

Evaluating Information Presentation Strategies for Spoken Dialogue Systems

Andi Winterboer



Doctor of Philosophy

Institute for Communicating and Collaborative Systems

School of Informatics

University of Edinburgh

2009

Abstract

A common task for spoken dialogue systems (SDS) is to help users select a suitable option (e.g., flight, hotel, restaurant) from the set of options available. When the number of options is small, they can simply be presented sequentially. However, as the number of options increases, the system must have strategies for helping users browse the space of available options.

In this thesis, I compare two approaches to information presentation in SDS: (1) the summarize and refine (SR) approach (Polifroni et al., 2003; Polifroni, 2008) in which the summaries are generated by clustering the options based on attributes that lead to the smallest number of clusters, and (2) the user-model based summarize and refine (UMSR) approach (Demberg, 2005; Demberg and Moore, 2006) which employs a user model to cluster options based on attributes that are relevant to the user and uses coherence markers (e.g., connectives, discourse cues, adverbials) to highlight the trade-offs among the presented items.

Prior work has shown that users prefer approaches to information presentation that take the user's preferences into account (e.g., Komatani et al., 2003; Walker et al., 2004; Demberg and Moore, 2006). However, due to the complexity of building a working end-to-end SDS, these studies employed an "overhearer" evaluation methodology, in which participants read or listened to pre-prepared dialogues, thus limiting evaluation criteria to users' perceptions (e.g., informativeness, overview of options, and so on).

In order to examine whether users prefer presentations based on UMSR when they were actively interacting with a dialogue system, and to measure the effectiveness and efficiency of the two approaches, I compared them in a Wizard-of-Oz experiment. I found that in terms of both task success and dialogue efficiency the UMSR approach was superior to the SR approach. In addition, I found that users also preferred presentations based on UMSR in the interactive mode.

SDS are typically developed for situations in which the user's hands and eyes are busy. I hypothesized that the benefits of pointing out relationships among options (i.e.,

trade-offs) in information presentation messages outweighs the costs of processing more complex sentences. To test this hypothesis, I performed two dual task experiments comparing the two approaches to information presentation in terms of their effect on cognitive load. Again, participants performed better with presentations based on the UMSR algorithm in terms of both dialogue efficiency and task success, and I found no detrimental effect on performance of the primary task.

Finally, I hypothesized that one of the main reasons why UMSR is more efficient is because it uses coherence markers to highlight relations (e.g., trade-offs) between options and attributes. To test this hypothesis, I performed an eye-tracking experiment in which participants read presentations with and without these linguistic devices, and answered evaluation and comparison questions to measure differences in item recall. In addition, I used reading times to examine comprehension differences between the two information presentation strategies. I found that the linguistic devices used in UMSR indeed facilitated item recall, with no penalty in terms of comprehension cost.

Thus, in this thesis I showed that an approach to information presentation that employs a user model and uses linguistic devices such as coherence markers to highlight trade-offs among the presented items improves information browsing. User studies demonstrated that this finding also applies to situations where users are performing another demanding task simultaneously.

Acknowledgements

First and foremost, I would like to express my gratitude to Johanna Moore, who has been a great supervisor. She was inspiring, supportive and always helpful. Working with her was both productive and enjoyable and I benefitted from her experience, knowledge, and kindness.

I would also like to thank my second advisor Fernanda Ferreira for being available for shortly arranged meetings, fruitful discussions and for providing me with helpful advice and excellent feedback on various thesis drafts.

Thanks also to my thesis committee consisting of Sebastian Moeller and Frank Keller for reading the thesis. Their comments and suggestions helped to make this thesis a much more complete and coherent piece of research.

I also benefitted from superb feedback from Ellen Bard and especially Jon Oberlander, who accompanied this research project almost from the beginning to the end.

I would like to thank a number of persons for their individual contributions: Joseph Eddy for helping me with the initial implementation of the Wizard tool that I used in some of the experiments; Vera Demberg for her initial implementation of the user model based summarize and refine approach and numerous fruitful discussions about various approaches to information presentation; Maria Wolters for her always useful comments; Robert Logie, Jean Carletta, and Richard Shillcock for their time; Martin Tietze for productive collaborations; John Hu, and especially the always inspiring and enthusiastic Cliff Nass, for the productive collaboration between Stanford University and the University of Edinburgh.

The School of Informatics at the University of Edinburgh provided an extremely inspiring and productive environment for carrying out my thesis research. I am very grateful to some School of Informatics staff members: Many thanks to David Dougal, Avril Heron, Andrew Finnie, Tamise Totterdell, Sarah White, Melissa Davis, Jenny Burness, Alyson Macdonald, and Ross Armstrong.

I very much enjoyed the four years in Edinburgh, they were among the best years in my life. This was due to the city, which I love endlessly, and also a number of people that made my time in Edinburgh so highly enjoyable. I would like to thank all of you (in no particular order): Michael Kaiser, Martin Tietze, Marc Heise, Abhishek Arun, Ivan Meza Ruiz, Jens Apel, Marc Andre Martel, Sebastian Bitzer, Sabrina Hsueh, Sebastian Riedel, Vera Demberg, and Verena Rieser.

I acknowledge support through the Edinburgh-Stanford Link.

I am indebted to my mother who always believed in me and supported me in a truly self-less way.

Finally, a special acknowledgement: To Wiebke, for her support, her patience, and being there whenever I needed her in any way. Without her, this thesis would not exist.

I love you!

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(Andi Winterboer)

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Objectives	5
1.3	Outline	6
2	Background - Information Presentation in SDS	8
2.1	Introduction	8
2.2	Spoken Dialogue Systems and Application Areas	8
2.3	Information Presentation	10
2.4	Recommender Systems and Techniques	10
2.5	Information Presentation Methods in current Spoken Dialogue Systems	13
2.5.1	The Enumerated List Approach	14
2.5.2	Generation of user-model based Recommendations	16
2.5.3	Refinement through Clustering and Summarization	21
2.5.4	Information Presentation using Interactive Questions Answering techniques	26
2.6	Adaptive dialogue strategies	27
2.7	Combining User Tailoring and Clustering	29
2.7.1	UMSR - System Architecture	29
2.7.2	Clustering	31
2.7.3	Building the Option Tree	33

2.7.4	Pruning the Tree Structure	35
2.7.5	Option Presentation	40
2.7.6	Turn Length	40
2.7.7	Referring to Sets of Options	42
2.7.8	Presenting Additional Attributes to Explain Trade-Offs	43
2.7.9	Summarizing Properties of Options	43
2.8	Dimensions for evaluating information presentation strategies	45
2.9	Conclusion	49
2.10	Open Questions	50
2.11	Hypotheses	51
3	Evaluating Task Success and User Perceptions in a WoZ-Experiment	55
3.1	The Wizard-of-Oz Paradigm	55
3.2	Wizard tool for comparing presentation strategies	56
3.3	Experiment participants and setup	58
3.4	Experimental procedure	58
3.5	Results	60
3.5.1	Dialogue efficiency and task success	60
3.5.2	User satisfaction ratings	61
3.6	Discussion	63
4	Related Work - Measuring Cognitive Load	64
4.1	Basic concepts of Cognitive Load	65
4.2	Relation between task demand and task performance according to Meister's model	68
4.3	Considerations regarding the Assessment of Cognitive Load	69
4.4	Measuring Cognitive Load with the dual-task methodology	70
4.5	Studies examining the interplay between driving and speech interfaces	71
4.6	Methods for Measuring Distraction	75

4.7	Basic concepts of Human memory	78
4.8	Working Memory and Priming	79
5	Evaluating Task Success, User Perceptions, and Cognitive Load	83
5.1	Experimental setup	85
5.1.1	Participants	87
5.1.2	User Profile and Flight Booking	87
5.2	Experimental procedure	88
5.3	Results	89
5.3.1	Manipulation Check	90
5.3.2	Dialogue Efficiency and task success	90
5.3.3	Driving Safety	91
5.3.4	Perceptions	92
5.4	Discussion	93
5.5	Conclusion	95
5.6	Revised Dual-Task Wizard-of-OZ Experiment	95
5.6.1	Modifications to UMSR algorithm	96
5.7	Results	100
5.7.1	Dialogue efficiency and task success	100
5.7.2	Driving safety	101
5.7.3	Perception of system and self	101
5.8	Discussion	102
5.9	Comparing UMSR with revised UMSR	103
5.10	Conclusion	104
5.11	What we have learned from the conducted experiments	105
6	Psycholinguistic background: Sentence Comprehension and Recall	107
6.1	Psycholinguistic background on the effect of coherence markers on recall and comprehension	112

6.2	Summary of results relevant for this research project	115
7	Evaluating textual and auditory comprehension and recall	117
7.1	Evaluating the effects of coherence markers on recall	118
7.1.1	Experimental setup and procedure	118
7.1.2	Results and discussion	120
7.2	Evaluating the effects of linguistic markers on comprehension of textual materials	122
7.2.1	Results	123
7.2.2	Discussion and conclusion of textual comprehension experiment	124
7.3	Evaluating the effects of linguistic markers on recall of written materials	125
7.3.1	Participants	126
7.3.2	Experimental setup and procedure	126
7.3.3	Results of web-based recall experiment	127
7.3.4	Discussion	128
8	Conclusions and future work	130
8.1	UMSR vs. SR - User preference and dialogue efficiency	130
8.2	Dual-task studies - effect of cognitive load?	131
8.3	Message length influence?	132
8.4	Do coherence markers facilitate recall?	134
8.5	Discussion	135
8.6	Future work	136
	Bibliography	138
A	Questionnaire - Dual-task experiments	154
A.1	Edinburgh experiment questionnaire	154
A.2	Stanford experiment questionnaire	156

B	Experimental materials used in the eye-tracking experiment	171
B.0.1	Hotels	171
B.0.2	Restaurants	172
B.0.3	Flights	172
B.0.4	Mobiles phone plans	173
B.0.5	Digital cameras	173
B.0.6	Notebooks	174
B.0.7	Refrigerators	175
B.0.8	Rental cars	175
B.0.9	MP3 Players	176
B.0.10	Monitors	177
B.0.11	Cinemas	177
B.0.12	Microwaves	178
B.0.13	Bookbags	179
B.0.14	Overcoats	179
B.0.15	Makeup palettes	180

List of Figures

2.1	Typical information presentation of the AT&T Communicator in the flight booking domain	15
2.2	Output generated by the MATCH system in the restaurant recommendation domain	17
2.3	Mapping of attribute values to utilities in the restaurant domain	18
2.4	Generated recommendation of the FLIGHTS system tailored for a student who is most interested in the price of a flight	20
2.5	Generated recommendation of the FLIGHTS system tailored for a business traveller who is most interested in flying business class with a direct flight not requiring a lay-over	20
2.6	Dialogue between a simulated User and the Polifroni system in the restaurant information domain	23
2.7	Example summary based on the hand-crafted SR approach with associative clustering in the restaurant domain from (Polifroni and Walker, 2008).	24
2.8	Example summary based on the hand-crafted UMSR approach (SR with user model approach) with associative clustering in the restaurant domain from (Polifroni and Walker, 2008).	25
2.9	Example dialogue with the explicit presentation of restaurant candidates	26
2.10	Example dialogue adding a constraint to refine the search	27
2.11	System Architecture.	31

2.12	Clustering options with agglomerative group-average clustering and labeling.	32
2.13	Option tree for business user	34
2.14	The shaded subtrees are the ones that are pruned. Subtrees 13-16 still need to be decided on at the stage shown here.	37
2.15	Diagram showing how the pruned option tree is mapped onto language. The tree on the right hand side corresponds to the example trees in Figures 2.13 and 2.14.	41
2.16	The option tree is cut into subtrees which determine turn length.	41
2.17	Example dialogue between the Demberg system and a frequent flyer	44
3.1	Web-based wizard user interface	57
4.1	Hypothetical relationship between task demand and performance based on Meister, 1976	68
5.1	STISIM Drive TM simulation system used in the dual-task experiments at Stanford University's CHIME lab	86
5.2	First-round presentation with both SR- and UMSR-based systems	88
5.3	Presentation based on original UMSR algorithm	97
5.4	Presentation based on SR algorithm	97
5.5	Presentation based on revised UMSR algorithm	98
5.6	Presentation of two flight alternatives with revised UMSR	99
7.1	<i>Example for experiment material without coherence markers.</i>	119
7.2	<i>Example for experiment material with coherence markers, where coherence markers are indicated in bold.</i>	119
7.3	<i>The three types of evaluation questions with examples:</i>	120
7.4	<i>Average recall on a scale from 0 to 1 for the three recall questions, $p < .05$ indicated with "*",</i>	121

List of Tables

1.1	<i>Communicator data: General contributions of system and user</i>	2
1.2	<i>System contributions: Requesting and presenting information</i>	2
3.1	<i>Average number of turns per booking, dialogue duration (for booking two flights, system time plus user time) with SR and UMSR and how often the "best" flight was chosen. Significance levels: $p < .05$, indicated with "*", $p < .01$, indicated with "**", $p < .001$, indicated with "***".</i>	61
3.2	<i>Answers to the 4 user satisfaction/evaluation questions (on a scale from 1-7), $p < .05$ indicated with "*").</i>	62
5.1	<i>Performance of easy vs. difficult-driving, $p < .001$, indicated with "***"</i>	90
5.2	<i>Average number of turns per booking, dialogue duration (for booking two flights, system time plus user time) with SR and UMSR and how often the "best" flight was chosen. Significance levels: $p < .01$ indicated with "**", $p < .001$ indicated with "***".</i>	91
5.3	<i>Average number of turns per booking, dialogue duration (for booking two flights, system time plus user time) with SR and UMSR and how often the "best" flight was chosen. Significance levels: $p < .05$ indicated with "*", $p < .01$ indicated with "**".</i>	100
5.4	<i>Comparison of dialogue efficiency results obtained with UMSR vs. the revised UMSR II, $p < .05$ indicated with "*".</i>	103

7.1 *Eye-tracking data per IA (first pass reading times, remaining time reading times, number of passes, regressions out and in) for messages with and without coherence markers* 123

7.2 *Average recall on a scale from 0 to 1 for the three questions - lab experiment (lab) vs. web based experiment results (web), t-Test, “*” indicates a significant difference with $p < .05$, significance between underlined values and values in italic* 128

Chapter 1

Introduction

1.1 Motivation

Spoken dialogue systems (SDS) are computer-based systems developed to provide natural and efficient access to information and carry out simple tasks using speech as the main interaction mode. Example applications include travel information and reservation, weather forecast information, product recommendation and comparison.

In this thesis, we examine approaches to content selection and information presentation in spoken dialogue systems to facilitate information retrieval. More specifically, we experimentally study the effect of information presentation strategies on user perception, task success, dialogue efficiency, recall, and cognitive load.

Although much research has been conducted on the information gathering phase of spoken dialogue systems, relatively little attention has been devoted to information presentation. However, the DARPA Communicator evaluation showed that task duration is negatively correlated with user satisfaction ($r = -0.31, p < .001$, Walker et al., 2001). Moreover, an analysis of the Communicator corpus consisting of approximately 2000 dialogues with nine different spoken dialogue systems (see Table 1.1), found that 69% of the dialogue when measured in time, and 91% when measured in words, is due to the system producing utterances (Moore, 2006).

Table 1.1: *Communicator data: General contributions of system and user*

	System	User
Utterances	56%	44%
Time	69%	31%
Words	91%	9%

A closer look reveals that the majority of the system time (54%) is spent on the information presentation phase of the dialogue (see Table 1.2). Therefore, the information presentation phase is the main contributor to dialogue duration. Hence, we believe there is potentially a large pay-off for improving information presentation in spoken dialogue systems.

Table 1.2: *System contributions: Requesting and presenting information*

System	Requesting info	Presenting info	Other
Utterances	43%	25%	32%
Time	31%	54%	15%
Words	28%	50%	22%

Although different approaches to information presentation have been proposed, the evaluations have mainly focused on users' perceptions of the quality of the information presented: e.g., informativeness, overview of options, and so on (see Walker et al., 2004; Demberg and Moore, 2006). To our knowledge, no studies have been performed comparing the effects of different information presentation strategies on task success, and, consequently, we performed experiments to evaluate this effect.

Furthermore, spoken dialogue systems are often intended for situations where the user's hands and eyes are busy performing another task. Applications such as spoken dialogue systems for disabled users who have physical difficulties operating conventional input devices, or voice services to be used in cars are examples of situations

where traditional graphical user interfaces are not practical. In this thesis, I will consider an additional dimension which has demonstrated its relevance in the areas of instructional design (Seufert and Brunken, 2006), usability research (Schultz et al., 2007) and generally in the area of human-computer interaction: cognitive load. While developers of human-computer interfaces generally strive to design interfaces which are as easy to use and comprehend as possible in order to ease processing and increase usability for the user, thus avoiding cognitive load, the role of cognitive load increases in relevance if another task, e.g., walking, driving, or the manual manipulation of the surroundings, which also requires the users' attention and cognitive processing, is performed simultaneously.

To our knowledge, although there have been many claims about the cognitive load that different information presentation strategies place on users (Walker et al., 2004; Moore et al., 2004; Kruijff-Korbayova et al., 2006), there has been no systematic empirical study of these claims, especially in terms of assessing the cognitive load that different presentation strategies place on users.

For example, in developing algorithms for presenting recommendations tailored to the user's interests, Moore et al. (2004) were attempting to address the overload created by information presentation strategies that simply enumerate options, without effectively supporting users in making decisions about complex options. Likewise, Kruijff-Korbayova et al. (2006) present a general framework for scheduling different presentation modes and modalities to take user's cognitive load into account when deciding which information to present when. Nevertheless, there has been no adequate evaluation comparing the information presentation strategies for their effect on cognitive load.

Thus, we set out to examine the effects of two recent approaches to information presentation on cognitive load. In this thesis, we develop strategies to present complex information to users in situations where interacting with the dialogue system may not be their primary task. We carry out experiments to gather empirical data which will

help us to better understand how people process information in contexts where their cognitive resources are split between tasks.

We compare two previously introduced approaches to information presentation: The summarize and refine approach (SR) to information presentation, developed by Polifroni et al. (2003) and later extended by Chung (2004) and Polifroni (2008), will be compared to the user-model based summarize and refine approach (UMSR), recently presented by Demberg (2005); Demberg and Moore (2006). We hypothesize that the UMSR approach, which explicitly points out trade-offs and highlights relations between different options, will place less cognitive burden on the user in comparison to SR, because when conversing with a dialogue system adopting the SR strategy to information presentation users must compute and compare the presented items and their attributes mentally. On the one side, UMSR creates longer, potentially more complex sentences which increase demands on language processing in comparison with SR, which selects attributes that partition the data into the minimal number of clusters, so that a concise summary can be presented to the user to refine. On the other hand, UMSR-based presentations employ coherence markers (e.g., discourse cues, connectives, and adverbials such as *but*, *however*, *moreover*, *only*, *just*, *also*) in order to highlight specific properties of and relations between items presented to the user, thus potentially facilitating processing and recall of the presented information.

We compare these particular information presentation strategies because a) they are recently introduced state-of-the-art approaches, b) they use interesting techniques to facilitate information browsing (SR) plus user-modeling (UMSR), and c) they can be implemented relatively easily. These approaches share some properties but can be distinguished by their different views on the data they use for their computations: The SR approach exclusively uses the attribute value pairs of the database with no knowledge regarding whether they are of interest to the user, whereas UMSR always takes into account the user (model) and generates recommendations with the belief that the presented items are most suitable for the specific user given the user model.

These approaches to information presentation allow us to assess the contribution of the user model for constructing summaries and, in addition, they allow us assess the contribution of coherence markers to comprehension and recall of options.

The expected insights of our research may also be applicable to situations where no other task is likely to interfere with the dialogue task, as our overall goal is to study the dimension of cognitive load in dialogue systems. In general, we aim to answer the question of how to present information in a spoken dialogue system in a way that effectively facilitates its comprehension even under the condition of a concurrent secondary task.

1.2 Objectives

In this thesis, I perform experimental studies which address the current lacuna. I investigate whether users 1) prefer and 2) perform better with the recently developed user-model based summarize and refine approach (UMSR) to information presentation (see Demberg, 2005; Demberg and Moore, 2006) than they do with a system employing the summarize and refine approach (SR, Polifroni et al., 2003; Polifroni, 2008). To evaluate these approaches, we compare them in user studies designed to assess their effects on:

- task efficiency,
- task effectiveness,
- user perception,
- cognitive load, and
- user recall of information.

Task efficiency will be measured by traditional dialogue system measures, such as dialogue duration and the number of dialogue turns to achieve a task.

Task effectiveness will be measured by looking at how well the deployed information presentation strategies support users in choosing the "best" option (typically the option that best matches their user profile) from the set of available options. Furthermore, questionnaire data will reveal how the participants perceived the information presentation strategies.

To examine the cognitive load of different presentation strategies I make use of two fundamentally different assessment methods: a) the dual-task paradigm, a procedure that requires an individual to perform two tasks simultaneously in order to compare performance with single-task conditions, and b) reading task studies conducted with an eye-tracker to assess the complexity of processing the examined materials. Reading times are considered to shed light on on-line discourse processing/comprehension (see Haviland and Clark, 1974, for example).

Finally, I examine whether the use of coherence markers in presentation messages facilitates user recall and comprehension of information using eye-tracking measures and comprehension questions in a reading task experiment. We hypothesize that coherence markers highlighting differences between options and making trade-offs explicit ease processing and recall of the presented information.

1.3 Outline

The remainder of the thesis is structured as follows: The second chapter gives an overview of the architecture of (typical) SDSs and introduces current approaches to information presentation. In the third chapter, the research questions and hypotheses are highlighted. Then, in the fourth chapter a Wizard-of-Oz experiment is described comparing two approaches to information presentation. In the fifth chapter, considerations concerning the assessment of cognitive load and relevant methods based on the literature are introduced. The sixth chapter describes two experiments comparing information presentation strategies in dual-task studies. The seventh chapter then reviews

(psycholinguistic) literature on sentence comprehension. The eighth chapter describes three user studies, an eye-tracking experiment, an additional web-based reading task study and an auditory recall experiment. Then, in the ninth chapter I summarize the contributions and findings of the thesis and present suggestions for future work.

Chapter 2

Background - Information

Presentation in SDS

2.1 Introduction

This chapter provides a brief overview of current research on information presentation in spoken dialogue systems. I start with a general overview of spoken dialogue systems, a brief summary of the system architecture of a typical SDS, and provide an example of the basic processing stages of a single conversational cycle. Next, I review typical application areas for SDS with some general considerations regarding the most common areas. I conclude with an account of recommender systems, specifically designed for allowing users to make well informed choices.

2.2 Spoken Dialogue Systems and Application Areas

A spoken dialogue system provides a natural language interface for conversations between users and a computer and typically consists of a speech recognizer, a parser (or keyword spotter), a natural language understanding module, a dialogue manager,

a natural language generation component, and a speech synthesizer. A more detailed overview of these modules can be found in Jurafsky and Martin (2008).

Usually, the basic stages of processing in a single conversational cycle are (see Toney, 2007): (i) the user's utterance in the form of a speech signal is sampled and processed by the speech recognizer; (ii) the recognizer passes a list of potential sentences (hypotheses), with associated confidence levels, to the language understanding component (NLU); (iii) the NLU component examines these hypotheses and decides on a meaning that can be usefully interpreted by the dialogue manager; (iv) the dialogue manager analyzes the parser's output in the context of the dialogue as a whole and then decides on the most appropriate response, possibly retrieving information from the database in the process; (v) this response is converted into a complete sentence by the language generator; (vi) finally, the speech synthesizer translates the text from the language generator into spoken language.

An SDS typically provides access to a computer-based application such as a database or an expert system. Spoken dialogue systems have been deployed, for example, as speech interfaces

for information retrieval and/or browsing allowing users to retrieve tourist and weather information from underlying databases and to make travel, restaurant, cinema or theatre bookings (e.g., Levin et al., 2000; Walker et al., 2004; Moore et al., 2004; Demberg and Moore, 2006).

for tutorial and expert systems allowing users to converse with an expert system in order to learn new skills or improve old ones (e.g., Ai et al., 2006; Callaway et al., 2007).

for intelligent assistance systems which allow users to engage in other activities (e.g., driving) while simultaneously conversing with the SDS (e.g., Becker et al., 2006; Vargas et al., 2006).

Conversations between user and system consist of a number of system and user turns. In a simple information retrieval task, an interaction may be completed in only a few turns whereas in other cases (or domains), such as tutorial dialogues, conversations may last for hours. Although spoken dialogue systems may cover a wide range of applications, here we focus on the information presentation phase of SDS for information retrieval and/or browsing.

2.3 Information Presentation

Information presentation plays an important role in all of these application areas, but this role is more decisive in some application areas than in others. In particular, in some domains the results of user queries cannot be presented sequentially in spoken form because there are too many options matching the query. Thus, strategies are required for presenting users with information that is useful for them. In this thesis we compare two information presentation strategies for spoken dialogue systems helping users to select a suitable option from a large set of options.

Recommender systems are a research area that traditionally deals with filtering and presenting information items that are likely of interest to the user. In the following section I give an overview of recommender systems and their underlying data filtering techniques.

2.4 Recommender Systems and Techniques

As information becomes abundant, and its access more and more important, we face the problem of choosing among all the available alternatives. The term *recommendation systems* describes computational aides that guide users through interesting and useful objects in a large space of possible options (Burke, 2002). Typically, research on recommender systems focuses on recommendation techniques and algorithms to find

the most useful set of items for the user from all options available and is less concerned with information presentation. The following section introduces existing recommendation techniques which are also applied in the spoken dialogue systems reviewed in this thesis (see Section 2.5).

In general, recommender systems consist of background data, which is the information that the system possesses before the recommendation process begins, input data, which is the information that users must communicate to the system in order to generate recommendations, and an algorithm combining background and input data to arrive at its suggestion. Burke (2002) proposed that there are at least five distinctive recommendation techniques.

He distinguishes collaborative, content-based, demographic, utility-based, and, finally, knowledge-based recommendation techniques. In addition, in hybrid recommendation systems, two or more recommendation techniques are combined. The most prominent recommendation techniques are the collaborative and the content-based recommenders.

Collaborative recommenders aggregate ratings or recommendations of objects, identify commonalities between users on the basis of their ratings, and generate new recommendations based on inter-user comparisons. Recommenders using collaborative filtering (CF) generate personalized recommendations, e.g., predictions of how a user may like an item, based on the assumption that users who agreed in the past, i.e., users whose opinion correlated in the past, will also agree in the future. The input for CF algorithms are *rating matrices* containing user profiles represented by *rating vectors*, i.e., lists of user ratings on a set of items. Therefore, collaborative recommenders necessarily require available user profiles that capture the past rating histories of users to generate, first, a neighborhood of K users having the highest degree of similarity with the active user and, second, a prediction for a specific item by computing a weighted average of the ratings of the other users in the neighborhood on this item (Berkovsky et al., 2007). This type of recommender can often be found in commercial

product recommender applications, for example on www.amazon.com. There, users are informed both graphically and textually that

“Customers who bought items in your shopping cart also bought:”

Content-based recommenders recommend an item to a user based upon a description of the item and a profile of the user’s interests (Pazzani and Billsus, 2007). At the beginning, a movie recommendation system, for instance, requires a database containing all the available movies and their attributes including genre, directors, actors etc. To select a set of promising movie recommendations, the system matches those data against the learned preferences of the user. Aside from collaborative recommender systems, content-based systems are probably the most common. Popular examples for content-based recommenders include the music recommenders www.pandora.com and www.mystrands.com that using categories suggest new music to people based on music the user liked before.

Demographic recommender systems aim to categorize the user on personal attributes and make recommendations based on demographic classes. Demographic techniques form “people-to-people” correlations like collaborative ones, but use different data. The benefit of a demographic approach is that it may not require a history of user ratings of the type needed by collaborative and content-based techniques, but it clusters users based on demographic data and tailors recommendations based on information about other users in that cluster.

Utility-based and **knowledge-based recommenders** do not attempt to build long-term generalizations about their users, but rather base their advice on an evaluation of the match between a user’s need and the set of options available. Utility-based recommenders make suggestions based on a computation of the utility of each object for the user. Here, the central problem is how to create a utility function for each user. The user profile therefore is the utility function that the system has derived for the user, and the system employs constraint satisfaction techniques to locate the best match. The

benefit of utility-based recommendation is that it can factor non-product attributes, such as vendor reliability and product availability, into the utility computation, making it possible for example to trade off price against delivery schedule for a user who has an immediate need.

In addition to the discussed recommenders, there is the special case of *conversational recommenders* (e.g., Thompson et al., 2004; McCarthy et al., 2005). In these systems, a structured human-computer dialogue guides users through the set of available options. Oftentimes, conversational recommendation systems make use of knowledge-based, content-based, or collaborative filtering to find and suggest items that satisfy user queries. Knowledge-based (sometimes in combination with content-based) recommendation techniques are used in some dialogue systems featuring user-model based algorithms discussed in this thesis. These systems support users in finding the most desired item(s) as determined from a model of their preferences (Viappiani et al., 2007).

2.5 Information Presentation Methods in current Spoken Dialogue Systems

Next, I introduce recent work on information presentation in spoken dialogue systems. These approaches all deal with the problem of presenting options to users so that they may chose among them.

The different approaches vary in the form in which they present the information, and in the way they select from the available options. Some of the described techniques are more appropriate for presenting a large number of options because they enable users to easily narrow down the potentially huge number of initially available options to a manageable number. In contrast, other strategies seem more advisable when the number of options available is already reduced through a preselection process (or when there are generally fewer options) as they are more appropriate for presenting a smaller

number of options. In those cases, they allow the system to present the trade-offs between different options explicitly to ease the decision making process.

2.5.1 The Enumerated List Approach

In traditional dialogue systems, information is typically presented by a simple enumeration of a set of different options and their relevant attributes. Thus, the system presents a (long) list of results sequentially one item after the other. At each step, the user can ask to hear another option, or attempt to refine the search by providing new constraints. In cases where various options are available and lots of information is to be presented, this might easily lead to a form of presentation which is complex and difficult to comprehend. In addition, some relevant options for the specific user might not be included at all because the list is confined to a certain number of results.

The enumerated list approach was used in the AT&T DARPA (Defense Advanced Research Projects Agency) Communicator Dialogue System (Levin et al., 2000). The main application for the DARPA Communicator project was a travel planning task, which included booking multi-leg flights, and making hotel and car arrangements.

In this system, if the retrieval of the user's query resulted in multiple flights, they were sorted, by default, by price. The system provided the user with a brief summary of the number of flights found together with information about the first flight in the list. The users had the option of selecting the presented flight or browsing through the list of flights, using commands such as "next option", "the fifth option", etc. Furthermore, users were able to filter the list of flights by providing additional constraints such as a specific airline, a different departure time and so on. The enumerated list approach is particularly advantageous if there is only a small set of options to present. In that case, this approach to information presentation presents all remaining options in a structured and obvious way. Thus, users will always feel they have heard about all options, increasing their confidence in being able to make a well informed choice.

System: I found 9 round trips from Hartford to Orlando on October 31st and returning Hartford on November 6th. For option 1, I have an outbound flight with US Air departing at 7 am and arriving at 9:49 am non-stop. A return flight is with US Air departing at 10:30 am and arriving at 12:28 am with 1 stopover. Ticket price is 165 pounds. Please say “next option”, or, “flight details”, or, “I’ll take it”.

User: Next option.

System: For option 2, I have an outbound flight with US Air departing at 3:30 pm and arriving at 6:19 pm non-stop. A return flight is with US Air departing at 8:25 pm and arriving at 12:28 am with 1 stopover. Ticket price is 168 pounds. Please say “next option”, or, “flight details”, or, “I’ll take it”.

