UPPSALA
UNIVERSITET

User response strategies to reprompting in a call routing service

Nils Hagberg

Uppsala University
Department of Linguistics and Philology
Språkteknologiprogrammet
(Language Technology Programme)
Bachelor's thesis in Computational Linguistics, 15 credits

December 23, 2010

Supervisors:
Joakim Nivre – Uppsala University
Robert Eklund – Voice Provider
Håkan Jonsson – Voice Provider

Abstract

In this bachelor's thesis I have investigated users' responses to two commonly used reprompts in a dialogue system. The aim has been to increase knowledge on users' behavior in order to develop more efficient dialogue. Data consisted of calls to a data collection application for a call routing service being developed by Voice Provider. 3 679 calls were selected from a total of 14 989 calls. The selected calls were calls that consisted of initial utterances that had not been understood and that had subsequently been reprompted at random with either a follow-up prompt or a repair prompt. Differences in users' responses were measured with utterance features that reflected effects that were predicted from prompt wordings. The result was that the differences in wording of reprompts affected utterances' length, similarity and content as well as the users' willingness to interact as had been predicted. The most striking and least anticipated result was a difference in users who expressed unwillingness to interact with the system. The result of the study also indicates that there are general user behaviors in a reprompting dialogue regardless of prompt wording. The conclusion was consequently that fine tuning of reprompts can enhance or diminish effects on user's response, but that there is an underlying behavior in reprompt dialogues that is difficult to alter. I suggest that future research should focus on ways to understand all types of responses with the intention of adapting to how users express themselves instead of focusing on how to shape users' input.

Sammanfattning

I denna kandidatuppsats har jag undersökt användares gensvar på två ofta använda återpromptar i ett dialogsystem. Syftet har varit att bidra till ökad kunskap om användarens beteende för att kunna utveckla mer effektiva dialoger. Data utgjordes av telefonsamtal till en datainsamlingsapplikation för ett röststyrt samtalsstyrningssystem under utveckling av Voice Provider. Från 14 989 insamlade samtal valdes 3 673 samtal ut som bestod av icke förstådda initiala yttranden där användaren slumpvis hade återpromptats med antingen en reparationsfråga eller en uppföljningsfråga. Skillnader i gensvar mellan promptalternativen mättes med hjälp av utvalda kännetecken av yttrandena, som återspeglade förväntade skillnader i formulering av promptalternativen. Resultatet blev att formuleringen av promptalternativen hade förväntad effekt på längd, likhet och innehåll av användarens gensvar, samt på perplexitet uttryckt av användaren. Det tydligaste och minst förutsedda resultatet var en skillnad i användarens uttryck för ovilja att interagera med systemet. Resultatet av undersökningen pekade även på att det finns generella beteenden hos användare vid en återpromptning oavsett frasering av återprompten. Slutsatsen blev följaktligen att finjustering av promptformulering har en förstärkande eller förminskande effekt på användares gensvar, men att ett grundbeteende hos användare svårligen ändras. Jag föreslår att framtida utveckling bör fokusera på sätt att förstå alla typer av svar och anpassa sig efter hur användaren uttrycker sig istället för att fokusera på sätt att forma användares tal.

Contents

Preface

This bachelor thesis has been conducted at Voice Provider Sweden AB. In parallel with writing the thesis I have taken part in creating and evaluating the call routing application that was the subject of the study. I have mainly contributed by transcribing speech and developing the semantic tagset. In addition I have written a Java program to extract information from collected calls.

Acknowledgements

I would like to express my profound gratitude to Voice Provider CEO Bengt Persson and CTO Per Sautermeister. First for providing me with the opportunity to write this thesis and secondly for their support, patience and determination that was crucial for the completion of the thesis. I thank Kronofogdemyndigheten for granting me access to their customer service data. My sincere gratitude to my supervisors at Voice Provider Robert Eklund and Håkan Jonsson. I thank Robert especially for his thorough research, extensive comments and his valuable insight in the scientific process. I thank Håkan especially for his knowledgeable and insightful analysis and comments on the results, and for his extensive knowledge in dialogue design. I thank Robert and Håkan both for their time and effort, but most for their sincere interest, enthusiasm and determination which have been a great support for me. I thank Kristian Ronge at Voice Provider for his tech support and his keen interest in discussing the small details as well as the not always relevant details. I thank Joakim Nivre at Uppsala University for his sensible support and advice. And for a valuable discussion of the statistical analysis. I would like to further acknowledge the staff not mentioned at Voice Provider for their acceptance, patience and positive mindset.

1 Introduction

Historically, automated telephone service applications have adopted controlled prompting strategies with *directed* questions or *menu* strategies that confine the user to utter digit strings or isolated words. Advances in speech recognition technology have enabled development of high accuracy automated telephone service applications utilizing natural language technology. These applications allow *open* prompting strategies where callers may express their reason for calling in their own words, using continuous speech.

A common use of automatic speech recognition (ASR) and natural language understanding (NLU) in automated telephone services is in call routing applications, where the task is to get sufficient information from the user to route the call to an appropriate destination or customer agent. Other applications include form filling and a wide range of information services.

An advantage of an *open* strategy using natural language is that, despite an increase in size and complexity of the service, the user interface remains simple. Typically a welcoming message followed by an open question of the type "How may I help you?" (Gorin et al., 1997).

The increased freedom for users to use their own words, instead of replying to the service or choosing an option from a menu, highlights the problem of understanding what they say. To construct Context Free Grammars (CFG) for such a task, extensive manual work and linguistic competence is required. Additionally their rigid structure does not apply well to spoken dialogue. Statistical methods for language modeling have proven far more capable of understanding natural speech.

When designing dialogue (or voice user interfaces), the designer must consider requirements from its client or from users (in this case callers) which may have conflict the desire to design effective dialogue. An example of this conflict is a common desire from clients to keep the introduction short, despite that Williams and Witt (2004) and others have found that longer greetings produce less confusion in the initial utterance.

The accuracy of the speech recognizer is of course of great importance as well as the quality of the users' first utterance which have been the subject of several studies. However, non-understandings will always occur and for this reason the application must have effective error handling.

Dialogues with many turns and complex structure increase the risk of error. In order to maintain a short and simple dialogue while still extracting enough detailed information to route the call, the dialogue needs to be effective. User responses that provide no information, no new information or in other ways do not answer the question prompted to the user can be considered wasted dialogue turns, which will complicate the interaction.

Statistical approaches do not eliminate the risk of mis- and non-understandings caused by users who do not produce well-formed speech or user utterances that do not correspond to the language model/grammar. To recover from these errors is vital in order to maintain dialogue and route the call. This thesis will focus on the reprompting of first utterances that are not understood (see chapter 3).

This thesis has been carried out at Voice Provider Sweden AB, under the supervision of two dialogue designers. The study performed in the thesis has been conducted in parallel with the development of a natural language call routing application for the Swedish Enforcement Agency (Kronofogdemyndigheten), who has authorized my use of their data. The entire data collection procedure was performed between January 2010 and June 2010.

1.1 Purpose of the thesis

This thesis will seek to answer the question: Does wording of a reprompt have an effect on users' response? Moreover I will investigate what part of the wording that might have an effect. In addition I will seek to find indications of users' interaction strategies by investigating users' response – what in the response is a function of prompt wording and what is a function of the users' interaction strategies?

The purpose of the thesis is to contribute to the knowledge required in the dialogue design process of natural language call routing applications of the type: "How may I help you?". The aim is to contribute to a deeper understanding of the effect of two common recovery (re)prompting strategies from dialogue situations where the user initially has not been understood by the call routing system. With increased understanding of user behavior in an error handling situation, the dialogue designer can construct the dialogue to avoid silent or non-informative utterances, or to know when to bail out from the dialogue.

The focus of this thesis is two common recovery strategies from a real data collection for a call routing application in the context of when the first user utterance has not been understood.

In the study users have been prompted with two different reprompts:

follow-up
: A short request to provide additional information, not signaling to users that any misunderstanding have taken place and with no further instruction of expected form of user input.

reject
: A longer request that signaled to users that the application had not understood the previous utterance, instructed users of expected form of the input and repeated the initial main question.

The properties of users' response that will be further investigated are expected effects from prompt wording differences, namely that the two prompt alternatives will have an effect on:

Utterance length

Utterance content

User willingness to interact with the application

Utterance similarity

User confusion

This knowledge may not necessarily have a direct impact on user task completion but might shorten the dialogue and make the interaction process more effective, possibly increasing user satisfaction.

1.2 Outline of the thesis

The remaining chapters in the thesis is structured as follows. Chapter 2 describes spoken dialogue systems in general and natural language call routing in particular. Chapter 3 describes the general approach. Chapter 3 is structured into three sections. Section 3.1 account for the domain of the study, section 3.2 describes the data that was used and section 3.3 describes the specifics of the setup of the experiment. In chapter 4 I will describe the result and evaluate the hypotheses that the two prompt variations resulted in different user responses. In chapter 5 my results will be discussed. Based on that discussion I will further discuss the results in the light of users' interaction strategies. Finally, in chapter 6 I will present what conclusions was drawn, the applicabilty of those conslusions and propose the direction of future studies.

2 Background

2.1 Spoken dialogue systems

Spoken dialogue systems include a variety of different systems where humans and computers interact in turns, using speech. The challenges for spoken dialogue systems are mainly automatic speech recognition (ASR), natural language understanding (NLU) and natural language generation (NLG).

The speech recognizer outputs one or more interpretation hypotheses that the natural language understanding module processes and respond to via the language generation module.

The possible combinations of phonemes are practically infinite. Speech recognizers therefore include a language model to constrain the possible interpretations. Language models range from more complex statistical models using Markov chains to the most simple context free grammars.

Spoken dialogue systems can be divided into *user-*, *system-* or *mixed-initiative* systems and are typically described as either *tool–like* or *anthropomorphic*. Edlund et al. (2006) argue for a view on the human–computer interaction as guided by an interactional metaphor, e.g. an interface metaphor or a human metaphor.

