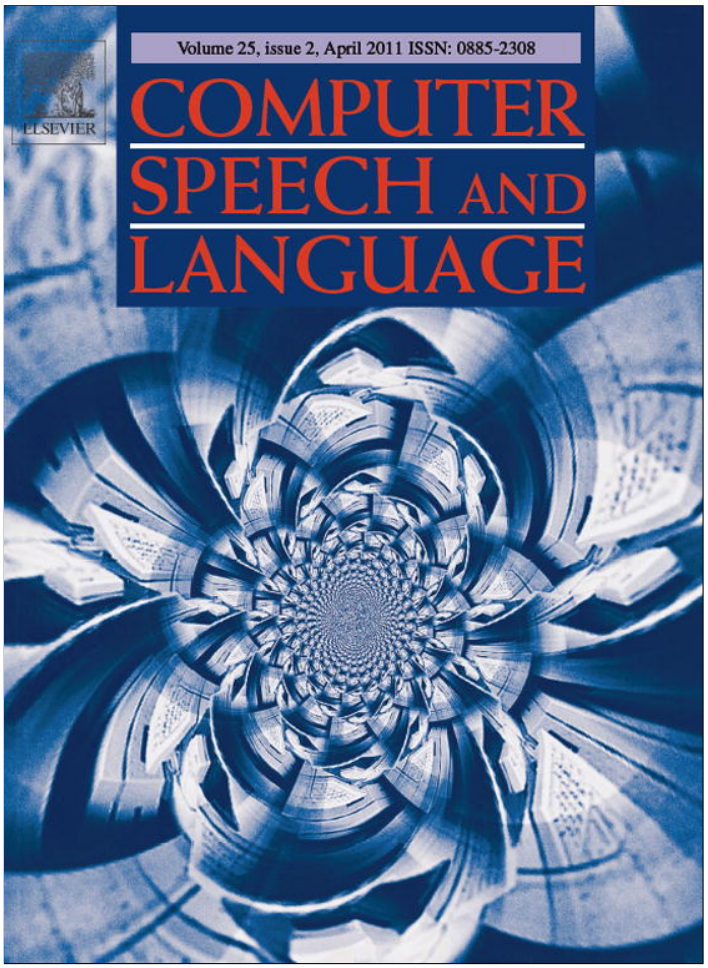

Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language 25 (2011) 175–191
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

The user model-based summarize and refine approach improves information presentation in spoken dialog systems

Andi K. Winterboer [a], Martin I. Tietze [b], Maria K. Wolters [b,*], Johanna D. Moore [b]

[a] University of Amsterdam, Human-Computer Studies Lab, Netherlands
[b] University of Edinburgh, Institute for Communicating and Collaborative Systems, United Kingdom

Available online 7 May 2010

Abstract

A common task for spoken dialog systems (SDS) is to help users select a suitable option (e.g., flight, hotel, and restaurant) from the set of options available. As the number of options increases, the system must have strategies for generating summaries that enable the user to browse the option space efficiently and successfully. In the user-model based summarize and refine approach (UMSR, Demberg and Moore, 2006), options are clustered to maximize utility with respect to a user model, and linguistic devices such as discourse cues and adverbials are used to highlight the trade-offs among the presented items. In a Wizard-of-Oz experiment, we show that the UMSR approach leads to improvements in task success, efficiency, and user satisfaction compared to an approach that clusters the available options to maximize coverage of the domain (Polifroni et al., 2003). In both a laboratory experiment and a web-based experimental paradigm employing the Amazon Mechanical Turk platform, we show that the discourse cues in UMSR summaries help users compare different options and choose between options, even though they do not improve verbatim recall. This effect was observed for both written and spoken stimuli.

Keywords: Information presentation; Spoken dialog systems; User modeling; Discourse markers

1. Introduction

Many spoken dialog systems (SDS) aim to offer efficient and intuitive access to applications and services. Within this context, a common task is to help users select a suitable option (e.g., flight, hotel, and 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 summarizing the options to enable the user to browse the option space.

In this paper, we evaluate two recent approaches to information presentation in SDS, the *summarize and refine (SR)* approach (Polifroni et al., 2003) and the *user-model based summarize and refine (UMSR)* approach (Demberg and Moore, 2006). SR generates summaries by clustering the options to maximize coverage of the domain. The goal is to provide an efficient path through the database. The only way to compare and contrast clusters is by size—anything else requires a model that describes how important certain attributes or attribute/value pairs are to the user. UMSR is

* Corresponding author. Tel.: +44 131 650 6542.
E-mail addresses: A.K.Winterboer@uva.nl (A.K. Winterboer), MTietze@inf.ed.ac.uk (M.I. Tietze), maria.wolters@ed.ac.uk (M.K. Wolters), J.Moore@ed.ac.uk (J.D. Moore).
URLs: http://awinterboer.com (A.K. Winterboer), http://homepages.inf.ed.ac.uk/mwolters (M.K. Wolters), http://homepages.inf.ed.ac.uk/jmoore (J.D. Moore).

