrase may be used differently for different CA’s, thinking of positions, surrounding words and so on. Another way is to study which cue phrases are most common for each CA, and how they are used within that CA in terms of position and surrounding context. All this studies may give valuable cues on specific patterns for how cue phrases can be generated in spoken dialogue systems.

Acoustic features for cue phrases need to be studied for proper generation. Same cue phrases may have different functions. Is there any difference in how they are generated, or are they perceived differently because of its surrounding context? One example is the connective word “men” (Eng: but). In the DEAL corpus it is annotated mostly as Connective Contrast, but also as Connective Additive. Is it contextual or acoustic cues that make them perceived in different ways? Or maybe both? How would a cue phrase be perceived if it was generated in the “wrong” context?

The data collection can be used also to study other features of spoken language, such as disfluencies, non-lexical features, but also haggling strategies for the system agent.

7.1.1 Responsives

More research may also be done on the responsive cue phrases. In this master’s project, the responsives have been studied in terms of acoustic features such as duration and pitch shape. Still there are more acoustic features for possible analyses; time-based features such as duration of pauses before and after the responsive, pitch and intensity with normalized values, etc. The speech spectrogram can be studied to find exact extracted phonemes in the responsive stimuli. These can also be studied with time-based features to find how long the different phonemes are. A long generation of the first letter “j” in “ja” (Eng: yes) may be a feature of hesitation and thus
belongs to cue phrases of Responsive Dispreference. More features may give more evidence on characteristic features for different functions of responsives. In this project “ja” (Eng: yes) was studied, but listening tests with other responsive cue phrases can also be performed. Using a dialogue system, there may be situations where the system only responds with a single responsive, without needing a whole sentence. Then the system needs to generate the “right” responsive if different acoustic realizations are interpreted differently. How would a responsive be perceived if it was generated in the “wrong” context?

7.2 Optional experiments

The data collection is a good foundation for further research on cue phrases. However, further experiments may be interesting to perform as complement. It may be interesting to study how cue phrases are perceived if used in dialogue systems and if they affect how the users behave using the system. A Wizard-of-oz test with a real person playing the system agent can be used to evaluate how future generation of cue phrases affect how users perceive the system and how they behave when speaking to the system.

The DEAL dialogues were also recorded with a video camera. Studies of the video and sound recording together can give information on how facial expressions and movement of face and body are combined with cue phrases. Results from this may be added to the agent’s facial behaviour in DEAL, to further express what the system wants to say. To continue with the responsive cue phrases, this could also give facial cues on how to generate the responsives for the user to perceive them in the intended way.
8 Acknowledgement

This Master’s thesis has been carried out at the Royal Institute of Technology (KTH), in the department of Speech, Music and Hearing in Stockholm, Sweden, between January and June in 2008.

To carry out this project I received help from several people. I would like to take the opportunity to give my special thanks to my supervisor Anna Hjalmarsson for giving me valuable advices and assistance throughout my work. Thank you for all encouragement and for your patience with all my questions. In addition, the data collection and annotation was performed together with Anna. I would also like to thank my examiner Rolf Carlson for his support. Furthermore I would like to thank Joakim Gustafson for helping me to extract the synthesised stimuli from their original. The dialogue recording was performed together with Anna and Kahl Hellmer and I would like to thank all who participated in the role-play. Also, thanks to all who participated in the pilot test and in the final listening test.

At last, I would like to thank my family and friends for great support.
9 References


Books


10 Appendix

10.1 Questionnaire before recording

Fp nummer: _____

1. Kön: Man ( ) Kvinna ( )

2. Jag är ___ år gammal.

3. Har du några erfarenheter av att förhandla om pris?

   NEJ ( )

   JA ( )

3a. Förhandlade du om pris utifrån ett försäljarperspektiv?

   ( ) ja   ( ) nej

   Beskriv kortfattat tillfället och eventuell strategi som du använde:

3b. Förhandlade du om pris utifrån ett kundperspektiv?

   ( ) ja   ( ) nej

   Beskriv kortfattat tillfället och eventuell strategi som du använde:
10.2 Instructions/scenarios

Kund

Instruktioner

Du kommer att vara med om 2 inspelningar. För varje inspelning får du ett nytt scenario beskrevit för dig.

Scenario
I

+ Tänk på att en glad försäljare ofta kan vara mer benägen att gå med på dina krav/sänka sina krav (Smicker).
eller

+ Undersök varorna noga så att du inte blir lurad och köper något som inte fungerar (Defekta varor).

II
Du har nyligen köpt ett renoveringsobjekt. Ett fint gammalt hus där både byggnaden och trädgården behöver en ordentlig renovering. Du har inga verktyg och innan du går och köper nya i den dyra järnhandeln har du 350 kronor att spendera på loppmarknaden. Försök att få minst 3 olika saker till ditt nya hus.