User: Next option

System: ...

Figure 2.1: Typical information presentation of the AT&T Communicator in the flight booking domain

The example in Figure 2.1, however, demonstrates the possible difficulties users face with this approach in situations when there are many different alternatives to choose from. Specifically, there are four main problems:

1. The list of options has too many options and too many features to keep in mind for comparison. Because this strategy always presents the same attributes for each option, it is difficult to keep track of which attribute values apply to which option. And even if users comprehend each of the options as they are presented, remembering all features of all the presented options in order to compare them places a large memory burden on the user.
2. Any long list risks creating a unsatisfyingly long dialogue.

3. The system offers no support for selecting from among the options. If there are many available options to be presented, and there is no preselection on the part of the presentation module, users are likely to be overwhelmed simply by the number of available options.
4. There are situations where users might be busy with another demanding task at the same time, making it even more difficult to select the most suitable option, especially, if two or more options are almost identically suitable.

Moreover, as Walker et al. (2001)'s evaluation of nine spoken dialogue systems in the DARPA Communicator domain showed, the information presentation phase of dialogues is one of the main contributors to dialogue duration and task duration is negatively correlated with user satisfaction. One feasible way to shorten dialogue duration is for the system to select a small subset of relevant options and present only these in detail. Another way to handle information presentation is to support users in refining their queries. The enumerated list approach does neither. This may lead users to make suboptimal decisions. For example, if the flight best matching their interests is not within the, say, top ten, presented options it may be very unlikely to be chosen.

In summary, presenting information by a simple enumeration of the available options is problematic for spoken dialogue systems when there are a large number of different alternatives to present.

2.5.2 Generation of user-model based Recommendations

In one of the first approaches taking into account the above mentioned insights from the DARPA communicator spoken dialogue system evaluation, Walker et al. (2004) used a user model in their multimodal *collaborative recommendation system* MATCH¹ in order to identify a small number of options that best match the user's preferences (UM

¹Multimodal Access To City Help

approach). A restaurant recommendation example generated by the MATCH system is given in Figure 2.2.

System: Among the selected restaurants, the following offer exceptional overall value.

Babbo's price is 60 dollars. It has superb food quality, excellent service and excellent decor. Il Mulino's price is 65 dollars. It has superb food quality, excellent service and very good decor. Uguale's price is 33 dollars. It has excellent food quality, very good service and good decor.

Figure 2.2: Output generated by the MATCH system in the restaurant recommendation domain

The MATCH system followed (Carenini and Moore, 2001) who used multi-attribute decision-theoretic models of user preferences to determine the options and attributes that are most relevant to mention when generating recommendations tailored to a specific user. MATCH extended this approach to content selection to generate summaries and comparisons among options, thus showing how the user model can be used to determine which options to mention, as well as the attributes that the user will find most relevant to choosing among them.

In MATCH, the top-level objective is to select a good restaurant. User interviews and data collection along with an analysis of online restaurant databases indicated that six attributes contribute to this objective: the quantitative attributes food quality, cost, decor, and service; and the categorical attributes food type and neighborhood. These attributes are structured into the one-level tree. The second step is to transform the real-domain values of attributes x into single-dimension cardinal utilities $u(x)$ such that the highest attribute value is mapped to 100, the lowest attribute value to 0, and the others to values in the interval 0-100. This is necessary to normalize the values of the different attributes. The vector of $u(x)$ values are aggregated into a scalar in order to determine the overall utility U_h of each option h . The final step of decision model

construction is the assignment of specific weights w_k to each attribute k . Attribute weights are user-specific, reflecting individual preferences about trade-offs between options in the domain,

Attribute	Range of values	Mapping of values to cardinal util.
Food quality, Service, Decor	0 – 30	value x 3 1/3
Cost	0 – 90	100 – (10/9 x value)
Food type, neighborhood	e.g. Italian, West Village	Top values listed by user are mapped to 90, bottom ones to 10 and all others to 50

Figure 2.3: Mapping of attribute values to utilities in the restaurant domain

The systems strategy for making a recommendation is to select the best option (based on overall utility) and provide convincing reasons for the user to choose it (based on weighted attribute values).

In the evaluation of the MATCH system, Walker et al. (2004) showed that tailoring of recommendations increases argument effectiveness and leads to greater user satisfaction. In addition, tailored recommendations were judged significantly better in terms of information quality than generic recommendations. Information quality is measured by users' response to the question: "Systems' response is easy to understand and provides exactly the information I am interested in when choosing a restaurant." Furthermore, users preferred the system in terms of ranking confidence, which was measured by the users' response to the statement: "The recommended restaurant is somewhere I would like to go."

In MATCH, a user model was used in order to select attributes and options that are most relevant for the specific user. However, options were presented using templates. Therefore, there was no variation in discourse or sentence structure. Moreover, the system was evaluated exclusively with textual and visual information presentation.

In the FLIGHTS² system, Moore et al. (2004) also followed Carenini and Moore (2001) in applying decision-theoretic models of user preferences to generate tailored descriptions of the most relevant available options. Such preference models enable systems to present information in ways that are both more concise and more tailored to the user's interests. In contrast to MATCH, the FLIGHTS system lets the user model affect all levels of natural language generation. For example, in the content selection step, the system decides what flights and attributes to present to users. The discourse planning phase determines the order of content as well as discourse relations (e.g., contrasts).

In addition, the information in the user model was exploited to select referring expressions that highlight attributes relevant to the user (e.g., "the cheapest flight" to a user concerned about price vs. "a KLM flight" to a user wishing to fly on KLM), and to signal discourse relations (e.g., contrast) with appropriate intonational and discourse cues (e.g., *although*, *but*, *because*), and scalar adjectives (e.g., *good price*). The result should be a more coherent and natural description. As a result, it was argued, users would find the information easy to understand and remember.

Figure 2.4 provides an example presentation generated by the FLIGHTS system for a student user. For obvious reasons, the price of a flight is likely to be the most relevant attribute in the user model of students, whereas they, for instance, are seldom interested in finding a flight with a specific airline. In contrast, this information could be very relevant for a frequent flyer participating in an airline bonus scheme.

Figure 2.5 presents a recommendation based on the same query but generated for a business traveller taking into account that the business traveller prefers flying business class, preferably without a lay-over and on a specific airline.

This prior work demonstrated that the user-based approach can concisely present a relatively small number of options and takes into account that users prefer a recommendation tailored to their user model in comparison to a generic one (Walker et al.,

²Fancy Linguistically Informed Generation of Highly Tailored Speech

User: I would like to book a flight from Paris to Edinburgh arriving at around 4 pm.

System: There is a direct flight on BMI with a good price - it arrives at four ten pm and costs a hundred and twelve pounds. The cheapest flight is on Ryanair - it arrives at twelve forty-five pm and costs just fifty pounds, but it requires a connection in Dublin.

Figure 2.4: Generated recommendation of the FLIGHTS system tailored for a student who is most interested in the price of a flight

User: I would like to book a flight from Paris to Edinburgh arriving at around 4 pm.

System: You can fly business class on British Airways, arriving at four twenty p.m., but you'd need to connect in London Heathrow. There is a direct flight on BMI, arriving at four ten p.m., but there's no availability in business class.

Figure 2.5: Generated recommendation of the FLIGHTS system tailored for a business traveller who is most interested in flying business class with a direct flight not requiring a lay-over

2004). In addition, the FLIGHTS user-model based approach points out the ways in which those options satisfy user preferences and presents trade-offs explicitly. Certainly, it can be seen as an appropriate strategy for dialogue systems if there are only a small number of options to present, either because the number of options is generally limited or because users can supply sufficient constraints to narrow down a large set before querying the database of options.

However, there are several limitations to this approach. First of all, it does not scale up to presenting a large number of options. When there are hundreds of options to consider (e.g., when choosing among consumer products, movies, or restaurants) there may be many options which are close in score. Additionally, users may not be able to provide constraints until they hear more information about the space of possible

options. This brings up a second problem with the user-model based approach, namely that it does not provide an overview of the option space to the users, because possible options scoring below an initially specified threshold are not mentioned. Consequently, users might miss out on options they would have chosen if they had heard about them. These last two problems may reduce user confidence in the system, because users may have the perception that the system is not telling them about all of the available options. This may ultimately lead to a decrease in user satisfaction.

2.5.3 Refinement through Clustering and Summarization

Another approach developed by Polifroni et al. (2003) structures large datasets for summarization and successive refinement (SR approach). Recommendations are based on attribute clusters which are sensitive to the data subset relevant in the current dialogue context. Thus, the system supports the user by dividing the large number of options into a small number of clusters that share attributes. Then, the system summarizes the clusters based on their attributes and presents the summaries to the user. For large data sets, the system selects attributes that partition the data into the minimal number of clusters, so that a concise summary can be presented to the user to refine.

In the SR approach, the prompts presented to the user, and the order in which they appear, are determined at run-time based on an algorithm that computes the most useful set of attributes, as dictated by the current data subset. Thus, the dialogue flow is not predetermined. It depends predominantly on the user constraints and the existing data. In addition, the implemented clustering algorithms are domain independent and thus able to cope with the changes to the underlying database. Furthermore, the SR approach solves the problem of systems presenting only the relevant options and not displaying the whole information space to users - which could lead to decreased user confidence - by giving the users the possibility of changing their constraints and preferences at any time. This is particularly useful when users do not know the structure of the underlying database.

However, there are also limitations to this approach. First, long paths in the dialogue structure may occur because the system does not have knowledge of the user's preferences. Hence, the option clusters might contain irrelevant entities which must be filtered out successively with each refinement step.

Another issue concerns the problematic exploration of trade-offs. If at least one option satisfies all requirements, this option can be found efficiently with the described strategy. In a case where there is no such "optimal" option, the system does not point out trade-offs among alternatives. For example, a user might want a flight that is both cheap and direct, while there are only expensive direct and cheap indirect flights. With the SR approach, users have to ask for cheap flights and direct flights separately. Thus, they have to explore different refinement paths successively.

Finally, the attribute that suggests the next user constraint might be suboptimal. The system selects the attribute to use in suggesting the next constraint on its ability to partition the data set into the smallest number of clusters. In the case that the attribute that is best for summarization is not of interest to this particular user, the duration of the dialogue is unnecessarily increased, and the user may be less satisfied with the system. Polifroni et al. (2003)'s approach was later extended by Chung (2004) and Polifroni (2008), who proposed a constraint relaxation strategy for coping with queries that are too restrictive to be satisfied by any option.

Figure 2.6 shows an example dialogue between a simulated user and the Polifroni system where the system first generates a summary of the different attribute values of the current option set based on the constraints of the user, and then suggests a category to refine.

Recently, Polifroni and Walker (2006b) proposed using machine learning techniques for automatically determining both summary attributes and system initiative for dialogue system responses. They have implemented these techniques in a restaurant recommendation system for London, based on the above introduced "summarize and refine" approach. In comparison to the initial SR implementation (Polifroni et al.,

Simulator: What restaurants do you have?

System: I have found 983 restaurants. Most of them are located in Boston and Cambridge. There are 32 choices for cuisine. I also have information about price range.

Simulator: Okay, tell me about the ones in Boston.

System: I have found 401 restaurants in Boston. There are 29 choices for cuisine.

Simulator: Do you have any that serve seafood?

System: I have found 19 seafood restaurants. They are predominantly in Back Bay, the North End, South Boston and the South End.

Simulator: In the South End.

System: I have found 3 restaurants. Grill Fish on Columbus Avenue, McCormick & Schmick's on Columbus Avenue and Joseph's on High Street.

Figure 2.6: Dialogue between a simulated User and the Polifroni system in the restaurant information domain

2003), where attribute clusters were determined in advance heuristically, the new implementation (see also Polifroni, 2008) determined the attribute clusters at runtime.

Polifroni and Walker (2006a) asked participants to rate hand-crafted summary statements about restaurants in London. A five-point scale was used to provide judgments in response to the statement: "This response contains information I would find useful when choosing a restaurant." The recommendations to be rated were different in the following ways: They were constructed considering different attribute ranking methods (user model versus refiner), clustering methods (association rules versus single attributes), and set sizes (large versus small).

The evaluation showed a main effect for set size with summaries describing small datasets getting higher scores. Additionally, a significant interaction between attribute

ranking method and clustering method was found. The two highest scoring summary types used (1) no association rules and no user model ranking (average score: 3.4) and (2) association rules with user model ranking (again average score: 3.4).

Optionally provided user comments suggested that associations among attributes are important to users, but only when those attributes are tailored to their preferences. Also, while in general users ranked summaries constructed for small datasets higher, if a user model was used, users give higher ratings to summaries for large datasets. With small datasets, users preferred summaries that did not utilize user model information.

Very recently, Polifroni and Walker (2008) conducted a study where they compared the initial summarize and refine strategy (SR) with a strategy that takes into account a user model (the criterion for choosing which attributes to include is based on the user's valuation instead of the diversity of the values of an attribute). Then, they also generated both the SR version and the user modelling version in a "single value" and a "associative" mode. There is an "association" between attributes, if all the options that are in a cluster because they have the same property X, also have property Y. In this case, both X and Y are mentioned in the summary. Figure 2.7 shows an example of a presentation based on the SR approach with associated clustering and Figure 2.8 shows an example message based on the hand-crafted UMSR approach with associative clustering. Both are based on the same user query. Polifroni and Walker (2008) use an association algorithm to determine which attributes are associated.

S: *I know of 35 restaurants in London serving Indian food. There are 3 medium-priced restaurants in Mayfair and 3 inexpensive ones in Soho. There are also 2 expensive ones in Chelsea.*

Figure 2.7: Example summary based on the hand-crafted SR approach with associative clustering in the restaurant domain from (Polifroni and Walker, 2008).

S: *I know of 35 restaurants in London serving Indian food. There are 4 medium-priced restaurants with good food quality and 10 with medium food quality. There are also 4 that are inexpensive but have poor food quality.*

Figure 2.8: Example summary based on the hand-crafted UMSR approach (SR with user model approach) with associative clustering in the restaurant domain from (Polifroni and Walker, 2008).

One major result of the Polifroni and Walker (2008) user study was that users preferred system utterances that were generated using a user model more when there was a large number of options available, presumably because it helps them narrow down the search space. In addition, they found that users who were not familiar with the available options (e.g., tourists in a foreign city searching the restaurant domain) appreciate tailored summaries more than users who are very familiar with the data and know more precisely what they are looking for – such users preferred a simple refinement strategy or refinement with association over a user model based recommendation.

In comparison to the initial SR implementation, Polifroni did not follow a particular attribute order he considered useful for the specific domain (restaurants) in the 2008 version of SR. Originally, the order of the attributes the system presented was predetermined based on their considered relevance in the domain.

An important consideration here is that these results were obtained off-line. The experiment presented single summaries on a web page, we do not know whether they are transferable to the spoken dialogue context. Participants were not in an interactive setting where they successively refined options and made a choice. The presented recommendation only represented a hand-crafted version of the first turn of a conversation with a dialogue system. Moreover, neither task duration nor task success was considered in the evaluation.

2.5.4 Information Presentation using Interactive Questions Answering techniques

In order to present ontological or factual information to users who are engaged in another task, Varges et al. (2006) used a knowledge manager to provide the required information, a content optimizer to regulate the amount of presented information, and a (response) generator that realizes the selected contents using a hybrid rule-based/machine learning-based approach. This dialogue system operates in the MP3 player and restaurant domains. The system features a constraint relaxation algorithm which can modify constraints (relax, tighten, add or remove) depending on the number of items in the result set, the system ontology, and (potentially) information from a user model.

User:	I want to find an inexpensive Japanese restaurant that takes reservations.
System:	I found 9 inexpensive restaurants that take reservations. Here are the first few: Ginza Japanese Restaurant, Oki Sushi Cafe, and Yona Sushi. Should I continue?

Figure 2.9: Example dialogue with the explicit presentation of restaurant candidates

The example presented in Figure 2.9 demonstrates that in this approach to information presentation a predetermined number of potentially relevant options are always presented, no matter how many remaining options there are. If there are tens or even hundreds of remaining options, this could be distracting especially when another task is being conducted at the same time.

If the list of results is very long, such as in the example shown in Figure 2.10, a constraint to refine the search is added from a list of constraints that is used up as the dialogue progresses. However, the list is domain dependent and created on the basis of perceived importance for the domain. This can be problematic if the user has specific needs and does not look for restaurants that can be found by means of

the typical ranking algorithm. For example, if the user is looking for a restaurant serving vegetarian food in Bleaker Street and the system tells the user there are 14 vegetarian restaurants in Bleaker Street and then asks the user whether she prefers a specific cuisine type (vegetarian is not considered a cuisine type in this system), the system will keep asking for the cuisine type until the constraint is satisfied, thus the user has to adopt to the system and not the other way around.

User:	I need a restaurant with a five star food rating and excellent service.
System:	I found 425 restaurants with a five star rating and excellent service, for example El Capitan Fish Market, K-Nine Country Club, and Kaiserhof German American. You could try to look for cuisine type.

Figure 2.10: Example dialogue adding a constraint to refine the search

In their experiments, Vargas et al. found that making suggestions about additional refinements or possible relaxations of constraints were particularly useful when the user query matched either many or none of the available options. For such cases, they found that fewer dialogue turns were necessary in a system that proactively suggested refinements and relaxations.

2.6 Adaptive dialogue strategies

Chu et al. (2007) developed a dialogue manager that is capable of determining dialogue strategies that are most appropriate for the current dialogue context. Their system uses data about the user's experience and dialogue performance history with the system to determine which dialogue strategy (system initiative, user initiative, or mixed initiative) is most likely to be successful. For instance, when the system does not understand the user's response, the system may re-attempt the question using the same dialogue strategy or change to a different dialogue strategy, depending on the user's experi-

ence and how many times the current dialogue strategy has failed. However, in this approach dialogue success is measured by whether the user's response to a system utterance has been recognized and/or understood. The main intention behind this work therefore seems to be to find the best possible error handling strategy rather than to actually take into account the current dialogue context, which, for instance, could be a demanding secondary task involving high cognitive load. All the same, obtaining information about the success of specific dialogue strategies in past conversations is generally a good starting point for research on the appropriateness of those strategies in cognitively demanding situations as well. It could turn out, for instance, that in such situations a user prefers always to be presented with explicit confirmations.

Bohus and Rudnicky (2007) present a similar approach considering knowledge acquired online during the conversation by discovering, eliciting and leveraging natural patterns that occur in interactions to learn dialogue strategies that are particularly successful for a specific user in a specific situation. This approach could be beneficial to develop rules that can estimate the performance of certain dialogue strategies in specific situations based on the dialogue history. For instance, it may be the case that a specific user performs better with explicit confirmation strategies, whereas another user is more successful with implicit confirmations. In this case such an adaptive dialogue manager could take the dialogue history into account and mainly use the preferred or more successful confirmation strategy.

Developing a dialogue system capable of adapting to the workload of the user in the current situation would potentially involve the consideration of external factors affecting the user's performance. For example, in an in-car scenario frequent use of the brake pedal could act as an indicator of situations in which the system should reduce the amount of information presented up to the point where the system would not generate any output at all for safety reasons.

2.7 Combining User Tailoring and Clustering

Finally, in a very recent line of research Demberg (2005) combined the benefits of the user-model based (UM, Moore et al., 2004) and the summarize and refine approach (SR, Polifroni et al., 2003). The user model reduces the dialogue duration by considering only options that are relevant to the user. When the number of relevant items exceeds a manageable number, the UMSR approach builds a cluster-based tree structure ordering the options for stepwise refinement based on the ranking of attributes in the user model. The effectiveness of the tree-structure, which directs the dialogue flow, is enhanced by taking the users preferences into account. Furthermore, trade-offs between alternative options are presented explicitly in order to provide the user with a better overview of the option space. In addition, to give users confidence that they are being presented with all relevant options, a brief account of all the remaining options is also provided.

All three problems of the summarize and refine approach are addressed in the UMSR approach. When a user model is available, it allows the system to specify which options and which attributes of options are likely to be of interest to the specific user based on user preferences. Then, the system can identify compelling options to include, and delete irrelevant options from the refinement structure, leading to shorter refinement paths. Furthermore, the user model enables the system to determine the trade-offs among options, which can then be presented explicitly. The user model also allows the identification of the most relevant attribute at each stage in the refinement process. In the following section, the UMSR approach is explained in detail based on a soon to be submitted publication: Demberg et al. (2009).

2.7.1 UMSR - System Architecture

An overview of the UMSR system's pipeline architecture is given in Figure 2.11. After speech recognition and natural language understanding, the first step in natural

language generation is the *content selection and structuring* step. This step is primarily responsible for deciding what information should be communicated in the system's response, and structuring the information based on the user query, user model and data base entries. The core part of this step is the tree building and pruning algorithm which structures the entities into the tree and selects the entities that should be mentioned.

The *text planning* step takes the pruned option tree as input and transforms it into natural language. First, it decides on the content of one dialogue turn, and how to structure the argument. In domains that aim to recommend items to users (i.e., in product recommendation), the ordering can be arranged to increase the effectiveness of a recommendation or argumentation (Carenini and Moore, 2001). The content planner of the system is implemented in the schema-based AI planning language O-PLAN (Currie and Tate, 1991). The resulting content plan is the input to the subsequent sentence planning step. The sentence planning component performs lexical choice, aggregation, and constructs alternative logical forms. These logical forms are combined into a single packed representation, which is then sent to OpenCCG, a CCG-based realizer (White et al., 2007). The realizer transforms the logical forms to natural language sentences, makes the final choice on structure using a statistical n-gram model, and adds intonation to support the theme/rheme structure of the utterance.

The system components are implemented as agents in the Open Agent Architecture (OAA) framework (Martin et al., 1998). All modules are implemented as agents, whose communication is managed by the DIPPER dialogue manager agent, that calls the different agents and stores the intermediate results from each component.

This approach to information presentation concerns mainly the content structuring and selection step of the system. It consists of three major steps: clustering, building the option tree and pruning. The first step in the content structuring algorithm is to cluster the values of each attribute in order to group them in a way that labels such as "cheap", "moderate", "expensive" can be assigned to values of continuous categories like *price*. This clustering also enables easier summarization of options later on.

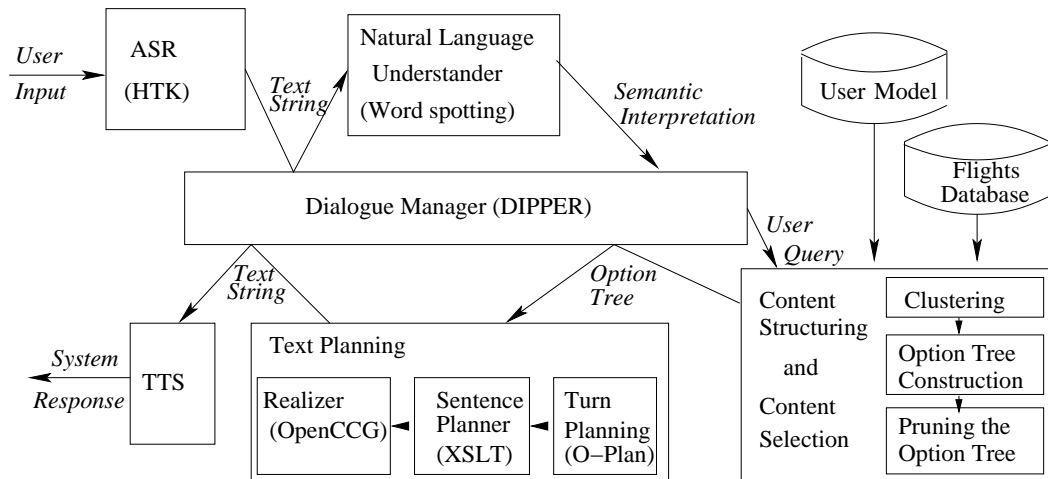


Figure 2.11: System Architecture.

Next, the system constructs the *option tree*. Each branch of the tree describes a possible refinement path and will thus direct the dialogue flow. The construction of the option tree is driven by three factors, the user model, the data base and the attribute value clustering. The resulting option tree determines how different options relate to one another, and which ones are most attractive for the user. After the option tree structure has been constructed, it is pruned based on the information from the user model which enables the system to distinguish between options that are likely to be compelling to the user and those that are not. At this point, the content selection and structuring process is complete, and the option presentation phase follows, which consists of determining turn length and deciding on realizations for the information that is to be conveyed. The content presentation component of the system is an adaptation and extension of the work of the FLIGHTS system (Moore et al., 2004).

2.7.2 Clustering

The UMSR based dialogue system uses agglomerative group-average clustering to automatically group values for each attribute; comparable to the algorithm described in Polifroni and Walker (2008). The algorithm begins by assigning each unique attribute

value to its own cluster, and successively merging those clusters whose means are most similar.

For example, see Figure 2.12 where the prices from six flights are displayed as dots on the price axis. In the first step, each flight is in its own cluster (represented as a circle around the dots). In the second step, the clusters of the two flights with the most similar prices are merged. This procedure continues until a stopping criterion is met. In our implementation, we stop when we have reduced the number of clusters to three. These clusters are then assigned predefined labels, e.g., cheap, average-price, expensive for the price attribute. This clustering is used to group similar attribute values together and is only performed once for each request (in the airtravel domain, a request corresponds to one origin-destination pair for a specific date). Categorical values are clustered using the user's valuation: For example, airlines are clustered into a group of preferred airlines, dispreferred airlines and airlines the user does not-care about.

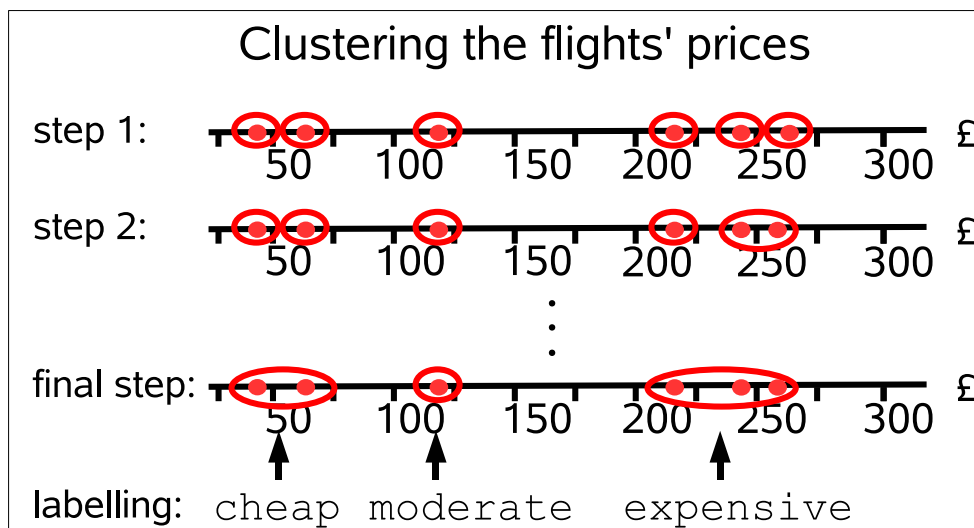


Figure 2.12: Clustering options with agglomerative group-average clustering and labelling.

Clustering allows the algorithm to assess the similarity of options, i.e., instead of talking about the “£51 flight” and the “£48 flight”, the system would refer to the “cheap flights”. This leads to more efficient summarizations and enables the system

to avoid presenting many options that are very similar in all respects. Furthermore, the clustering process enables the system to assign labels that are sensitive to the other options in the data base. For example, a £300 flight is assigned the label *cheap* if it is a flight from Edinburgh to Los Angeles (because most other flights in the database are more costly) but *expensive* if it is from Edinburgh to Amsterdam (for which there are many cheaper flights in the data base).

2.7.3 Building the Option Tree

The tree building algorithm arranges the available options into a tree structure, see Figure 2.13. Every branching point in the tree corresponds to a choice (e.g., between economy vs. business class flights). The nodes of the option tree are labeled with a specific value and attribute (e.g., fare class: *economy*) and correspond to sets of options: see for example Figure 2.13, where the root of the tree contains all options, and its left child contains all flights with seats in economy class. The children contain complementary subsets of these options (i.e., all *direct* economy class flights vs. all *indirect* economy class flights). Leaf-nodes correspond either to a single flight or to a set of flights, where for each attribute of an option, the values are either the same, or fall within the same cluster (e.g., prices of all these flights are moderate, they all require one connection, they are all economy class, etc.).

To maximize the efficiency and effectiveness of the dialogue, the dialogue structure is tailored to the user based on her user model. Users' ranking of attribute importance is crucial for dialogue efficiency. If an irrelevant criterion is used as the branching criterion high up in the tree, interesting trade-offs would risk being scattered across the different branches of the tree. For example, it would be suboptimal to ask a business user to make a choice about cheap vs. expensive flights first, if she does not care about this aspect, and would then have to try to identify interesting flights among both the cheap and the expensive flights. The algorithm chooses the attribute that has the highest

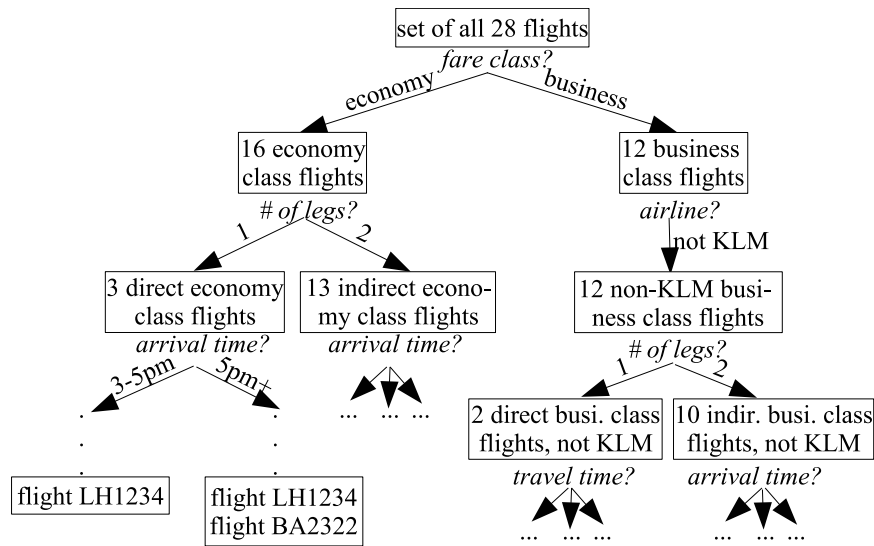


Figure 2.13: Option tree for business user

weight according to the user model as the the branching criterion for the first level of the tree. For the business user, this would be fare-class.

The next decision concerns the attributes that are second most important, such as the number of legs required, and so on. The system therefore constructs the tree such that it presents the criteria which are most relevant for the specific user first, and leaves less relevant criteria for later in the dialogue (i.e., further down in the tree). The advantage of this ordering is that it minimizes the probability that the user needs to backtrack.

A special case occurs when an attribute is homogeneous for all options in an option set (for instance if none or all of the business class flights happened to be on the user's preferred airline). In that case, a unary node is inserted regardless of the rank of its attribute (see for example the right subtree with the attribute `airline`, which is inserted far up in the tree despite its low rank, in Figure 2.13). This special case allows for more efficient summarization, e.g., “None of the business class flights are on KLM.” instead of having to say this in subsequent dialogue turns for each of the business flights that the user explores.