0885-2308/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2010.04.003

based on such a user model. Options are not clustered to maximize coverage, but to maximize utility with respect to a user model. Based on the model we can point out trade-offs between choices explicitly using linguistic devices such as discourse cues or discourse adverbials. An initial study using an "overhearer" evaluation methodology showed that the UMSR approach may significantly improve the user's overview of the available options, increase the user's confidence in having heard about all relevant options, allow the user to find a satisfactory option more quickly, and increase overall user satisfaction, when compared to the SR approach (Demberg and Moore, 2006; Demberg et al., submitted for publication). However, in "overhearer" experiments, human raters do not interact directly with the system; instead, they listen to or read canned dialogs.

In order to validate the results of Demberg and Moore (2006), we conducted a Wizard-of-Oz study where users interacted both with SR and UMSR systems (Winterboer and Moore, 2007). We found that presenting relevant options with UMSR significantly decreases dialog duration while increasing task success and user satisfaction. The Wizard-of-Oz experiment is reviewed in detail in Section 3.

Having established that the UMSR approach is superior to the SR approach, we wished to gain a deeper understanding of why this is the case. It could be because UMSR uses attributes that reflect user preferences. Alternatively, it could be that UMSR points out trade-offs explicitly using discourse markers. In this paper, we concentrate on the effect of using *discourse markers*. Discourse markers are linguistic devices that signal relations between syntactic and discourse units. For our purposes, we focus on markers of comparison and contrast that relate options and attributes. We assessed the effect of discourse markers on recall and comprehension in a large web-based listening experiment using Amazon's micro-task platform Mechanical Turk (AMT). The messages were short summaries of three options. All summaries were presented using synthetic speech, since this is the output users of real-world SDS will hear. Participants were exposed to the two main current speech synthesis technologies, unit selection (Hunt and Black, 1996) and HMM-based synthesis (Zen et al., 2009). For half the participants, questions were presented in written form, for the other half, questions were presented in spoken form.

The materials used in the listening experiment had been developed for a lab-based eye tracking study that used written stimuli (Winterboer et al., 2008; Winterboer, 2009). Reusing the materials had two main advantages: it enabled us to validate the web-based experimental paradigm by replicating the original reading experiment on-line, and it allowed us to compare the results of the listening experiment to the outcome of the eye-tracking study. We found that results were broadly similar across modalities and experiment environments. As predicted by the literature, discourse markers of comparison and contrast help users compare different options, but they do not improve verbatim recall.

The remainder of this paper is structured as follows. In Section 2, we present the UMSR approach to information summarisation. In Section 3, we summarize the Wizard-of-Oz study comparing SR and UMSR. The eye-tracking experiment from which the stimuli were taken is summarized in Section 4. Section 5 presents the results of the web-based listening experiment. The results of a pilot experiment validating our experimental methodology are reported in Section 5.2. We conclude in Section 6 by reflecting on the implications of our results for the design of spoken dialog systems.

2. Background

2.1. Recent approaches to information presentation

The usability of a spoken dialog interface depends not only on its ability to recognise and interpret user utterances correctly, but also on its ability to present information in ways that users can understand and remember. This has led to increasing interest in information presentation among SDS researchers and recently, a number of approaches to information presentation have been proposed. Intelligent summarization algorithms are particularly important when users need to choose one of many options that potentially satisfy the current query. In the summarize-and-refine (SR) approach (Polifroni et al., 2003; Chung, 2004; Polifroni and Walker, 2008), the set of available options O is described using attributes that cover a large number of possible choices with a small number of attribute values. The algorithm proceeds as follows. First, all attributes are selected that cover more than $M\%$ of O with less than N values. Next, the members of O are clustered according to the top 2–3 attributes determined in the first step. The largest of the resulting clusters are then selected for presentation. In the original implementation of the SR approach, summaries used all attributes that covered more than 82% of the set of options with less than 4 attributes (Polifroni et al., 2003), in more recent implementations, these values are determined at run-time (Polifroni and Walker, 2008).

	SR dialog	*UMSR dialog*
User:	I would like to go from San Francisco to Prague on June 14th.	
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.	There are no direct flights from San Francisco to Prague, but I found 11 flights with availability in business class. 2 of these are on KLM.
User:	I would like a business class flight, please.	Do you have anything arriving before 3 p.m.?
System:	I found 11 business class flights from San Francisco to Prague. All these flights require a connection. There are flights in the expensive and moderate price range. I also have information about arrival times.	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 2 p.m. with a total travel time of 18 hours and 25 minutes. The second flight arrives at 2:45 p.m. with a total travel time of only 14 hours.
User:	I'd like to arrive before 3 p.m.	I'd like the second flight, please.