+ Tänk på att en glad försäljare ofta kan vara mer benägen att gå med på dina krav/sänka sina krav (Smicker).
eller

+ Undersök varorna noga så att du inte blir lurad och köper något som inte fungerar (Defekta varor).

III
Du har en ung släkting som snart fyller år. Han har nyligen börjat lära sig klocka i skolan och du tänkte att det vore roligt att ge honom en klocka i present så att han kan öva hemma. Du är ute efter både en analog och en digital klocka för bästa möjliga inlärning för din släkting. Du har 250 kr till ditt förfogande. Tänk på att en glad försäljare ofta kan vara mer benägen att gå med på dina krav/sänka sina krav (Smicker).

66
Försäljare

Instruktioner


Du kommer att vara med om 4 inspelningar. För varje inspelnings får du ett nytt scenario beskrevet för dig.

Några tips (om du vill) innan du börjar:
  o Börja med att dubbla inköpspriset.
  o Försök få en vinst på minst 10%.
  o Om kunden ger skambud, eller använder ”fula” strategier ska du inte sänka priset utan försöka argumentera på annat sätt, till exempel påpeka varans förrådförsörjning.
  o När kunden börjar nära sig det pris du tänkt dig kan du börja gå ner i pris.
  o Om kunden argumenterar väl, bör du övervaga att sänka priset.

Scenario

Defekta varor
Du är på gott humör och resonabel. Om kunderna hittar fel eller brister hos dina prylar är du villig att sänka priset.

Smacker
Affärerna har gått dåligt på sistone och du är på dåligt humör. Dessutom är allt fler kunder otrevliga och du tycker inte ditt jobb är lika roligt som tidigare. Ibland är allt som krävs lite småprat för att du ska bli glad och mer villig att ge kunderna ett bra pris.
10.3 Questionnaire after recording

FÖRSÄLJARE

Försöksperson nummer: ___

Beskriv vilka/ vilken strategi du använde för att få kunden att betala så mycket så möjligt:

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

Beskriv kunden (tex. kön, ålder, humör, vilken typ av varor han/hon var intresserad av):

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

Vilken strategi/vilka strategier hade kunden för att få dig att sänka priset?

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

KUND

Försöksperson nummer: ___

Beskriv vilka/ vilken strategi du använde för att få ett billigt pris:

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

Beskriv försäljaren (tex. kön, ålder, humör, typ av varor till försäljning):

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

Vilken strategi/vilka strategier hade försäljaren för att få dig att betala ett högt pris?

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

68
10.4 Annotation scheme for cue phrases in DEAL corpus

Kategorier: Additive Connective
Additive Contrastive
Additive Alternative
Responsive
Responsive New Information
Responsive Dispreference
Response Eliciting
Repair Correction
Modifying
Referring

Nedan förklaras dessa mer ingående:

Konnektorer: Knyter ihop yttrande med tidigare diskurs, vilken konnektor som används påverkar hur yttrandet ska tolkas.

**Person 1:** jag tänker inte betala för en produkt som är defekt
**Person 2:** men vi är ju ändå på en lojmarknad

Konnektor är här ordet 'men'. Om konnektorn 'och' placeras dit istället kan meningen med yttrandet tolkas på ett annat sätt.

Konnektorer kan förekomma i slutet, mitten och början av turer för att knyta ihop med vad föregående talare sa, eller knyta ihop delar av sitt eget tal. En exakt definition av hur en konnektiv diskursmarkör skiljer sig från ett ”och” eller ”men” som inte är en diskursmarkör är inte helt enkelt men den ”hörs” snarare än ”syns”. En diskursmarkör är snarare en fras som knyter ihop ”chunkar” av tal än enstaka ord. Tex:

Inte dm: Jag vill ha en grön OCH blå klocka
Dm: Jag vill ha en grön klocka …. ELLER blå skulle den kunna vara också

Ett sätt att se det är att tänka bort konnektorn för att se att dess bortfall inte gör skillnad på det egentliga innehållet, (fast ett byta av konnektor fortfarande ändrar hur yttrandet ska tolkas). Detta är vagt uttryckt och inte helt korrekt eftersom diskursmarkören faktiskt har en inverkan på det som sägs, men på en liten högre nivå än andra ord.

Inte dm: Den är röd **OCH** grön
JÄ dm: Den är röd **OCH** jag tycker att vi ska köpa den

OCH kan i andra yttrandet tas bort utan att det egentliga innehållet ändras. Ett byte av OCH till exempelvis MEN ger dock yttrandet en annan betydelse. Vilken konnektor som ska användas beror inte enbart på vilket ord som används som markör, utan också sammanhanget har betydelse.