2.2 Natural language call routing

Call routing is a common task for spoken dialogue applications where the task is to direct users, in this case callers, to a given destination. Historically the call routing task has been handled with touch-tone menus but they are increasingly being replaced with systems capable of handling natural language. These systems were pioneered by AT&T's HMIHY system (Gorin et al., 1997) and Bell Labs Canada's system named Emily (Hafner, 2004).

Dialogue design strategies originate from touch tone menu interfaces. Advances in ASR have enabled design approaches that mimic human–human interaction, either with open ended prompts such as "How may I help you?", or more directed, system-initiated dialogue such as "Are you a registered customer?". The different approaches in dialogue design vary as a function of domain, the users' task, and the magnitude of the call routing problem.

Menu-based approaches have been inherited from touch tone-applications and often suffer from a rigid hierarchical structure. Menu interfaces introduce a *mapping problem* to the users where they need to map their task to an alternative in the top menu, a task aggravated by the transient nature of speech.

The *open* dialogue strategy enables users to state their reason for calling

using continuous speech. This shifts the *mapping problem* from the user to the application. When comparing touch-tone menu applications and call routing applications using speech recognition, Suhm et al. (2002) found that users both preferred using speech and were more often routed to the correct destination.

Figure 2.1 illustrates a natural language spoken dialogue system that is essentially made of a speech recognizer, a dialogue manager and an interface to the user. The speech recognizer applies constraints from a language model on the phoneme recognition to produce speech result hypotheses. The speech result typically enters the dialogue manager via a mapping function from text string to a more abstract representation, a process often referred to as natural language understanding. The dialogue manager will often operate by filling slots that will guide the system if it has enough information to route the call, or if it needs to reprompt the user.

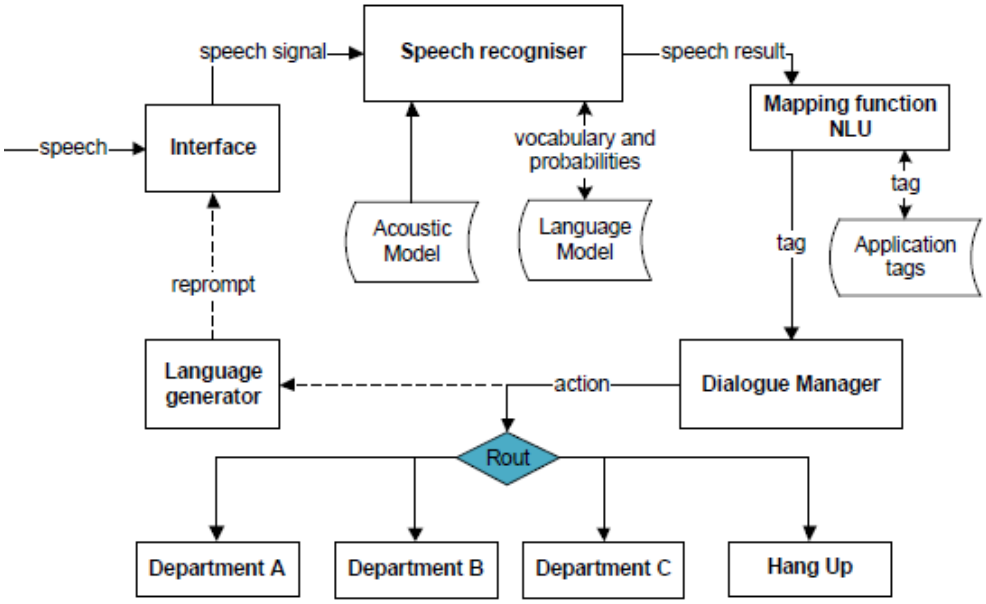

Figure 2.1: Sketch of a call routing application.

2.2.1 Error handling

Reprompts follow when users need to provide additional information, either because the preceding utterance was not understood, was misunderstood or because the preceding utterance did not contain enough information. Reprompts that followed utterances with too little information typically take the form of prompts that requests additional or clarifying information (a *follow-up* prompt or *clarification* prompt). Mis- or non-understood utterances typically lead to prompts where the question is somehow repeated (a *reject* prompt). Handling of the errors are crucial for the performance of the service, especially when seeking to improve an existing service.

In terms of error handling in the dialogue the difference between mis- and non-understandings is mainly that an occurrence of a misunderstanding is not

noticed by the application and might not be noticed by the user, whereas a non-understanding is directly noticed by the application and demands immediate repair. For a more in-depth discussion of error handling and the difference of mis- and non-understandings, see Skantze (2007).

Applications typically separate non-recognized utterances into those that *i*) lack speech entirely; *ii*) those with too much speech; or *iii*) those simply not recognized (or lie below a pre-set recognition confidence threshold). This enables the application to reprompt users differently for different non-understandings.

There is a great variation of (combinable) strategies to recover from a non-understanding. Examples of strategies presented below are quick typical examples and by no means an exhaustive representation.

To signal error: *I'm sorry...*

To provide feedback to the user what has been understood (i.e. nothing): *I didn't quite understand*

To instruct the user: *Using your own words, tell me briefly...*

To provide examples of what to say: *You can for example say "I want to book a flight", "refund" or "I want luggage information."*

To keep face (not provide any feedback): *Say something more about...*

Repeating the main question *... How may I help you?*

Rephrasing the main question *Tell me how I can be of assistance.*

Informing of the expected input *This is a voice controlled service, tell me...*

Reprompts do not necessarily vary only in wording but also vary in duration or emotion.

2.3 Building a natural language call routing application

Boyce (2008) describes the process of building a natural language call routing application and states that hand crafting grammars for natural speech is exceptionally difficult due to the infinitely flexible use of spoken language. A statistical approach with statistical language models (SLM) is far more suitable.

In order to create SLMs, data need to be collected and orthographically annotated. Following the collection and annotation of calls, the call flows and application code are developed as well as the SLM and NLU module (sometimes referred to as SSM). In the NLU module utterance strings can be mapped directly to system reprompting and routing actions or utterance strings can be mapped to semantic representations of the utterance. The semantic representations of the utterance can then be mapped to dialouge actions (such as a reprompt or a route) in the dialogue manager. A two-layered mapping from user utterance to application behavior is described in Boye and Wirén (2007). The non-trivial task of mapping utterance strings to semantic representations of the utterance meaning is typically managed by neural networks or by developing a robust parser grammar (RPG).

The final steps in the making of a natural language call routing application is testing and post-deployment tuning of the application.

2.3.1 Common data collection methodology

User data are collected with the scientific purpose of understanding user behavior, or with the more pragmatic purpose of predicting user behaviour. In both cases it is paramount that the data are a representative sample of the population.

A Wizard-of-Oz data collection application is made to be perceived as an authentic automated application but is in fact operated by a human, a wizard. Wirén et al. (2007) describe Wizard-of-Oz data collections as being "regarded as a superior (though not unproblematic) method of collecting high-quality, machine-directed speech data in the absence of a runnable application."

With a Wizard-of-Oz data collection methodology, collecting data of sufficient size to train a speech recognizer is more costly than simpler collection methodologies that do not involve wizards.

A perhaps more important feature of the data collection methodology is whether it is used *in-service* or in a laboratory setting. *In-service* methodologies will benefit from the use of real users with authentic tasks but may suffer from the increased difficulty of performing post-experiment interviews.

2.4 Previous studies

Boyce (2008) points out that that even small changes in wording of reprompts have an effect on responses. The claim is based on a study using data from AT&T's call routing application, described in Gorin et al. (1997). In the study the wording of reprompts varied in terms of how instructive they became.

Instructive prompt: "I'm sorry. Please briefly tell me how I may help you?"

Short prompt: "I'm sorry. How may I help you?"

Table 2.1 displays results from Boyce (2008), showing that when users were reprompted with the "instructive" prompt users' responses were labeled as "productive" to a greater extent (11 percentage points) than for users reprompted with the "short" reprompt.

Table 2.1: Percentage of responses to reprompts by category from a study by Boyce (2008). The category "Productive" is a term used by Boyce (2008) as a summarization of the categories: Rephrased, Shorter, Fewer Ideas and Changed Request

Category of Response	Short Reprompt (% of responses)	Instructive Reprompt (% of responses)
Same	42%	34%
Rephrased	6%	11%
Shorter	14%	18%
Fewer Ideas	8%	6%
Longer	14%	11%
More Ideas	12%	12%
Changed Request	4%	8%
Productive	32%	43%

McInnes et al. (1999) compared three opening prompts for a banking call routing application:

Open strategy: *"Main Menu – How can I help you?"*

Mid strategy: *"Main Menu – Which service do you require?"*

Closed strategy: *"Main Menu – Please say 'help' or the name of the service you require."*

McInnes et al. (1999) found that when the possibility to say "help" was explicitly mentioned, users overwhelmingly said "help" (70%). When it was not mentioned, none of the callers said this (0%). The open and mid prompts both yielded longer responses but the open prompt generated fewer keywords and fewer silent utterances. McInnes et al. (1999) (page 164) concluded that "there were significantly more silences overall in the mid condition (0.9 per participant) than in the open condition (0.3). [. . .] This suggests that an open ended prompt such as 'How can I help you?' encourages people to say something even if they do not know exactly what the options are, whereas a specific prompt such as 'Which service do you require?' may cause users to remain silent if they are not sure of the exact option name."

Suhm et al. (2002) found that commonly adopted menu interfaces generated confusion, and were a source of error in systems. The authors concluded that a directed strategy was not suitable for larger applications. In addition they also found that users preferred using speech instead of using touch tone.

Sheeder and Balogh (2003) (page 110) compared different variations of open prompts. They experimented with placing examples of what to say preceding or succeeding the main query, and using a "natural" or "keyword" strategy for those examples. The option with examples placed preceding the main question resulted in higher rates of task completion. However, only 13% of the users waited to listen for the examples succeeding the main query. The study found no significant differences in efficiency (number of turns), number of words or length and no correlation between utterance length and task completion. The study reveals that the preceding option generated fewer errors (e.g. silent utterances or rejections).