Fig. 1. Dialogue between a simulated user and both an SR system and a UMSR system in the flight booking domain.

A sample dialog between a (simulated) user and an SR-type system is shown in Fig. 1, which reproduces the first two system–user interactions from Fig. 1 of Polifroni and Walker (2008). The system responses are based on the generation system used in CRUISER (Becker et al., 2006; Weng et al., 2006). In the first summary, the system has identified three potentially relevant clusters based on two attributes, cuisine and price. In the second summary, the relevant attributes are cuisine and neighbourhood. The system only reports the number of restaurants in the database.

Demberg and Moore (2006) identified three main limitations of the SR approach: attributes may be irrelevant to the user's priorities, information important to the user may only emerge at a late stage in the dialog, and trade-offs are difficult to weigh when there is no optimal option. Take for example the SR dialog in Fig. 1, which is taken from Tables 1 and 2 (Winterboer and Moore, 2007). Here, we have a user who prefers to fly business class, is also concerned about arrival time, travel time, and total number of connections, and would prefer to fly on KLM if possible. In the first response, the SR model mentions an irrelevant attribute, price. In the following turn, the user needs to narrow down the choice to business class flights. Next, the user specifies the next relevant priority, arrival time. At this stage, the user still has had no opportunity to enquire about KLM flights, or to mention that they prefer short travel times. Instead, the system keeps talking about price range, an attribute that is not important to our business user. There is no mechanism within standard SR for indicating trade-offs between choices because the only principled way of comparing two clusters is by the number of options they cover.

Table 1
Task success, efficiency, and user ratings (rating scale: 1–7).

Variable	System		Sig.
	SR	UMSR	
% Optimal flights booked	73.53%	91.18%	$p < 0.05$
Average no. dialog turns	14.53	10.53	$p < 0.001$
Perceived ease of comprehension	5.27	5.79	$p < 0.05$
Overview of options	4.85	5.18	n.s.
Relevance of options	3.76	4.00	n.s.
Perceived efficiency	4.86	5.63	$p < 0.005$

Table 2
Recall, comparison, and evaluation of read UMSR summaries (percent correct).

Task	Method	Marker	Mean	SD	p (Marker)	p (Method)
Recall	Lab	Absent	82.14	17.37		
	Lab	Present	79.05	20.28	<0.46	
	Web	Absent	79.17	14.58		
	Web	Present	81.55	17.10	<0.485	<0.65
Comparison	Lab	Absent	65.00	19.76		
	Lab	Present	70.24	19.73	<0.02	
	Web	Absent	67.26	18.11		
	Web	Present	78.57	17.37	<0.15	<0.08
Evaluation	Lab	Absent	81.19	22.08		
	Lab	Present	86.43	18.86	<0.42	
	Web	Absent	75.00	23.17		
	Web	Present	80.36	19.67	<0.12	<0.04

All three limitations identified by Demberg and Moore (2006) can be addressed if user preferences are taken into account when generating summaries. The user-model (UM) based approach employs a user model and decision theory techniques to identify and present a small number of options that best match the user's preferences (Moore et al., 2004; Walker et al., 2004). Although the UM approach to information presentation enables a concise presentation of a small number of options, highlighting the ways in which those options satisfy user preferences, it does not scale up to presenting a large number of options. When there are many potentially relevant options to consider (e.g., all flights from London to New York), there may be a large number of options that fit the user model, making it difficult to narrow down the number of options. In addition, users may not be able to provide constraints until they are presented with more details concerning the option space. Moreover, the UM approach does not provide an overview of the available options, because options scoring below a specified threshold or below a certain rank are not mentioned. Therefore, there is a risk that users might miss out on options they would have chosen if they had heard about them. The last two problems may reduce user confidence in the system, which may ultimately lead to decreased user satisfaction.

The user-model based summarize and refine (UMSR) approach to information presentation was devised in an attempt to combine the benefits of the UM and SR approaches (Demberg and Moore, 2006). It employs a user model to reduce dialog 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 which orders the options for stepwise refinement based on the ranking of attributes in the user model. Trade-offs between alternative options are presented explicitly, which helps the user assess the relative benefits of each potential choice. In addition, a brief account of all the remaining (irrelevant) options is also provided. Thus, users are provided with a complete overview of the option space, which gives users confidence that they are being presented with all possible options. Thus, the UMSR approach maintains the benefits of user tailoring, while allowing for presentation of large numbers of options in an order reflecting user preferences. The third column in Fig. 1 shows how UMSR leads to shorter, more efficient dialogs. In the first system response, the system only talks about business class flights. Thus, the user can proceed straight to the next important aspect, arrival time. Since both KLM flights fit the specified arrival time, the final options can be summarized. The trade-off between arrival time and travel time is pointed out using the discourse marker "only". Price is not mentioned.