In cases where several attributes have the same rank in the user model, UMSR follows the SR approach Polifroni et al. (2003); Polifroni and Walker (2008). The algorithm selects the attribute that partitions the data into the smallest number of sub-clusters. Consider again the tree in Figure 2.13: `number-of-legs` creates only two sub-clusters for the data set (`direct` and `indirect`) and is therefore further up in the tree than `arrival-time`, which splits the set of economy class flights into three subsets (“before 3pm”, “3pm to 5pm”, “after 5pm” for a user whose preferred arrival time is “by 5 pm”).

The tree building algorithm constitutes one of the main differences between the UMSR and SR algorithm’s refinement processes. The SR system chooses the attribute which partitions the data into the smallest set of unique groups for summarization is chosen, whereas our UMSR system takes the ranking of attributes in the user model into account.

2.7.4 Pruning the Tree Structure

After the tree building step, the tree contains all the options in the data base. This tree can potentially be quite large and navigating through it would be very pain-staking for the user. At this point, the user model comes into play again: since the system already knows which options are relevant to the user (and which ones are not), it can prune the option tree to retain only options that it classifies as being useful to the user.

To determine the relevance of options, we define the notion of “dominance”. *Dominant* options are those for which there is no other option in the data set that is better on all attributes. A *dominated* option is in all respects equal to or worse than some other option in the relevant partition of the database; it should therefore not be of interest to any rational user.

Pruning dominated options is crucial to the structuring process. The algorithm uses information from the user model to prune all dominated options. The pruning

algorithm operates directly on the option tree, and exploits the tree structure in order to efficiently determine dominance relations.

The first step of the algorithm is to *order*³ the tree such that the best options are leftmost. The algorithm then traverses the tree in depth-first order and *generates constraints* during this process. These constraints encode the properties that other options would need to satisfy in order to be classified as not being dominated by any of the options seen so far. A branch must fulfill the constraints that apply to it, otherwise it is *pruned*. If an option (or a cluster of options) satisfies a constraint, the property that satisfied the constraint is marked as the options' *justification*. If some, but not all, of the constraints can be satisfied by an option, the constraints are *propagated* to the options that are further to the right in the ordered option tree. Once all the dominated options have been pruned from the option tree, there is a *homogeneity check* to ensure that attributes which have the same value among a set of options are annotated at a node that is a common ancestor of all of these options.

Tree Ordering In the crucially important first step of the pruning algorithm the tree is ordered. In this step, the available options are ordered and arranged such that the best option of every node becomes that node's leftmost leaf. For example, the tree in Figure 2.13 is not ordered, because the business user prefers flying business class over economy class. Therefore, the two subtrees under the root node need to be exchanged, see Figure 2.14. The total ordering is enforced by sorting the attribute values from best to worst in each node.

Constraint Generation After ordering the tree, the globally best option is described by the leftmost branch in the option tree. In our example in Figure 2.14, this is flight LH1554, in node 6. If the globally best option in node 6 was perfect (i.e., if it was exactly what the user was looking for), the option in node 6 would dominate all other options, and the rest of the tree would be pruned. However,

³Alternatively, the tree construction algorithm can be designed to insert all options such that the resulting tree is already ordered.

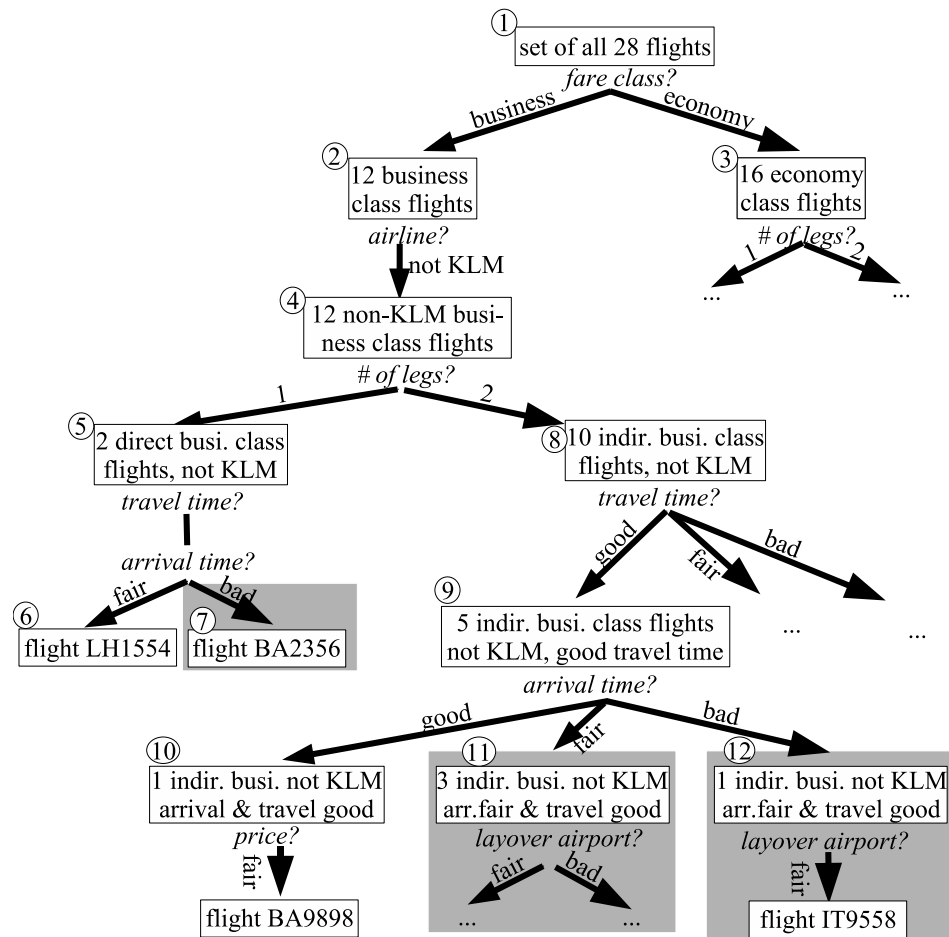


Figure 2.14: The shaded subtrees are the ones that are pruned. Subtrees 13-16 still need to be decided on at the stage shown here.

if there is an aspect of the globally best option which does not match the user's ideal, the user will have to make some kind of trade-off. This is what happens in the example, because the arrival time of the flight in node 6 was only classified as "fair" but not as "good", while there exist some connections with arrival times that were classified as "good". A flight with a good arrival time constitutes a possibly interesting alternative. In order to find such an option and filter out the others, the constraint `arrival-time:good` is generated.

Pruning Options from the Tree When node 7 is reached by the depth-first traversing algorithm, a constraint (`arrival-time:good`) has been generated by node 6. Node 7 does not satisfy this constraint; this means that it is dominated at node 6

and therefore is pruned from the option tree (as indicated by shading in Figure 2.14).

Constraint Propagation Once the status of a node's children has been determined, any unsatisfied constraints that were generated by the child nodes are propagated to the parent. In the example, the constraint generated by node 6 is propagated up to parent node 5. The sibling of node 5, node 8 is then tested against the constraint `arrival-time:good`. Since there is no information about arrival time available at node 8, the constraint is passed down to its leftmost child (node 9). If that child node does not have information about arrival time, the constraint is passed down again. The constraint is passed down to node 10, and we find that this flight satisfies the constraint. Next, the constraint generation step is repeated. Flight BA9898 generates the constraint `price:good` because its own price is only classified as `fair`. Nodes 11 and 12, both constraints `arrival-time:good` and `price:good` have to be satisfied, which they can not. They are therefore pruned. The depth-first traversal continues through the tree trying to find options that satisfy the constraints.

Note that the constraints allow for efficient pruning: it is not necessary to look at the exact instances or properties of nodes 11 and 12 or their children. All that must be done is to determine which properties are relevant to the constraints because the tree is ordered. This allows us to conclude that all options in a specific subtree are dominated by the options in branches to the right of that subtree.

Justifications An important by-product of the pruning algorithm is the identification of attributes that make an option cluster compelling with respect to alternative clusters. For example, the flights in node 10 were considered compelling because they had a *better arrival time* than the flight in node 6. In UMSR, such an attribute is called the “justification” for a cluster, as it justifies its existence, i.e.,

it is the reason it is not pruned from the tree. Node 6 in turn is kept in the tree because it is the leftmost child, which means that its attribute values best match the user's preferences. Its compelling property when compared to the flights in node 8 is that it is *direct* (i.e., number of legs = 1). The default justification for a node is the attribute value on which the branch is based (e.g., fare class for node 2 in Figure 2.14). It is used for nodes on the leftmost branch. Justifications are used by the generation algorithm to present trade-offs between alternative options explicitly (see Section 2.7.7).

The reasons why options have been pruned from the tree are also registered. These reasons contain information about which constraints the options *failed* to satisfy; in our example, the flight in node 7 is deleted because of its bad arrival time. These pruning reasons are later used to provide information for the summarization of poor options whose function it is to give the user a better overview of the option space (e.g., "All other flights arrive too late or are more expensive."). To keep summaries about irrelevant options short, we back off to a default statement "or are undesirable in some other way." if these options are very heterogeneous.

Homogeneity Check After deleting branches from the option tree, it may be the case that several options have the same attribute value, but are located in different branches in the tree. For example, imagine there are three economy class flights, two direct ones (1 leg) and one which requires a connection (2 legs). Among the two direct ones, one has a good price, and the other one is more expensive. The 2-leg flight also has a good price. If the more expensive direct flight is pruned, both of the remaining options have a good price. This property should therefore be above the `number-of-legs` branching level in the tree. This is important for efficient information presentation and summarization of options.

2.7.5 Option Presentation

The user model also comes into play when determining the wording of the option presentation. Because the system has a model of the user's preferences, it can effectively compare and contrast alternatives by highlighting compelling aspects of an option (e.g., a *direct* flight, the KLM flight), and by acknowledging drawbacks using linguistic markers (e.g., *but*, *however*, *although*), intonation and comparatives (e.g., the *cheapest* flight, the *only* KLM flight). For the options that were considered unattractive for the particular user, it can provide an overview to cover the option space (e.g., *All other flights arrive later than 3pm.*).

Figure 2.15 shows how the nodes in the pruned option tree translate to the system's utterances. The different design decisions underlying sentence planning and realization will be explained in the following sections.

2.7.6 Turn Length

In a spoken dialogue system, it is important not to present too much information in a single turn in order to keep the memory load on the user manageable. Thus, a system based on the UMSR approach to information presentation aims at presenting no more than two or maximally three options at once. However, the pruned option tree sometimes contains more than this critical number of options, and thus needs to be broken down into smaller entities. We thus cut the pruned option tree into several smaller dialog-turn-sized subtrees. Typically not all of these subtrees will be presented, but only the ones between the root of the tree and the chosen subset of flights that the user wishes to hear more about.

In addition to determining the number of options to present in a turn, the system must decide about which properties to present. Arguably, mentioning too many properties of options will also lead to memory overload, which may ultimately reduce user satisfaction. While keeping this in mind, the algorithm needs to provide enough infor-

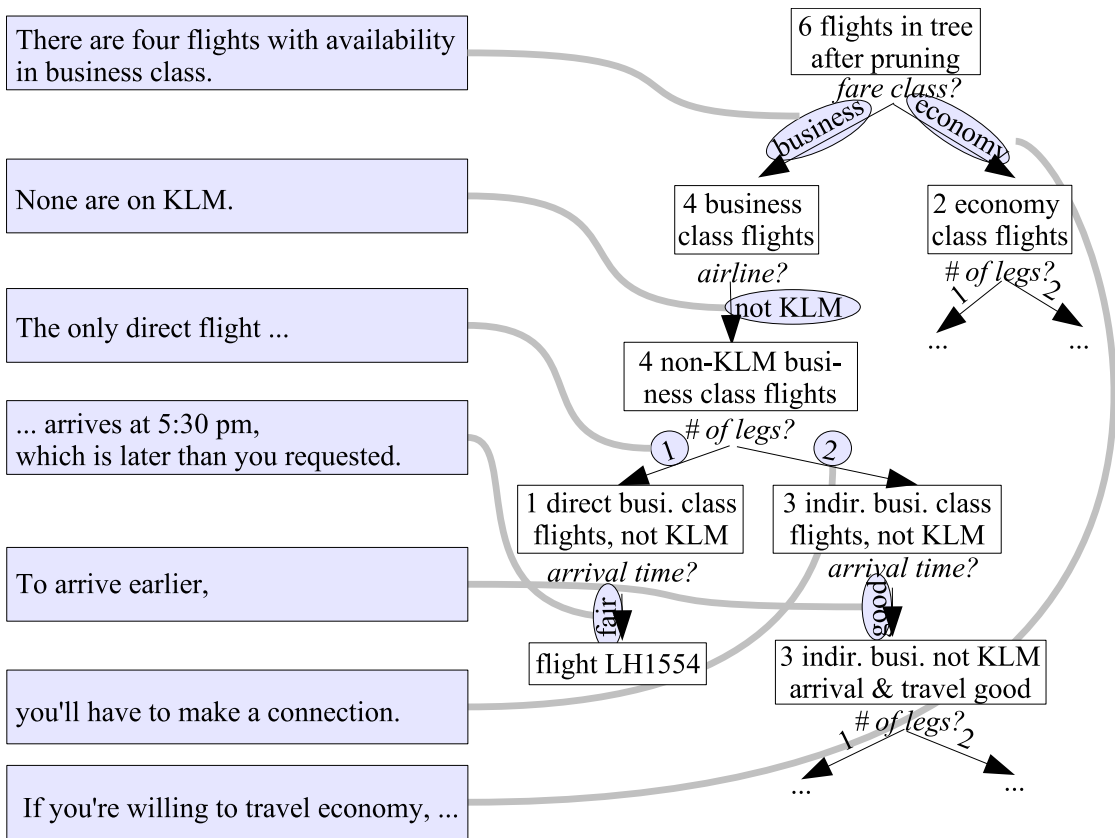


Figure 2.15: Diagram showing how the pruned option tree is mapped onto language. The tree on the right hand side corresponds to the example trees in Figures 2.13 and 2.14.

information to fully account for what constitutes the trade-off and thus give the reasons for why an option is potentially relevant.

In order to segment the pruned tree into turn-sized subtrees, we chose a very simple heuristic segmentation algorithm. The heuristic cut-off point is visualised in Figure 2.16, and defined as “no deeper than two branching⁴ nodes and their children”.

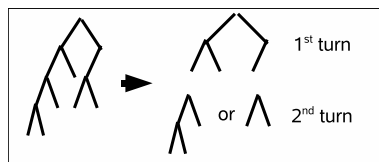


Figure 2.16: The option tree is cut into subtrees which determine turn length.

⁴Branching nodes as opposed to unary nodes. For example, in Figure 2.13, the unary node in the right subtree would not count as a separate level

The above heuristic produces a set of options with a limited size to be presented in a single turn. The target size is two to three options. In practice, there are typically about three or less options in any two branching levels left after pruning. Two layers were chosen in order to allow for informative trade-offs: If information from only one layer was available at any time, it would not be possible to contrast the most relevant advantages and disadvantages of alternative options in comparison, which is required to make explicit trade-offs. At the end of the turn, the user is expected to make a choice indicating which of the options she would like to hear more about.

2.7.7 Referring to Sets of Options

Each branch in the pruned tree corresponds to a set of options. These options should be referred to in an effective way. This is done by taking into account both the dialogue structure (i.e., structure of the argumentation) and the user's interest: The description of a set of options is based on their justification. For example, the justification of the flights in the left branch of the tree in Figure 2.15 is their fare class. Therefore, they are described as flights "with availability in business class". On the other hand, the justification for the indirect business class flights is that they have an arrival time that matches the user query better. They are thus referred to by their justification "to arrive earlier".

If a node is justified by several attributes, only one of them is selected for reference. If one of these multiple justifications is a contextually salient attribute, this one is preferred over the justifications that are not salient. For example, if a node is justified both by its arrival time and its price, it would be referred to by the price attribute in a context that just mentioned the price of another flight as being expensive:

“[... but it costs 1000£.]*context* A [cheaper]*salient* flight...”

If none of the attributes are particularly salient, the options in the cluster are referred to by the highest ranked attribute, i.e., arrival time in this example.

2.7.8 Presenting Additional Attributes to Explain Trade-Offs

In order to present trade-offs between options, it is necessary to provide information about the properties of options that constitute that trade-off. Any of these additional properties, which are not already mentioned as part of the referring expression, are ordered to optimize coherence. First, all positive attributes are enumerated and contrasted against all average or negative attributes. These negative attributes, which are presented last, are then salient and will be used in the description of an alternative option.

2.7.9 Summarizing Properties of Options

When talking about a set of flights that are in the same cluster (e.g. because they have “good” arrival times), the specific attribute values of the options in this cluster may vary (e.g., one flight might arrive at 3 pm and the other at 3:30 pm). In that case, it is necessary to generate a summarizing expression for these attribute values.

There are three main cases to be distinguished:

1. The continuous values for the attributes `price`, `arrival-time`, etc. must be summarized, as they may differ in their values even if they are in the same cluster. One way to summarize them is to use an expression that reflects their value range, e.g., “between x and y”, or that contrasts them to some previously mentioned value, e.g., “earlier” or “The *cheapest* flights.”
2. For discrete-valued attributes with a small number of possible values, for example, `number-of-legs` and `fare-class`, summarization is not an issue, because when homogeneous for a cluster, the attribute values of the options are identical.
3. The third group are attributes with categorical values, e.g., “airline”. If there are no more than three different values, we summarize using quantifications such as “none/all/both of them”, as done in Polifroni et al. (2003), for example. If

the values are more diverse, the user model comes back into play to produce a tailored summary based on user preferences. For a user who prefers to fly on KLM and a cluster which contains no KLM flights, as in Figure 2.15, we generate “None are on KLM.”, which takes into account the user’s preference of liking KLM and is more concise than mentioning all airline names.

Following is an example dialogue between a system using the UMSR approach (Demberg and Moore, 2006) and a frequent flyer. The frequent flyer is particularly interested in flying on a specific airline (in this case, KLM), and is also concerned about the arrival time, whereas the price is not relevant.

User:	Hello. I would like to book a flight from Edinburgh to Brussels for Thursday, 27th July, arriving at about noon. I prefer KLM flights.
System:	I found three flights with a good price. Two of them are on KLM. They both cost 111 pounds, but require a connection in Amsterdam. If you want to fly direct, there’s a BMI flight arriving Brussels at 9:40 am that costs 107 pounds. All other flights arrive after noon.
User:	Please tell me more about the KLM flights.
System:	One of them arrives at 11:20 am but has a travel-time of 4hours 15 minutes. The other takes just 3hours 10 minutes but arrives at 4:40 pm.

Figure 2.17: Example dialogue between the Demberg system and a frequent flyer

In order to evaluate the newly introduced approach to information presentation, a laboratory study comparing the UMSR and SR presentation approaches was carried out (Demberg and Moore, 2006). In this experiment, participants were asked to read and evaluate transcripts of six dialogue pairs. Each pair consisted of one dialogue adopting the SR approach to information presentation and one using the UMSR approach. The four criteria used for evaluation were:

Q1: Understandability “Did the system give the information in a way that was easy to understand?”

Q2: Overview of options “Did the system give you a good overview of the available options?”

Q3: Relevance of options “Do you think there may be flights that are better options for the user that the system did not tell her about?”

Q4: Efficiency “How quickly did the system allow the user to find the optimal flight?”

The within-participants laboratory study with 38 participants showed no difference between the systems in terms of understandability. However, on all the three other criteria, namely overview of options, relevance of options, and efficiency, the UMSR approach was rated significantly better. In sum, the evaluation of UMSR demonstrated that integrating a user model to the content structuring techniques utilized in the SR approach allows the system and user to navigate through a large set of options. Moreover, such a combined information presentation approach enables the explicit presentation of trade-offs. This resulted in this experiment in increased overall user satisfaction, a better overview of options, and increased user confidence in the system.

2.8 Dimensions for evaluating information presentation strategies

There are multiple dimensions of variations to be considered when studying information presentation strategies. There are three main dimensions which are discussed in this chapter.

concise vs. not concise To test whether differences in conciseness influence argument effectiveness, Carenini and Moore (2006) devised and implemented an evaluation framework in which the effectiveness of evaluative arguments can be mea-

sured with real users. Participants were asked to act as “decision-makers” in a selection task. This role involved selecting a subset of preferred objects (e.g., houses) out of a set of possible alternatives by considering trade-offs among multiple objectives (e.g., house location, house quality) and by evaluating the objects with respect to their values for a set of primitive attributes (e.g., distance from work, size of the garden). The experimental framework assumes that a model of the user’s preferences (AMVF) has been previously acquired from the user, to assure a reliable initial model. The specific task comprised two subtasks. At the start of the first subtask the user is presented with information about a set of alternatives. Then, she is asked to select a subset of n preferred alternatives and to order them by preference in what is called a “Hot List”. Then, the User Model Refiner refines the initial model, making any adjustments necessary to make the model as consistent as possible with the preferences that the user expressed by creating her Hot List. This refinement process produces a Refined Model of the User’s Preferences by heuristically adjusting the model weights.

In the second subtask the user is presented with an evaluative argument about a new instance (not included in the initial set of alternatives), and she is asked whether she wants to include it in her Hot List. This new instance was designed to have an overall utility between the utilities of the top two options in the user’s hot list. If the user’s answer is affirmative, she has to decide where to place the new instance in the ordered Hot List. When the user decides to stop exploring, and can thus be assumed to be satisfied with the selections in the Hot List, measures related to the argument’s effectiveness can be assessed. Finally, the user fills out a questionnaire about her attitudes and beliefs about the new instance and the decision task. Measures of argument effectiveness are obtained from the record of the user’s interaction with the system and from user self-reports in the final questionnaire.

The experiment focused on the empirical questions related to two assumptions; namely, whether recommendations based on a user model led to more effective than non-tailored recommendations and what is the most effective degree of conciseness for evaluative arguments. To test the first assumption, Carenini and Moore (2006) compared the effectiveness of arguments tailored to the user's AMVF with the effectiveness of arguments tailored to a default AMVF, for whom all aspects of a house are equally important (i.e., all the weights in the AMVF are equal). To test the second assumption, as a preliminary attempt to determine an optimal level of conciseness for evaluative arguments, the authors compared the effectiveness of arguments generated by their argument generator at two different levels of conciseness.

(Carenini and Moore, 2006), by comparing the four different experimental conditions

- No-Argument (NA) - no evaluative argument, only information about the new house under discussion),
- Tailored-Concise (TC) - evaluative argument about the new house tailored to their preferences and at a level of conciseness that Carenini and Moore (2006) hypothesized to be optimal),
- Non-Tailored-Concise (NTC) - an evaluation of the new house that, instead of being tailored to their preferences, is tailored to the preferences of a default average user),
- and Tailored-Verbose (TV) - evaluation of the new house tailored to their preferences, but at a level of conciseness that they hypothesized to be too low),

demonstrated that differences in conciseness significantly influence argument effectiveness. Z-scores⁵ were used as the primary measure of argument effectiveness. They are also sometimes called “standard scores”, and are especially useful when comparing the relative standings of items from distributions with different means and/or different standard deviations. The satisfaction z-scores were supposed to precisely and concisely integrate all the measures of behavioral intentions and attitude change. To summarize the results, the satisfaction z-scores obtained in the experiment provide support for the above-mentioned hypotheses. Arguments generated for the TC condition had greater satisfaction z-scores than arguments generated for the TV, NTC and NA conditions. The difference in effectiveness between arguments generated in the TC condition and arguments generated in the TV condition was statistically significant ($p < 0.05$, TC had greater z-scores), while the differences between the other two conditions TC vs. NTC and NA only approached significance ($p < 0.1$).

In another experimental study in the restaurant information domain, Whittaker et al. (2003) showed that users are indeed sensitive to conciseness and that there is a correspondence between algorithmic conciseness and user judgments of conciseness, meaning that algorithmic control over useful conciseness can be achieved. In addition, they found that users judged “recommendations” to be more concise than “comparisons” of options.

tailored vs. not tailored In a study on user tailoring in regard to generating and presenting effective evaluative arguments, Carenini and Moore (2001) found that all previous approaches acted on the assumption that effective evaluative arguments should be constructed considering the values and preferences of the audience towards the information presented. To empirically evaluate this claim they compared the effectiveness of arguments that were tailored vs. non-tailored to a

⁵subject’s self-reported satisfaction with the new house, with respect to the self-reported satisfaction with the other houses

(quantitative) model of the user preferences and showed that tailored arguments were significantly more effective than non-tailored arguments. In addition, users generally prefer responses generated using their own model over responses generated using a randomly chosen model of another user (Walker et al., 2002).

list style presentation vs. pointing out trade-offs The two previously mentioned dimensions concerned decisions about a natural language generation system's content planning phase of natural language generation (NLG) (Reiter, 1994). The third dimension is concerned with how the information is presented in natural language and is called realization phase in the NLG paradigm. In this thesis, I am interested whether users perform better (in terms of comprehension and recall) with the "list style" approach to information presentation that is most commonly used in current spoken dialogue systems (see the DARPA communicator and MATCH dialogue systems, Levin et al., 2000; Walker et al., 2004, for instance), or with the type of presentations that uses coherence markers (e.g., *but, however, moreover, only, just, also, etc.*) in order to highlight specific properties of and relations between the presented items.

2.9 Conclusion

In this chapter I gave an overview about a number of previously proposed information presentation strategies in spoken dialogue systems. I specifically looked at systems that help users navigate through a large number of options/items in order to present them with the information they are looking for. Having reviewed the most relevant existing approaches to information presentation, I will focus on the SR and UMSR approaches to information presentation because a) they are recently introduced state-of-the-art approaches, b) they use interesting new techniques to facilitate information browsing (SR) plus user-modeling (UMSR), and c) they can be implemented relatively easily.

We carry out experiments to determine how easy it is for users to comprehend and to recall information when presented using different strategies. The obtained results will undoubtedly help us to answer questions about cognitive load, since the more complex a sentence is the more cognitive resources are required to comprehend it. Results of our experiments will provide valuable guidelines to developers of intelligent dialogue systems, which support the user by adapting to the actually occurring workload or the users' cognitive capabilities (for instance, older users).

2.10 Open Questions

Results from the work reported in this chapter suggest that tailored recommendations based on a user model may be one suitable way to address the problem of sequential information presentation tending to overload users. We aim to test our hypothesis that users are likely to prefer messages containing sentences that are more complex but point out trade-offs between different options explicitly, rather than (potentially) simpler sentences which do not explicitly point out the trade-offs and contrasts involved. Therefore, I will implement different presentation methods and compare them by experimental validation.

The results of the experiments aimed at evaluating the UMSR approach were based on a non-interactive experiment asking participants to rate presentations based on the SR and UMSR approaches to information presentation presented as dialogue transcripts. However, research on spoken dialogue systems is ultimately aiming to facilitate real interactions between humans and machines. Thus, there still remain three questions:

1. Would users still prefer UMSR when they are actively interacting with the systems?
2. Will users perform better with UMSR in comparison with SR in terms of task success and dialogue efficiency?

3. Would UMSR show an advantage when participants are performing a demanding secondary task at the same time?

To answer these questions, a series of experiments were conducted.

2.11 Hypotheses

A reading-task experiment was conducted showing that participants preferred system responses based on the UMSR approach to information presentation (Demberg and Moore, 2006). We believe that the two main reasons for participants' preferring UMSR over SR are that a) UMSR utilizes information from a user model to present only relevant options and b) UMSR presents trade-offs between options explicitly making it easier to compute the differences between options mentally. Since we base these hypotheses so far only on a reading-task experiment where participants judged written dialogues, we are planning to compare SR and UMSR in an experiment where participants actually interact with a system deploying the two strategies to information presentation. Thus, we aim to experimentally test **Hypothesis 1a** in an interactive Wizard-of-Oz experiment, which is described in Chapter 3 - Evaluating Task Success and User Perceptions in a WoZ-Experiment:

Hypothesis 1a: Users will prefer and perform better in terms of task efficiency and task success when they interact with a system that uses the UMSR approach to information presentation when compared to a system employing the SR approach.

We consider cognitive load an important factor in spoken dialogue system research because spoken dialogue systems are often intended for situations where the user's hands and eyes are busy performing another task. Because cognitive resources are limited, developing SDS that adapt to users' cognitive load would be a big step forward. In order to find out about the level of cognitive load that different approaches to

information presentation place on users, I plan to conduct experiments evaluating the effect of cognitive load on task efficiency, task effectiveness, user perception, and user recall of information. One method of measuring cognitive load involves conducting dual-task studies, as highlighted in Chapter 4 - Related Work - Measuring Cognitive Load. In order to investigate whether the UMSR approach to information presentation also shows an advantage when participants are performing a demanding secondary task at the same time, I conducted a Wizard-of-Oz experiment.

I generally assume that UMSR makes it easier to mentally compute trade-offs and differences between options because it uses linguistic devices (e.g., connectives, lexical cue phrases, and adverbials) that highlight specific properties of and relations between items presented to the user. This hypothesis is mainly based on psycholinguistic findings that are detailed in Chapter 6. To test **Hypothesis 1b**, I will make use of the dual-task paradigm and conduct a Wizard-of-Oz experiment in which participants drive a simulated car while simultaneously conversing with a dialogue system presenting information according to SR or UMSR (see Chapter 5 - Evaluating Task Success, User Perceptions, and Cognitive Load):

Hypothesis 1b: Users that are performing another (demanding) task simultaneously will also benefit from a system employing the UMSR approach in comparison with a system using the SR approach to information presentation.

Designing the best possible experimental materials is critical for conducting interactive Wizard-of-Oz experiments requiring users to perform two tasks simultaneously. Flaws in the materials of the first dual-task study highlighting the relevance of using concise experimental materials make it necessary to conduct a second dual-task study with materials considering balancing message length and information density per dialogue turn. Without balancing message length, task performance of the presentation method deploying longer messages will be negatively affected. For this experiment, described in Section 5.6 - Revised Dual-Task Wizard-of-OZ Experiment, I developed the following hypothesis:

Hypothesis 2: Concise messages are more effective than verbose ones in situations where users have to divide their attention between two or more stimuli. Even though more turns may be required to complete the task, users will perform better (dialogue task performance, secondary task performance) when interacting with a system that presents concise messages in comparison with a system deploying more verbose presentations.

Finally, partly derived from the considerations presented at the end of Chapter 2.5 regarding the potential variations in presenting options, our experimental results, and based on psycholinguistic findings concerning online sentence comprehension presented in Chapter 6 - Evaluating textual and auditory comprehension and recall, we developed and tested (see Chapter 6) **Hypothesis 3:**

Hypothesis 3: Once the system arrives at a point where there is a manageable number of items to present within a single turn, messages that make trade-offs between items explicit will facilitate recall in comparison with a system that presents the remaining items as a list (as does the MATCH system, Walker et al., 2004). At the same time, messages that point out trade-offs and contrasts between options will not negatively affect message comprehension.

While the first hypotheses covered the part of the dialogue that narrows down the number of initially available options, this last hypothesis covers the last step of a typical information seeking dialogue: how best to present a manageable number of options for the user to choose between facilitating comprehension and recall of the presented items and their properties.

In this thesis, we aim to understand different levels of cognitive load placed on users by SR and UMSR dialogue strategies performing user studies. Ultimately, the goal is to empirically study how to adapt to differences in cognitive load. On the one hand, UMSR should place lower cognitive load on users because this approach to information presentation explicitly points out trade-offs between options. We assume

that this requires less cognitive processing because it is easier for users to build a situation model. On the other hand, the longer and potentially more complex sentences used in UMSR should place higher cognitive load on users.

To address the first of the open questions that remained after the first UMSR evaluation (see Section 2.10), namely whether users would still prefer UMSR when they are interacting with a spoken dialogue system, we conducted a Wizard-of-Oz experiment with a simulated dialogue system (Winterboer and Moore, 2007).