Sheeder and Balogh (2003) also found that users interacting with the system that prompted the user with examples of what to say using keywords responded with keywords not fitting to their task. This suggests that if users are presented with a fix set of keywords the *mapping problem* of those keywords to the user task was a source of error, as have been shown for other highly constrained dialogue systems (Carpenter and Chu-Carroll, 1998).

In summary Sheeder and Balogh (2003) concluded: "call-routing applications generally will benefit from rethinking the currently employed strategies. Namely, that providing callers with examples that they are likely to hear (i.e., prior to the initial solicitation for a request) and focusing on imparting a sense of the expected form of the request, as opposed to an arbitrary semantic category label, should be the normative prompting strategy."

In a laboratory study on how users interact with a spoken dialogue system with a limited grammar, Tomko and Rosenfeld (2004) found that users readily adapted to the limitations of the system, using information from rejections and confirmations.

Williams and Witt (2004) studied dialogue strategies for a call routing application across two domains. Three dialogue strategies were compared:

Open *"How May I Help You?"*

Menu A list of choices

Directed Multiple question pairs. *"Do you have an account number?" [positive reply] "What is your account number?"*

The authors also varied the greeting prompts with three different types:

- Greeting
- Greeting + Earcon ("sound logo")
- Greeting + Earcon + persona & service recently changed notice

They found the followin.

- The longest greeting produced the least confusion in users' first utterance.
- Directed strategies which most closely model human–human conversations resulted in higher task completion rates than an open strategy.
- Most of the failures in the main menu strategy were due to selecting the wrong item from the main menu.
- Most of the failures with the open strategy occurred when the caller waited for the delayed help that ensued the open prompt. The delayed help had the same structure as the menu approach; most failures here were again due to selection of the wrong menu item.
- There were fewer routable responses when the initial prompt as well as the reprompt adopted a menu strategy.
- Domain had a significant effect on routability, and interaction between domains and prompt type had a significant effect on confusion levels.

3 Method

The main question to answer is if wording of a reprompt has an effect on the subsequent utterance. More specifically, does a reject prompt and a follow-up prompt differently affect the utterance with regard to the following expected effects from prompt wording differences.

utterance length

utterance content

user willingness to interact with the application

utterance similarity

user confusion

The setup of the experiment was to extract utterance features that corresponded to the features listed above, from calls to a data collection application. The extracted features was used to test the null hypothesis that differing reprompts did not have an effect on the succeeding utterance on any of the above mentioned utterance properties.

3.1 Domain

The source in this study are calls to Kronofogdemyndigheten (the Swedish Enforcement Authority's) customer service center. The callers are actual users of the customer service, including debtors, creditors, debt collection companies, various public authority officials, auditors or private persons. The tasks that callers intended to carry out in the customer service center consisted of a variety of the different tasks normally handled by a contact center. A task could e.g. be debtors that requested extracts of their current debts or employees seeking information regarding a bankruptcy, etc.

3.2 Data collection

The data collection was not designed by me for scientific purposes but was designed and set up by Voice Provider with the intention of collecting data for a real application. The data collection was set up *in-service* to resemble the real application to be developed and deployed. However the resemblance was only superficial. The data collection application posed as an authentic natural

language call routing application in order to collect authentic user behavior. The data collection application is described in detail in section 3.2.4.

The collected calls were subsequently transcribed manually by myself using a transcription scheme where all occurrences (not durations) of the following were inserted in the transcribed text:

- Filled pauses
- Fragments and repairs
- Extra linguistic sounds (such as laughter)
- Breath noises
- Speech directed to other locutors than the application

Numerals (e.g. etthundratjugofem ("onehundredtwentyfive")) were split into smaller constituent parts, while compound words were not split. Personal information about callers was censored from the transcriptions.

3.2.1 Pilot data collection

The main data collection was preceded by a pilot data collection. The pilot data collection provided necessary data to create the main data collection. The pilot data collection was very simple and did not try to recognize any speech but simply prompted the user with an initial question and a reprompt before transferring the call to a customer service agent. The pilot data collection resulted in 1502 calls with a total of 2300 user utterances, that were transcribed by myself using the transcription scheme previously described.

Common words, phrases and filler phrases from the pilot data collection was used to construct a simple CFG that was used by the main data collection. A stop word list was also constructed from the 250 most frequent words in the pilot data collection by manually removing words with specific semantic meaning until 186 stop words remained (see 7).

3.2.2 Semantic tag set

A semantic tagset was manually created with inspiration from the transcribed calls from the data collection. The tagset was designed to enable capturing of the content of utterances using a three-slotted semantic tag. The structure of the three-slotted semantic corresponds to tags described in Boye and Wirén (2007). Each tag consisted of three slots and represented different parts of the utterance content. The *Meta* slot represents phrases that concerned context e.g.: `greeting`, `introduction-as-authority` or `disconfirm`. The *Intent* slot represents the action of the utterance, typically a verb phrase, e.g.: `want-information`, `order` or `dispute`. The *Object* slot represent the object of the utterance typically a noun phrase, e.g.: `debt`, `tax-debt` or `e-services`. See figure 3.1 for examples of utterances with corresponding tags.

The semantic tagset was, as well as the whole data collection, created with a pragmatical approach. The consequences of this approach is discussed further in section 4.2.

"I want to know my current debt" –

`Meta:unknown_Intent:want-information_Object:debt`

"Hello I'm calling from the police" –

`Meta:introduction-as-authority_Intent:unknown_Object:unknown`

"No I want to can I I visited your homepage and" –

`Meta:disconfirm_Intent:unknown_Object:e-services`

Figure 3.1: Examples of utterances with corresponding semantic tag

The dialogue manager decides on appropriate action from speech recognition results complemented with semantic tags.

3.2.3 Previous touch-tone menu

The final call routing application replaced two existing services: a call routing touch-tone menu consisting of four menu alternatives and a manually operated switch.

3.2.4 Data collection application

Approximately 10% of the calls to the two existing touch tone menus were directed to the (main) data collection application where users interacted in one to four turns before being routed to a customer service agent. In total 14 989 calls were collected during a period of six weeks.

The speech recognizer for the main data collection application was trained on the transcribed calls from the pilot data collection. The dialogue manager made use of a rudimentary *Context Free Grammar* constructed by myself from analyzes of phrasings and word frequencies from the pilot data collection.

The CFG covered greetings and commonly used filler phrases followed by a set of keywords and/or keyword phrases. A filler phrase could e.g. be: "I'm calling about..." or "Would like to...". The ambition of the CFG was to cover the most common and most straightforward ways of formulating the users' reason for call, in order to direct those calls to an operator immediately. Users formulating less straightforward and less common reasons for call were kept longer in the data collection application and were prompted to state their reasons for call more briefly or to use other words.

The dialogue in the data collection application (illustrated in figure 3.2) consisted of two main dialogue states, main and follow-up. In the dialogue states users were prompted and their speech was recorded. In the main state users were prompted with an open style question of the type "How may I help you?". If the speech recognizer returned a speech result with a sufficient confidence level, the dialogue manager would use the rudimentary CFG to send users to one of four different follow-up states, or to directly connect the call to a customer service agent.

The difference of the four follow-up states was that they prompted users differently and used different error handling. From the main dialogue state, users with initial utterances lacking a speech result were reprompted with a

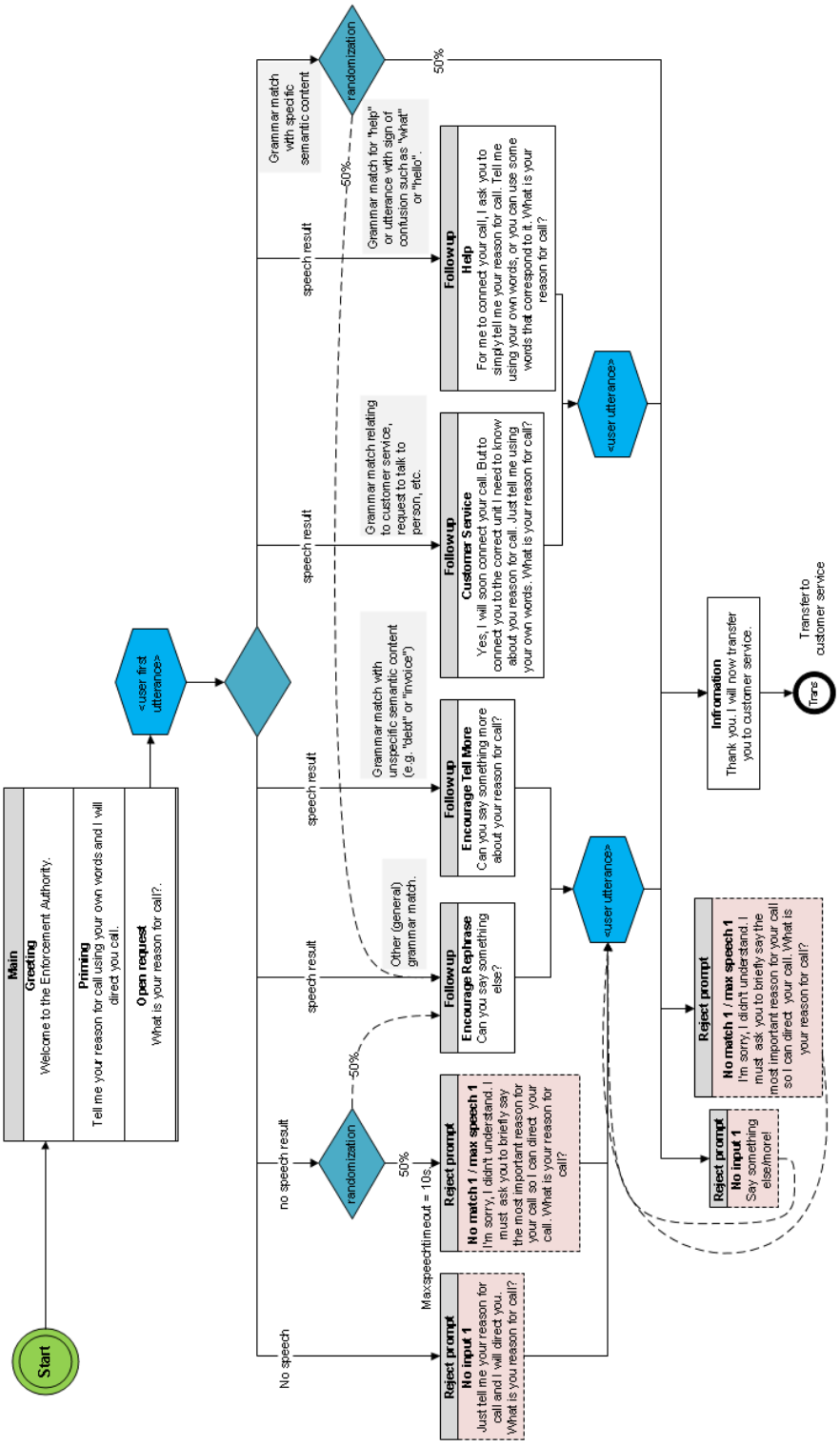

Figure 3.2: Illustration of dialogue design in data collection application

reject prompt. Users with initial utterances that contained no speech at all were prompted with an instructive prompt. In each dialogue turn the user had a max response time of ten seconds, but the response time was never exceeded.