Additive Connective

"att" "och" "så" "alltså" "förresten" "då" "för att" "eftersom"

Lägger till information som hör ihop med tidigare diskurs och kan även vara slutsatser utifrån det som sagts tidigare.
Additive Contrastive  "fast"  "men"  "alltså"  "i alla fall"
"i och för sig"
Ger ett medgivande, kontrast eller förnekar tidigare diskurs.

Additive Alternative  "eller"  "istället"  "annars"
Ger ett alternativ till något i tidigare diskurs.

Responsiver:  Att ge respons på att man lyssnar och kanske följer med i dialogen. Chunkas så stort som möjligt, om man inte hör att responserna har olika funktion.
Svar på direkta frågor är INTE diskursmarkörer:
Ex:  - Vill du ha gröt
- Ja

Responsive  "m"  "ja"  "jo"  "nej"  "exakt"  "precis"
Man visar att man lyssnar, följer med i dialogen. Inleder ofta turer.

Responsive New Information  "aha"  "jaha"  "okej"  "eller hur"
Att visa att man tar emot information som är ny, eller att man verkligen lyssnat på och förstått eller är lite förvånad över det som sagt.

Responsive Dispreference  "ja"  "i och för sig"
Visar en osäkerhet om man håller med om det som sägs, alternativ inte håller med.

Ex:  - Vad vill du ha för den då
  - det är väl bättre om du säljer den
  - du tänker köpa den eller
  - du kommer ihåg det va

Yttranden som vill ha ett svar tillbaka. I de sista tre yttranden, om markören tas bort, dvs. "väl"  "eller"  "va" så uppfattas inte längre yttrandet som en fråga.

Response Eliciting  "va"  "väl"  "eller"  "då"  "förstår du"
Alternativt hela yttranden:  "kan det va nått"
"vad säger du om det"  "Den här menar du".

Reparerande:  Markörer som används vid felsägningar eller när man letar efter ord. Här markeras inte ordsökande markörer som "eh", då de istället markeras upp som
fyllda pauser, som disfluenser. Däremot ska ord som korrigeras en felsägning märkas upp.

ex: - jag vill ha en blå boll **nej jag menar** en röd

**Repair Correction**  "nej"  "jag menar"  "eller"  "alltså"

**Nyanserande:** Ord som färger yttrandet och påverkar hur vi väljer att uppfatta det som sägs, men som kan tas bort utan att det egentliga innehållet ändras.
Typifierande: "liksom"  "typ"  "sån där".
Garderande: "i alla fall"  "atminstone"  "i och för sig"  "kanske".
Evaluerande: "faktiskt"  "allvarligt talat"  "så".
Epistemiskt: "så"  "väl"  "nog"  "ju". Som påverkar sanningshalten i det som sägs.
Andra: "om man säger så"  "menar jag"  "tycker jag"  "ju"  "kan man tycka" som subjektiverar yttrandet.

Inte dm: Jag tycker om gröna bollar

Denna grupp innehåller också yttranden som ytterligare poängterar vad som sägts:
"du kan ju inte göra så **förstå du**".
"den är ju inte så användbar **det kan man inte säga**"

**Modifying**  "ju"  "så"  "faktiskt"  "kanske"  "väl"  "nog"
"tror jag"  "menar jag"  "förstå du"

**Refererande:** Yttranden som pekar på/refererar till tidigare diskurs

**Referring**  "som sagt"  "som jag sa tidigare"
"vad var det hon hette **nu igen**"
"vad var det vi sa"
10.5 Annotation scheme for disfluencies and communicative acts