Chapter 3

Evaluating Task Success and User Perceptions in a WoZ-Experiment

The main motivation for conducting the Wizard-of-Oz experiment described in the following was to examine whether we would obtain the same pattern of results that were found in the “reading task” experiment (presented in Section 2.7) when participants actually interacted with a (simulated) dialogue system. In this study, the UMSR and SR approaches to information presentation are compared in terms of their impact on task success, dialogue efficiency, and user satisfaction.

3.1 The Wizard-of-Oz Paradigm

In human-computer interaction research, a Wizard-of-Oz experiment (WoZ, Dahlback et al., 1993) describes a research experiment in which participants interact with a computer system that they believe to be autonomous, but which is actually being operated or at least partially operated by an unseen human being. For example, a participant in an experiment may think she is communicating with a computer using a speech interface, whereas the participant’s words are actually being secretly entered into the computer by a person in another room (a “wizard”). Usually, the missing system function-

alities the wizard provides may be implemented in later versions of the system. Here, automatic speech recognition and natural language understanding were performed by the wizard, since the performances of these components are considered the bottlenecks of most SDSs. Typically, the goal of a WoZ experiment is to observe the use and effectiveness of a user interface by studying participants, rather than to measure the quality of the entire system.

3.2 Wizard tool for comparing presentation strategies

We considered the flight booking domain to be very suitable for our research due to the many options and alternatives which have to be presented to users, and thus we made use of a modified version of the existing FLIGHTS spoken dialogue system (Moore et al., 2004, introduced in Section 2.5.2). Although the actual dialogue system was not suitable because it was not robust enough to carry out experiments with a large number of participants, we implemented a database-driven Web interface which automatically generated system responses based on either the summarize and refine (SR) or the user-model based summarize and refine (UMSR) strategy to presenting information.

The wizard sat in a separate room and performed speech recognition and natural language understanding. The wizard used drop-down menus to perform stepwise queries upon request from the participants until the user found a satisfactory flight and booked it. Technically, this was done by an HTML-based interface which is connected to an SQL database containing actual flight information as provided by airlines. JavaScript and PHP were used in order to dynamically change the content of the pull-down menus according to the associated database entries and to generate text strings based on the two presentation strategies.

Figure 3.1 shows an example summary generated using the UMSR approach. To generate the presented summary, the wizard selects the relevant flight route (“San Francisco to Prague”, in this case) and enters the preferred arrival time information (in this

Select Flight Filtering Criteria

From - To: SFO (San Francisco) - PRG (Prague) ▼

Approximate Arrival Time: 14:00 ▼

Fare Class: Overview ▼

Search direct flights only? Direct Flights Only | Layovers Okay

Airline: Any Airline ▼

Approximate Departure Time: Any Time ▼

Price Range: Any Price Range ▼

I found one KLM flight with availability in business class that will get you there on time.

It arrives at 2:20 p.m., with a total travel time of just 14 hours and requires a connection in Amsterdam.

All other flights take longer.

Would you like to book the flight?

Figure 3.1: Web-based wizard user interface

example “2 pm”), both provided by the user. The tool then checks the database entries fitting the preferences and constraints and returns a text string with information about available flights. There are four different city pairs to choose from and approximately 25 flights per origin-destination combination. The following attributes can be found in the database for each flight: flight number (a unique identifier), airline name, the flight’s fare class, price, departure time, arrival time, layover airport (if any), layover time, and total travel time. If the participants asked for information the wizard was not able to give them using only the information contained in the automatically generated flight summary, a second window could be opened containing information about all remaining flights in the database view. The generated textual information provided by the Web interface was copied-and-pasted to Speechify™, a text-to-speech application

provided by Nuance Communications, Inc. All participants heard a synthetic voice of their own gender.

3.3 Experiment participants and setup

A total of 34 participants, mostly students of the University of Edinburgh, were paid to participate in the experiment. The average age of the 17 female and 17 male participants was 24 years. All participants were naive to the purpose of the experiment. The experiment was conducted in rooms of the University of Edinburgh. Participants sat in front of a desk equipped with a laptop computer, two microphones, and small speakers. On the wall opposite the one the participants were facing sat the wizard, hidden behind a visual protection screen preventing the participants from seeing or hearing the wizard during the experiment. The wizard's laptop computer was connected to the speakers and the microphones on the participant's desk via long cables running on the floor along the walls of the room in order to not attract attention.

3.4 Experimental procedure

Each participant was directly led to a chair in front of a table facing a wall. Then, they were asked to read the instructions on the laptop computer's screen explaining that they would be booking four flights with a spoken dialogue system. In order to enable reliable and rigorous comparisons, all participants were briefed to act as a business traveler for the flight booking task. In descending order of importance, the business traveler 1) prefers flying *business class*, 2) is concerned about *arrival time*, *travel time*, and *number of stops*, and 3) wants to fly on *KLM* if possible. In addition, the participants received detailed instructions regarding the two flights to be booked in the first part of the experiment mentioning the reasons for flying to the destination for the business traveler. To make the booking process more realistic, the four routes (i.e., pairs

of cities) were carefully chosen in order to guarantee that each participant experienced four different scenarios: 1. no KLM flight was available, 2. one KLM flight matched all the criteria, 3. one KLM flight in business class was available but required a connection, and 4. one KLM flight was found but it was in economy class. The order in which the four flights were booked was randomized to counter-balance possible order effects. The order of the information presentation strategy used was rotated as well. The participants booked two flights in the first part and two flights in the second part of the experiment. Half of the participants obtained flight information presented from the system adopting the SR approach; the other half received search results presented with the UMSR approach. The opposite approach was used in the second part of the experiment.

The experimental phase in this study consisted of two major steps. In Step 1, the participant was informed that she would interact with a “flight information system” to book a total of four flights. She was requested to pretend that she was “a business traveler” and then learned about the details of the persona she was to adopt. At the same time, she received instructions on booking the first two flights, including a short story explaining the business traveler’s motivation to travel to the specific destination at the specified time.

In the second step, the wizard started the conversation with the first system utterance: “This is the flight information system. I’m now connected to the network. Would you like to book a flight?” A conversation began as soon as the participant responded to this prompt. The wizard performed database queries and converted textual output into synthetic speech. After confirming the booking of the second flight, the participant received a questionnaire containing the evaluation questions that were used in (Demberg and Moore, 2006), repeated here for convenience:

- Q_1 : Did the system give the information in a way that was easy to understand?,
- Q_2 : Did the system give you a good overview of the available options?,

- *Q*₃: Do you think there may be flights that are better options for the user that the system did not tell her about?, and
- *Q*₄: How quickly did the system allow the user to find the optimal flight?.

Next, participants received instructions on booking two more flights. However, this time they received system utterances based on a different presentation method, i.e., participants receiving SR-based presentations for the first two flights received UMSR-based presentations for the next two flights and vice versa. After completing the last of the four flights, the participant again received a questionnaire to provide judgments on the four criteria introduced above. Then, the participant was debriefed, paid, thanked, and discharged.

3.5 Results

Dialogues were recorded and analyzed. Data captured by the questionnaires were tabulated and analyzed in SPSS. For the questionnaire data, seven-point category scales were used to allow for more fine-grained ratings in comparison with the previous experiment (Demberg and Moore, 2006)(the questionnaire can be found in the appendix of this thesis).

3.5.1 Dialogue efficiency and task success

Overall, there was a highly significant difference in the number of dialogue turns each participant required for booking a flight when the system adopted the SR approach in comparison to the system adopting the UMSR approach to information presentation (as shown in Table 3.1). Participants using UMSR took significantly fewer turns than when using the SR-based system.

In addition, there was a highly significant difference in the average dialogue duration between bookings made with the UMSR system vs. SR system. When the system

Table 3.1: Average number of turns per booking, dialogue duration (for booking two flights, system time plus user time) with SR and UMSR and how often the "best" flight was chosen. Significance levels: $p < .05$, indicated with "*", $p < .01$, indicated with "**", $p < .001$, indicated with "***".

	SR system	UMSR system
Dialogue turns***	14.53	10.53
Dialogue duration***	391.65	252.55
Best flights*	50/68 (73.53%)	62/68 (91.18%)

used presentations based on the UMSR approach, participants were able to complete their task in less time.

We also counted how often the flight "best" matching the business traveler's profile was chosen, in order to test the hypothesis that participants would be more likely to select the best flight when the UMSR approach was used. The results are presented in Table 3.1 and show that there is again a significant difference. Potentially, 68 "best" flights could be booked with each system. However, with presentations based on the SR approach only 50 "best" flights were booked in comparison to 62 with presentations based on UMSR.

Overall, the average flight booking dialogue with a system giving recommendations based on UMSR took considerably less time and required fewer dialogue turns. In addition, users selected the best available flight significantly more frequently. Thus, in terms of both dialogue efficiency and task success, the UMSR approach outperformed the SR approach in this interactive experiment.

3.5.2 User satisfaction ratings

In the questionnaire data, presented in Table 3.2, we found a general preference for UMSR-based recommendations on all four evaluation criteria. However, only differ-

ences between answers to the first (“Did the system give the information in a way that was easy to understand?”), and last question (“How quickly did the system allow the user to find the optimal flight?”) were statistically significant.

Table 3.2: Answers to the 4 user satisfaction/evaluation questions (on a scale from 1-7), $p < .05$ indicated with “*”).

	SR	UMSR
Q_1 - Understandability*	5.27	5.79
Q_2 - Overview of options	4.85	5.18
Q_3 - Relevance of options	3.76	4.00
Q_4 - Efficiency**	4.86	5.63

We also investigated whether there was a correlation between overall user satisfaction (the mean of the ratings on the four user satisfaction scales per participant) and dialogue duration. We hypothesized that there should be a close correlation meaning that shorter dialogue duration translates in higher user satisfaction. We found that there is a very weak correlation which, however, is just not significant, $r = 0.242$ (34), p (two-tailed) = .056. We did not find a correlation between dialogue duration and one of the user satisfaction questions either.

Based on the audio recordings of the experiment we believe that the exceptionally longer dialogue duration of participants booking flights with SR is mainly due to them having to explore many dialogue paths requiring considerably more dialogue turns than participants booking flights with UMSR. The switching and backtracking not only takes time, but also requires concentration and consumes cognitive load to recall which of the heard options is most suitable.

3.6 Discussion

The results of the previously conducted experiments, asking participants to evaluate presentations based on SR and UMSR presented as dialogue transcripts (Demberg and Moore, 2006) or as sound files where the participants “overhear” the dialogues (Moore, personal communication) demonstrated a clear preference for UMSR. In the experiment described here we again found a general preference for presentations based on the UMSR approach to information presentation. However, in this experiment where users actively interacted with the system, we also found that the UMSR approach outperforms the SR approach in terms of task success and dialogue efficiency.

This brings us back to the open questions formulated in Section 2.10. We have demonstrated that presentations based on the UMSR approach were rated higher in an experiment where users *interact* with a dialogue system. In Section 1.1 we highlighted the role of cognitive load when evaluating information presentation strategies: First, because although there have been many claims about the cognitive load that different information presentation strategies place on users (Walker et al., 2004; Moore et al., 2004; Kruijff-Korbayova et al., 2006), there has been no systematic empirical study of these claims, especially in terms of assessing the cognitive load that different presentation strategies place on users. Second, because SDS are often intended for situations where the user’s hands and eyes are busy performing another task and the role of cognitive load increases in relevance if another task, e.g., walking, driving, or the manual manipulation of the surroundings, which also requires the users’ attention and cognitive processing, is performed simultaneously.

To evaluate the amount of cognitive load that information presentation strategies place on users, a familiarization with the concept of cognitive load is essential. Cognitive load is a term that refers to the load on working memory during problem solving, thinking and reasoning (including perception, memory, language, etc). The following chapter attempts to clarify the concept of and describes methods to measure cognitive load.

Chapter 4

Related Work - Measuring Cognitive Load

Human beings do not have an infinite capacity to process information (Reed, 1996). Approaches to information presentation developed by computational linguists typically do not take cognitive load into account despite the fact that every task uses some resource and despite the fact that dialogue systems are often designed to be used while users are performing other concurrent tasks. We assume that cognitive load is the amount of mental resource needed to perform a given task (Cohen et al., 2004). Hence, if a second task is conducted during the performance of the main task and the demands of the two tasks exceed the available resource, performance of at least one task will suffer. Of course, this should be avoided in situations where task performance is vitally important, e.g., when driving a car.

Here, we examine the cognitive load that different approaches to information presentation impose on their hearers. Unlike other approaches aiming for “naturalness” in spoken dialogue systems (Stent, 2001, for example), we will mainly focus on the imposed cognitive load of presentation methods. More specifically, we aim at finding out whether there is a difference regarding the comprehension complexity of presentation methods that point out trade-offs explicitly (which leads to potentially more complex

sentences) as opposed to the presentations used in the MATCH (Walker et al., 2004) or SR systems (Polifroni et al., 2003) consisting of simple repetitive sentence structures. Therefore, we carry out a series of experiments to gather data which may help us to better understand how humans process information in contexts/situations where not all of their cognitive resources are available due to another demanding task.

Although there are several other scenarios imaginable where users have to deal with another task while simultaneously conversing with a dialogue system, we consider the in-car scenario as particularly interesting because driving is a very common yet complex task and automotive manufacturers are increasingly interested in putting voice services (e.g., navigation services, voice controlled MP3 players, air-conditioning) into their automobiles. Driving involves the continuous multitasking of different subprocesses utilizing the driver's cognition, perception, and motor movements. Also, it offers multiple continuous performance measures, if conducted with a sophisticated driving simulator. The measures include proximal aspects of driving skill, such as steering, braking, and moving the accelerator, as well as the distal consequences of these activities, including maintenance of lane position, following distance, and acceleration. This broad range of required skills makes driving very suitable for examining how humans execute dialogue tasks while simultaneously dealing with another task.

4.1 Basic concepts of Cognitive Load

In order to understand cognitive (work)load, some basic concepts need to be introduced. Here, I partly follow the PhD thesis by De Waard (1996). Although he predominantly aims at defining workload and related terms in relation to driving as the main task, his remarks are valid for other cognitively demanding situations as well. De Waard argues that workload is the specification of the amount of information processing capacity that is used for task performance. This is consistent with O'Donnell and Eggemeier (1986) who define workload as that portion of the operator's limited

capacity that is actually required to perform a particular task. Workload measurement is therefore the specification of the amount of capacity used.

Wickens (1992) and Norman and Bobrow (1975) define *capacity* as the maximum or upper limit of processing capability, while *resources* represent the mental effort supplied to improve processing efficiency. They also distinguish capacity as the upper limit of capability and resources as the amount of processing facilities allocated. The relation between resource allocation and task performance is supposed to be linear, until the moment when all resources are invested. From that point on, no more resources can be invested and task performance will remain stable.

There are three competing theoretical views regarding an operator's capability to perform two different tasks simultaneously. According to "single channel theory" (e.g., Broadbent, 1958; Kahnemann, 1973), an operator can, in somewhat simplified terms, perform only one task at a time. This assumption is supported by experiments showing that although attention can be distributed among several inputs, conscious and focused attention is solely allocated to a single task (Pashler and Johnston, 1998). Therefore, the simultaneous handling of information always leads to a decrease in task performance.

In contrast, according to "multiple resources theory" (e.g., Treisman and Davies, 1973; Wickens, 1984), the human cognitive system has separate resources or channels for different kinds of tasks. Therefore, following this theory, a human operator can consequently perform two tasks simultaneously, provided that these tasks use different resources of the operator, such as vision and hearing. Additionally, central resources are supposed, which are required for the performance of almost all tasks. An overlap in resource requirement, e.g., the performance of two auditory tasks, soon requires full auditory capacity use. In that case, performance on both tasks will be affected. In general, tasks that require different resources, e.g., a visual task combined with an auditory task, will not directly interfere with each other and performance of either task

may remain unaffected, provided there is no performance decrement caused by central resource use.

The concept of multiple resources is connected to four dimensions in Wickens (2002)'s theory. The first dimension is the processing *stage*, i.e., perception (including encoding), central, and response processing. The second dimension is *modality* of input and response. The auditory, visual and tactile modalities draw upon different resources and cross-modal time-sharing can be better performed than intramodal timesharing. Listening to someone and watching something at the same time associate better than listening to two things at the same time. The third dimension is the processing *code*. The processing code can be either visual or spatial. Finally, the fourth dimension is the processing *channel* (focal vs. ambient).

According to "connectionist control architecture" (e.g., Schneider and Detweiler, 1988), an operator can function both according to "single channel theory" and "multiple resource theory". The operator's experience of the tasks involved, separately and combined, will determine if the two tasks interfere with each other.

Kantowitz (1987) has proposed a differentiation between complexity and difficulty as a property of, respectively, the task in isolation versus the interaction between task and individual. He argues that workload depends upon the individual, and owing to the interaction between operator and task structure, the same task demands do not result in an equal level of workload for all individuals. Directly related to demand is (task) *complexity*. Complexity increases with an increase in the number of stages of processing required to perform a task. Task demand and complexity are mainly external, but both depend upon (subjective) goals set for task performance. *Difficulty* of a task is related to the processing effort (amount of resources) that is required by the individual for task performance, and is dependent upon context, state, capacity and strategy or policy of allocation of resources.

4.2 Relation between task demand and task performance according to Meister's model

A relation between task demand and task performance has been described by Meister (1976), who defined three regions, region A, B and C. Region A is described as low operator workload with high performance. An increase in demands does not lead to performance decrements. In region B the level of performance declines with increased task demands. So, region B is the region where performance decreases with increases in demand, and increases in workload. In region C extreme levels of load have diminished performance to a minimum level, and performance remains at this minimum level with further increases in demand.

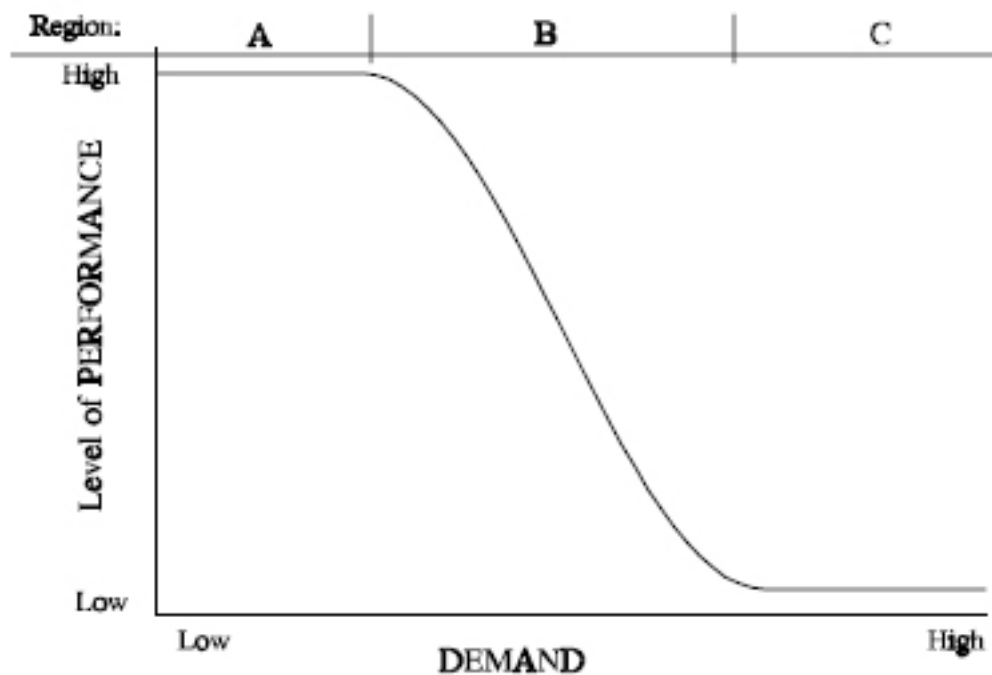


Figure 4.1: Hypothetical relationship between task demand and performance based on Meister, 1976

According to this model, a primary-task workload measure, i.e., a measure of performance, will only be sensitive to variations in levels of workload in region B. In region A performance remains stable and is independent of variations in demand, while

in region C performance will remain at a minimum level, independent of demand. Other measures, e.g., self-report measures of workload, may be sensitive in region B and may clearly reveal overload in the C-region, while they need not be sensitive in region A. While extreme levels of load resulting in overload can be situated in the C-region, it is not clear where the domain of underload is.

4.3 Considerations regarding the Assessment of Cognitive Load

According to O'Donnell and Eggemeier (1986) there are three distinct types of workload measurements: subjective (i.e., self-report) measures, performance measures and physiological measures. Performance measures can be split into three categories again: primary-task performance measures, secondary-task performance measures and reference tasks.

Of the above mentioned measurements, self-report measures have always been particularly popular because they are relatively easy to accomplish, inexpensive and have proven to deliver relatively reliable results. In addition, maybe no one is able to provide a more accurate judgment with respect to experienced mental workload than the person concerned. A frequently used rating scale is the NASA Task Load Index (NASA TLX, Hart and Staveland, 1988), which requires ratings on several subscales (experienced mental demand, frustration etc.) to be made. The summarized ratings are then used to obtain an overall workload assessment.

In contrast to self-report measures, performance measures on the primary task objectively assess the performance of each individual under the same conditions, for instance, the number of errors made, the speed of performance or the reaction time. Outside the laboratory, primary-task performance is, by its nature, very task specific. There is not one prevalent primary-task measure, although all primary-task measures are speed or accuracy measures.

O'Donnell and Eggemeier (1986) argue that primary-task performance is a measure of the overall effectiveness of man-machine interaction. However, there are some limitations to this statement. Primary-task performance diminishes outside the A region according to Meister's model, while a constant performance in the A region does not necessarily reflect low operator workload. No performance differences between two operators can be determined, although one can be "at the limit of his capability", while the other is capable of performing an additional task, without any change in primary-task performance level. Therefore, it seems to be necessary to combine primary-task performance and other workload measures in order to draw valid conclusions about man-machine interaction.

4.4 Measuring Cognitive Load with the dual-task methodology

In general, dual-task studies provide a suitable basis for examining the effects of two simultaneously conducted tasks, such as communicating with a device (e.g., mobile phone, spoken dialogue system) and engaging in an activity (e.g., walking, driving, typing), on one another.

Two paradigms can generally be applied to dual-task performance according to O'Donnell and Eggemeier (1986) and De Waard (1996). First, the "Loading Task Paradigm", where the instruction is to maintain secondary task performance even if decrements in primary-task performance occur, while at the same time the workload shifts from region A to region B (according to Meister's model), so that primary-task performance measures can be used as indicators of workload.

Second, in the "Subsidiary Task Paradigm" participants are instructed to maintain the primary task. Hence, secondary task performance varies with difficulty and indicates "spare capacity", provided the secondary task is sufficiently demanding. Brown and Poulton (1961) state that spare capacity is a concept that is used frequently in

dual-task performance, assuming a total undifferentiated capacity that is available to perform all tasks. When single task performance is unaffected, the unused capacity is called spare capacity, and is theoretically available for a secondary task.

The choice of secondary task is difficult in tasks approaching everyday performance. Car driving, for example, is to a large extent automated, and has a visual component. The usefulness of a secondary auditory digit-addition task, for example, is therefore not completely clear. It is possible that performance on the latter task reflects central resource use. However, the extent to which performance of the primary task makes use of central resources is not clear in advance. The use of secondary tasks in applied environments is more complex than in laboratory experiments, and for this reason caution is required. However, precisely because it is a common task whose difficulty can be easily influenced and because it offers multiple performance measures, driving appears to be a suitable secondary task for studying the impact of cognitive load on dialogue performance in dual-task experiments.

Finally, there are workload measures derived from the user's physiology. Different physiological measures have been found to be differentially sensitive to either global arousal or activation level (e.g., pupil diameter, Kahnemann, 1973), or to be sensitive to specific stages in information processing (e.g., the evoked cortical brain potential, Meijman and O' Hanlon, 1984). The advantage of physiological responses is that they do not require an overt response by the operator, and most cognitive tasks do not require overt behavior.

4.5 Studies examining the interplay between driving and speech interfaces

Several previous studies, examining the role of verbal tasks in the in-car domain, have shown that, for example

- manual manipulation of equipment (e.g., dialing the phone or adjusting the radio (Briem and Hedman, 1995) in a car,
- cell-phone conversations (Strayer and Johnston, 2001) while driving,
- the usage of speech-based email systems (Young et al., 2003) during driving,
- and too much visual information (Dybkjaer and Bernsen, 2001)

may distract a driver. Other studies (partly collected in Kubose et al., 2006) examined real and simulated driving performance especially regarding concurrently performed verbal tasks and showed that they result in

- more glances away from the road (Jenness et al., 2002),
- increased reaction time to breaking events (Irwin et al., 2000),
- increased subjective mental workload (Haigney et al., 2000),
- decreased detection of changes in the visual environment (McCarley et al., 2004),
- and a smaller window of gaze, with glances more concentrated towards the center of the field of vision and reduced glances to side mirrors and speedometer (Recarte and Nunes, 2000).

Cohen et al. (2004) summarize some general design principles concerning the development of speech applications. They emphasize that systems using only auditory interfaces particularly challenge human memory and attention because they present information serially and non-persistently. They present guidelines for minimizing the cognitive load to avoid a design which, for instance, requires the user to hold too many items in short-term memory. As possible solutions to overcome the problem of cognitive load in speech interfaces they advise the establishment of a small number of universal commands (easing the participant's memory access to and processing of those fundamental commands), the consideration of consistency throughout the system (e.g., regarding dialogue strategies and grammar coverage), and the application

of context setting (meaning, for example, the usage of metaphors to navigate the users through the system, where a metaphor is a familiar object or schema that is used to help facilitate understanding in another domain). Furthermore, they suggest that designers of voice interfaces to be used while driving should consider and accommodate for situations where the drivers' attention is completely allocated by their surroundings. In such situations the users need to be in control over the pacing of the interaction and must be able to stop the interaction completely if necessary.

Geutner et al. (2002) describe a Wizard-of-Oz experiment in the context of the VICO¹ project, which aimed to develop an intelligent conversational agent enabling ubiquitous natural interaction between humans and digital devices and services. The experimental study was carried out to examine which utterances a car driver would use in order to solve various given tasks within a driving simulator environment. The responses of their VICO system were predefined or slightly varied (on-the-fly generated) sentences, which were then synthesized by a text-to-speech system. The results of this experiment indicate that the presence of the speech-based co-driver was generally found to be very pleasant by the participants. Furthermore, the results with the human wizards revealed that natural language interaction using "human-like" conversation was clearly preferred over command-and-control input.

A comparable Wizard-of-Oz experiment was conducted by Cheng et al. (2004). It was carried out in order to gather human-computer dialogs in particularly stressful conditions. They asked their participants to drive a simulated car and, for instance, to simultaneously collaborate with the in-car speech based computer interface to solve several tasks. The focus here was on speech recognition rather than on dialogue design, although this was also considered. In their data collection they concentrated on the tasks of navigation and operating in-car Mp3 players. In the pilot sessions, they found that drivers tended to use disfluent and distracted speech when focusing on driving.

¹Virtual Intelligent Co-Driver

In studies to assess the suitability of verbal versus graphical feedback for in-car dialogue systems, Dybkjaer and Bernsen (2001) found (with the above-mentioned VICO system) that the need for text output is fairly limited in this domain. Some of the participants in their study, which compared textual output displayed on a built-in display with oral output, mentioned that they would prefer not to have to use the display at all while driving, whereas others simply did not pay attention to what was being displayed. This study demonstrated that drivers generally prefer oral information in comparison with text-based presentations.

Finally, Kruijff-Korbayova et al. (2005) and Becker et al. (2006) recently presented an experimental setup for collecting data via a Wizard-of-Oz environment with the help of a driving simulator for the EU project TALK². In their experiments they also gathered interaction data for a music player. Wizards were asked to choose between different presentation modalities (either speech-only or multimodal) in order to select the most appropriate one for the user. Their evaluation showed that participants reported too much information was often displayed, which they felt sometimes distracted them. Moreover, participants mentioned that they would prefer (more) oral rather than textual feedback, especially while driving. In Becker et al. (2006) a revised version of the multimodal in-car spoken dialogue system SAMMIE is presented. SAMMIE supports speech-centered multimodal access to an MP3-player application including search and browsing, as well as composition and modification of playlists. It supports mixed-initiative interaction, with particular emphasis on multimodal turn-planning and natural language generation to produce output adapted to the context, including the driver's attentional state with respect to the primary driving task. However, to date the developers have not performed an adequate experiment or evaluation to present results regarding the cognitive load or attentional state. A formal usability evaluation of the system's baseline version in a simulated environment was carried out with overall positive results. Again, although users could choose between different

²Tools for Ambient Linguistic Knowledge

modalities, about 70% of the subjects chose speech when they had the choice, and less than 10% changed the modality during the task.

The results of the experiments by (Dybkjaer and Bernsen, 2001; Kruijff-Korbayova et al., 2005; Becker et al., 2006) reveal that there is strong evidence that users prefer speech as the main interaction mode within the in-car environment in comparison to graphical interaction devices. Certainly, one of the reasons is that driving is mainly a visual task and that further visual attention is required for the display which then distracts drivers by using the already allocated visual processing resources. This is part of the motivation for studying spoken dialogue systems for in-car applications.

4.6 Methods for Measuring Distraction

The above-mentioned studies have not or at least not mainly considered the aspects of cognitive load caused by the different information presentation techniques in their evaluation. Instead, they focused on the examination of questions such as: Which modality do wizards think users prefer under conditions of heavy stress? (Kruijff-Korbayova et al., 2005), or Do users prefer textual or oral information presentation? (Dybkjaer and Bernsen, 2001). In contrast, our focus is studying the effect of information presentation strategies on cognitive load. In particular, we examine information presentation strategies in situations where the conversation occurs in connection with another demanding task. To distinguish the different presentation methods with regard to their cognitive load, we have to be able to specify cognitive load. In the following, some of the methods used to measure the distraction affecting drivers are introduced.

There are various measurement techniques and measurements concerning driver distraction. In Young et al. (2003), the following techniques are introduced:

On-road and test track studies These studies are very realistic and compare the driving performance while drivers interact with the in-car technologies against a baseline measure, usually driving without interacting with the device. Major

drawbacks of this method are that it is time consuming and highly expensive. In addition, it is dangerous for the driver and thus seldom used.

Eye glance monitoring studies The eye glance technique measures visual behavior by recording the frequency and duration of eye glances at particular objects in the driver's visual field (Farber and Scott, 2000).

The visual occlusion method This method measures the visual demand of a device. The method makes the assumption that drivers only need to observe the road part of the time and the rest of the time is available for other purposes, such as interacting with in-car devices. To test the driver, his vision is partially or fully occluded through the use of a shield/visor or another similar device that opens and shuts at various time intervals. The aim is to simulate an on-road situation where the driver is interacting with a device while driving. The phase where the vision is occluded simulates the time he is looking on the road, while the open phase represents the time he is looking at the in-car device. If a task can be carried out using only short, periodic glances it is classified as "chunkable" and therefore suitable for the in-car use.

The peripheral detection task This method was invented by van Winsum et al. (1999) to measure both driver's mental workload and visual distraction. Participants are asked to perform a series of tasks while detecting and responding to targets appearing in the periphery. As drivers become more distracted by the primary task, they respond slower and fail to detect more targets. Winsum et al. found that peripheral detection task (PDT) performance provides a suitable measure of how distracting the primary task is; but it is also applied to measure the level of distraction caused by in-car devices. As Martens and van Winsum (1999) found in their detailed study, PDT is a valid and sensitive method for measuring increases in driver workload and driver distraction.