There was a risk that the data collection application and the rudimentary CFG it used could skew the language model that the speech recognizer for the real application eventually will use. To remedy that risk, users with utterances that lacked a speech result were randomly reprompted with either a reject or a repair prompt. Initial utterances with a speech result that generated a grammar match with very specific content would normally be transferred directly to a customer service agent without being asked a follow-up question. Instead those calls were randomly reprompted with either a follow-up prompt or transferred to a customer service agent.

3.3 Experimental design

The randomization implemented in the data collection application was used in the setup of the experiment. Comparison of different prompts was only possible when users had been prompted with either prompt *x* or prompt *y* at random, and not when users were prompted as a function of their previous utterance. Utterances without a speech recognition result would normally be prompted with a reject prompt, but were now randomly prompted with either a reject prompt or a follow-up prompt.

The reject prompt:

Reject: *"Ursäkta, jag förstod inte. Jag måste be dig att helt kort säga det viktigaste om ditt ärende, så kan jag hjälpa dig vidare. Vad gäller ditt ärende?"* (I'm sorry, I didn't understand. I must ask you to briefly say the most important reason for your call so I can direct your call. What is your reason for call?)

The reject prompt signaled to users that the application had not understood the previous utterance, then instructed users of the expected form of the input and finally repeated the initial main question.

The follow-up prompt:

follow-up: *"Kan du säga något annat?"* (Can you say something else)

The follow-up prompt was a short request to provide additional information that did not signal to users that any misunderstanding had taken place and did not further instruct the user of the expected form of user input.

The calls that were used in the experiment traversed the dialogue tree in exactly two separate paths, displayed in figure 3.3. The two paths consisted of calls were the first utterance hade been rejected by the application and that was reprompted with either the reject or the follow-up prompt. Calls might have included several more subsequent turns but only the first and second turns were used.

In total, 14 989 number of calls were collected in the main data collection, from which 3 673 calls matched the requirements used in the experiment. 1 859 calls had been prompted with the reject prompt, in the following chapters

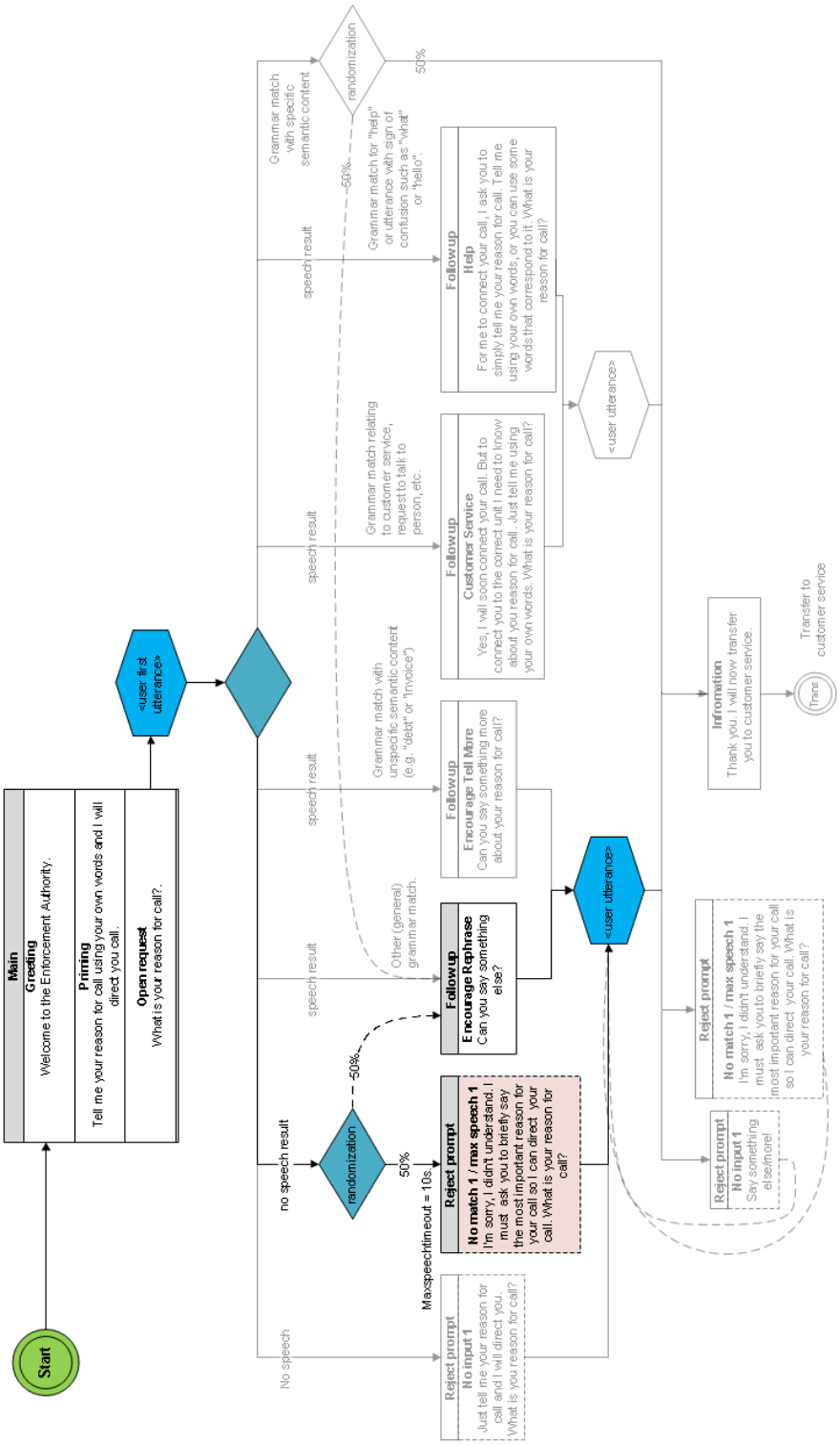

Figure 3.3: Illustration of the segment of the data collection application that was used for the thesis

refered to as: *reject*. 1 814 calls had been prompted with the follow-up prompt, in the following chapters referred to as: *follow-up*.

3.3.1 Extracting utterance parameters

The effects on user response to be investigated were selected primarily from my hypotheses of expected differences as a result of prompt wordings, namely that:

Requests to users to express their reason for call more briefly would impact *utterance length* and *content*.

Requests to users to say something else would impact *content* and *similarity*.

Signaling error would impact *willingness to interact*.

Instructing users of expected form would impact *confusion*

Repeating the initial question would impact *confusion* and *similarity*.

Formulating request as a yes/no question would impact *willingness to interact*.

The effects on user response to be investigated were selected secondly from the notion of distinctly separated strategies employed by users that I had formed when transcribing calls. The five interaction strategies observed when transcribing calls were the following: Users that. . .

verbose strategy: . . . talk/babble with little concern of not being understood.

menu strategy: . . . utter typical one-word menu items identical to touch-tone systems.

keyword strategy: . . . stacks keywords.

unresponsive strategy: . . . utter silent, defiant or repeated utterances.

responsive strategy: . . . respond to prompts with careful concern of not repeating previous errors.

These expected effects were measured with automatically extracted utterance features. Utterance features originated from three different sources:

application platform log files

manual transcriptions

semantic tags of transcribed speech.

3.3.2 Extraction from application platform log files

Application platform log files contain, for each call and each state, information about application settings (such as barge-in, a feature to allow users to interupt the system propmts), speech recognition outcome (such as rejected or no speech detected), speech recognition result (including a confidence score) and dialogue information (such as prompts played or utterance length). From the application log files the following features was extracted:

Recognition outcome: no speech / reject / recognition match

Dialogue information: hang-up

Note that only the outcome ("success") of the speech recognition was extracted and not the speech recognition result (the actual interpretation hypotheses).

When processing the recognition outcome and dialogue information, the following utterance features were derived:

`recognitionOutcome`: String value "reject", "no speech" or "recognized".

`isHangUp`: boolean value of the application platform's signal that a call had been terminated by the caller.

3.3.3 Extraction from manual transcriptions

Manual transcriptions contain a string representation of the utterance including filled pauses, word fragments, words, extra linguistic noise, noise and side speech. From the manual transcriptions the following features were extracted:

Manual transcription String representation including transcription notation, fragments and filled pauses.

Cleaned transcription String representation excluding transcription notation, fragments and filled pauses.

In the processing of the manual transcriptions information, manual transcription and cleaned transcription, the following utterance features were derived:

`word`: space separated character sequence.

`numberOfWords`: integer value of `words` in `cleanedTranscription`

`content word`: string `word` $\notin$ stop words list. (See appendix)

`numberOfContentWords`: integer value of content words $\in$ `cleanedTranscription`

`contentWordset`: a set of all content words contained in `cleanedTranscription`.