<table>
<thead>
<tr>
<th>Disfluency</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filled_pause</td>
<td>Filled pauses between words and within words, e.g. “bluuuuuuu” and “eh”</td>
</tr>
<tr>
<td>Pause</td>
<td>Long pauses within utterances. Often obvious that the utterance will continue.</td>
</tr>
<tr>
<td>Repeat</td>
<td>Repetitions of words or phrases, e.g. “if if it”, “yes yes”, “I could I could” and “b because of”</td>
</tr>
<tr>
<td>Correction</td>
<td>Corrections of what was previously said, e.g. “I’m w* driving home”. Also restarts with corrections, e.g. “I want I could”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Communicative act (CA)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confirmation</td>
<td>All sorts of confirmations, such as “m”, “yes”, “okay, the green one”. Utterances were what said is confirmed, or what was said is repeated and maybe agreed on, but no new information has been added.</td>
</tr>
<tr>
<td>Social</td>
<td>Social utterances dealing with things other than the objects in the store or the prices (does not include phrases of politeness).</td>
</tr>
<tr>
<td>Polite</td>
<td>Polite utterances, e.g. “thank you” and “that was nice of you”.</td>
</tr>
<tr>
<td>Greeting</td>
<td>E.g. “hello” and “hi”.</td>
</tr>
<tr>
<td>Fragmental</td>
<td>Utterances not completed, that complicate the conveyed interpretation of the utterances e.g. “and L.”.</td>
</tr>
<tr>
<td>Discourse_meta</td>
<td>Utterances concerning the main topic, but not explicitly the objects and prices, e.g. “what did we say” and “shall we continue”.</td>
</tr>
<tr>
<td>Object_request</td>
<td>Implicit and explicit requests for objects, e.g. “I would like to have a blue clock” and “do you have any nice products”.</td>
</tr>
<tr>
<td>Object_claim</td>
<td>Descriptions, statements or assertions of an object.</td>
</tr>
<tr>
<td>Object_suggestion</td>
<td>Give suggestions to the customer on what items to buy, e.g. “what do you think about this clock?” and “maybe you want them both?”</td>
</tr>
<tr>
<td>Object_presentation</td>
<td>Present objects for the buyer, e.g. “here!”, “here is a nice clock”, “what do you think about this blue teddybear?”</td>
</tr>
<tr>
<td>Request_for_objectspecification</td>
<td>Ask the customer if he/she needs help and what he/she wants to buy, e.g. “okej, what can I help you with?” and “what are you looking for?”</td>
</tr>
<tr>
<td>Point_out_object_of_interest</td>
<td>When the customer shows interest to a certain object or a certain category of objects, e.g. “I am interested in that clock” and “I can see that you have both a red and a green doll”</td>
</tr>
<tr>
<td>Price_request</td>
<td>Questions concerning the price.</td>
</tr>
<tr>
<td>Price_suggestion</td>
<td>Utterances suggesting a price for an object, e.g. “what about 100 crowns”</td>
</tr>
<tr>
<td>Price_haggle</td>
<td>Different tactics for trying to raise or lower the price of an object, without talking about the shortage or excellence of the object because then it is marked with object_claim. E.g. “I think that is far too expensive” and “I can’t reduce the price any more, I have my children to think of”.</td>
</tr>
<tr>
<td>Money_request</td>
<td>Requests for the money, e.g. “are you not going to pay?” and “give me the money”.</td>
</tr>
<tr>
<td>Deal_request</td>
<td>“shall we make it a deal”</td>
</tr>
<tr>
<td>Deal</td>
<td>Agreement of a deal, e.g. “yes, let’s make this a deal”, “yes, I take that”.</td>
</tr>
</tbody>
</table>
10.6 Cue phrases in DEAL corpus

Connective Additive

Connective Contrastive

Connective Alternative

Response Eliciting

Modifying
10.7 Instructions to the listening test

INSTRUKTIONER

Detta test görs inom mitt examensarbete vid KTH, på institutionen TMH (Tal, Musik och Hörsel). Din uppgift är att lyssna på ett antal ljudfiler och sedan avgöra vilken kategori du tycker att dessa tillhör. Dina svar och personuppgifter behandlas anonymt. Läs följande instruktioner nogal!

1. Du kommer få lyssna till ett antal ljudfiler, varav alla säger ordet ”ja” på olika vis.
2. Din uppgift går ut på att lyssna på en ljudfil i taget och välja vilken av fyra kategorier du tycker att ordet tillhör. Du kan lyssna på ljudfilen så många gånger du vill.
4. Testet innehåller 46 st ljudfiler med en människas röst som du ska bedöma. I slutet förekommer även 7 st ljudfiler med syntesröst. En del ”ja” förekommer utan kontext och andra ”ja” förekommer i en kontext. När ”ja” förekommer i en kontext så är det i formen:

   Talare 1: Säger något
   Talare 2: ”ja”

Det är det ”ja” talare 2 säger som du ska bedöma och för det välja en passande kategori.
Kanske kommer du tycka att det fattas någon kategori. Försök ändå att välja den kategori du tycker passar bäst.

[Ljudspelare]

Definition av kategorierna

När vi talar med varandra så ger vi varandra respons för att visa att vi lyssnar eller för att visa att vi inte håller med om vad som sägs. Vi ger också respons när vi får en fråga. De ’ja’ som du ska bedöma är responsen av olika slag. Du kan välja mellan nedanstående kategorier:

Bekräftande respons: ”ja” som bekräftar att man lyssnar och/eller hänger med på och accepterar vad den talande sagt.

Nyhetsmottagande respons: ”ja” som visar att man fått reda på något nytt, eller visar att man tagit emot information som man verkligen lyssnat på och förstått.

Tvekande respons: ”ja” som visar att man inte riktigt håller med om det som den andra har sagt eller ställer sig tveksam till det.

Ja som svar: ”ja” som svar på en fråga som kan besvaras med antingen ”ja” eller ”nej”

[Till test]