The 15-second Rule This standard, established by the Society of Automotive Engineers (SAE), determines a design limit for the total time required to feed information into navigation systems while the vehicle is in motion. Although initially created for the evaluation of navigation systems, it can also be used to evaluate the distraction caused by other in-car devices. To be precise it is stated: “All navigation functions that are accessible by the driver while the vehicle is in motion, shall have a statistically measured total task time of less than 15 seconds” (Farber and Scott, 2000). If, for example, a car driver can complete a task within 15 seconds or less in a stationary vehicle, then this task is also suitable for in-car use and can be made available to drivers while the vehicle is moving as well. However, there are some concerns because an evaluation of the 15 second rule by Tijerina et al. (2000) showed that the rule, even though effective in identifying the most distracting tasks, does not work better than, for example, a 30-second rule. In addition, they stated that the rule would firstly not take into account the “chunking” of the tasks and secondly it fails to address the issues of speed maintenance and object detection. Finally they noted that there are no baselines against which to compare driving performance while completing a task.

Driving simulator studies When high-fidelity driving simulators are used, they offer a relatively realistic driving environment without the costs and risks involved in the use of on-road and test track studies. Other advantages are the safe use of different in-car devices while driving, and the ease of examining various driving performance measures simultaneously. Additionally, the experimenter can easily adjust the difficulty of the driving task in some advanced simulators to observe the changing impact of the increase in difficulty on the performance of the two tasks respectively. However, one major drawback of driving simulators is the awareness of the test participants that possible driving errors have almost no consequences (Goodman et al., 1997) and therefore the amount of cognitive resources they devote to performing concurrent tasks while using the simula-

tor may differ significantly from their behavior in real cars where a moment of inattention may cause serious accidents. Still, driving simulator studies are very suitable for observing the impact of in-car devices on driving performance.

Dual-task studies Young et al. (2003) stated that Dual-task studies assess the effects of performing one task on the performance of another concurrent task. In the context of driver distraction, these studies normally analyze the effects of using an in-car device or engaging in an activity on driving performance.

Furthermore, Young and colleagues argue that their review of the literature suggests that using a range of distraction measurement techniques, rather than a single technique, would be appropriate in evaluating Human-Machine Interface (HMI) design concerning in-car systems. Which technique to prefer in a specific situation depends, of course, on the particular aspect of HMI to be assessed and on the form of distraction that affects the driver by that aspect of the interface.

According to the US National Highway Traffic Safety Administration (NHTSA, 2000) there are four distinct types of driver distraction: visual, auditory, physical and cognitive distraction. In our studies, users interact with a spoken dialogue system, and thus we are dealing both with auditory distraction, occurring when drivers momentarily or continually focus their attention on sounds or auditory signals rather than on the road environment, as well as with cognitive distraction, which includes any thoughts that absorb the driver's attention to the point where they are unable to navigate through the road network safely and their reaction time is reduced.

4.7 Basic concepts of Human memory

In order to understand how humans handle newly perceived information it is necessary to understand basic concepts of human memory. In general, memory is the ability of an organism to store, retain, and subsequently retrieve information. There are several ways of classifying memories, based on duration, nature and retrieval of information.

From an information processing perspective there are three main stages in the formation and retrieval of memory:

- Encoding (processing and combining of received information)
- Storage (creation of a permanent record of the encoded information)
- Retrieval/Recall (calling back the stored information in response to some cue for use in some process or activity)

A basic and generally accepted classification of memory is based on the duration of memory retention, and identifies three distinct types of memory: *sensory memory*, *short-term memory*, and *long-term memory*.

Sensory memory corresponds approximately to the initial moment that an item is perceived. Some of the information in the sensory area proceeds to the sensory store, which is referred to as short-term memory. Sensory memory is characterized by the duration of memory retention from milliseconds to seconds and short-term memory from seconds to minutes. Information in short-term memory can be held there indefinitely as long as it is rehearsed, and the typical cause for its loss is that it is displaced by the presence of other, new information that has been attended to. Generally, however, short-term memory is considered to be a temporary resting place and information is held there for approximately 30 seconds to two minutes.

Just as the sensory and short-term memory systems are associated with the process of encoding or registering information in memory, the long-term memory system is associated with the processes of storage and retrieval of information from memory. Long-term memory storage is considered to be relatively permanent.

4.8 Working Memory and Priming

Long-term memory consists of two systems - declarative and nondeclarative. Declarative memory can be further delineated into the episodic and semantic systems. The

nondeclarative system includes procedural learning and priming. Priming (proposed by Squire, 1992) is a subsystem of nondeclarative memory which, according to Baddeley (1996), refers to the phenomenon that once an object has been perceived or processed, it can be more easily perceived or processed the next time it is encountered. Furthermore, such nonconscious effects of prior experience seem to be an important component of functioning in everyday life, influencing, for example, the particular ideas or words that come to mind and the effects of prior practice on performance.

This particular property makes priming interesting for our research concerning spoken dialogue systems. According to Bock (1986), structural priming is a tendency to reuse previously heard or produced sentence structures, phenomena called comprehension-production and production-production priming, respectively. But structural priming can also be long-lasting, suggesting that it is a form of implicit learning that is shared between production and comprehension. Since we aim to develop spoken messages that are as easy to remember as possible, we should try to take advantage of these implicit learning mechanisms. In general, priming, in this case the reuse of sentence structures, could possibly ease the users' comprehension efforts since the processor can focus on "new items" almost exclusively.

Additionally, the term *working memory*, introduced as a concept by Baddeley and Hitch (1974), is used to refer to the short-term store needed for certain mental tasks - it is not a synonym for short-term memory, since it is defined not in terms of duration, but rather in terms of purpose. Some theories consider working memory to be the combination of short-term memory and some attentional control. For instance, when we are asked to mentally multiply two figures, we have to perform a series of simple calculations (additions and multiplications) to arrive at the final answer. The ability to store the information regarding the instructions and intermediate results is what is referred to as working memory.

There are two subsystems of working memory: (1) verbal working and (2) visual working memory. In addition, working memory contains a main controller or central

executive that interprets information we have just been presented with and integrates it with information already stored in long-term memory (Baddeley, 1996). It is said that we can think of working memory as remembering what we are doing while doing it.

When designing speech interfaces, various problems concerned with human memory have to be tackled. For example, it was demonstrated by Miller (1956) that a human being is not able to recall more than 7 ± 2 options in a verbal menu due to limitations of their short-term memory. Longer menus therefore require a corresponding higher concentration and could lead to overload. However, even remembering this relatively small number of items was only possible in a laboratory setting in the underlying experiments where no one and nothing distracted the participants. In situations where another demanding task is being performed by users concurrently (in the “real world”) they might not even achieve this number.

In order to retain more than 7 ± 2 single numbers in working memory, the numbers must be chunked. That is, they must be grouped together so that several single numbers are organized into one “conceptual” chunk. For example, the single numbers seven, one and four could be chunked into one number, 714. Thus, if we were presented with a series of numbers to remember, we could likely recall more if we “chunked” them into groups of twos or threes. In order to keep information in short-term memory, we must continue to actively process it (Broadbent, 1975).

Particularly interesting against the background of our expectation that more and more older people will become users of spoken dialogue systems are the findings of a study by Zajicek and Morrissey (2001). They found that older adults could retain fewer options in memory than younger adults when examining the optimal number of function key options that can be presented verbally to older people. In contrast, in a recent experimental study examining the number of options presented in an Interactive Voice Response system, Pineau et al. (2003) found no significant advantage of presenting fewer options. Obviously, general design guidelines for complex systems, such as

dialogue systems, should ideally first be evaluated and experimentally validated in the context of real conversations.

Chapter 5

Evaluating Task Success, User Perceptions, and Cognitive Load

Now that we have shown that the general findings of the first UMSR evaluation can be replicated in an interactive scenario where users are actually engaged in a conversation with the system and that the USMR approach outperforms the SR approach in terms of task success and dialogue efficiency in such an interactive scenario (in Chapter 3), we focus on the question of whether the UMSR approach to information presentation also shows an advantage when participants are performing a demanding secondary task at the same time. That is, we want to show that UMSR leads to better dialogue performance and higher task success even when conducted with a simultaneously performed second task. Therefore, we experimentally evaluated and validated **Hypothesis 1b** (see Section 2.11) claiming that users would also prefer and perform better with UMSR in comparison to SR in terms of dialogue task efficiency and task success if they perform a (demanding) secondary task at the same time.

We were particularly interested in studying the interplay between a range of information presentation strategies and users who are confronted with the varying demand of an additional task. Therefore, we carried out experiments with users driving on a simulated driving course while conversing with a dialogue system. We chose the (sim-

ulated) car driving scenario because driving is a demanding task and soaks up cognitive resources allowing us to measure the effect of information presentation strategies on cognitive load.

However, car driving is generally a dynamic control activity in a continuously changing environment. Using a driving simulator to measure cognitive load has many advantages (e.g., see chapter 4), because driving is a complex task with processes at at least three different hierarchical levels (the *strategic* level, the *maneuvering* level, and the *control* level, according to De Waard (1996)), and interruptions and changes in task-related workload can happen at any time, either in the middle of a task unit, or at unit boundaries. Booking flights, too, has different hierarchical levels and it is not entirely clear how interruptions of the two tasks on different levels will affect one another.

Presenting information to drivers requires the consideration of the distractive factor imposed by communicating with the SDS. This is particularly true when road conditions are unfavorable and require a large proportion of cognitive resources. Based on the rationale behind the UMSR approach, one would expect that, compared to an SR-based spoken dialogue system, a UMSR-based system should A) be more efficient, B) cause fewer harmful distractions to drivers, and C) lead to pleasant user experience, especially under difficult driving conditions. To test these hypotheses (that were developed for this particular experiment and are not to be confused with the hypotheses developed in Section 2.11), the following laboratory experiment was designed and conducted.

For this experiment, aiming to find out about users' behavior in cognitively demanding situations, we conducted a new Wizard-of-Oz experiment, where users drove a simulated car while at the same time conversing with a dialogue system. We followed the Wizard-of-Oz approach that Geutner et al. (2002) (see Section 4.5) chose for their experiments with the VICO system. In their study, they gathered speech data by asking participants to solve tasks using a multimodal dialogue system while simul-

taneously driving in a driving simulator. However, instead of actually conversing with the dialogue system they talked with a human wizard in another room who acted as if the system responded.

We applied the “Subsidiary Task Paradigm”, explained in Section 4.4, which is, within the psychological attention and memory research, an essential part of the dual task paradigm (which is itself a part of the divided attention research).

5.1 Experimental setup

The experiments were performed using the STISIM DriveTM simulation system by SystemTech in use within the CHIME/CARSITE research lab at Stanford University. This simulation environment allows for various measurements, e.g., lane deviations and reaction times. The CARSITE lab performs research aimed at improving the safety and overall experience of driving through human-computer interfaces.

The STISIM DriveTM simulation system was installed on a desktop computer and displayed on a wall-sized back-projection screen. Participants sat in a car seat with actual instruments similar to those present in a real car’s dashboard, used an authentic driving wheel, and a gear-stick. A total of four courses with two levels of difficulty were used to vary the driving-related cognitive load affecting the participants.

For the experiments, two different routes varying considerably in difficulty were created. The participants were then asked to drive both routes with and without talking to the dialogue system. We assumed that the more challenging the course is, the more likely it is to influence the performance of the participants regarding the simultaneously conducted second task - conversing with the dialogue system.

Each course contained four sequential sections: a residential area, a small city, a country highway, and a big city. Posted speed limits ranged from 25 mph to 55 mph. It took approximately 16 minutes to finish each course when driving in accordance with posted speed limits. In order to achieve a realistic driving environment there were



Figure 5.1: STISIM DriveTM simulation system used in the dual-task experiments at Stanford University's CHIME lab

trees, mountains, residential houses, commercial buildings, stop signs, traffic lights, pedestrians, cyclists, pets, parked vehicles, and running vehicles in both directions depending on the sections.

Compared to the easy course, the difficult course had three times as many vehicles, cyclists, and pedestrians, as well as sharp curves, two foggy sections, a construction site, slopes of various degrees, and several vehicles that behaved in dangerous ways (e.g., speeding). Pilot-tests showed that the difficult course was harder to drive than the easy course in terms of effects on actual and perceived driving performance. Additionally, a short demo course was used to familiarize participants with the simulator before the start of the actual experiment. The demo course required the participants to drive for about five minutes in a residential area.

The simulator kept track of the participant's driving performance in terms of numbers of collisions, speeding tickets, traffic light and stop sign violations, and minor driving errors including centerline crossing and road edge excursion.

5.1.1 Participants

A total of 32 students from Stanford University were paid to participate in the study. All participants were licensed drivers and had previous driving experience. Students with prior exposure to the driving simulator were excluded; gender was balanced across conditions.

5.1.2 User Profile and Flight Booking

To be able to make reliable and rigorous comparisons, participants were asked to use the same business traveler's profile for the flight-booking task (the same profile that was used in the Wizard-of-Oz experiment presented in Chapter 3). Recall that the business traveler most importantly prefers flying business class. Second, she is concerned about arrival time, travel time, and number of stops. Finally, she wants to fly on KLM if possible. Each participant drove on two experimental courses and booked four different one-way flights. Prior to each round of driving, participants received detailed instructions on the two flights to be booked. The following is an excerpt of the instructions:

New York to Frankfurt: You're going from New York to Frankfurt departing on July 5. You'd like to arrive in the late morning so that you can make it to a meeting that begins at 2 pm.

To make the booking process more realistic, the four routes (i.e., pairs of arrival and departure cities of the flights) were again carefully chosen so that each participant experienced four different scenarios (the same scenarios that were used in the previous experiment, repeated here for the reader's convenience): 1. no KLM flight was available, 2. one KLM flight matched all the criteria, 3. one KLM flight was available but required a connection, and 4. one KLM flight was found but did not have business class availability. The order in which the four flights were booked was rotated to counter-balance possible order effects. The following two examples offer a side-by-side comparison of first-round presentations for this persona (see Figure 5.2):

User: I'd like to book a flight from New York to Frankfurt, please.

SR: I found 23 flights from New York to Frankfurt. There are direct flights as well as flights that require a connection. I also have information about fare classes.

UMSR: I found 6 direct business flights from New York to Frankfurt. None are on KLM. However, if you're willing to make a connection, there is a business class KLM flight arriving at 1:35 p.m., connecting in Amsterdam.

Figure 5.2: First-round presentation with both SR- and UMSR-based systems

5.2 Experimental procedure

The participants were randomly assigned to the “easy course” or the “difficult course” condition. The order of each participant’s two courses was also randomized. During the first round of experimental driving, half of the participants received flight information presented with the SR approach; the other half heard search results presented with the UMSR approach. Participants were presented information with the opposite approach during the second round of experimental driving.

Before the experimental phase, participants took a test drive on the demo course to familiarize themselves with the simulator. The experimental phase that followed consisted of three major steps. In step 1, the participant was informed that she would talk to an “in-car information system” to book flights while driving. She was instructed to assume the persona of the business traveler for the booking tasks. At the same time, she received instructions on booking the first two flights.

In step 2, the participant drove on the first experimental course. Shortly after she passed the residential area (roughly after three minutes), a short beep was played, followed by the first utterance from the system saying that “This is the in-car information system. I’m now connected to the network. Would you like to book a flight?” A conversation began as soon as the participant responded to this prompt from the wizard

sitting in a neighboring room. Via a wireless connection, the wizard monitored all audio events around the driving simulator, performed database queries, and converted textual output into synthetic speech on a laptop computer. The synthetic speech utterances were transmitted to a pair of speakers by the back-projection screen. After confirming the booking of the first flight, the wizard offered help to book the second one. The participant continued driving to finish the course after both flights were successfully booked.

In step 3, the experimenter returned to the lab and administered a questionnaire (see appendix) that asked the participant to evaluate the “in-car information system” during the interaction, and the driving course. Once the participant indicated that she was ready for the second round of driving, Steps 1 through 3 were repeated, with different flights to book, and a different method of information presentation, i.e., SR participants in Round 1 used UMSR in Round 2, and vice versa. Upon completing the second questionnaire, the participant was debriefed, paid, thanked, and discharged.

5.3 Results

Dialogues were recorded and transcribed; data captured by the driving simulator and the questionnaires was tabulated and analyzed in SPSS. Factor analyses were performed for all questionnaire items to extract reliable and meaningful indices. All indices are reliable with Cronbach’s alpha values ranging from .65 to .92. Ten-point category scales were used unless noted otherwise. The ten-point scales were meant to capture subtle variations and to avoid a middle point that often encourages “satisficing” ?. A series of SPSS repeated-measure ANOVAs were conducted, followed by post-hoc analyses when necessary.

5.3.1 Manipulation Check

The manipulation of driving course difficulty was successful. Specifically, although the average *number of collision accidents* was quite low, participants driving the difficult course had significantly more accidents than those driving the easy course (see Table 5.1. This was also true for the average *number of minor driving errors*, including center-line crossing and road edge excursions. No difference was found in terms of stop sign and traffic light violations, and number of speeding tickets. Moreover, easy-driving participants rated their courses as much *easier* than did difficult-driving participants.

Table 5.1: *Performance of easy vs. difficult-driving, $p < .001$, indicated with “***”*

	accidents	# of minor driv. errors	ratings
Difficult driving	0.82***	2.19***	5.92***
Easy driving	0.0***	0.60***	7.63***

5.3.2 Dialogue Efficiency and task success

Significant differences were observed between dialogues with the SR-based system and those with the UMSR-based system. The results are shown in Table 5.2. In general, participants took fewer dialogue turns when the system adopted the UMSR approach than when it utilized SR. The average duration of dialogues (in seconds) was also shorter when the system adopted the UMSR than when it used the SR approach. These results support Hypothesis A) (i.e., UMSR is more efficient than SR).

To assess task success, we also counted again how often the flight “best” matching the business traveler’s profile was chosen. Of the 64 flights that were booked with the SR-based system, the most suitable flight was booked in approximately 53% of the cases. In comparison, the participants booked the most suitable flight in roughly

Table 5.2: Average number of turns per booking, dialogue duration (for booking two flights, system time plus user time) with SR and UMSR and how often the "best" flight was chosen. Significance levels: $p < .01$ indicated with "***", $p < .001$ indicated with "****".

	SR system	UMSR system
Dialogue turns***	16.44	11.80
Dialogue duration**	457	379
Best flights	34/64 (53.125%)	38/64 (59.375%)

60% of the cases with the UMSR-based system. However, this was not a significant difference.

5.3.3 Driving Safety

Unexpectedly, participants had significantly more minor errors when the system adopted the UMSR approach than when it used the SR approach, $F(1, 30) = 6.08$, $p < .05$, $M_{SR} = 1.09$, $M_{UMSR} = 1.69$, however, this appears to be driven by the difference observed among easy-driving participants. Therefore, Hypothesis B), claiming that UMSR causes fewer harmful distractions to drivers, was not supported. In fact, the reverse was true for easy-driving participants. However, the participants' average number of minor errors was less than one, thus having little negative impact on driving safety. Therefore, the very small number of accidents demonstrates that participants were concentrating on the driving task. Moreover, the difference in driving performance on the easy vs. the difficult courses indicates that the difficult courses require more cognitive resources than the easy ones - which is exactly what we aimed to achieve.

5.3.4 Perceptions

There seemed to be a cross-over interaction between driving condition and the style of information presentation on the participants' perception of how much *fun* the system was, $F(1, 30) = 7.24, p < .05$, but *post hoc* analyses suggest that the difference was only significant for easy-driving participants. That is, easy-driving participants thought that the UMSR approach was more fun to use.

Answers to the four questions/scales used in the previous study by Demberg and Moore (2006) were also analyzed. The only significant result was that participants thought that UMSR was more likely than SR to overlook better options (seven-point scales), $F(1, 30) = 5.33, p < .05, M_{SR} = 3.94, M_{UMSR} = 4.68$, but this difference was primarily observed among difficult-driving participants.

Overall, the participants perceived themselves more *positively*¹ when the system adopted the SR approach to presenting search results, $F(1, 30) = 9.65, p < .01$. This main effect appeared to be driven by the difference observed among difficult-driving participants. An interaction of the presentation style and driving condition was found on participants' self-reported *friendliness*², $F(1, 30) = 7.44, p < .05$, yet *post hoc* analyses indicate that only the difference observed among easy-driving participants was significant. That means, easy-driving participants thought they were friendlier when the system adopted the UMSR style than when it adopted the SR approach. The above subjective findings were mixed and only partly supported Hypothesis C) claiming that UMSR leads to pleasant user experience.

Finally, a comparison of participants' self-reported usual driving behavior and in-experiment driving behavior shows an interaction between course difficulty and presentation style, $F(1, 30) = 6.25, p < .05$. Specifically, easy-driving participants reported that they had reduced offensive driving (suggesting more cautious driving) when the system had adopted the SR approach, and had increased offensive driving when it

¹This index is composed of 10 items including *competent, powerful, skilled, successful, and intelligent*.

²This index is composed of items such as *cooperative, friendly, and polite*.

presented information in the UMSR style. There was also an expected main effect of driving condition, such that difficult-driving participants drove more cautiously than did easy-course participants.

5.4 Discussion

Although there was a slight increase in minor driving errors when the system used the UMSR approach as opposed to the SR approach, the general finding is that a voice browsing system based on UMSR is more efficient than one that is based on SR. This is consistent with the findings of Demberg and Moore (2006) and Winterboer and Moore (2007), and provides behavioral evidence supporting the UMSR approach.

However, improved dialogue efficiency with a spoken dialogue system does not necessarily lead to positive subjective user experience. In our study, when driving conditions were difficult and demanded a great deal of attention, the SR approach was preferred despite the high efficiency of UMSR. For example, whereas participants in the previous studies believed that UMSR provides a better overview of the available options than does SR (Demberg and Moore, 2006; Winterboer and Moore, 2007), participants of this dual-task experiment thought otherwise when driving conditions were unfavorable. Findings like this unequivocally highlight the importance of context of use in usability testing, and prompt researchers to identify problems with interface design.

A further examination of transcribed dialogue data helped us uncover a potentially critical flaw with our current UMSR simulation: for one of the four city pairs, the system generated an extremely long first-round presentation with the user-model based summarize and refine approach followed by details of three flights. Moreover, there were unnecessary pieces of information in that long presentation. Even though the presentation was based on the user model, the large amount of information nonetheless placed a cognitive burden on our participants especially when driving-related cognitive

load was already heavy. In addition, a close examination of the UMSR presentations' wording revealed that there were other presentations that contained redundant information and were generally more verbose than the used SR presentation.

The key conclusion here is that the decision about which information presentation method should be deployed in an in-car spoken dialogue system is critically dependent on the type of driving that is required. Although there was a slight increase in minor driving errors when using the UMSR system as opposed to the SR system, the general finding is that a information presentation strategy based on UMSR is more efficient than SR when the cognitive load on the driver is low. This is consistent with the findings of Demberg and Moore (2006), which suggest that when the users' complete attention can be devoted to comprehending the recommendations, the UMSR system is superior.

Conversely, when the driver must pay a great deal of attention to the road, a (relatively) simple SR strategy seem to be preferable, mainly due to the generally shorter message length. Interestingly, drivers seemed to be intuitively aware of the trade-offs between the naturalness of the UMSR system and its cognitive demands: The UMSR system was considered to provide friendliness and encouraged caution when road conditions were difficult. The critical issue here seems to be that the UMSR approach tends to present more complex sentences because it explicitly points out trade-offs using contrastive coherence markers (e.g., *but, however, although, ...*) and in the original Demberg and Moore (2006) algorithm, UMSR always presents N (typically two) levels (plus children) of the option tree containing all potentially relevant options for the user. However, this also means that presentations based on the UMSR approach to information presentation are longer than messages based on SR. In this experiment, we identified the longer message length of UMSR, partly caused by redundant information, as the confounding factor.

The key challenge, then, is to utilize the strength of UMSR systems without burdening the user with lengthy interactions. If this goal can be achieved, the presentation

of information in the car can be made both safer and more efficient. On the basis of these results we determined that it was necessary to perform further experiments which controlled the amount of content presented across the two conditions.

5.5 Conclusion

This experiment showed that it is crucial to design the presentation messages very carefully. Even though the UMSR approach performed better on some aspects, the expected positive results are not reflected in the overall outcome. After thoroughly analyzing the setup and the procedure of the study, we found considerable room for improvement, especially concerning the way we designed the UMSR presentations. Therefore, we decided to run a second experiment with an identical setup in order to test **Hypothesis 2** (Users perform better with concise messages). For this study, we improved the realization of the UMSR algorithm by using generally shorter messages, shorter sentences, and reducing the overall number of alternatives and corresponding options mentioned in each presentation. We carefully redesigned the UMSR algorithm to ensure that the same number of information units were presented in each condition as in the SR approach.

5.6 Revised Dual-Task Wizard-of-OZ Experiment

In this second experiment (Winterboer et al., 2007), we used a modified UMSR algorithm producing concise messages. The main motivation was to balance the message length in the two conditions in order to ensure that participants in both the UMSR and SR conditions would be presented with the same amount of information in each turn. In all other aspects (experimental setup, procedure, etc.), this second dual-task study resembled the first one (Hu et al., 2007). Because the UMSR approach was successful for drivers of the easy courses (in terms of dialogue efficiency, task success and partly

in terms of user satisfaction) in the first experiment despite the confounds, participants were asked to drive exclusively on the difficult courses in this second study. In the following section, we describe, first, how we balanced the message length and, second, how we modified UMSR to present concise messages.

5.6.1 Modifications to UMSR algorithm

In order to balance message length between the two conditions, we reduced the amount of information presented in each turn. Therefore, we did not present the complete length of the tree branches of the option tree (described in Section 2.7.6) as we did in the previous version of UMSR deployed in the first experiment. Instead, we split the tree into several smaller trees if otherwise the message would become too verbose. More precisely, in the initial Demberg and Moore (2006) implementation of UMSR, a heuristic cut-off point (no deeper than two branching nodes and their children) is used. Although UMSR considers only options that are relevant to the user, we found that following the original implementation sometimes led to long messages. To provide the user with a better overview of the option space, trade-offs between alternative options are presented explicitly, oftentimes leading to relatively long sentences. In addition, to give users confidence that they are being presented with all relevant options, a brief account of the remaining (irrelevant) options is also provided. Thus, the user is given an overview of the whole option space. However, if, because many options are considered relevant, the first two branching nodes and children contain large numbers of items, this leads to very long messages. For example, see Figure 5.3 showing an actual example from the first experiment.

Certainly, comprehending this message and recalling the presented information is difficult. This is especially true if another task is being conducted simultaneously and the task's performance is crucially important. In contrast, the same first-round presentation for flights from San Francisco to Prague based on the SR algorithm produced a considerably shorter message, see Figure 5.4. However, recall that the SR algorithm

System: I found 11 business class flights arriving in Prague, but you will have to make a connection. Two of the flights are on KLM. The first flight leaves at 3:20 p.m. and arrives at 2:20 p.m. with a total travel time of only 15 hours. It costs 4574 dollars and you will have to make a connection in Amsterdam.

The second flight leaves at 11:20 a.m. and arrives at 2:45 p.m. with a total travel time of 18 hours and 25 minutes. It also costs 4574 dollars and you will have to make a connection in Amsterdam as well. Would you like to book one of the KLM flights?

Figure 5.3: Presentation based on original UMSR algorithm

does not know about the user's preferences, for instance, that the user prefers flying business class. Therefore, its search space consists of all database entries for the relevant origin-destination pair on that particular date, unless the user has already provided a more specific query.

System: I found 21 flights from San Francisco to Prague. All these flights require a connection. There are flights available in economy, business and first class. I also have information about price range.

Figure 5.4: Presentation based on SR algorithm

To tackle these problems, the implemented algorithm was revised so that the option tree which is responsible for determining the dialogue flow and for the content selection was cut into smaller trees. Whereas in the prior implementation the tree was cut after a maximum of two branching nodes and their corresponding children, in the new implementation the tree was cut after one branching node plus children. In addition, we modified the algorithm to ensure that no more than two flights were ever presented in detail. If there are more than two remaining flights, we exclusively present attributes

that distinguish the flights (“The 3 direct flights are on Continental, Lufthansa, and Delta. They arrive at 9:55 a.m., 10:15 a.m., and 11:15 a.m.”).

Finally, we compared the SR and UMSR implementations to make sure that at each step, the turn length and information density would be roughly the same for both conditions. After readjusting some parameters in the algorithm we found this to be true for the majority of the cases.

Figure 5.5 shows an example with the revised UMSR algorithm. The presented pieces/units of information are marked as bold.

System: There are **no direct flights** from **San Francisco to Prague** but I found **11 connecting flights** with availability in **business class**. **2 of these** are on **KLM**.

Figure 5.5: Presentation based on revised UMSR algorithm

Of course, sticking to such a precise number can cause problems if the number of initially available flights is reduced to two. If there are only two flights that match the user’s query, the system should present all relevant details of these flights in order to allow the user an informed choice. However, even in this case the system did not necessarily have to present all attributes of the remaining flights, because in order to get to the stage where only two flights are available, the user must have already provided some details. Thus, we avoided presenting all the already obtained information a second time. Furthermore, we did not present all available flight information. For example, departure time was only presented if the user explicitly asked for it.

Figure 5.6 shows a message where only two flights remain that satisfy the user query. We find seven pieces of information in this presentation. However, this is supposed to be the upper bound, and we aimed to create messages that people are able to comprehend even if they are performing a demanding secondary task simultaneously.

A good example of how we reduced the pieces of information presented in each message are messages about time. Imagine a user who would like to book a flight

System: There are **2 business class flights** from **San Francisco to Prague** on **KLM** that will get you there on time. The **first flight arrives at 1:15 p.m.** with a **total travel time of 18 hours and 25 minutes**. The **second flight arrives at 2:20 p.m.** with a **total travel time of only 14 hours**. Would you like to book one of the KLM flights?

Figure 5.6: Presentation of two flight alternatives with revised UMSR

arriving in Frankfurt by 11 a.m., flying from New York. She would say, for example: “I’d like to book a flight from New York to Frankfurt arriving at around 11 a.m.” The system could now respond by presenting information about all flights arriving at that time frame in Frankfurt, originating in New York, and add a precise time frame to the sentence, informing the user about the time frame it used for its search, e.g., from 9 a.m. to 11:30 a.m. However, the user already knows that her query contained a specific time and that she instructed the system to look for flights arriving around 11 a.m. Therefore, we added an imprecise “that will get you there on time” to the text string with information about the number of available flights. Thus, we hoped to further ease the processing of the generated presentations because we avoided mentioning specific times.

In the next section we present results of the second WoZ dual-task study with the revised UMSR algorithm. Recall that in all other aspects (experimental setup, procedure), the second study resembled the first one (Hu et al., 2007) except that this time participants drove exclusively on the difficult driving course because we found that driving on the difficult driving course led to the observed differences between systems in the first dual-task experiment. 16 students of Stanford University participated in this experiment. This time, as already mentioned, participants drove exclusively on the difficult course.

5.7 Results

Dialogues were again recorded, transcribed, and analyzed. Data captured by the driving simulator and the questionnaires were tabulated and analyzed in SPSS.

5.7.1 Dialogue efficiency and task success

The mean number of turns each participant required for booking a flight with the system adopting the UMSR strategy (as shown in Table 5.3) remained relatively unaffected by the conducted modifications. The slight increase in number of turns can be explained by the shorter turn length which necessarily resulted in a higher number of required turns. Still, participants using UMSR took significantly fewer turns than when using the SR-based search system ($p < .05$, indicated with a “*” below).