`sameContentWordset`: boolean value of `contentWordset` $\equiv$ the previous utterance `contentWordSet`

`conciseness`: integer 0–1

$$\frac{numberOfContentWords}{numberOfWords}$$

`sameUtterance`: boolean value of the utterance `cleanedTranscription` = the previous utterance `cleanedTranscription`

`utteranceSuperset`: boolean value if `contentWordset` $\subseteq$ previous utterance `contentWordset`.

`utteranceSubset`: boolean value if `contentWordset` $\supseteq$ previous utterance `contentWordset`

`similarContent`: boolean value of $|$ `contentW wordset` $\cap$ previous utterance `contentWordset` $| > 0$

`isNegated`: boolean value if `cleanedTranscrition` initial word is. “nej”, “nä” (“no” or “nay”)

3.3.4 Extraction from semantic tags of transcribed speech

The manual transcriptions were tagged with semantic tags using a robust parsing algorithm. The semantic tag of the speech result was available for the application’s dialogue manager but was not logged in the application platform’s log files, and therefore was not used as an utterance feature. The utterance features utilizing information from semantic tags instead used semantic tags assigned to the manual transcriptions, however using the same RPG. From the manual transcriptions’ semantic tags the following features was extracted:

`SemanticTag`: The semantic tag.

The semantic tagging information of the transcribed speech was processed together with the semantic tagset and the following utterance features were derived:

`tag`: string representation of only the intention and object slot value. The intention of the feature was to capture the content of the utterance. The meta slot was in that respect more similar to filler phrases which, for the semantic tagging, were disregarded.

`tagBranchId`: integer identification number of unique top branch in the tagset tree structure for the object slot. Intent tags were not ordered hierarchically and were disregarded.

`tagDepth`: integer value for this semantic tag's object slot's position in the tagset's hierarchical structure. `TagDepth` could take the value of 0–4, where leaves in the tagset equaled `tagDepth`=4. `TagDepth` was gradually reduced by one for each level ending at the root, tag "unknown" with `depthDalue`=0. Intent tags were not ordered hierarchically and were disregarded.

`tagSimilarity`: boolean value if `tagBranchId`= previous `tagBranchId` and `tag` =! unknown.

`supertag`: boolean value of current `tagDepth` $\subseteq$ previous `tagDepth` and `tagSimilarity`=true.

`subtag`: boolean value of current `tagDepth` $\supseteq$ previous `tagDepth` and `tagSimilarity`=true.

`same tag`: boolean value if `tag`=previous utterance `tag` and `tag`!=unknown.

3.3.5 Feature Extraction Tool

A program was written by myself in Java to extract information from the three main sources, mentioned in previous sections. The extracted information was complemented with internal resources consisting of the semantic tagset and a stop word list (see Appendix A). The Java program traversed the manual transcriptions file, and for each unique call parsed its corresponding log file. From the log file the call was recreated, supplemented with manual transcriptions and semantic tags. The recreated calls were processed to include utterance features derived from the extracted information and from internal resources, described in more detail below.

The extracted utterance features were intended to be measures of the utterance effects that were to be investigated:

- Requests to users to express their reason for call more briefly would impact *utterance length* measured by `numberOfWords`, and impact *content* measured by `numberOfContentWords` and `conciseness`.
- Requests to users to say something else would impact *content* and *similarity* measured by. `numberOfContentWords` and `sameUtterance`.
- Signaling error would impact *willingness to interact* measured by `isHangUp`.
- Instructing users of expected form would impact *confusion* measured by `recognitionOutcome=noSpeech`.
- Repeating the initial question would impact *confusion* and *similarity* measured by `recognitionOutcome=noSpeech` and `sameUtterance`.
- Formulating request as a yes/no question would impact *willingness to interact* measured by `isNegated`

A number of the derived utterance features were never used. Features that were extracted from the semantic tagset was intended to measure utterance content. However, it became increasingly clear that the semantic tags were not suitable as indicators of the specificality of the utterance because they had been designed rather for practical use than for being a correct hierarchical representation of the utterance meaning. Therefore the utterances' semantic taggings were completely disregarded.

There were two different types of similarity measures: `UtteranceSuper-/Subset` and `utteranceSimilarity` that were both

similarity measures of utterance content fidelity. `SameUtterance` was a similarity measure indicating the specific user behavior of repeating the exact same utterance that had just been rejected by the system. Tag similarity measures had already been ruled out. For clarity I decided to use only one similarity measure. I decided to use the measure that also indicated an unresponsive strategy, namely `sameUtterance`.

`NumberOfWords` was chosen as an indication of utterance length in favor of e.g. total number of tokens, syllables or length in milliseconds because it is an established measure. `Conciseness` was chosen as both a complement to utterance length as well as a measure for the `keyword` strategy.

4 Results

In the following sections the results of the experiment is presented. In the final section the hypothesis testing is summarized.

4.1 Utterance length

Table 4.1 displays `numberOfWords` mean in the second turn as a function of prompt wording. There is a decrease in utterance length, measured by average number of words, from the first to the second utterance. The decrease is present for both prompt alternatives but was greater for the *reject* prompt with utterance length decreasing from 6.20 to 3.53 (43%) and lesser for *follow-up*, which decreased from 6.10 to 4.05 (34%).

Table 4.1: `numberOfWords` mean as a function of prompt wording. p is calculated with a t-test of the null hypothesis that there is no difference between prompt alternatives in the second turn.

	Turn 1			Turn 2		
	reject	follow-up	all calls	reject	follow-up	p
`numberOfWords`	6.20	6.10	3.54	3.73	4.54	0.00000008

If the baseline for comparison is not the investigated prompt alternatives' first utterances but instead the first utterances of all calls, then the results is that the second utterance length for the investigated prompt alternatives is quite similar and actually slightly more verbose than the baseline for comparison.

In figure 4.1 utterances are broken down to display second turn utterance length as a function of first turn utterance length. The values inside the bars show that *reject* and *follow-up* calls with 1–2 words in the first utterance had in fact increased in length in the second utterance.

The prompt alternatives *reject* and *follow-up* display differences in that, in general, *follow-up* used more words and that *follow-up* had very modest decreases of words for first utterances with 1–6 words. After >6 words the decrease seems to level out for both prompt wordings. One should keep in mind that the amount of data is greater for calls with fewer words, therefore the small differences in the leftmost bars are statistically more certain than the greater variations in the rightmost bars, which are more uncertain.

Figure 4.2 illustrates the distribution of number of calls over utterance length in the first or second turn. Figure 4.2 illustrates a great increase of single-word calls from the first to the second turn. For *reject* the increase was from 286 to 671 calls (135%), *follow-up* had an increase from 307 to 614 calls

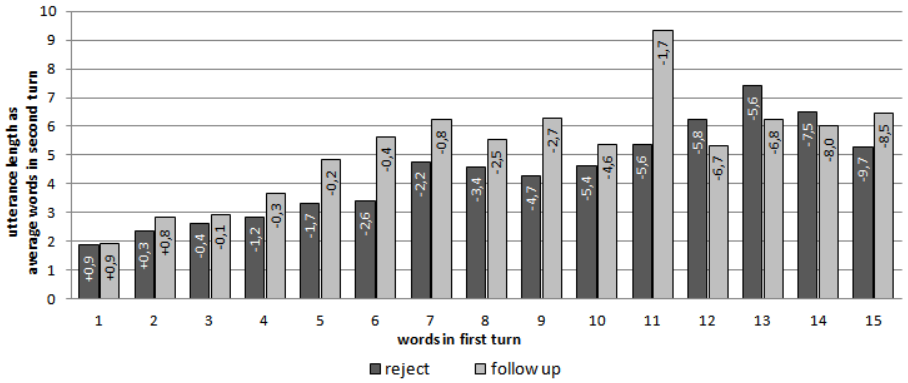

Figure 4.1: Second utterance length (average number of words) for calls with different number of words in the first turn. Values inside bars denote increase or decrease of `numberOfWords`.

(100%). Utterances with 2-4 words in the second utterance showed only modest changes from the first utterance. This leads one to believe that the great increase of single-word utterances was made at the expense of utterances with >4 words.

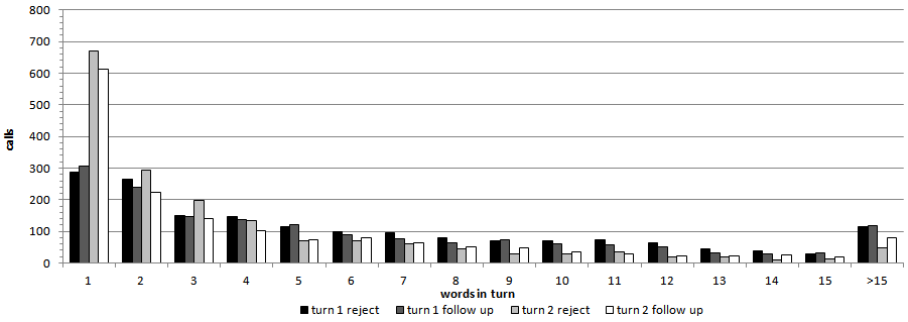

Figure 4.2: Total amount of calls for prompt alternatives (*reject* and *follow-up*) and turn number (fist or second turn) as a function of number of words.

Table 4.2 (p. 28) displays the decrease in `numberOfContentWords`. The equivalent decrease in `numberOfWords` as well as in `numberOfContentWords`, illustrated by the consistent `conciseness`, indicates that users have removed content and reduced utterance length without changing the conciseness of their utterance.

4.2 Utterance content

`numberOfContentWords` for the first and second turn displayed in table 4.2 show similarities with `numberOfWords` (displayed in figure 4.1). `numberOfContentWords` in the second turn decrease for both prompt alternatives but decrease more for *reject*.

`conciseness`, displayed in table 4.2 does not vary between turns or between prompt alternatives.

Table 4.2: `numberOfContentWords` mean and `conciseness` mean as a function of turn and prompt alternative. *p* is calculated with a *t*-test with the null hypothesis that there is no difference between prompt alternatives in the second turn.