Table 5.3: *Average number of turns per booking, dialogue duration (for booking two flights, system time plus user time) with SR and UMSR and how often the “best” flight was chosen. Significance levels: $p < .05$ indicated with “*”, $p < .01$ indicated with “**”.*

	SR system	UMSR system
Dialogue turns*	16.06	12.94
Dialogue duration**	423	323
Best flights*	19/32 (59.375%)	26/32 (81.25%)

Although average dialogue duration for SR as well as for UMSR are reduced in comparison with the first dual-task WoZ experiment, the significant difference between duration of UMSR and SR remains roughly the same ($p < .01$). The relatively big difference in dialogue duration between participants using SR in the first experiment (on average 457 seconds for booking two flights) and participants using SR in the second dual-task study (on average 423 seconds) can only be attributed to a general performance difference between participants.

To assess task success, we counted how often the flight “best” matching the business traveler’s profile was chosen. Of the 32 flights that were booked with the SR-based system, the most suitable flight was booked in only approximately 60% of the cases. In comparison, the participants booked the most suitable flight in about 80% (significant difference, $p < .05$) of the cases with the UMSR-based system.

In sum, the average flight booking dialogue with a system based on UMSR had a considerably shorter dialogue duration and required fewer dialogue turns. Moreover, UMSR enabled the user to select the best available flight in more cases. Thus, information access with the UMSR approach is more efficient than with the SR approach, even when participants are simultaneously performing a difficult secondary task.

5.7.2 Driving safety

Whereas we found in the first WoZ dual-task experiment that participants had significantly more minor driving errors when the system adopted the UMSR approach than when it utilized the SR approach, this time there were no observable differences between the driving performance of participants in the two conditions in terms of numbers of collisions, speeding tickets, traffic light or stop sign violations, or minor driving errors.

5.7.3 Perception of system and self

In the data obtained from the questionnaire, we found no significant differences between UMSR or SR concerning the participant’s perception of the system, the driving course and self. Answers to the four questions (concerning understandability, overview of options, relevance of options, and efficiency) used in the previous studies (Demberg and Moore, 2006; Winterboer and Moore, 2007) were also analyzed. The answers to all four questions concerning the UMSR presentations achieved higher scores than they

did in the evaluation of the first experiment. Nevertheless, no significant difference between the UMSR-based and the SR-based system could be found.

5.8 Discussion

The results of previous studies, asking participants to evaluate presentations based on SR and UMSR presented as dialogue transcripts (Demberg and Moore, 2006) or as sound files where the participants “overhear” the dialogue [Moore, personal communication] demonstrated a clear preference for UMSR. The same pattern of preferences were found in our first Wizard-of-Oz experiment (Winterboer and Moore, 2007) in which users actually interacted with (what they thought was) a spoken dialogue system.

In the first dual-task experiment, however, the results were twofold. Participants driving on the easy courses seemed to prefer presentations based on the UMSR approach to information presentation, whereas participants driving on the difficult driving courses preferred SR. In the current study, no significant differences on the four user satisfaction questions were found. Recall, all participants drove exclusively on the difficult driving courses in this study. However, the evaluation questions were asked at the end of a very long list of evaluation questions about the participants’ perception of the in-car system, the driving course, and themselves (see questionnaire in appendix). The sheer number of questions may have affected participants’ motivation for answering them accurately. In addition, in contrast to the previous study where participants rated dialogue transcripts (Demberg and Moore, 2006), participants in this experiment were actively interacting with the spoken dialogue system while conducting another very demanding task simultaneously. In such conditions, participants may be more concerned with completing both tasks, and less able to make subtle distinctions between systems.

However, with the refined UMSR approach, there were no significant differences in the number of driving errors between UMSR and SR. This shows that in prior experiments the confounding factor was the length of the UMSR presentations (rather than the user-model controlling the choice of attributes) making it difficult for the participants to comprehend the presentations, especially in unfavorable driving conditions involving high cognitive workload. Therefore, it was necessary to run the follow-up study with a modified UMSR algorithm controlling for turn length and information density. In addition, dialogue duration was significantly shorter with the refined UMSR approach, and users were more likely to pick the best option. Thus we see that the refined UMSR approach is equivalent to SR in terms of user satisfaction and driving safety, but better in terms of task success and dialogue duration.

5.9 Comparing UMSR with revised UMSR

We performed a post hoc analyses comparing the experiment results of participants that used the previously deployed version of UMSR with the revised version producing concise messages. In particular, we looked at the number of dialogue turns, the average dialogue duration per flight booking, the number of words the system presented to the user during one flight booking task, and task success (how often was the “best” flight available booked?), see Table 5.4.

Table 5.4: *Comparison of dialogue efficiency results obtained with UMSR vs. the revised UMSR II, $p < .05$ indicated with “*”.*

	UMSR	UMSR II
Dialogue turns	11.8	12.6
Dialogue duration	179.5	165.7
Words*	246.1	186.3
Best flights*	22/32 (68.75%)	26/32 (81.25%)

The results demonstrate that the revision was successful insofar as fewer words were presented by the system to the user while at the same time dialogue duration decreased and, most importantly, on average more often the flight best matching the user profile was selected.

Moreover, the modifications to UMSR were successful regarding the occurred driving errors. Although participants had slightly more minor driving errors (center line crossings, road edge excursions, etc.) with 2.79 (UMSR II) against 2.50 (UMSR), more speeding errors (2.73 versus 3.13) as well as accidents (0.88 versus 0.60) were made with the previous version of UMSR. However, none of the differences regarding driving errors are statistically significant.

5.10 Conclusion

In line with results from previous experiments we found that in terms of task efficiency the user-model based summarize and refine (UMSR) approach clearly outperforms the summarize and refine (SR) approach, and enables more effective information access. In contrast to previous experiments where participants focused solely on the flight booking task (Demberg and Moore, 2006), we have shown that this finding also applies to situations with another highly demanding task conducted simultaneously. **Hypothesis 1a** as well as (partly) **1b**, and in particular **Hypothesis 2** (see Section 2.11) were supported by these results. Participants achieved better task efficiency without any detrimental effects in terms of driving safety. In order to examine the impact of the secondary task on dialogue task performance, another experiment comparing the two presentation strategies without an additional driving task was conducted.

5.11 What we have learned from the conducted experiments

In this section, I briefly summarize the findings that cleared the way for the research carried out in this thesis. Then, the findings that were made during the research project are outlined.

1. In the evaluation of the MATCH system it was found that users prefer a recommendation tailored to their user model in comparison to a generic one (Walker et al., 2004).
2. Participants in a laboratory study rated presentations based on the UMSR approach to information presentation better in terms of overview of options, relevance of options, and efficiency in comparison to presentations based on SR (evaluating transcripts of dialogues, see Demberg and Moore, 2006).
3. In an interactive Wizard-of Oz experiment, the UMSR approach outperformed SR in terms of task success and dialogue efficiency. In addition, we found a general preference for UMSR-based recommendations on all four evaluation criteria (Winterboer and Moore, 2007).
4. We again observed significant differences between dialogues with the SR-based system and those with the UMSR-based system in terms of task success and dialogue efficiency a first dual-task Wizard-of-Oz study. However, we did not find differences in user perceptions of self or system. But on the evaluation questionnaires many questions were asked before the four evaluation questions appeared that were most interesting to us. Also, we found that when driving on the difficult courses, drivers performed slightly better on the driving task with the SR-based system (Hu et al., 2007).

5. In a second dual-task WoZ study with a revised UMSR algorithm utilizing more concise UMSR presentations in comparison to the first dual-task WoZ study, UMSR again outperformed SR in terms of task success and dialogue efficiency. Moreover, there were no differences concerning the driving task performance between the systems. Finally, there was a consistent trend favoring presentations based on UMSR (Winterboer et al., 2007).

Results of our experiments generally support **Hypothesis 1a**, but only partly support **Hypothesis 1b** (see Section 2.11). However, we have shown that concise messages are indeed more effective in situations where users have to divide their attention between two or more stimuli in comparison to more verbose ones while at the same time there was no difference in terms of driving performance. Therefore, **Hypothesis 2** was supported by the experimental results.

Even though we have made progress, there are still some remaining questions. First, we plan to tackle **Hypothesis 3**: Will users recall more items when the remaining items are presented as a list or when the messages use coherence markers (such as *but*, *however*, *only*, *also*, *just*, etc.) in order to make trade-offs between items explicit? Furthermore, we are interested in examining whether there are comprehension differences between the two different approaches to information presentation.

Chapter 6

Psycholinguistic background:

Sentence Comprehension and Recall

Thus far, current research efforts on information presentation in spoken dialogue systems were introduced. In doing so, it became clear that in order to be able to actually design auditory presentations that fulfill our demands for presentations that are both easy to remember and easy to comprehend, it is also necessary to take current research on sentence comprehension into account. In this chapter, general insights and findings from (online) sentence comprehension research relevant to understanding how humans process textual input are summarized. Research on the differences between listening and reading comprehension suggests that findings from reading research can also be applied to spoken stimuli, due to the commonality of processing between the two modalities (Sinatra, 1990),

Most of the findings mentioned here are based either on reading time measures derived from eye-movement research or on self-paced reading time measures. In general, it is assumed that reading times and comprehension are closely related and that eye-tracking provides the necessary measures to shed light on the underlying process of comprehending sentences on-line.

When developing their early model of text comprehension based on reader's eye movements, Just and Carpenter (1980) made two assumptions: the immediacy and the eye-mind assumptions.

1. A word is a unit of processing, and processing occurs immediately and completely at the time the word is encountered.
2. Gaze duration, which is the summed duration of consecutive fixations on one word before the reader's eyes leave that word, reflects processing time of that particular word.

Recent research casts some doubts on Just and Carpenter's assumptions. Although it appears that some aspects of lexical, syntactic, and semantic processing do (largely) respect both assumptions, many aspects of sentence interpretation are somewhat delayed. Thus, it seems that we need to relax the immediacy hypothesis because some aspects of processing simply take more time than the eye is prepared to wait.

One major advantage of eye movement measures in contrast to other comprehension measures is that they are a standard feature of normal silent reading, and, moreover, the reading rates and levels of comprehension attained in this way are indistinguishable from those that occur in the absence of eye-tracking (see Rayner et al., 1998, for example).

Interpreting an expression, regardless of whether it is written or spoken, requires integrating it into an evolving discourse model. To accomplish this task, ambiguities need to be resolved, references need to be fixed, and inferences need to be drawn to align local and global aspects of the discourse. In addition to lexical and syntactic constraints, comprehenders must draw upon pragmatic knowledge (Pylkkinen and McElree, 2006). The importance of high-level constraints has been illustrated by findings showing that comprehenders sometimes adopt a pragmatically plausible interpretation even if it is incongruent with lexical and syntactic constraints. Similarly, it was shown by Ferreira and colleagues (Ferreira et al., 2002; Ferreira, 2003) that comprehenders

often fail to accurately interpret surprisingly simple and common sentences, apparently opting for shallow forms of processing that are “good enough” for some purposes. They conclude that the goal of the comprehension system might be to deliver an interpretation that is just good enough to allow the production system to generate an appropriate response, since it is the response that is overt and determines the success of the participants’ joint activity. These findings indicate that language stimuli, like stimuli in other domains, can be processed to different depths depending on task demands and subjective criteria.

In general, studies examining the effects of the encoding task on memory for sentence pairs varying in causal relatedness indicate that the results produce an inverted-U shaped recall function. The findings in a nutshell (according to Duffy et al. (1990)): Readers show lower recall for texts for which coherence cannot be established, higher recall for texts for which coherence can be established with effort, and finally lower recall for texts for which coherence is easily established. This seems to suggest that when designing presentation methods, developers should avoid both messages where there is a very high (causal) relatedness between sentences and those where there is (almost) no relation at all.

One very robust effect in the reading literature is the *word frequency* effect which predicts that it takes longer to process a low-frequency word than a high-frequency word. Accordingly, eye-movement studies have shown that readers look longer at low-frequency words than high-frequency words (see Rayner, 1998, for a review). We can therefore infer that people perform some part of lexical access while fixating on a word and that rare words are more difficult to access than frequent words. An additional variable that affects fixation time on a word is word predictability. In this regard, Ehrlich and Rayner (1981) found that words that are constrained by preceding context are fixated for less time and skipped more often than unconstrained words. This *predictability* effect has now been replicated a number of times (see Inhoff, 1984; Rayner et al., 2001, for instance). Other factors influencing how long it takes to access the

lexical entry for a word and incorporate the new lexical information into the structural and conceptual representations the reader is constructing for the sentence are the word's length and ease of integration into the sentence (Pollatsek and Rayner, 1990). The same factors, the word's length, frequency, predictability, and ease of integration into the sentence also influence whether the eyes fixate on a word and, if so, how long the fixation is maintained (Just and Carpenter, 1980; Rayner et al., 1998).

If we apply these findings to our experiments, we can act on the assumption that users use their lexical, syntactic and even aspects of their semantic knowledge immediately when a new word appears in order to integrate it into the context, a phenomenon which is referred to as *incrementality*, since people appear to compute the grammatical structure of sentences incrementally. Moreover, the processor makes use of the sentence constraints and its knowledge regarding how the (acoustic) input is likely to unfold in order to anticipate possible ways in which the input might continue. Thus people eventually make use of information like plausibility in choosing an analysis.

In the psycholinguistic literature it is often assumed that working memory plays an important role in sentence processing. For example, according to the *shared resources account*, individual differences in working memory capacity as assessed by the reading span test (e.g., Daneman and Carpenter, 1980) should affect sentence processing. In contrast, the *dedicated resources account* (see Caplan and Waters, 1999, for example) claims that the working memory resources used for sentence processing are different from resources used by other forms of processing, so this test should not predict sentence processing effects.

Largely as a result of Gibson (1998), there has been a resurgence of interest in the relation between working memory and sentence processing difficulties. Gibson provided an account of processing complexity that at the same time sought to explain some issues in ambiguity resolution. He proposed the *Syntactic Prediction Locality Theory* (SPLT), which claims that two factors contribute to sentence complexity: storage costs, and integration costs (both drawing on the same pool of working memory

resources). Storage costs occur when there is a dependency between two syntactic elements in a sentence and the first element has to be stored in memory before it can be integrated with the later element. Integration costs occur when this integration happens and a syntactic prediction is satisfied. SPLT claims that both costs are influenced by locality (or distance), defined as the number of new discourse referents that are being processed. Thus, both storage and integration costs increase as more new discourse referents are processed since the prediction of a syntactic dependency was made at the first linguistic element. The locality aspect of SPLT also accounts for recency effects in attachment ambiguities by predicting that integration costs are larger when there is a dependency between two distant syntactic elements than between two local elements, thus providing an independent motivation for recency preferences. Therefore, a recency, or locality, preference occurs because shorter dependencies involve less memory costs than longer ones (all other things being equal). However, the theory's main contribution is in explaining processing cost in (largely) unambiguous sentences containing syntactic dependencies. An additional finding in Gibson's theory is the greater complexity of object-extracted relative clauses (OEC)

“The reporter who the photographer sent to the editor hoped for a good story.”

as compared with subject-extracted relative clauses (SEC):

“The reporter who sent the photographer to the editor hoped for a good story.”

In a variant of SPLT, the Dependency Locality Theory (DLT) (Gibson, 2000), Gibson shows that his theory (associating (1) increasing structural integration cost with the distance of attachment, and (2) storage cost with predicted syntactic categories) provides a unified theory of a large array of disparate processing phenomena.

Following the new avenue of research on the investigation of more naturalistic language (e.g., dialogue), there may be a closer link between comprehension and production (Pickering and Van Gompel, 2006). Dialogue involves tightly coupled production

and comprehension, suggesting that people may straightforwardly access information that is common to both processes. One interesting case in which production may be implicated is when the comprehender predicts upcoming structure. In individual sentences there is good evidence that individual words can be predicted (Van Berkum et al., 2005, for example). Similarly, comprehenders may also predict grammatical structure, and possibly use the production system to generate those predictions.

Likewise, when summarizing the results of three eye-tracking experiments testing the hypothesis that statistical information in the form of transitional probabilities has an influence on eye fixations during reading, McDonald and Shillcock (2004) suggest that indeed lexical statistical information is exploited by the processor during reading in order to facilitate the processing of upcoming words in the unfolding text. Thus, they argue, the on-line formation of lexical predictions is a functional (and perhaps inevitable) component of normal reading, and more generally, language comprehension. Evidence from eye-tracking and other experimental paradigms appears to suggest that readers are able to exploit context-dependent and context-independent statistical knowledge in order to anticipate the upcoming words.

6.1 Psycholinguistic background on the effect of coherence markers on recall and comprehension

Research on the differences between listening and reading comprehension suggests that findings from reading research can also be applied to spoken stimuli, due to the commonality of processing between the two modalities (e.g., Sinatra, 1990), and because it is generally assumed that the same general principles emerging from research in this field apply to both written and spoken messages (Just and Carpenter, 1984).

When Britton et al. (1982) examined the effect of linguistic markers on on-line text processing in a dual-task study, they found no effect of signaling on the amount of information readers recalled in a free recall task and yet linguistic/relational mark-

ing led to a faster average secondary task reaction time. The authors concluded that the marked version, which included words and phrases such as *therefore* or *consequently* (antecedent-consequent relations), *likewise* or *similarly* (comparison-contrast relations), *in addition* or *taken together* (collection relations), and *in particular* or *for example* (description relations) that cue, or signal, important ideas and relationships among those ideas, requires less cognitive processing capacity than the implicit version. Readers are supposed to have less trouble in inferring the relations between ideas when the signals are present. If they are not present, readers have to infer the relations between idea units to construct an adequate representation. As these inferences use cognitive capacity, readers need more time to react to a secondary task. The subsequently conducted free recall test did not reveal any significant differences between the with- and without-signaling conditions. These results suggest that relational markers guide the construction of the reader's mental representation of the text because they provide explicit information about the relations between segments. We found similar results in the two dual-task studies we conducted, where performance on the driving task did not differ between users interacting with a system using coherence markers, and those interacting with a system that did not use coherence markers, but differed significantly on the dialogue task. Users using a system that deployed coherence markers performed considerably better on the dialogue task.

This hypothesis was tested by Haberlandt (1982) who used a reading task paradigm to investigate the on-line effect of linguistic markers. The presence of connectives such as *however* was varied. Target sentences with connectives were read faster than those without connectives. However, results for free-recall measures failed to demonstrate a facilitative effect on content recall due to the presence of connectives.

Sanders and Noordman (2000) conducted an experiment using reading, verification and free recall to examine two crucial aspects of the structure of expository texts: the type of coherence relation between segments and the linguistic marking of the relations by means of signaling phrases. Similarly to Haberlandt (1982), they found

that linguistic markers facilitate the encoding of the coherence relation between two text segments. In particular, they highlight that markers lead to subsequent segments being processed faster, but do not play a role in recall.

More recently, Sanders et al. (2007) conducted a study focusing on the influence of connectives and lexical markers on text comprehension. They found evidence for a positive influence of linguistic markers of coherence on text comprehension. However, they focused on different markers than we do: on causal markers (e.g., “because”, “therefore”) and on specific lexical signals (e.g., “for that reason”, “on the other hand”).

For us, the most interesting findings concern the role of linguistic coherence markers. Their study indicated that markers expressing the relation between a text segment and the preceding context lead to faster processing of that segment and, in addition, the experiment showed that the faster processing of information following a coherence marker does not negatively affect their reproduction. Consequently, an online representation may be constructed faster with the aid of connectives/coherence markers, but may not necessarily lead to enhanced off-line recall performance.

In a recent meta-review Ben-Anath (2005) reviewed empirical research studying the role of connectives in the interpretation of coherence relations so as to facilitate the construction of a text representation. She concludes that although dialogue between cognitive studies and linguistic theory is necessary in order to elucidate a variety of issues such as semantic distinctions, the role of text genre, communicative context, and reader characteristics, in sum, a definitive assessment of the effectiveness of connectives in terms of communicative meaning distinctions remains tentative. Nevertheless, the findings from the reviewed research demonstrate that connectives do not merely signal the existence of thematic relations. Rather, connectives and their modulating effect reactivate a preceding clause that leads to the construction of a coherent relevant relation. Connectives serve as linguistic devices that provide procedural knowledge

that constrains the multiple contextual effects generated in the process of interpretation.

Thus, in sum on-line data from empirical studies suggests that the presence of coherence markers that explicitly point out similarities and differences among the options facilitates processing (Britton et al., 1982; Haberlandt, 1982; Ben-Anath, 2005; Sanders et al., 2007) and improves integration of information (Kamalski et al., 2008). However, because researchers in these studies found mixed results regarding content recall, it is difficult to say whether the information that is processed is understood equally well. It may well be that the quicker the information is processed, the less completely it is processed. Thus, we were interested whether we would be able to observe differences in reading times indicating differences in terms of comprehension.

6.2 Summary of results relevant for this research project

The following summarizes the findings from memory and psycholinguistic research that we consider relevant for the proposed research project.

- Messages where there is a very high (causal) relatedness between sentences and those where there is (almost) no relation at all should be avoided because it was found that there is lower recall for texts for which coherence cannot be established, higher recall for texts for which coherence can be established with effort, and finally lower recall for texts for which coherence is easily established.
- Various factors, such as the word's length, frequency, predictability, and ease of integration into the sentence influence the reader's eye fixations and therefore affect comprehension difficulty. Research seems to suggest that this finding can also be applied to spoken language comprehension (Sinatra, 1990).
- The processor makes use of statistical, lexical, syntactic, semantic and world knowledge to anticipate possible ways in which the input might continue.

- It is assumed that working memory plays an important role in an individual's sentence processing capabilities.
- According to Gibson (1998, 2000), both storage and integration costs increase the more new discourse referents that have been processed since the prediction of a syntactic dependency is made at the first linguistic element.
- There is strong evidence that the presence of coherence markers that explicitly point out similarities and differences among the options facilitates processing, but whether the same is true for recall is not quite so clear.

Chapter 7

Evaluating textual and auditory comprehension and recall

In this chapter, we present a comprehension and recall experiment (Winterboer et al., 2008) designed to examine the trade-off between reusing sentence structures, and employing varied sentence structures containing coherence markers, such as connectives and adverbials. The motivation for this study is that on the one hand, we know that syntactic priming and simple sentences eases comprehension (e.g., Bock, 1986), see Chapter 4.8. On the other hand, reading comprehension experiments indicate that coherence markers such as connectives and lexical cue phrases help the reader structure the presented information (e.g., Sanders and Noordman, 2000; Louwerse, 2001). Also, well structured information is usually easier to recall. Work in this area might serve both what people think they like and what actually helps them, i.e., using devices such as coherence markers that help the reader structure information increases both user satisfaction and recall.

7.1 Evaluating the effects of coherence markers on recall

Spoken dialogue systems have traditionally used simple templates for natural language realization to present options (of e.g., flights, restaurants, hotel rooms) and their attributes to users (e.g., Levin et al., 2000; Walker et al., 2004). Recently, however, researchers have proposed approaches to information presentation that use coherence markers (e.g., lexical cue phrases, connectives, and adverbials such as *but*, *however*, *moreover*, *only*, *just*, *also*, etc.) in order to highlight specific properties of and relations between items presented to the user, e.g., associations (Polifroni and Walker, 2006b) and contrasts (Winterboer and Moore, 2007).

Previous research indicates that coherence markers such as connectives facilitate comprehension (see, e.g., Ben-Anath, 2005, for a literature review). However, to our knowledge, no empirical validation has been performed to test whether these coherence markers have an effect on comprehension and recall in information presentation messages.

7.1.1 Experimental setup and procedure

In order to test whether there are differences in recall due to the usage of coherence markers, we performed a within-participants reading experiment comparing item recall for experimental material presented with or without coherence markers.

A total of 24 participants, native English speakers and mostly students of the University of Edinburgh, were paid to participate in the study. They were naive to the purpose of the experiment but were told that they were about to be presented with information about a number of consumer products and that they were supposed to answer questions about these.

Each participant read 14 short texts describing consumer products from 14 domains, see Figure 7.1 and Figure 7.2 for examples. The domains were chosen to

Messina's price is £22. It has very good food quality, attentive service, and decent décor.

Ray's price is £34. It has very good food quality, excellent service, and impressive décor.

Alhambra's price is £16. It has good food quality, bad service, and plain décor.

Figure 7.1: *Example for experiment material without coherence markers.*

guarantee that each participants would find familiar and less familiar domains. All experimental materials can be found in Appendix 8.6. The texts are the type of presentation typically produced by spoken dialogue systems designed to help users select an entity from a small set of available options. In the SR and UMSR approaches we have described in previous chapters, this type of presentation typically occurs once the user has provided enough constraints to narrow the available options to a small number.

Messina's price is £22. It has very good food quality, attentive service, and decent décor.

Ray's price is £34. It **also** has very good food quality, **but** excellent service, and **moreover** impressive décor.

Alhambra's price is **only** £16. It has good food quality, **but** bad service, and **only** plain décor.

Figure 7.2: *Example for experiment material with coherence markers, where coherence markers are indicated in bold.*

There were two types of texts, one containing coherence markers to point out similarities and differences among the options, and one without coherence markers. Each participant read seven texts of each type, alternating between types. Ordering of both the domains and text types was controlled for. We took particular care to add coherence markers without modifying the propositions in any other way.

The experiment took approximately 40 minutes. First, participants were seated in front of a monitor which displayed the instructions and were requested to read and sign a consent form telling them about the experiment. Second, when they were ready, the eye-tracker was adjusted and calibrated. We used the SR Research Experiment Builder software and an EyeLink II eye-tracker¹ to design the experiment and present the materials. In each trial, participants read a text with or without coherence markers, which was presented for up to 45 seconds on the screen, and pressed “enter” when they were finished reading. Figures 7.1 and 7.2 show example messages without and with coherence markers.

We used examples from 14 different domains (Rental cars, fridges, book bags, Mp3 players, hotels, digital cameras, flights, mobile phone plans, restaurants, make-up palettes, laptops, monitors, microwaves, cinemas) thus guaranteeing that almost everyone experienced familiar and unfamiliar item domains. After reading each text, participants were presented with a series of three questions, which they had to answer one after the other; examples of each type of recall question are given in Figure 7.3. After a question was presented, the participant pressed “enter” again to be prompted to type in an answer.

1. Verbatim questions: e.g., *Which restaurant's price is £34?*
2. Comparison questions: e.g., *Which restaurant is the cheapest?*
3. Evaluation questions: e.g., *Which restaurant would you like to go to and why?*

Figure 7.3: *The three types of evaluation questions with examples:*

7.1.2 Results and discussion

Overall, we found a consistent numerical trend indicating that items in messages containing coherence markers could be recalled more easily (see Figure 7.1.2). In partic-

¹<http://www.eyelinkinfo.com/>

ular, answers to comparison questions were correctly recalled significantly more often when coherence markers were present (recall data is provided on a scale from 0 to 1). We used the 0 to 1 scale, because its outcomes are more transparent to the reader in comparison to the actual fractions derived from the computations of correctly recalled items divided by maximally correctly recalled items. In addition, we used a chi squared test receiving similar results.

	w/o linguistic markers	with linguistic markers
Verbatim question	0.79	0.82
Comparison question*	0.68	0.79
Evaluation question	0.73	0.81

Figure 7.4: Average recall on a scale from 0 to 1 for the three recall questions, $p < .05$ indicated with “*”,

In the user study, we found that using coherence markers (e.g., connectives, adverbials, and discourse cues) indeed facilitates the recall and discrimination of information presented on a screen in a reading experiment. However, we only found this result for the comparison question. This could have to do with comparison questions benefitting more from the used lexical markers that make it easier to discriminate between options in comparison to the message version without linguistic markers. If it is the case that presenting spoken information with coherence markers makes the information easier to discriminate and recall, both developers and users of dialogue systems would benefit. Ultimately, what developers care most about is to support users in choosing the best available option, which in turn should lead to increased user satisfaction with the system.

Because we used an eye-tracking setup for our experiment there was additional data to be analyzed, which may reveal comprehension differences between the two approaches to information presentation.

7.2 Evaluating the effects of linguistic markers on comprehension of textual materials

We used an eye-tracker in this reading experiment in order to measure reading times, because reading times are considered to shed light on on-line discourse processing and comprehension (see e.g., Haviland and Clark, 1974). We assumed that it would take more time to read presentation messages containing coherence markers because a) there were more words in the messages (on average 76.9 words in the presentations without coherence markers vs. 81.6 words in presentations with coherence markers), and b) because typically messages with coherence markers are more complex and therefore assumed to be more difficult to process. Greater processing difficulty usually means longer reading times.

In this experiment, readers eye-movements were monitored while reading sentences with and without coherence markers. In order to provide even participants' whose reading speed is slower than average with sufficient time for reading and comprehending the presented messages, the message presentation duration was approximately one second per presented word, or until participants used the "enter" button on the keyboard. This is based on findings from our own research indicating that a slow reader requires a reading time of just under one second per word when reading a text carefully for optimal comprehension.

In general, reading the messages containing coherence markers generally took slightly longer, with participants reading messages containing coherence markers taking 37.93 seconds per message on average, and reading messages without coherence markers taking 35.28 seconds on average. The question is, however, whether this difference can be attributed exclusively to the number of additional words in the messages with coherence markers, or whether readers also required more time to process the presentation's content because the presence of coherence markers increases sentence

complexity. Sentence complexity might increase with the introduction of coherence markers, which in turn increases reading times.

In order to answer this question, we compared the reading times of only small areas within the presentation messages. These interest areas (IA) were located directly (one word) after the (potential) location of the coherence marker. In total, we identified 46 IAs within the 14 presentations, each one consisting of two words or around nine characters on average.

7.2.1 Results

	FPRT	RT	NoP	RegrIn	RegrOut
with markers	473.83	1055.56	3.639	0.430	0.322
w/o markers	510.24	1150.70	3.567	0.494	0.350
	t = -1.511	t = -0.820	t = 0.625	t = -1.002	t = -0.519
	p = 0.131	p = 0.412	p = 0.5321	p = 0.3164	p = 0.6039

Table 7.1: *Eye-tracking data per IA (first pass reading times, remaining time reading times, number of passes, regressions out and in) for messages with and without coherence markers*

The results of the different reading time measures, established with linear mixed-effects model (LME) analyses in R² (R Development Core Team, 2005) (see Table 7.1), do not reveal any significant differences between the two conditions, although, surprisingly, IAs had a numerically shorter reading time when coherence markers were used. In this repeated measures design experiment, participant, IA, and item were random-effect factors and the fixed-effect factor was whether the presentation contained coherence markers. We compared first pass (FPRT, measure of early processing) and remaining pass (RT, measure of late processing) reading times per IA, the mean number of passes (NoP), and regressions in (RegrIn) and out (RegrOut) of the IA.

²www.r-project.org

First pass reading time is defined as the sum of the durations of the fixations the reader makes in a given part of a sentence (a word or a group of words) from the time he/she enters in it until he/she leaves it forwards or backwards for the first time. It is generally assumed that this measure is sensitive to early processes in the comprehension of a sentence, such as syntactic parsing and the early integration of information. On the other hand, total reading time is defined as the sum of the duration of all the fixations made in a given part (a word or a group of words) of a sentence. It is assumed that this measure is sensitive to the later processes involved in the comprehension of sentences, such as re-analysis and discourse integration (Rayner et al., 1989; Espino et al., 2005). The number of passes indicate how often a certain part of a sentence was read and regressions in and out an IA establish how often readers' eyes fixated on the IA making an eye-movement from a position before or after the IA.

7.2.2 Discussion and conclusion of textual comprehension experiment

Although sentences containing coherence markers are more complex and thus should incur longer reading times, our analyses do not show any differences in reading times for the words directly following the coherence markers. The differences in the overall reading times noted above are therefore due to the additional words (the presence of the coherence markers) and not caused by differences in sentence complexity or increased effort towards the marked parts of the text.

The combination of eye-tracking and recall data seems to provide a relatively clear picture: Although sentences with coherence markers took more time to read, this is exclusively due to the additional words and not caused by a difference in sentence complexity. In addition, we found that using coherence markers indeed facilitates recall of the presented information and makes the presented information more memorable. The results of this experiment therefore confirm **Hypothesis 3** (see Section 2.11). Interestingly, the supposedly more difficult questions requiring more process-

ing, for instance a comparison of different presented options, achieved better recall in the version containing linguistic markers.

We conducted another experiment, this time web-based, in order to examine whether the same pattern of results could be observed in an environment that is more natural and convenient for participants in comparison to the eye-tracking lab.

7.3 Evaluating the effects of linguistic markers on recall of written materials

We carried out a web-based user study both in order to verify the results obtained in the previous recall experiments and in order to test whether results obtained from casual website users are comparable to those obtained from laboratory participants who focus exclusively on performing the experiment in the lab (Tietze et al., 2009). Thus, we recruited native English speakers online to carry out the same experiment previously conducted in the lab. For this experiment, we used Amazon.com's *Mechanical Turk*³ environment - a web based micro-task platform that allows researchers and developers to create and upload *Human Intelligence Tasks* (HITs) on the web. HITs are generally small tasks such as information filtering, feedback on pictures and texts, or anything that requires human intelligence.

Amazon.com's Mechanical Turk is advantageous because it attracts many visitors due to its affiliation with the well established Amazon website and thus eases recruitment of new participants especially from outside the student population. In addition, conducting experiments online significantly reduces the effort involved in data collection for the experimenter. Moreover, the website allows for convenient payment for both participants and the experimenter. For these reasons, Mechanical Turk has recently been used in a number of language experiments (e.g., Kaisser et al., 2008; Nakov, 2008; Kittur et al., 2008).

³<https://www.mturk.com/mturk/>

7.3.1 Participants

60 participants read the same materials that were used in the laboratory recall and comprehension experiment (Winterboer et al., 2008). Mechanical Turk allows the experimenter to place a restriction on participant location (only users from the US were allowed to participate in an attempt to ensure English language skills), or the number of trials (each participants was only allowed to participate once). However, one cannot balance gender or control for age and literacy reliably, as user provided data cannot be verified. Also, one does not know whether participants are conducting another task simultaneously, or are otherwise distracted.

We paid \$ 2.50 for participation, which was, given that we expected the experiment to last less than 30 minutes, considerably more than participants would receive for most other tasks available on Amazon.com's Mechanical Turk website. We hoped that the higher reward would encourage participants to take the task more seriously.

7.3.2 Experimental setup and procedure

In order to resemble the Experiment Builder interface that was used in the eye-tracking experiment as closely as possible in terms of the general "look and feel", a web based interface was implemented using Adobe's Flash format. We chose the widely used Flash format because it can be integrated into the Mechanical Turk environment easily and allows for tighter control on user behavior in comparison with standard HTML pages. For example, we made it impossible for users to reread the presented information once they read the corresponding question. With standard HTML users would have been able to use their browser's back button to do just that. The experiment was then made available to potential users on Amazon's Mechanical Turk website. The procedure was otherwise exactly the same as in the previous laboratory recall experiment.

Again, each participant read 14 short texts describing consumer products from 14 domains. Again, there were two types of texts, one containing coherence markers to point out similarities and differences among the options, and one without those coherence markers. Each participant read seven texts of each type, alternating between types. Ordering of both the domains and the text type was controlled for. We also took particular care to add coherence markers without modifying the propositions in any other way.

In each trial, participants read a text presented on the screen, and pressed enter when they were finished reading. They were then presented with a series of three questions, which they had to answer one after the other. After a question was presented, the participant pressed enter to be prompted to type in an answer to that question.

7.3.3 Results of web-based recall experiment

The first thing we noticed when evaluating the data was that it took only a couple of hours from making the tasks available on the Mechanical Turk website to receiving the results. In addition, we learnt from the submitted answers that the general answer quality was comparable to answers obtained in the laboratory-based eye-tracking experiment - the average answer quality was very high.

Of the 60 participants we rejected three straightaway for answering less than 50 percent of the questions. The answers of four participants were not included because they either only took one quarter of the average time (meaning that they more or less guessed at the answers) or because they participated twice which we could not let happen, because learning the materials and questions would influence the results. Thus, we eventually based our data analysis on the answers of 53 participants. Three different experimenters independently assessed the correctness of the recalled information.

Overall, we found close similarities between results obtained in the eye-tracking experiment and those obtained in the web-based experiment. This was slightly surprising given that participants performing the eye-tracking experiment were concentrating

on the reading task exclusively while, in theory, the participants of the web-based experiment could do whatever they wanted during the experiment. In addition, participants in the web study were not necessarily students. Since students are typically younger and used to focusing on intellectual tasks in comparison with the standard population, one might expect them to achieve higher results in recall and comprehension tasks.

Table 7.2: Average recall on a scale from 0 to 1 for the three questions - lab experiment (lab) vs. web based experiment results (web), *t*-Test, “*” indicates a significant difference with $p < .05$, significance between underlined values and values in italic

	w/o mark. (lab)	w/o mark. (web)	w/ mark. (lab)	w/ mark. (web)
Verbatim Q.	0.79	0.83	0.82	0.80
Comp Q.*	<u>0.68*</u>	0.62	<u>0.79*</u>	0.81
Evaluation Q.*	0.73	<i>0.83*</i>	0.81	<i>0.88*</i>

The results of the web study show that the general answer quality was comparable to answers obtained in the laboratory-based recall experiment. Average recall rate was nearly identical with 0.76 (web-based) and 0.77 (lab-based), respectively. In addition, the average answer time was also almost identical, approximately 23 minutes (web-based) and 26 minutes (lab-based) on average per participant, respectively.

However, we did not find an effect on the comparison question. Instead, this time the difference between the two conditions was significant in terms of correct answers to the evaluation question. Thus, we again found that using coherence markers facilitates recall of information.

7.3.4 Discussion

Taken together, we found a small but significant effect of coherence markers on recall again supporting **Hypothesis 3**, see Section 2.11. The combination of eye-tracking

and recall data seems to provide a relatively clear picture: Although sentences with coherence markers took more time to read, this is exclusively due to the additional words and not caused by differences in the construction of the internal representation. While these findings are in line with results from psycholinguistics which demonstrated that coherence markers may improve comprehension and recall (Britton et al., 1982; Haberlandt, 1982; Ben-Anath, 2005), given the small effect, it does not fully explain the improvements in terms of task effectiveness found in information presentation for spoken dialogue systems (Hu et al., 2007; Winterboer et al., 2007; Winterboer and Moore, 2007).

We additionally validated the results using participants recruited online. The similar results show that this web-based method is applicable to the evaluation of written language materials and adds further strength to its establishment as an alternative to lab-based experiments.

Nonetheless, in real-world spoken dialogue systems users are presented with information about different options auditorily. Listening to auditory stimuli should be more difficult than reading the same stimuli, because readers can always re-read a problematic word or sentence, whereas auditory stimuli are presented sequentially and are transient. However, research on the differences between reading and listening comprehension seems to suggest that the findings found in reading can also be applied to spoken stimuli due to the commonality of processing between the two modalities (Just and Carpenter, 1984; Sinatra, 1990).

Chapter 8

Conclusions and future work

In this thesis, I examined approaches to content selection and information presentation in spoken dialogue systems. To be more specific, I experimentally studied the effect of information presentation strategies on user perception, task success, dialogue efficiency, recall, and cognitive load conducting interactive Wizard-of-Oz, dual-task, eye-tracking, and web-based experiments. Information presentation is a crucial area of spoken dialogue system research helping users to deal with and browse through the potentially large space of available options in a world of information abundance.

In this chapter, I summarize the contributions of the thesis following the hypotheses formulated in Section 2.11 and suggest directions for future work describing how the work conducted for this thesis can be extended.

8.1 UMSR vs. SR - User preference and dialogue efficiency

The first hypothesis of this thesis was that users would a) prefer and b) perform better with the recently developed user-model based summarize and refine approach (UMSR, see Demberg, 2005; Demberg and Moore, 2006) to information presentation in comparison to a system deploying the summarize and refine approach (SR, Polifroni et al.,

2003; Chung, 2004). The results of the previously conducted experiments, asking participants to evaluate presentations based on SR and UMSR presented as dialogue transcripts (Demberg and Moore, 2006) or as sound files where the participants “overhear” the dialogues (Moore, personal communication), demonstrated a clear preference for UMSR.

We performed a within-subjects user study with 34 users comparing the SR and UMSR approaches to information presentation in terms of their effect on dialogue efficiency and task success. In contrast to previous experiments, participants in this experiment were actively interacting with the dialogue system. In this user study, we found a general preference for presentations based on the UMSR approach to information presentation. Furthermore, we also found that the UMSR approach outperforms the SR approach in terms of task success and dialogue efficiency. With UMSR, users choose the flight that best matches their user profile more often than with SR and, in addition, booking flights with UMSR takes less time and less dialogue turns in comparison with SR.

8.2 Dual-task studies - effect of cognitive load?

Our second hypothesis was that users who are performing another (demanding) task simultaneously will also benefit from a system employing the UMSR approach in comparison with a system using the SR approach to information presentation. Accordingly, we conducted a dual-task studies to examine the effect of the two previously introduced information presentation strategies on cognitive load (Hu et al., 2007). We used driving as the secondary task varying in difficulty to measure how well participants performed on both tasks.

Although there was a slight increase in minor driving errors when the system used the UMSR approach as opposed to the SR approach, the general finding was that a voice browsing system based on UMSR is more efficient than one that is based on SR.

This is consistent with the findings of Demberg and Moore (2006) and Winterboer and Moore (2007), and provides behavioral evidence supporting the UMSR approach. In terms of task efficiency, UMSR again outperformed SR.

However, improved dialogue efficiency with a spoken dialogue system does not necessarily lead to positive subjective user experience. In our study, when driving conditions were difficult and demanded a great deal of attention, the SR approach was preferred despite the high efficiency of UMSR. For example, whereas participants in the previous studies believed that UMSR provides a better overview of the available options than does SR (Demberg and Moore, 2006; Winterboer and Moore, 2007), participants of this dual-task experiment thought otherwise when driving conditions were unfavorable. Taken together, the results of this dual-task experiment were twofold. Participants driving on the easy courses seemed to prefer presentations based on the UMSR approach to information presentation, whereas participants driving on the difficult driving courses preferred SR.

As we uncovered a potentially critical flaw with our initial UMSR simulation, we revised the UMSR algorithm with the main goal of balancing message length between UMSR and SR presentations.

8.3 Message length influence?

The third hypothesis was that concise messages are more effective than verbose ones in situations where users have to divide their attention between two or more stimuli, that is, users will perform better in terms of both dialogue task and secondary task performance when interacting with a system that presents concise messages in comparison with a system deploying more verbose presentations. To examine the message length influence, we modified the UMSR approach to information presentation and ran a second dual-task experiment. In this experiment, all participants drove exclusively on the difficult driving courses.

In this second dual-task study with a revised UMSR approach, no significant differences on the four user satisfaction questions were found. However, the evaluation questions were asked at the end of a list of 85 evaluation questions about the participants' perception of the in-car system, the driving course, and themselves. The sheer number of questions may have affected participants' motivation or ability to for answer them accurately.

In addition, in contrast to the previous study where participants rated dialogue transcripts Demberg and Moore (2006), participants in the dual-task experiments were actively interacting with the spoken dialogue system while conducting another very demanding task simultaneously. In such conditions, participants may be more concerned with completing both tasks, and less able to make subtle distinctions between systems. However, unlike in the first dual-task study, there were no significant differences in the number of driving errors between UMSR and SR with the refined UMSR approach. This shows that in prior experiments the confounding factor was the length of the UMSR presentations (rather than the user-model controlling the choice of attributes) making it difficult for the participants to comprehend the presentations, especially in unfavorable driving conditions involving high cognitive workload.

Therefore, it was necessary to run the follow-up study with a modified UMSR algorithm controlling for turn length and information density. In addition, dialogue duration was again significantly shorter with the refined UMSR approach, and users were more likely to pick the best available option based on their user profile. Thus we see that the refined UMSR approach is equivalent to SR in terms of user satisfaction and driving safety, but better in terms of task success and dialogue duration.

We also performed a post hoc analysis comparing the experiment results of participants that used the previously deployed version of UMSR with the revised version producing concise messages. The results demonstrate that the revisions were successful insofar as fewer words were presented by the system to the user while at the same

time dialogue duration decreased and, most importantly, on average more often the flight best matching the user profile was selected.

8.4 Do coherence markers facilitate recall?

Finally, we formulated the hypothesis that once the user has provided enough constraints to narrow the available options to a small number of items to present within a single turn, messages that make trade-offs between items explicit will facilitate recall in comparison with a system that presents the remaining items as a list (as in the Communicator dialogues and the MATCH systems Levin et al., 2000; Walker et al., 2004, , for instance). At the same time, messages that point out trade-offs and contrasts between options will not negatively affect message comprehension.

In order to test whether there are differences in recall, we performed a within-participants reading experiment with 24 participants comparing item recall for experimental material presented with or without coherence markers. We used an eye-tracker in this reading experiment in order to measure reading times, because reading times are considered to shed light on on-line discourse processing/comprehension (see e.g., Haviland and Clark, 1974). Although sentences containing coherence markers are more complex and thus should incur longer reading times, our analyses do not show any differences in reading times for the words directly following the coherence markers.

Taken together, we found a small but significant effect of coherence markers on recall. The combination of eye-tracking and recall data seems to provide a relatively clear picture: Although sentences with coherence markers took more time to read, this is exclusively due to the additional words and not caused by differences in the construction of the internal representation. While these findings are in line with results from psycholinguistics which demonstrated that coherence markers may improve comprehension and recall (Britton et al., 1982; Haberlandt, 1982; Ben-Anath, 2005),

given the small effect, it does not fully explain the improvements in terms of task effectiveness found in information presentation for spoken dialogue systems (Hu et al., 2007; Winterboer et al., 2007; Winterboer and Moore, 2007).

We additionally validated the results using participants recruited online. Overall, we found close similarities between results obtained in the eye-tracking experiment and those obtained in the web-based experiment.

8.5 Discussion

The results of the described experiments have contributed to answering the research questions and hypotheses formulated in Section 2.11. They show that an information presentation strategy that takes a model of the user's preferences into account to cluster options based on attributes that are relevant to the user and uses coherence markers (e.g., discourse cues, adverbials) to highlight the trade-offs among the presented items (UMSR) is both preferred by users and leads to higher task efficiency in comparison with a system that enables information browsing by creating summaries that are generated by clustering the options based on attributes that lead to the smallest number of clusters (SR).

Furthermore, our experiments show that UMSR is also more effective than SR when users are performing a demanding secondary task simultaneously. Moreover, we found that with a revised UMSR dialogue efficiency increased even more and participants also chose the flights that best matched their user preferences more often than with SR.

Finally, we hypothesized that one of the main reasons why UMSR is more efficient is because it uses coherence markers to highlight relations (e.g., trade-offs) between options and attributes. Thus, we performed an eye-tracking experiment in which participants read presentations with and without these coherence markers, and answered evaluation and comparison questions to measure differences in recall. In addition, we

used reading times to examine comprehension differences between the two information presentation strategies. We found that the lexical devices used in our experiment material, which was based on UMSR, indeed facilitated item recall, with no penalty in terms of comprehension cost. The recall part of these results was additionally validated with a web-based user study in which we obtained similar results.

8.6 Future work

The line of work described in this thesis points towards several possible directions for future work. For instance, I would like to examine whether the findings obtained in the non-interactive experiments with written and auditory materials can be replicated using a real-world spoken dialogue system. Real-world dialogue systems have to deal with many problems that can be avoided in lab-based Wizard-of-Oz experiments, such as the ones we performed (potential problems are noisiness, speech recognition errors, system robustness, and so on).

Another potential line of future research concerns the development of a machine-learning algorithm for learning preferences and dialogue patterns. This way, each time the user collaborates with the system to solve a given task, such as booking a flight or finding a restaurant, the user model is updated accordingly. Clearly, with the data provided, algorithms could be integrated to take into account user behavior and user preferences based on the data thus adapting to the specific user making future conversations more effective.

When evaluating the eye-tracking experiment we found that adding coherence markers explicitly pointing out trade-offs between options (which introduces more words) imposes no significant difference in processing. The standard finding is that user satisfaction is inversely correlated with interaction duration (see Section 1.1). However, we have evidence that extra words improve recall with no significant penalty in processing time. It is therefore tempting to distinguish “bad length” from “good

length”. If the time spent on processing coherence markers is somehow “good”, because it means that average recall increases, the question arises as to whether this will be reflected in user satisfaction. In particular, while we still expect the correlation between satisfaction and duration to be inverse, we might expect it to be “weaker” than in the standard case. It would be interesting to evaluate the effect of message length on user satisfaction in another user study.

Finally, the concise presentations of both SR and the revised UMSR approach to information presentation appear to be well suited for information browsing applications deployed on mobile devices with limited screen space. Currently, to my knowledge, no mobile phone/device makes use of the kind of user model-based information presentation approach deployed in the described experiments. But when time and screen space is limited, and there are many potentially relevant items available, applying intelligent techniques to guide users through the huge information space is both necessary and adequate. This area is particularly interesting because there are already a huge number of mobile devices available that could be used to provide implicit and explicit feedback alike in such a recommender system about how users arrive at an item in the space of options. Eventually, after booking the flight, eating at the restaurant, etc. users could then rate the quality of the item. Those ratings then could be aggregated and integrated for future recommendations.

Bibliography

- Ai, H., Litman, D. J., Forbes-Riley, K., Rotaru, M., Tetreault, J., and Purandare, A. (2006). Using system and user performance features to improve emotion detection in spoken tutoring dialogs. In *Proceedings of ICSLP (Interspeech '06)*, Pittsburgh, PA.
- Baayen, R. H. (2008). *Analyzing Linguistic Data - A practical introduction to statistics*. Cambridge University Press, Cambridge, England.
- Baddeley, A. (1996). The psychology of memory. In A.D. Baddeley, B. A. W. and Watts, F. N., editors, *Handbook of Memory Disorders*, pages 3–25. John Wiley and Sons, Chichester.
- Baddeley, A. D. and Hitch, G. (1974). Working memory. In Brower, G., editor, *The Psychology of Learning and Motivation*, pages 47–89. Academic Press, New York, USA.
- Becker, T., Blaylock, N., Gerstenberger, C., Kruijff-Korbayova, I., Korthauer, A., Pinkal, M., Pitz, M., Poller, P., and Schehl, J. (2006). Natural and intuitive multimodal dialogue for in-car applications: The SAMMIE system. In *Proceedings of the ECAI Sub-Conference on Prestigious Applications of Intelligent Systems (PAIS 2006)*.
- Ben-Anath, D. (2005). The role of connectives in text comprehension. *Working papers in TESOL and Applied Linguistics*, 5:1–27.

- Berkovsky, S., Eytani, Y., Kuflik, T., and Ricci, F. (2007). Enhancing privacy and preserving accuracy of distributed collaborative filtering. In *Proceedings of the ACM conference on Recommender Systems 2007*, pages 9–16, Minneapolis, MN.
- Bock, J. (1986). Syntactic persistence in language production. *Cognitive Psychology*, 18:355–387.
- Bohus, D. and Rudnicky, A. L. (2007). Implicitly-supervised learning in spoken dialogue interfaces: an application to the confidence annotation problem. In *Proceedings of 18th SIGdial Workshop on Discourse and Dialogue (SIGDial '07)*, pages 256–264, Antwerp, Belgium.
- Briem, V. and Hedman, L. (1995). Behavioural effects of mobile telephone use during simulated driving. *Ergonomics*, 38:2536–2562.
- Britton, B. K., Glynn, S. M., Mayer, B. J. F., and Penland, M. J. (1982). Effects of text structure on use of cognitive capacity during reading. *Journal of Educational Psychology*, 74:51–61.
- Broadbent, D. (1958). *Perception and Communication*. Pergamon, Oxford, England.
- Broadbent, D. (1975). The magic number seven after fifteen years. In *Studies in Long-Term Memory*, pages 3–18. Wiley.
- Brown, I. and Poulton, E. (1961). Measuring the spare 'mental' capacity of car drivers by a subsidiary task. *Ergonomics*, pages 35–40.
- Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370.
- Callaway, C. B., Dzikovska, M., Farrow, E., Marques-Pita, M., Matheson, C., and Moore, J. D. (2007). The beetle and beediff tutoring systems. In *Proceedings of the 2007 Workshop on Spoken Language Technology for Education (SLaTE)*, Farmington, Pennsylvania, USA.

- Caplan, D. and Waters, G. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, 22:77–126.
- Carenini, G. and Moore, J. D. (2001). An Empirical Study of the Influence of User Tailoring on Evaluative Argument Effectiveness. In Nebel, B., editor, *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI'01)*, pages 1307–1314, San Francisco, CA. Morgan Kaufmann Publishers, Inc.
- Carenini, G. and Moore, J. D. (2006). Generating and evaluating evaluative arguments. *Artificial Intelligence*, 11:925–952.
- Cheng, H., Bratt, H., Mishra, R., Shriberg, E., Upson, S., Chen, J., Weng, F., Peters, S., Cavedon, L., and Niekrasz, J. (2004). A Wizard-of-Oz framework for collecting spoken human-computer dialogs. In *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech '04)*.
- Chu, S.-W., O'Neill, I., and Hanna, P. (2007). Using multiple strategies to manage spoken dialogue. In *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech '07)*, pages 158–161, Antwerp, Belgium.
- Chung, G. (2004). Developing a flexible spoken dialog system using simulation. In *Proceedings of the 42th Annual Meeting of the Association for Computational Linguistics (ACL '04), Barcelona, Spain*, pages 63–70.
- Cohen, M., Giangola, J., and Balogh, J. (2004). *Voice User Interface Design*. Addison-Wesley Publishing Company.
- Currie, K. and Tate, A. (1991). O-plan the open planning architecture. *Artificial Intelligence*, 52:49–86.
- Dahlback, N., Jonsson, A., and Ahrenberg, L. (1993). Wizard of oz studies – why and how. In *Proceedings of the 1993 International Workshop on Intelligent User Interfaces, Session 7: Design & Evaluation*, pages 193–200.

- Daneman, M. and Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19:450–466.
- De Waard, D. (1996). *The Measurement of Drivers' Mental Workload*. Phd thesis, University of Groningen, Haren, The Netherlands: University of Groningen, Traffic Research Centre.
- Demberg, V. (2005). Information presentation in spoken dialogue systems. Msc Thesis, School of Informatics, Edinburgh, United Kingdom: University of Edinburgh.
- Demberg, V. and Moore, J. (2006). Information presentation in spoken dialogue systems. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL '06)*, Trento, Italy.
- Demberg, V., Winterboer, A., and Johanna, J. (2009). Information presentation in spoken dialogue systems. *Computational Linguistics*.
- Duffy, S., Shinjo, M., and Myers, J. (1990). The effects of encoding task in memory for sentence pairs varying in causal relatedness. *Journal of Memory and Language*, 29:27–42.
- Dybkjaer, L. and Bernsen, N. (2001). Exploring Natural Interaction in the Car. In *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue*, pages 75–79, Verona, Italy.
- Ehrlich, S. F. and Rayner, K. (1981). Contextual effects on word recognition and eye movements during reading. *Journal of Verbal Learning and Verbal Behaviour*, 20:641–655.
- Espino, O., Santamara, C., Meseguer, E., and Carreiras, M. (2005). Early and late processes in syllogistic reasoning: Evidence from eye-movements. *Cognition*, 98:1–9.

- Farber, E. Foley, J. and Scott, S. (2000). Visual attention design limits for its in-vehicle systems: The society of automotive engineers standard for mimiting visual distraction while driving.
- Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47:164–203.
- Ferreira, F., Ferraro, V., and Bailey, K. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11:11–15.
- Geutner, P., Steffens, F., and Manstetten, D. (2002). Design of the vico spoken dialogue system: Evaluation of user expectations by wizard-of-oz experiments. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation 2002 (LREC '02)*, Las Palmas, Spain.
- Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68:1–76.
- Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Miyashita, Y., Marantz, A., and O'Neil, W., editors, *Image, Language, Brain*, pages 95–126. MIT Press, Cambridge, MA.
- Goodman, M., Bents, F., Tijerina, L., Wierwille, W., Lerner, N., and Benel, D. (1997). An investigation on the safety implications on wireless communications in motor vehicles. Technical Report DOT HS 808-635, National Highway Traffic Safety Administration, Washington, DC.
- Haberlandt, K. (1982). Reader expectations in text comprehension. In Le Ny, J. F. and Kintsch, W., editors, *Language and language comprehension*, pages 239–249. North-Holland, Amsterdam, Netherlands.

- Haigney, D. E., Taylor, R. G., and Westerman, S. J. (2000). Concurrent mobile (cellular) phone use and driving performance: task demand characteristics and compensatory processes. *Transportation Research*, pages 113–121.
- Hart, S. and Staveland, L. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Human Mental Workload*, volume 1, pages 139–183. Elsevier Science, Amsterdam, North-Holland.
- Haviland, S. and Clark, H. (1974). What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behaviour*, 13:512–521.
- Hu, J., Winterboer, A., Nass, C., Moore, J., and Illowsky, R. (2007). Context & usability testing: User-modeled information presentation in easy and difficult driving conditions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*, San Jose, CA.
- Inhoff, A. (1984). Two stages of word processing during eye fixations in the reading of prose. *Journal of Verbal Learning and Verbal Behavior*, 23:612–624.
- Irwin, M., Fitzgerald, C., and Berg, W. P. (2000). Effect of the intensity of wireless telephone conversations on reaction time in a braking response. *Perceptual and Motor Skills*, 90:1130–1134.
- Jenness, J. W., Lattanzio, R. J., O'Toole, M., and Taylor, N. (2002). Voice-activated dialling or eating a cheeseburger: Which is more distracting during simulated driving? In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*, Pittsburgh, PA.
- Jurafsky, D. and Martin, J. (2008). *Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall, New Jersey, USA.

- Just, M. and Carpenter, P. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87:329–354.
- Just, M. and Carpenter, P. (1984). Reading skills and skilled reading in the comprehension of text. In Mandl, H., Stein, N., and Trabasso, T., editors, *Learning and comprehension of text*. Erlbaum, Hillsdale, NJ.
- Kahnemann, D. (1973). *Attention and Effort*. PhD thesis, University of Groningen.
- Kaisser, M., Hearst, M., and Lowe, J. (2008). Improving Search Result Quality by Customizing Summary Lengths. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, USA.
- Kamalski, J., Lentz, L., Sanders, T., and Zwaan, R. A. (2008). The forewarning effect of coherence markers in persuasive discourse: Evidence from persuasion and processing. *Discourse Processes*, 45:546–579.
- Kantowitz, B. (1987). Mental workload. In Hancock, P., editor, *Human Factors Psychology*, pages 81–121. North-Holland, Amsterdam.
- Kittur, A., Chi, E., and Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems (CHI '08)*, Florence, Italy.
- Komatani, K., Ueno, S., Kawahara, T., and Okuno, H. G. (2003). Flexible guidance generation using user model in spoken dialogue systems. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 256–263, Morristown, NJ, USA. Association for Computational Linguistics.
- Kruijff-Korbayova, I., Blaylock, N., Gerstenberger, C., Rieser, V., Becker, T., Kaisser, M., Poller, P., and Schehl, J. (2005). Presentation strategies for flexible multi-

modal interaction with a music player. In *Proceedings of DIALOR 9th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*, Nancy, France.

Kruijff-Korabayova, I., Amores, G., Blaylock, N., Ericsson, S., Perez, G., Georgila, K., Kaisser, M., Larsson, S., Lemon, O., Manchon, P., and Schehl, J. (2006). Deliverable 3.1: Extended information state modeling. Talk and Look: Tools for Ambient Linguistic Knowledge, IST-507802 Deliverable 3.1.

Kubose, T., Bock, J., Dell, G., Garnsey, S., Kramer, A., and Mayhugh, J. (2006). The effects of speech production and speech comprehension on simulated driving performance. *Applied Cognitive Psychology*, 20(1):43–63.

Levin, E., Pokrovsky, A., Bocchieri, E., Di Fabrizio, G., Biatov, K., Rahim, M., Walker, M., Ruscitti, P., Pieraccini, R., Lee, S., Narayanan, S., and Eckert, W. (2000). The At&t-darpa communicator mixed-initiative spoken dialog system. In *Proceedings of the 1st International Conference on Spoken Language Processing - (Interspeech '00)*, Beijing, China.

Louwerse, M. (2001). *PhD thesis*. PhD thesis, University of Edinburgh. Institute for Communicating and Collaborative Systems.

Martens, M. and van Winsum, W. (1999). Activation criteria and warning strategies: Driving simulator results. Technical report, TNO Human Factors Research Institute, TNO Soesterberg. IN-ARTE Project (TR4014), ID 7.1 Unpublished technical report.

Martin, D. L., Cheyer, A. J., and Moran, D. B. (1998). Building distributed software systems with the open agent architecture. In *Proceedings of the Third International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology*.

- McCarley, J., Vais, M., Pringle, H., Kramer, A., Irwin, D., and Strayer, D. (2004). Conversation disrupts scanning and change detection in complex visual scenes. *Human Factors*, 46:424–436.
- McCarthy, K., McGinty, L., Smyth, B., and Reilly, J. (2005). A live-user evaluation of incremental dynamic critiquing. In *Proceedings of the 6th International Conference on Case-Based Reasoning (ICCBR-05)*, pages 339–352, Chicago, USA.
- McDonald, S. and Shillcock, R. (2004). Lexical predictability effects on eye fixations during reading. In Carreiras, M. and Clifton, C., editors, *The On-line Study of Sentence Comprehension: Eyetracking, ERP and Beyond*, pages 77–94. Psychology Press, New York.
- Meijman, T. and O' Hanlon, J. (1984). Workload. an introduction to psychological theories and measurement methods. In P.J.D. Drenth, H. Thierry, P. W. and de Wolff, C., editors, *Handbook of Work and Organizational Psychology*, pages 257–288. Wiley, New York.
- Meister, D. (1976). *Behavioral foundations of system development*. Wiley, New York, USA.
- Miller, G. A. (1956). The magical number seven, plus or minus two; some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97.
- Moore, J. D. (2006). Natural language generation for information presentation. Talk at the 1st IEEE/ACL 2006 Workshop on Spoken Language Technology Workshop 2006 (SLT '06).
- Moore, J. D., Foster, M., Lemon, O., and White, M. (2004). Generating tailored, comparative descriptions in spoken dialogue. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS '04)*.