	Turn 1		Turn 2		
Feature	reject	follow-up	reject	follow-up	*p*
`numberOfContentWords`	3.43	3.39	2.23	2.62	0.0000000005
`conciseness`	0.55	0.56	0.60	0.58	0.59

`numberOfContentWords` mean was, similar to `numberOfWords` mean, broken down to be displayed as a function of number of first utterance content words, displayed in figure 4.3. The result was parallel to the result on average number of words: *i*) there was an increase in `numberOfContentWords` for utterances with one content word; *ii*) *follow-up* was more similar to its first utterance; and *iii*) the `numberOfContentWords` mean stabilizes at around four content words;

Table 4.2 shows that prompt wording did not produce any difference in average `conciseness`.

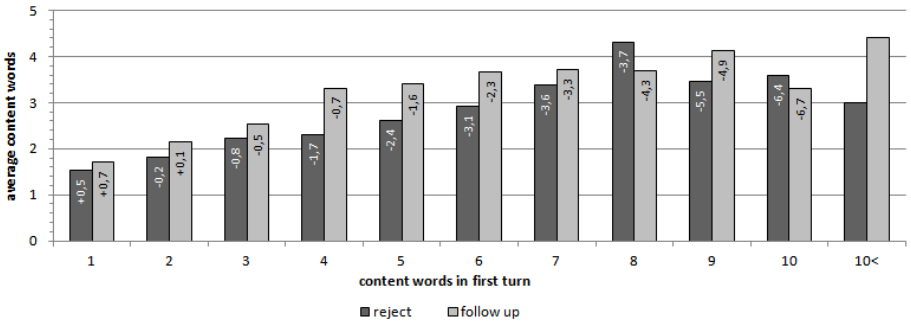

Figure 4.3: `numberOfContentWords` mean in the second turn for calls with different number of content words in the first turn. The spike at far right side of the table is caused by grouping together the tail into one category: >10 content words.

There are other results regarding differences in conciseness from the first and second turn that stand out; see figure 4.4. With a null hypothesis of expecting no change in conciseness from first to second utterance, then the conciseness in second turn would have the values depicted in figure 4.4 as "turn 1 approx. average". One can notice that there is a greater difference between first and second utterance `conciseness` for calls with "extreme" first turn conciseness. In other words, first turn utterances with low `conciseness` increase in conciseness in the second turn and first turn utterances with a high `conciseness` value decreases in the second turn. There are indications that, at least for this segment of calls, there is a stable `conciseness` value range of 0.45–0.75 that users move towards, regardless of the first utterance, and regardless of reprompt.

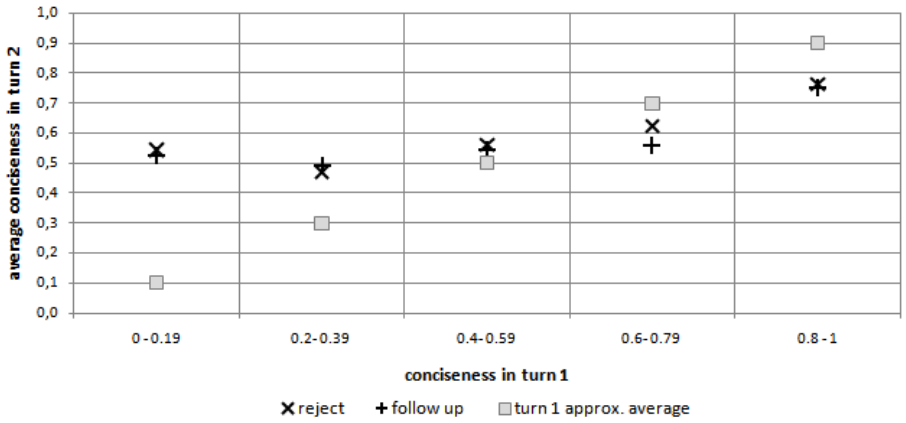

Figure 4.4: Average conciseness for varying conciseness in first turn.

4.3 Users' willingness to interact

Negated utterances were common for the *follow-up* prompt and practically non-existent for *reject*, illustrated in table 4.3. *follow-up* resulted in more hang-ups and significantly more silent responses.

Counting `isNegated`, `isHangUp` and `noSpeech` together it is clear that *follow-up* resulted in more uncoöperative responses. Users that answer a question which to its form may be a yes/no question but which is pragmatically obvious to be a request illustrates a demonstration by the users of an unwillingness to interact with the system.

Table 4.3: Utterance features in second turn. Significant differences in proportions were calculated with a two-tailed z-test of proportions.

feature	reject		follow-up		p<0.05
	f	%	f	%	yes/no
`isHangUp`	52	2.8%	66	3.6%	no
`noSpeech`	79	4.2%	188	10.4%	yes
`sameUtterance`	150	8.1%	54	3.0%	yes
`isNegated`	4	0.2%	238	13.1%	yes
Other	1574	84.7%	1268	69.9%	-
Sum	1859	100%	1814	100%	-

4.4 Utterance similarity

The *reject* prompt resulted in more similar utterances, displayed in table 4.3.

Figure 4.5 and 4.6 illustrate how users, for both prompt alternatives, changed strategy when they were not understood. The combined difference in changing or not changing strategy between the prompt alternatives is in total merely 1 percentage point. Looking instead at the previous utterance length, 90–95% of users with many words in their first turn utterances changed strategy to a single word or few word utterance. 59–79% of users with few words in their first turn utterance change strategy. 86–73% of users with single word first utterances changed strategy. In total 73% of users changed strategy.

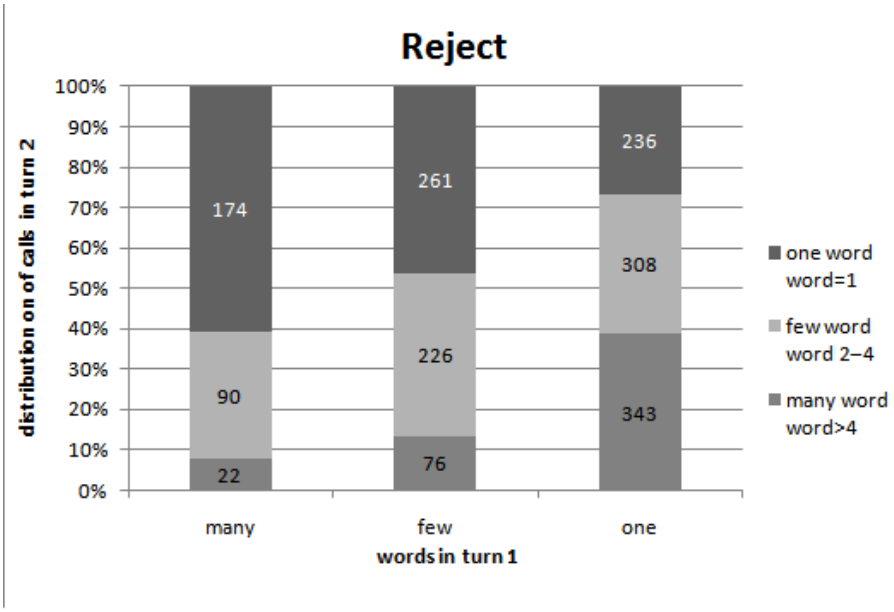

Figure 4.5: Distribution of single word (1), few (2–4) and many (>4) words in second turn as a function of the same in first turn for *reject*. The subdivision into these categories was made after examining results presented in figure 4.1.

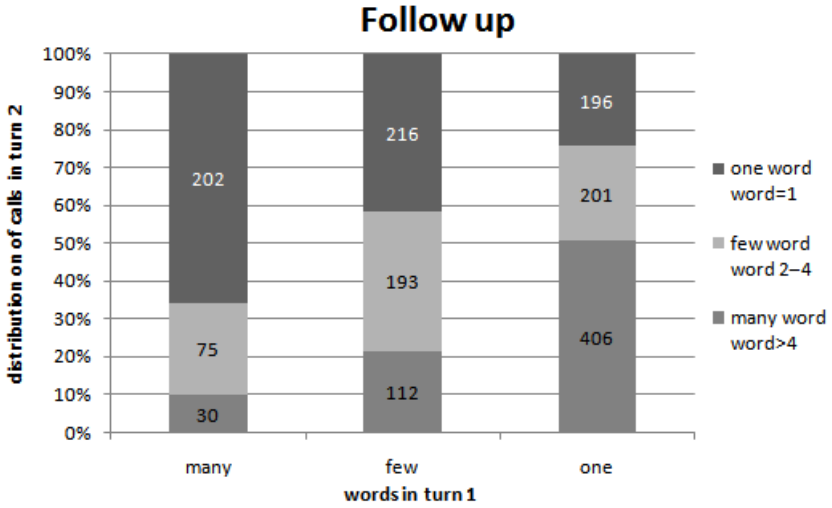

Figure 4.6: Distribution of single word (1), few (2–4) and many (>4) words in second turn as a function of the same in first turn *follow-up*. The subdivision into these categories was made after examining results presented in figure 4.1.

4.5 User confusion

follow-up resulted in more silent utterances (`noSpeech`), displayed in table 4.3. The difference in hang-up rate between the prompt alternatives was not statistically significant.

4.6 Hypothesis testing

The main question this thesis set out to answer was if wording of reprompts had *any* effect on succeeding utterances. The null hypothesis (H_0) was that there was no difference. H_1 was that wording of prompts has an effect on the succeeding utterance. Significance tests returned $p<0.05$ for 5 of 7 investigated effects. Thereby H_0 is rejected in favor of H_1.

The more elaborate hypotheses were that wording of reprompts had an effect on the selected features in the users' response. The result is displayed in table 4.4.