- Nakov, P. (2008). Paraphrasing verbs for noun compound interpretation. In *In Proceedings of the Workshop on Multiword Expressions (MWE '08), in conjunction with the Language Resources and Evaluation conference 2008 (LREC '08)*, Marrakech, Morocco.
- NHTSA (2000). Driver distraction expert working group meetings: Summary and proceedings. Available at: <http://www-nrd.nhtsa.dot.gov/departments/nrd-13/driver-distraction/Welcome.htm>.
- Norman, D. and Bobrow, D. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, 7:44–64.
- O'Donnell, R. and Eggemeier, F. (1986). Workload assessment methodology. In *Cognitive processes and performance*, volume 2, pages 42–49. Wiley, New York.
- Pashler, H. and Johnston, J. (1998). Attentional limitations in dual-task performance. In Pashler, H., editor, *Attention*. Psychology Press, Hove, UK.
- Pazzani, M. J. and Billsus, D. (2007). Content-based recommendation systems. In Peter Brusilovsky, Alfred Kobsa, W. N., editor, *The Adaptive Web*, volume 4321 of *Lecture Notes in Computer Science*, pages 325–341. Springer Verlag, Berlin, Germany.
- Pickering, M. and Van Gompel, R. (2006). Syntactic parsing. In Traxler, M. and Gernsbacher, M., editors, *Handbook of Psycholinguistics (2nd Ed)*. Elsevier, New York, USA.
- Pineau, J., Montemerlo, M., Pollack, M., Roy, N., and Thrun, S. (2003). Towards robotic assistants in nursing homes: Challenges and results. *Robotics and Autonomous Systems*, 42(3-4):271–281.
- Polifroni, J. (2008). *Information Browsing*. PhD thesis, University of Sheffield, UK.

- Polifroni, J., Chung, G., and Seneff, S. (2003). Towards automatic generation of mixed-initiative dialogue systems from web content. In *Proceedings of Eurospeech '03*, pages 193–196, Geneva, Switzerland. AAAI Press.
- Polifroni, J. and Walker, M. (2006a). An analysis of automatic content selection algorithms for spoken dialogue system summaries. In *Proceedings of the 1st IEEE/ACL 2006 Workshop on Spoken Language Technology (SLT '06)*, Aruba.
- Polifroni, J. and Walker, M. (2006b). Learning database content for spoken dialogue system design. In *Proceedings of the International Conference on Language Resources and Evaluation 2006 (LREC '06)*, Genoa, Italy.
- Polifroni, J. and Walker, M. (2008). Intensional summaries as cooperative responses in dialogue: Automation and evaluation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08: HLT)*, pages 479–487, Columbus, Ohio. Association for Computational Linguistics.
- Pollatsek, A. and Rayner, K. (1990). Eye movements and lexical access in reading. In Balota, D., Flores d'Arcais, G., and Rayner, K., editors, *Comprehension Processes in Reading*, pages 143–163. Erlbaum, Hillsdale, NJ.
- Pylkkinen, L. and McElree, B. (2006). The syntax-semantics interface: On-line composition of sentence meaning. In Traxler, M. and Gernsbacher, M., editors, *Handbook of Psycholinguistics (2nd Ed)*. Elsevier, New York, USA.
- R Development Core Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rayner, K. (1998). Eye movements in reading and information processing. *Psychological Bulletin*, 124:327–422.

- Rayner, K., Binder, K. S., Ashby, J., and Pollatsek, A. (2001). Eye movement control in reading: Word predictability has little influence on initial landing positions in words. *Vision Research*, 41:943–954.
- Rayner, K., Sereno, S., Morris, R., Schmauder, A., and Clifton, C. (1998). Eye movements and on-line language comprehension processes. *Language and Comprehension Processes*, 4:21–49.
- Rayner, K., Sereno, S. C., Morris, R. K., Schmauder, A. R., and Clifton, C. (1989). Eye movements and on-line language comprehension processes. *Language and Cognitive Processes*, 4:21–50.
- Recarte, M. A. and Nunes, L. M. (2000). Effects of verbal and spatial-imagery tasks on fixations while driving. *Journal of Experimental Psychology: Applied*, 6:31–43.
- Reed, S. K. (1996). *Cognition: Theory and applications (4th ed.)*. Brooks/Cole, Pacific Grove, CA, USA.
- Reiter, E. (1994). Has a consensus NL generation architecture appeared, and is it psycholinguistically plausible? In *Proceedings of the 7th International Workshop on Natural Language Generation (INLGW '94)*.
- Sanders, T., Land, J., and Mulder, G. (2007). Linguistic markers of coherence improve text comprehension in functional contexts. *Information Design Journal*, 15(3):219–235.
- Sanders, T. and Noordman, L. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, 29(1):37–60.
- Schneider, W. and Detweiler, M. (1988). The role of practice in dual task performance: toward workload modeling in a connectionist/control architecture. *Human Factors*, 30(5):539–566.

- Schultz, R., Peter, C., Blech, M., Voskamp, J., and Urban, B. (2007). Towards detecting cognitive load and emotions in usability studies using the RealEYES framework. In *Proceedings of the HCI International '07 Conference*, Beijing, China.
- Seufert, T. and Brunken, R. (2006). Cognitive load and the format of instructional aids for coherence formation. *Applied Cognitive Psychology*, 20:321–331.
- Sinatra, G. (1990). Convergence of listening and reading processing. *Reading Research Quarterly*, 25:115–130.
- Squire, L. R. (1992). Declarative and nondeclarative memory: multiple brain systems supporting learning and memory. *Journal of Cognitive Neuroscience*, 4:232–243.
- Stent, A. J. (2001). *Dialogue systems as conversational partners : Applying conversation acts theory to natural language generation for task-oriented mixed-initiative spoken dialogue*. PhD thesis, University of Rochester. Department of Computer Science.
- Strayer, D. L. and Johnston, W. A. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular phone. *Psychological Science*, 12:462–466.
- Thompson, C. A., Göker, M. H., and Langley, P. (2004). A personalized system for conversational recommendations. *Journal of Artificial Intelligence Research (JAIR)*, 21:393–428.
- Tietze, M. I., Winterboer, A., and Moore, J. D. (2009). The effect of linguistic devices in information presentation messages on comprehension and recall. In *Proceedings of the 12th European Workshop on Natural Language Generation*, Athens, Greece.
- Tijerina, L., Johnston, S., Parmer, E., Winterbottom, M. D., and Goodman, M. (2000). Driver distraction with wireless telecommunications and route guidance systems.

Technical Report DOT HS 809-069, National Highway Traffic Safety Administration.

- Toney, D. (2007). *Evolutionary Reinforcement Learning of Spoken Dialogue Strategies*. PhD thesis, University of Edinburgh.
- Treisman, A. and Davies, A. (1973). Dividing attention to ear and eye. *Attention and Performance*, 4:101–117.
- Van Berkum, J., Brown, C., Zwitserlood, P., Kooijman, V., and Hagoort, P. (2005). Anticipating upcoming words in discourse: Evidence from erps and reading times. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 31(3):443–466.
- van Winsum, W., Martens, M., and Herland, L. (1999). The effects of speech versus tactile driver support messages on workload, driver behaviour and user acceptance. Technical Report Report TM-99-C043, TNO Human Factors Research Institute, TNO Soesterberg.
- Varges, S., Weng, F., and Pon-Barry, H. (2006). Interactive question answering and constraint relaxation in spoken dialogue systems. In *Proceedings of the 7th SIG-dial Workshop on Discourse and Dialogue*, Sydney, Australia.
- Viappiani, P., Pu, P., and Faltings, B. (2007). Conversational recommenders with adaptive suggestions. In *Proceedings of the ACM conference on Recommender Systems (RecSys '07)*, Minneapolis, MN.
- Walker, M. A., Passonneau, R., and Boland, J. E. (2001). Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL '01)*.

- Walker, M. A., Whittaker, S., Stent, A., Maloor, P., Moore, J. D., Johnston, M., and Vasireddy, G. (2002). Speech plans: Generating evaluative responses in spoken dialogue. In *Proceedings of the International Natural Language Generation Conference (INLG '02)*, pages 73–80.
- Walker, M. A., Whittaker, S., Stent, A., Maloor, P., Moore, J. D., Johnston, M., and Vasireddy, G. (2004). Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28:811–840.
- White, M., Rajkumar, R., and Martin, S. (2007). Towards broad coverage surface realization with ccg. In *In Proc. of the Workshop on Using Corpora for NLG: Language Generation and Machine Translation (UCNLG+MT)*.
- Whittaker, S., Walker, M. A., and Maloor, P. (2003). Should I Tell All? An experiment on conciseness in spoken dialogue. In *Eurospeech: European Conference on Speech Processing*, pages 1685–1689, Geneva.
- Wickens, C. (1984). Processing resources in attention. In Parasuraman, R. and Davies, D. R., editors, *Varieties of Attention*, pages 63–102. Academic Press, London, UK.
- Wickens, C. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 32:150–177.
- Wickens, C. D. (1992). *Engineering psychology and human performance*. Harper Collins, New York.
- Winterboer, A., Hu, J., Moore, J. D., and Nass, C. I. (2007). The influence of user tailoring and cognitive load on user performance in spoken dialogue systems. In *Proceedings of the 8th International Conference on Spoken Language Processing (Interspeech '07)*, Antwerp, Belgium.

- Winterboer, A. and Moore, J. D. (2007). Evaluating information presentation strategies for spoken recommendations. In *Proceedings of the ACM conference on Recommender Systems (RecSys '07)*, Minneapolis, MN.
- Winterboer, A., Moore, J. D., and Ferreira, F. (2008). Do discourse cues facilitate recall in information presentation messages? In *Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech '08)*, Brisbane, Australia.
- Young, K., Regan, M., and Hammer, M. (2003). Driver distraction: A review of the literature. Technical report, Monash University Accident Research Centre.
- Zajicek, M. and Morrissey, W. (2001). Spoken message length for older adults. In *Proceedings of the 8th International Conference on Human-Computer Interactions (INTERACT '01)*, pages 789–790.

Appendix A

Questionnaire - Dual-task experiments

This appendix contains the questionnaires used in the experiments described in Chapters 3, 5, and 5.6. Furthermore, the materials used in the eye-tracking experiment are appended.

A.1 Edinburgh experiment questionnaire

In this appendix, I present the relevant part of the questionnaire used in the Wizard-of-Oz experiment described in Chapter 3. These evaluation questions regarding system understandability, the provided overview and relevance of options, and the system's efficiency are based on the ones used in the reading experiment evaluation by (Demberg and Moore, 2006).

A.2 Stanford experiment questionnaire

This appendix consists of the questionnaire used in both dual-task studies described in Chapters 5 (see Hu et al., 2007) and 5.6 (see Winterboer et al., 2007). This questionnaire was developed in Cliff Nass' CHIME Lab at Stanford University and the evaluation questions I used to compare the experimental results with previous results were added to the existing list of questions.

Instructions:

For this questionnaire, we would like you to think about your experience as a participant in this study.

There are no wrong answers; we are interested in your opinions.

Answer the questions in the order that they appear.

This questionnaire is completely anonymous. The experimenter will only identify you with a number.

Part A

1. How many years have you been driving?
 - a. 1 year or less
 - b. 2 years
 - c. 3 years
 - d. 4 years
 - e. 5 years or more
2. How many accidents have you been in?
 - a. None
 - b. 1 accident
 - c. 2 accidents
 - d. 3 accidents
 - e. 4 or more accidents
3. How many traffic tickets have you received?
 - a. None
 - b. 1 ticket
 - c. 2 tickets
 - d. 3 tickets
 - e. 4 or more tickets
4. On average, how many miles do you drive per week?
 - a. Less than 50 miles
 - b. 51~100 miles
 - c. 101~200 miles
 - d. 201~300 miles
 - e. More than 301 miles
5. Where do you most often drive?
 - a. City
 - b. Suburb
 - c. Country side
6. Have you used a cell phone while driving?
 - a. Yes
 - b. No
7. Do you like driving?
 - a. Yes
 - b. No
8. Do you drive in rush hour traffic?
 - a. Yes
 - b. No
9. Do you take long road trips?
 - a. Yes
 - b. No

Part B

1. Please complete the following statements about the profile (a business traveller) you used to book your flights:

- a. When I travel, my top concern is flying _____ class.
- b. My second most important concerns are _____, _____, and _____.
- c. Finally, I prefer _____ (airlines) because I collect bonus mileage.

2. Now recall the two flights you just booked:

- a. The first flight was from _____ to _____. My final choice was a _____ (direct or indirect) flight with _____ (airlines). It's _____ class and costs _____ dollars.
- b. The second flight was from _____ to _____. My final choice was a _____ (direct or indirect) flight with _____ (airlines). It's _____ class and costs _____ dollars.

3. Please circle the dot that best describes **how you felt while booking flights**.

active	• • • • • • • • • •	passive
dominant	• • • • • • • • • •	submissive
drained	• • • • • • • • • •	invigorated
excited	• • • • • • • • • •	calm
flexible	• • • • • • • • • •	inflexible
focused	• • • • • • • • • •	distracted
frustrated	• • • • • • • • • •	not frustrated
happy	• • • • • • • • • •	unhappy
ignorant	• • • • • • • • • •	knowledgeable
in control	• • • • • • • • • •	not in control
incompetent	• • • • • • • • • •	competent
negative	• • • • • • • • • •	positive
polite	• • • • • • • • • •	impolite
powerless	• • • • • • • • • •	powerful
productive	• • • • • • • • • •	unproductive
rigid	• • • • • • • • • •	not rigid
skilled	• • • • • • • • • •	unskilled
successful	• • • • • • • • • •	unsuccessful
tense	• • • • • • • • • •	relaxed
uncooperative	• • • • • • • • • •	cooperative

unfriendly • • • • • • • • • • friendly
unintelligent • • • • • • • • • • intelligent

4. Please circle the dot that best describes **how you felt when driving**.

active	• • • • • • • • • •	passive
dominant	• • • • • • • • • •	submissive
drained	• • • • • • • • • •	invigorated
excited	• • • • • • • • • •	calm
flexible	• • • • • • • • • •	inflexible
focused	• • • • • • • • • •	distracted
frustrated	• • • • • • • • • •	not frustrated
happy	• • • • • • • • • •	unhappy
ignorant	• • • • • • • • • •	knowledgeable
in control	• • • • • • • • • •	not in control
incompetent	• • • • • • • • • •	competent
negative	• • • • • • • • • •	positive
polite	• • • • • • • • • •	impolite
powerless	• • • • • • • • • •	powerful
productive	• • • • • • • • • •	unproductive
rigid	• • • • • • • • • •	not rigid
skilled	• • • • • • • • • •	unskilled
successful	• • • • • • • • • •	unsuccessful
tense	• • • • • • • • • •	relaxed
uncooperative	• • • • • • • • • •	cooperative
unfriendly	• • • • • • • • • •	friendly
unintelligent	• • • • • • • • • •	intelligent

5. How well do each of these adjectives describe **the information system**.

accurate	• • • • • • • • • •	inaccurate
active	• • • • • • • • • •	passive
annoying	• • • • • • • • • •	not annoying
bad	• • • • • • • • • •	good
boring	• • • • • • • • • •	exciting
demanding	• • • • • • • • • •	undemanding
difficult	• • • • • • • • • •	easy
distracting	• • • • • • • • • •	not distracting
dominant	• • • • • • • • • •	submissive
dull	• • • • • • • • • •	stimulating
effective	• • • • • • • • • •	ineffective
efficient	• • • • • • • • • •	inefficient
engaging	• • • • • • • • • •	not engaging
fun	• • • • • • • • • •	not fun
helpful	• • • • • • • • • •	unhelpful
incompetent	• • • • • • • • • •	competent
intelligent	• • • • • • • • • •	unintelligent
interesting	• • • • • • • • • •	uninteresting
knowledgeable	• • • • • • • • • •	ignorant
likeable	• • • • • • • • • •	dislikeable
pleasant	• • • • • • • • • •	unpleasant
powerful	• • • • • • • • • •	powerless
reasonable	• • • • • • • • • •	unreasonable
simple	• • • • • • • • • •	complicated
unfriendly	• • • • • • • • • •	friendly

6. Did the system give the information in a way that was easy to understand?

very hard to understand • • • • • • very easy to understand

7. Did the system give you a good overview of the available options?

very poor overview • • • • • • very good overview

8. Do you think there may be flights that are better options for you that the system did not tell you about?

I think that is very possible • • • • • • I feel the system gave a good
overview of all options that are
relevant for me

9. How quickly did the system allow you to find the optimal flight?

slowly • • • • • • quickly

10. How well do each of these adjectives describe **the course you drove**.

boring	• • • • • • • • • •	exciting
demanding	• • • • • • • • • •	undemanding
difficult	• • • • • • • • • •	easy
distracting	• • • • • • • • • •	not distracting
dull	• • • • • • • • • •	stimulating
simple	• • • • • • • • • •	complicated

11. How do you **normally** drive?

Defensively	• • • • • • • • • •	Offensively
-------------	---------------------	-------------

12. How did you drive **in this round**?

Defensively	• • • • • • • • • •	Offensively
-------------	---------------------	-------------

Part C

1. Once again, please complete the following statements about the profile (a business traveller) you used to book your flights:

- c. When I travel, my top concern is flying _____ class.
- d. My second most important concerns are _____, _____, and _____.
- e. Finally, I prefer _____ (airlines) because I collect bonus mileage.

2. Please circle the dot that best describes **how you felt while booking flights**.

active	• • • • • • • • • •	passive
dominant	• • • • • • • • • •	submissive
drained	• • • • • • • • • •	invigorated
excited	• • • • • • • • • •	calm
flexible	• • • • • • • • • •	inflexible
focused	• • • • • • • • • •	distracted
frustrated	• • • • • • • • • •	not frustrated
happy	• • • • • • • • • •	unhappy
ignorant	• • • • • • • • • •	knowledgeable
in control	• • • • • • • • • •	not in control
incompetent	• • • • • • • • • •	competent
negative	• • • • • • • • • •	positive
polite	• • • • • • • • • •	impolite
powerless	• • • • • • • • • •	powerful
productive	• • • • • • • • • •	unproductive
rigid	• • • • • • • • • •	not rigid
skilled	• • • • • • • • • •	unskilled
successful	• • • • • • • • • •	unsuccessful
tense	• • • • • • • • • •	relaxed
uncooperative	• • • • • • • • • •	cooperative
unfriendly	• • • • • • • • • •	friendly
unintelligent	• • • • • • • • • •	intelligent

3. Please circle the dot that best describes **how you felt when driving**.

active	• • • • • • • • • •	passive
dominant	• • • • • • • • • •	submissive
drained	• • • • • • • • • •	invigorated
excited	• • • • • • • • • •	calm
flexible	• • • • • • • • • •	inflexible
focused	• • • • • • • • • •	distracted
frustrated	• • • • • • • • • •	not frustrated
happy	• • • • • • • • • •	unhappy
ignorant	• • • • • • • • • •	knowledgeable
in control	• • • • • • • • • •	not in control
incompetent	• • • • • • • • • •	competent
negative	• • • • • • • • • •	positive
polite	• • • • • • • • • •	impolite
powerless	• • • • • • • • • •	powerful
productive	• • • • • • • • • •	unproductive
rigid	• • • • • • • • • •	not rigid
skilled	• • • • • • • • • •	unskilled
successful	• • • • • • • • • •	unsuccessful
tense	• • • • • • • • • •	relaxed
uncooperative	• • • • • • • • • •	cooperative
unfriendly	• • • • • • • • • •	friendly
unintelligent	• • • • • • • • • •	intelligent

5. Did the system give the information in a way that was easy to understand?

very hard to understand • • • • • • very easy to understand

6. Did the system give you a good overview of the available options?

very poor overview • • • • • • very good overview

7. Do you think there may be flights that are better options for you that the system did not tell you about?

I think that is very possible • • • • • • I feel the system gave a good
overview of all options that are
relevant for me

8. How quickly did the system allow you to find the optimal flight?

slowly • • • • • • quickly

9. How well do each of these adjectives describe **the course you drove**.

boring	• • • • • • • • • •	exciting
demanding	• • • • • • • • • •	undemanding
difficult	• • • • • • • • • •	easy
distracting	• • • • • • • • • •	not distracting
dull	• • • • • • • • • •	stimulating
simple	• • • • • • • • • •	complicated

10. How do you **normally** drive?

Defensively	• • • • • • • • • •	Offensively
-------------	---------------------	-------------

11. How did you drive **in this round**?

Defensively	• • • • • • • • • •	Offensively
-------------	---------------------	-------------

Age: _____

Gender: Male Female

Are you a native speaker of American English? Yes No

Appendix B

Experimental materials used in the eye-tracking experiment

On the following pages present the materials used in the eye-tracking experiment 7.

B.0.1 Hotels

The rate for a double room at the Worthington is £199. It is five stars, in a central location, and the rooms are beautifully decorated.

The rate for a double room at the Occidental is £132. It is four stars, in a very central location, and the rooms are decently decorated.

The rate for a double room at the Dorian is £189. It is four stars, in a suburban location, and the rooms are nicely decorated.

The rate for a double room at the Worthington is £199. It is five stars, in a central location, and the rooms are beautifully decorated.

The rate for a double room at the Occidental is just £132. It is four stars, in a very central location, but the room decoration is only decent.

The rate for a double room at the Dorian is £189. It is also four stars, but in a suburban

location, and the rooms are nicely decorated.

B.0.2 Restaurants

Messinas price is £22. It has very good food quality, attentive service, and decent décor.

Raymonds price is £34. It has very good food quality, excellent service, and impressive décor.

Alhambras price is £16. It has good food quality, bad service, and plain décor.

Messinas price is £22. It has very good food quality, attentive service, and decent décor.

Raymonds price is £34. It has also very good food quality, but excellent service, and moreover impressive décor.

Alhambras price is only £16. It has good food quality, but bad service, and only plain décor.

B.0.3 Flights

The first flight is on Northwest. It costs £167, the plane is a turboprop, and it is a direct flight.

The second flight is on Air Galapagos. It costs £149, the plane is a jet, and it requires a connection.

The third flight is on Royal Caribbean. It costs £109, the plane is a turboprop, and it requires a connection.

The first flight is on Northwest. It costs £167, the plane is a turboprop, and it is a direct flight.

The second flight is on Air Galapagos. It costs £149, and the plane is a jet. However, it requires a connection.

The third flight is on Royal Caribbean. It costs only £109, but the plane is a turboprop, and it requires a connection as well.

B.0.4 Mobiles phone plans

The first option is a contract with Viago. They offer a next generation mobile phone, 200 minutes call time and 100 free text messages for £39 per month.

The second option is a contract with MobileWorld. They offer an up-to-date mobile phone, 500 minutes call time and 500 free text messages for £29 per month.

The third option is a contract with Callstar. They offer a gadget-free mobile phone, 100 minutes call time and 100 free text messages for £16 per month.

The first option is a contract with Viago. They offer a next generation mobile phone, 200 minutes call time and 100 free text messages for £39 per month.

The second option is a contract with MobileWorld. They offer an up-to-date mobile phone, but 500 minutes call time as well as 500 free text messages for £29 per month.

The third option is a contract with Callstar. They just offer a gadget-free mobile phone, 100 minutes call time and 100 free text messages for only £16 per month.

B.0.5 Digital cameras

The first camera is a Nokota. It features a 2 inch LCD display, 7.1 megapixels, and 3x optical zoom for £180.

The second option is a Reica. It features a 2 inch LCD display, 6 megapixels, and 5x optical zoom for £200.

The third option is a Zuma. It features a 1.7 inch LCD display, 6 megapixels, and 3x optical zoom for £129.

The first camera is a Nokota. It features a 2 inch LCD display, 7.1 megapixels, and 3x optical zoom for £180.

The second option is a Reica. It features also 2 inch LCD display, but only 6 megapixels. However, it offers 5x optical zoom for £200.

The third option is a Zuma. It features just a 1.7 inch LCD display and also 6 megapixels, but just 3x optical zoom for only £129.

B.0.6 Notebooks

The first option is the Tashaba for £1000. It has an Intel CoreDuo processor, is equipped with one Gigabyte RAM and a 120 Gigabyte hard drive.

The second option is the ADD for £1249. It has an Intel Core2Duo processor, is equipped with two Gigabyte RAM and a 100 Gigabyte hard drive.

The third option is the Matsushita for £1099. It has an Intel CoreDuo processor, is equipped with one Gigabyte RAM, and a 80 Gigabyte hard drive.

The first option is the Tashaba for just £1000. It has an Intel CoreDuo processor, is equipped with one Gigabyte RAM and a 120 Gigabyte hard drive.

The second option is the ADD for £1249. It has an Intel Core2Duo processor, and is equipped with two Gigabyte RAM but only a 100 Gigabyte hard drive.

The third option is the Matsushita for £1099. It has an Intel CoreDuo processor, is equipped with just one Gigabyte RAM, and only a 80 Gigabyte hard drive.

B.0.7 Refrigerators

The first fridge is the Coldpoint. Its net fridge capacity is 6.4 cubic feet and the freezer can store up to 3.2 cubic feet. This fridge has an energy efficiency class of A+, no automatic fridge defrost, and costs £259.

The second fridge is the Teko. Its net fridge capacity is 4.7 cubic feet and the freezer can store up to 2.2 cubic feet. This fridge has an energy efficiency class of B, no automatic fridge defrost, and costs £219.

The third fridge is the Frosty. Its net fridge capacity is 7.9 cubic feet and the freezer can store up to 2.8 cubic feet. This fridge has an energy efficiency class of A, automatic fridge defrost, and costs £259.

The first fridge is the Coldpoint. Its net fridge capacity is 6.4 cubic feet and the freezer can store up to 3.2 cubic feet. This fridge has an energy efficiency class of A+, no automatic fridge defrost, and costs £259.

The second fridge is the Teko. Its net fridge capacity is only 4.7 cubic feet and the freezer can only store up to 2.2 cubic feet. This fridge has an energy efficiency class of B, also no automatic fridge defrost, but only costs £219.

The third fridge is the Frosty. Its net fridge capacity is 7.9 cubic feet and the freezer can store up to 2.8 cubic feet. This fridge has an energy efficiency class of A, automatic fridge defrost, and costs also £259.

B.0.8 Rental cars

Your first option is the Targus. It is a compact car for up to 4 passengers with automatic transmission. The car has no air conditioning, is subject to an insurance excess of 600 and costs £42 a day.

Your second option is the Bancia. It is an intermediate car for up to 5 passengers with

manual transmission. The car has air conditioning, is subject to an insurance excess of 800 and costs £75 a day.

Your third option is the Silhouette. It is an intermediate car for up to 5 passengers with automatic transmission. The car has air conditioning, is subject to an insurance excess of 800 and costs £68 a day.

Your first option is the Targus. It is a compact car for up to 4 passengers with automatic transmission. The car has no air conditioning, is subject to an insurance excess of 600 and costs £42 a day.

Your second option is the Bancia. It is an intermediate car for up to 5 passengers but only with manual transmission. The car has air conditioning, is subject to an insurance excess of 800 and costs £75 a day.

Your third option is the Silhouette. It is an intermediate car for up to 5 passengers, too, but with automatic transmission. The car has air conditioning, too, and is also subject to an insurance excess of 800, but costs just £68 a day.

B.0.9 MP3 Players

Hifistore sells the Podstar Giga for £129. It is a stylish MP3 player offering 8 Gigabyte storage, a 2 inch colour screen and a battery life of 24 hours.

Saturn sells the Genius Video for £159. It is a bulky looking MP3 player offering 30 Gigabyte storage, a 2.5 inch colour screen and a battery life of 14 hours.

Readings sell the Zone Player for £199. It is a nice looking MP3 player offering 60 Gigabyte storage, a 1.8 inch colour screen and a battery life of 20 hours.

Hifistore sells the Podstar Giga for £129. It is a stylish MP3 player offering 8 Gigabyte storage, a 2 inch colour screen and a battery life of 24 hours.

Saturn sells the Genius Video for £159. It is a bulky looking MP3 player but offers

30 Gigabyte storage and a 2.5 inch colour screen. However, its battery life is only 14 hours.

Readings sell the Zone Player for £199. It is a nice looking MP3 player offering 60 Gigabyte storage, only a 1.8 inch colour screen and a battery life of 20 hours.

B.0.10 Monitors

Your first option is the Sonyo. It costs £229 and has a 22 inch widescreen display with very good colour fidelity and contrast.

Your second option is the Viewmotion. It costs £199 and has a 19 inch 4:3 display with decent colour fidelity and contrast.

Your third option is the BNC. It costs £289 and has a 22 inch widescreen display with excellent colour fidelity and contrast.

Your first option is the Sonyo. It costs £229 and has a 22 inch widescreen display with very good colour fidelity and contrast.

Your second option is the Viewmotion. It costs only £199 but has just a 19 inch 4:3 display with only decent colour fidelity and contrast.

Your third option is the BNC. It costs £289 and has a 22 inch widescreen display with excellent colour fidelity and contrast.

B.0.11 Cinemas

The movie is shown at the Curzon. It is a midsize cinema close to the city centre with a standard sound system and the ticket price is £6.50.

The movie is also shown at the Screeneo. It is a small independent cinema close to the city centre with a decent sound system and the ticket price is £5.50.

The movie is also shown at the Moviepalace. It is a multiplex cinema in a suburban location with a state-of-the-art sound system and the ticket price is £5.

The movie is shown at the Curzon. It is a midsize cinema close to the city centre with a standard sound system and the ticket price is £6.50.

The movie is also shown at the Screeneo. It is a small independent cinema and also close to the city centre but with just a decent sound system and the ticket price is £5.50.

The movie is also shown at the Moviepalace. It is a multiplex cinema but in a suburban location. It has a state-of-the-art sound system and the ticket price is only £5.

B.0.12 Microwaves

The first microwave is from Heatstar and costs £30. Its capacity is 0.8 cubic feet, its wattage is 800 W and it does not feature a grill.

The second microwave is from Bellion and costs £44. Its capacity is 0.6 cubic feet, its wattage is 700 W and it does not feature a grill.

The third microwave is from Taiwoo and costs £59. Its capacity is 0.6 cubic feet, its wattage is 850 W and it does feature a grill.

The first microwave is from Heatstar and costs £30. Its capacity is 0.8 cubic feet, its wattage is 800 W and it does not feature a grill.

The second microwave is from Bellion and costs £44. Its capacity is only 0.6 cubic feet, its wattage is just 700 W and it does not feature a grill either.

The third microwave is from Taiwoo and costs £59. Its capacity is also only 0.6 cubic feet, but its wattage is 850 W and it does feature a grill.

B.0.13 Bookbags

The price of the Kipling bookbag is £125. It is made of canvas, has three inside compartments, and three outside pockets.

The price of the LLBean bookbag is £45. It is made of canvas, has four inside compartments, and two outside pockets.

The price of the Burberry bag is £450. It is made of sturdy nylon, has no inside compartments, and five outside compartments.

The price of the Kipling bookbag is £125. It is made of canvas, has only three inside compartments, but three outside pockets.

The price of the LLBean bookbag is just £45. It is also made of canvas, has four inside compartments, but only two outside pockets.

The price of the Burberry bag is £450. It is made of sturdy nylon, has no inside compartments, but five outside compartments.

B.0.14 Overcoats

The price of the Topshop overcoat is £165. It is charcoal grey, knee-length, and has a buckle belt.

The price of the Calvin Klein overcoat is £400. It is black, below the knee, and buttons down the front.

The price of the H&M overcoat is £75. It is black and grey checks, hip-length, and closes with a tie belt.

The price of the Topshop overcoat is £165. It is charcoal grey, knee-length, and has a buckle belt.

The price of the Calvin Klein overcoat is £400. It is black, below the knee, and buttons

down the front.

The price of the H&M overcoat is only £75. It is black and grey checks, hip-length, and closes with a tie belt.

B.0.15 Makeup palettes

The price of the Nars palette is £48. It contains three eyeshadows, three lipsticks, and no brushes.

The price of the MAC palette is £35. It contains two eyeshadows, two lipsticks, and two small brushes.

The price of the Benefit palette is £30. It contains three eyeshadows, one lipstick, and two full-sized brushes.

The price of the Nars palette is £48. It contains three eyeshadows, three lipsticks, but no brushes.

The price of the MAC palette is £35. It contains only two eyeshadows and lipsticks, but two small brushes.

The price of the Benefit palette is just £30. It contains three eyeshadows, one lipstick, and two full-sized brushes.