Table 4.4: Summary of significance tests on differences between second utterances with *follow-up* or *reject* on the selected features. Differences in `numberOfWords`, `numberOfContentWords` and `conciseness` were tested with a two-tailed *t*-test. Differences in proportions of `sameUtterance`, `isNegated`, `hangUp`, and `noSpeech` were tested for significance with a two-tailed *Z*-test.

utterance feature	$p<0.05$ yes/no
`numberOfWords` mean	yes
`numberOfContentWords` mean	yes
`conciseness`	no
`similar utterance`	yes
`isNegated`	yes
`hangUp`	no
`noSpeech`	yes

5 Discussion

As previously mentioned users were prompted with two different reprompts:

follow-up *"Kan du säga något annat?" (Can you say something else)* A short request to provide additional information, not signaling to users that any misunderstanding had taken place and with no further instruction of what form of user input was expected.

reject *"Ursäkta, jag förstod inte. Jag måste be dig att helt kort säga det viktigaste om ditt ärende, så kan jag hjälpa dig vidare. Vad gäller ditt ärende?" (I'm sorry, I didn't understand. I must ask you to briefly say the most important reason for your call so I can direct your call. What is your reason for call?)* The other prompt instead signaled to users that the application had not understood the previous utterance and then instructed users of expected form of the input, and repeated the initial main question.

This thesis started with the question if there was *any* effect on user response and continued with a more detailed picture of the effects on user response in terms of:

length

content

willingness to interact with the application

similarity

confusion

A straightforward approach to evaluate prompt efficiency would be to measure success of the speech recognition on the turn following the prompts in question. Such an approach would depend on high accuracy speech understanding, similar to that of a real application. Call data are from a data collection application with low speech recognition accuracy and a not fully developed robust parser grammar and consequently were not suitable for this efficiency evaluation.

One could argue for using "perfect speech recognition" available through manual transcriptions to perform efficiency evaluations. However, the result would not be fully applicable on a system using automatic speech recognition, which might never achieve accuracy similar to human speech recognition. The result is on the other hand interesting when not measuring success of the reprompt, but when trying to conclude what prompt alternative is preferable by looking at certain features in the responses themselves.

The data presented in chapter 4 support the hypothesis that wording of a reprompt has an effect on user responses. Impact of prompt wording is discussed in the next section and details of the effect it had is discussed in following subsections.

5.1 Impact of prompt wording

Table 4.1 illustrated a decrease in `numberOfWords` uttered in the second turn. The decrease was more significant for calls with the rejection prompt. However, the difference of the prompt wording alternatives is overshadowed by the general decrease of `numberOfWords`. The distribution of calls displayed in figure 4.2 shows a similar result, there is a difference between the two prompt wording alternatives, but it is overshadowed by the general differences in the first and second turn.

One should keep in mind that when comparing turns for calls with the specific prompts, the first turn utterance is limited to utterances that yielded those prompts, ignoring all utterances that did not yield a rejected "out-of-grammar" utterance. Second turn utterances contained all utterances without any selection.

A difference in utterances' `numberOfContentWords` was found for responses to the two prompt wordings, which was similar to result for utterances `numberOfWords` mean. But this difference was overshadowed by the greater difference between first and second turn utterances.

Williams and Witt (2004) concluded that "callers who provide more information initially but aren't understood remove content from their speech." Results in my study show that users who were not understood initially also removed content from their speech when the previous utterance had been long and added content when the previous utterance had been short.

There were also differences between prompt alternatives in the distribution of single-word utterances, few word utterances, and many word utterances in the second turn. Yet again the general trend outweighs the individual differences of the two prompts.

The most significant difference between the prompt alternatives was the difference in `isNegated`. Negated utterances were not present in the first turn and remained so in the second turn for *reject*, but not for *follow-up*, of which 13% of all responses were initiated by a negation.

Negated utterances where numbers increased from 0% to 13.1% for *follow up*, but did not increase for *reject* show that wording of prompts, can either strongly decrease user willingness to interact, or not affect it at all. One should keep in mind that the reason for users' unwillingness to interact might not be that they are unwilling to say something else about their reason for call, but that they are *unable* to do so.

User confusion measured as number of `noSpeech` utterances, also varied greatly between the prompt alternatives in the second turn. Users that did not produce any speech in a turn would not be prompted with any of the investigated prompt alternatives in the first turn and it was thus not possible to compare confusion levels in the second turn with confusion levels in the first.

Boyce (2008) describes similar experiments on call data from the "How May I Help You" system (Gorin et al., 1997). Boyce (2008) concluded that the two prompt variations "I'm sorry. Please briefly tell me how I may help you?" and "I'm sorry. How I may help you?", resulted in differences in user response length, similarity with previous utterance, amount of content/ideas and change of content.

I note a difference in similarity between results presented by Boyce (2008) and my results: 34–42% of users in Boyce (2008) repeat the exact same utterance, while only 3–8% of users in my study do the same. It is likely that the difference in similarity between the studies is the result of the different domains. The tasks performed by callers to AT&T's operator service are probably less complex than calls to an enforcement agency's customer service. Consistent with both my and the study by Boyce (2008) was the result that both prompt wordings that contained an instruction yielded shorter responses.

5.1.1 On utterance length

The prompt alternative *follow-up*, which prompted the user to say something else, did not produce an increased `numberOfWords` mean in second utterances. The prompt alternative *reject*, which prompted the user to be more brief, produced lower second utterance `numberOfWords` mean, compared to *follow-up*.

The mean length of first and second utterances in this study (6.1–6.2 initial and 3.5–4 second utterances) was short in comparison with the initial greetings reported in e.g. Boyce (2008) – 8.5–13 or Sheeder and Balogh (2003) – 8.3–9.6 . Tomko and Rosenfeld (2004) reported a mean word of 3–4.5 for all utterances in the dialogue. The experiment conducted by Tomko and Rosenfeld (2004) was designed to monitor user adaption to an automatic dialogue system. The similarity of second utterances `numberOfWords` mean in this study and the word mean reported by Tomko and Rosenfeld (2004) indicates that users in this study might have adapted to the dialogue system in a way similar to subjects in Tomko and Rosenfeld (2004) study.

The drastic increase (in the range of 100–135%) of single-word utterances in the second turn is probably in part due to the fact that most single word first utterances were recognized and therefore out of the scope of this thesis. Consequently the high percentage of single-word utterances may in fact more closely resemble the first turn average distribution.

In terms of differences between the prompt wordings my results point out a difference in that *reject* results in more single-word utterances and fewer longer utterances than *follow-up* does. However, the general pattern of users' behavior is consistent between prompts. The consistent behavior is a change in strategy from verbose utterances to single-word utterances and vice versa, and also in the decrease of words in the second utterance.

Sheeder and Balogh (2003) did not report any significant difference in utterance length (measured in seconds or in number of words) for subjects exposed to different initial prompts. Neither of the prompt alternatives did explicitly instruct the user to be brief. The prompts only differed in providing examples of what to say and how this was presented, either with a keyword or natural strategy. One could argue that the difference in results on length between this study and the study by Sheeder and Balogh (2003) point to the

fact that prompt length rather follows the wording of the instruction than the format of the instruction. However, additional studies are needed to verify this claim.

Williams and Witt (2004) pointed out that: "the proportion of two-slotted utterances is usually higher at the 1st utterance, indicating that callers who provide more information initially but aren't understood remove content from their speech."

My results and results in Williams and Witt (2004) show a similarity for utterance length, namely that in both studies users removed words when the initial utterance had not been understood.

5.1.2 On utterance content

My result for measuring content shows that the amount of content in the second turn was also reduced. Users did not only shorten their utterances but also removed actual content from the utterance. The level of conciseness did not increase indicating that users did not resort to a strategy of stacking content words.

In this thesis the *amount* of content and not the *actual* content is essentially what has been measured. It would have been interesting to have investigated exactly how the content in the second utterance changes compared to the first utterance. The original thought was to use the hierarchically structured semantic representation of the utterances to find out if users changed content to be more specific or more general, and if users completely changed topic or not. This approach was abandoned for several reasons.

The first reason was that the semantic tagging in the data application was performed with an incomplete parsing algorithm and an incomplete tagset. Utterances required retagging. The main reason for abandoning the use of semantic tags was that the specificity (level of depth) of the tags was not assigned equally among the branches in the tagset's tree structure. In other words: the tagset's hierarchical levels did not indicate specificity of the tags but was rather a result of grouping of tags, as discussed in section 3.3.4. It might also be important for utterances to have somewhat evenly distributed tags, otherwise common phrases might skew the results of a tag analysis. The second reason was that it was a problematic approach that the tagset was constructed using the same data that the tagset would be used upon.

Another approach would be to manually categorize if utterances have specific content or more general content. Still the effort required for manual classifications will not outweigh the benefits for any application other than purely scientific.

5.1.3 On users' willingness to interact

It is a common belief among dialogue designers that a great portions of users will opt out from the interaction if they are presented with a possibility to do so. McInnes et al. (1999) showed that when users were presented with the opportunity to say "help" they overwhelmingly did this. It seems that there is a percentage of users who, from the beginning, are unwilling to interact.

follow-up was phrased as a yes/no question, but anybody will identify the pragmatic meaning of the prompt as a request to elaborate their response, and not interpret it as a questioning of users' ability to say something else. The 13.1% of users who chose the latter interpretation are assumed to have done so to demonstrate their unwillingness to interact with the application.

By first glance one might think that for these 13.1% it was a wasted turn in the interaction dialogue. However, it might carry useful information to single out users whose unwillingness can result in an unsuccessful call, and might as well be handled by the application fallback. The unresponsiveness can on the other hand be interpreted as an indication that the users' reason for call is too complicated to be explained to the application, that users are not able to phrase it, or that they simply had already said everything relevant. These users might benefit from an instructive prompt, or transfer to a live agent.

Dialogue designers that do not intend to handle the portion of callers who will seize the opportunity to be "negativistic" should avoid using questions which can be intentionally erroneously interpreted.

5.1.4 On utterance similarity

Reject resulted in more identically phrased utterances (`similarUtterance`) than *follow up*. It could be explained by *reject's* shorter utterances which increase the probability of the utterance being identical to the previous utterance. But I argue that the increased probability can not constitute the entire difference between the prompt alternatives, where *reject* resulted in 158% (5.1 percentage units) more identical utterances compared to *follow up*.

The signaling of error and the repetition of the initial request in the *reject* prompt was a likely source for verbatim repetitions. The explicitly mentioned request by the *follow-up* prompt to users to say something *else* is likely to have contributed to the differing response to the prompts.

Similarity was also examined as `numberOfWords` categorized into three categories with one, few or many words. The results presented in section 4.4 show that 28% of users' output will be categorized in the same category as the previous utterance. Verbose utterances generated the least similar calls (5–10%), single-word utterances (14–27%) while the mid category (21–41%) had results on similarity that were less prominent.

It would be interesting to follow users longer in the call to see if they once more changed strategy when they were not understood. What relation is there between callers changing strategy and number of turns when they are not understood? Are there callers that will stick to their interaction strategy until either the application or the users gives up?

5.1.5 On user confusion

The more frequent silent utterances for *follow-up* indicate greater user confusion. *follow-up* resulted in more than twice as many `noSpeech` as *reject*. The big difference in silent utterances is likely to be a combined effect of a very short prompt and not repeating the request from the initial prompt.

5.2 User strategies

Do users interact with distinct interaction strategies? Different interaction strategies are visible in confirmation dialogues. Users consistently either explicitly confirm utterances or consistently remain silent. Edlund et al. (2006) discuss this phenomenon in the light of interaction metaphors and see it as users interact using either an interface metaphor or a human metaphor.

An interesting question for dialogue designers is if one should prime users and try to shape their output, or to accept how users interact and try to adapt. The example with silent confirmations mentioned above is usually dealt with by allowing implicit confirmations but making sure users can always return to past dialogue steps.

This study's result on reprompts was that, added together, users were silent, uttered the same utterance or uttered a negated utterance in 19.5% of calls for *reject* and 26.5% for *follow-up*. A typical view of these utterances is that they are errors and they will most likely lead to a reprompt. One could instead view any response to the prompt as a valid response and design specific actions for each response type.

Rejected utterances can be either silent utterances, utterances exceeding a set time limit, or utterances without interpretation. Rejected utterances without interpretations are difficult to handle because the speech recognizer simply does not know what was said. However, the speech recognizer can still output results on utterance durations which the dialogue manager could use to select a proper reprompt.

According to my results, users who utter a long initial utterance that is rejected by the application will shift to a short utterance. Therefore reprompts on long rejected utterances may not need to ask users to be more brief. Reprompts on shorter utterances run the risk of being to verbose and may benefit from reprompts that request users to be brief.

Tomko and Rosenfeld (2004) found that even in user-initiated dialogue systems users adapted to (or mirrored) the system. They reported that "rejections and confirmations had more of an effect on shaping user input than the introductory information".

I did not find indications of users talking with a "keyword strategy" of stacking content words. However, when examining the distribution of calls with different number of words (figure 4.1) there seemed to be a distinct portion of users interacting with the application using single-word utterances. Compared to the previous utterance there was a large increase in single-word utterances. As discussed in section 5.1.1, this could be an effect of the fact that not recognized utterances are more likely to be non single-word utterances. Therefore the number of single-word utterances in the initial turn that was prompted with either of the investigated prompt alternatives would be lower than for other utterances.

By visually examining the relative values inside the bars in figure 4.1 I categorized second utterances into three categories depending on whether the initial utterance was: single word, between two and four, and more than four words. At approximately three words there was a shift from increasing or providing same length to reducing the utterance length. Using this categorization length similarity was illustrated in figure 4.5 and 4.6. The close to inverted

results for single word utterances and long utterances show that utterance responses differ greatly depending on previous input length.

6 Conclusion

Based on results presented in this thesis I conclude that wording of a reprompt can enhance or reduce the effects that reprompting have on users' responses. Specifically it can affect the succeeding utterance length, amount of content, similarity and confusion. The more general user behavior in the dialogue is more consistent between the two prompt wordings. Users that were prompted with either of the prompt alternatives:

- shifted strategy from verbose to single-word utterances
- shifted strategy from single word to verbose utterances
- did not resort to stacking keywords

The prompt that explicitly requested users to be brief shortened the length of succeeding utterances, but it did not cause users to be more concise. The result of the prompt that was phrased in a way that it could be intentionally erroneously interperated was that 13.1% of succeeding utterances did that intentional erroneous interpretation. The prompt that requested from users to say "something else" caused less utterances that were verbatim repetitions of the previous utterance than the prompt the repeated the initial question. The prompt that signaled error was followed by fewer silent utterances than the other prompt was. The shorter prompt that also changed the question resulted in more confusion.

The results summarized above point to the fact that wording of reprompts affects utterances in a way that can be predicted from the wordings. The conclusion is that careful wording of reprompts can be used to avoid confusion or user unwillingness to interact. Wording of reprompts can also be used to shape the relative length of user utterances, but their effect on the more general length of utterances is small.

6.1 Applicability

Williams and Witt (2004) discuss the importance of domain and state that domain cause differences in absolute values on certain utterance features. However the conclusions in this thesis are drawn rather from the relative values that indicated user behaviour or prompt differences than absolute values. Therefore I argue that the results and subsequent conclusions are highly applicable on other domains.

I consider it probable that the effects of the different prompt wordings are consistent not only between domain but also between what tasks the users are performning and between different dialogue histories.

However I propose that dialogue designers do not consider results in this thesis applicable on other utterances than the non-understood utterances that was the subject of the study.

This thesis is a testament to the usability of commercial data for scientific purposes. Commercial data is an untapped resource of rich authentic material.

6.2 Further research

The significant differences between various studies and this study demand more comprehensive studies covering a wider range of prompt wordings and combinations of initial prompts and reprompts, in order to get a more complete picture. Data are being collected and annotated for commercial purposes in large data collections, but lack control of variables and lack randomized variation. However, the use of this kind of data might require ethical considerations.

Future studies of the effects of reprompts will benefit from looking into the actual content and meaning of utterances, and not just the form. Further investigations into how users change interaction strategies should look at several turns in the dialogue, including successful turns and turns with multiple errors.

With the goal of creating user adaptable applications future studies may need to concentrate on methods for recognizing distinct interaction behaviors. Research into efficent dialog management may then need to shift from trying to find the single most efficient prompt to finding the most efficent prompt for separate interaction behaviors. Research would then move away from the effort to shape user input with carefully worded prompts or to shape user input by using interaction metaphores in the dialogue design. Instead the focus would be on how applications can adapt to the users' style of interaction.

7 Appendix

A stopwords

och
i
att
som
det
en
på
av
är
den
för
med
till
de
inte
om
ett
han
var
jag
sig
men
kan
man
så
från
vi
eller
hon
hade
när
vid
under
också
där
skulle
sin
vara
vad
ut
nu
då
detta
här
än
finns
efter
upp
över
du
mycket
hur
mot
bara
får
in
genom
ska
ha
även
skall
något
utan
mig
honom
mellan
två
få
någon
kunde
hans
inom
denna
dem
sedan
många
måste
blir
allt
några
dessa
sina
fick
kom
fram
blev
säger
varit
sitt
se
sätt
vill
ju
bli
ta
kanske
enligt
själv
sa
henne
oss
annat
hos
ur
åt
tid
väl
redan
min
gör
vilket
dock
därför
varje

bör
hennes
annan
aldrig
ner
dag
ofta
fall
bland
just
alltid
fanns
samt
ser
alltså
deras
vet
ni
eftersom
vissa
gång
dessutom
tog
igen
ännu
ändå
vår
längre
ge
exempel
dess
tar
medan
först
innan
ville
särskilt
själva
ligger
nog
säga
mitt
blivit
vidare
vilka
mest
dig
liv
sett
sådan
trots
ibland
bort
stod
låg
kvar
nästan
sådana
framför
våra
inget
vilken
båda
varför
tror
fast
runt
visar
därmed
gjort
viss
sen

Bibliography

S. J. Boyce. User interface design for natural language systems From research to reality. In *Human Factors and Voice Interactive Systems*, pages 43–80. Springer US, 2 edition, 2008.

J. Boye and M. Wirén. Multi-slot semantics for natural-language call routing systems. In *NAACL-HLT '07: Proceedings of the Workshop on Bridging the Gap*, pages 68–75, Morristown NJ USA, 2007. Association for Computational Linguistics.

B. Carpenter and J. Chu-Carroll. Natural language call routing: A robust self-organizing approach. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP), Sydney*, 1998.

J. Edlund, M. Heldner, and J. Gustafson. Two faces of spoken dialogue systems. In *Interspeech 2006 - ICSLP: Satellite Workshop Dialogue on Dialogues: Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems*, 2006.

A. L. Gorin, G. Riccardi, and J. H. Wright. How may I help you? *Speech Communication*, 23(1-2):113–127, 1997.

K. Hafner. A voice with personality, just trying to help. World Wide Web electronic publication, 2004. URL `http://www.nytimes.com/2004/09/09/technology/circuits/09emil.html`.

F.R. McInnes, I.A. Nairn, d.J. Attwater, and M.A. Jack. Effects of prompt style on user responses to an automated banking service using word spotting. *BT Technical Journal*, 7(1):160–171, 1999.

T. Sheeder and J. Balogh. Say it like you mean it: Priming for structure in caller responses to a spoken dialog system. *International Journal of Speech Technology*, 6(2):103–111, 2003.

G. Skantze. *Error Handling in Spoken Dialogue Systems Managing Uncertainty Grounding and Miscommunication*. PhD thesis, KTH Department of Speech Music and Hearing, 2007.

B. Suhm, J. Bers, D. McCarthy, B. Freeman, D. Getty, K. Godfrey, and P. Peterson. A comparative study of speech in the call center: Natural language call routing vs. touch-tone menus. In *Proceedings of the SIGCHI conference on human factors in computing systems: Changing our world, changing ourselves.*, pages 283–290, 2002.

S. Tomko and R. Rosenfeld. Shaping spoken input in user-initiative systems. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP-Interspeech)*, pages 2825–2828, 2004.

J. D. Williams and S. M. Witt. A comparison of dialog strategies for call routing. *International Journal of Speech Technology*, 7(1):9–24, 2004.

M. Wirén, R. Eklund, F Engberg, and J Westermark. Experiences of an in-service wizard-of-oz data collection for the deployment of a call-routing application. In *NAACL: Proc. Bridging the gap: Academic and industrial research in dialog technology.*, NAACL-HLT '07, pages 56–63, 2007.