In their initial comparison of UMSR and SR, Demberg and Moore (2006) asked human judges to read six pairs of SR and UMSR dialogs and rate them on the following four criteria:

1. perceived ease of comprehension ("Did the system give the information in a way that was easy to understand?"),
2. overview of options ("Did the system give the user a good overview of the available options?"),
3. 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?"), and
4. perceived efficiency ("How quickly did the system allow the user to find the optimal flight?").

The human judges found the UMSR and SR approaches equally easy to understand and rated the UMSR approach significantly more highly on all other criteria. The study was replicated using an *overhearer* technique where the judges listen to scripted dialogs between a user and a simulated spoken dialog system (Demberg et al., submitted for publication). In these dialogs, system utterances were generated by a speech synthesis system, while the part of the user was played by a male and a female researcher. Again, judges showed a significant preference for dialogs in which the UMSR approach to information presentation was used.

Despite these encouraging results, the judges' perceptions need to be validated further by an experiment where users interact with real SR and UMSR systems. Section 3 describes a WOZ experiment that confirmed the results of Demberg and Moore (2006) and Demberg et al. (submitted for publication).

2.2. Why discourse markers?

A key feature of the UMSR approach is that trade-offs between options are pointed out explicitly using discourse markers. Although a user model is necessary for computing trade-offs between options, these trade-offs can be expressed in a range of different surface forms. In the classical UMSR algorithm as defined by Demberg and Moore (2006), trade-offs are signaled using a specific subset of discourse markers, namely connectives and discourse adverbials that signal contrast and similarity. Discourse markers allow us to express differences in utility in a clear, succinct, and effective way. Consider for example the statement "The first flight arrives at 2 p.m. with a total travel time of 18 h and 25 min. The second flight arrives at 2:45 p.m. with a total travel time of only 14 h." from the second system turn in Fig. 1. If we wanted to express the information contained in *only* using, e.g., predicative constructions, the resulting statement would be longer and clunkier, e.g., "The first flight arrives at 2 p.m. with a total travel time of 18 h and 25 min. The second flight arrives at 2:45 p.m. with a total travel time of 14 h. The advantage of the first flight is its earlier arrival time, the advantage of the second flight is its shorter travel time."

The psycholinguistic literature on discourse cues suggests that they may decrease reading times (Haberlandt, 1982) and speed up the processing of information immediately following the marker (Sanders and Noordman, 2000). Connectives in particular have been shown to aid comprehension of written text and integration of information because they provide explicit information about the relations between discourse segments (Ben-Anath, 2005; Kamalski et al., 2008; Britton et al., 1982; Millis and Just, 1994). The effect of discourse markers on the semantic and pragmatic representation of discourse is less clear (Degand and Sanders, 2002; Sanders et al., 2007). While experiments using free recall tend to show no effect of discourse markers on linguistic representations (e.g., Haberlandt, 1982), experiments that use a task such as question answering might demonstrate a clear benefit of discourse markers (McNamara and Kintsch, 1996; Degand and Sanders, 2002).

In order to assess the effect of discourse markers on the mental representation of UMSR stimuli, we used three tasks, a verbatim recall task, a comparison task, and an evaluation task. Since speed of processing is difficult to measure using spoken stimuli, we measured reading times for summaries with and without discourse markers in an eye-tracking experiment (Winterboer et al., 2008; Winterboer, 2009). This eye-tracking study served as the template for our web-based listening experiments. We not only reused the summaries and tasks that had been prepared for that study, but we also replicated it on the platform selected for our web experiments, AMT, in order to validate our experimental setup.

3. Experiment 1: summarize and refine versus user-model based summarize and refine

3.1. Aims and design

This experiment was designed to determine whether UMSR outperformed SR when participants interacted with both UMSR and SR systems instead of judging read or acted dialogs. Specifically, we hypothesised that a recommender system that generated summaries with the UMSR method would have a higher usability than recommender systems where summaries were generated using the SR method. Our implementation of the SR approach is based on the simple refiner algorithm described in Polifroni (2007) and Polifroni and Walker (2008).

Following ISO 9241 (ISO, 1998), we assessed three aspects of usability, effectiveness (task success), efficiency (dialog duration), and user satisfaction (subjective user ratings). In order to isolate the effect of the information presentation algorithm on dialog system usability, we employed a Wizard-of-Oz (WoZ) methodology, which has been widely used to test components of complex natural-language systems (Dahlbäck et al., 1993). A human wizard

simulated speech recognition and natural language understanding and controlled the dialog, while the system output was generated automatically and presented to the user using text-to-speech.

3.2. Participants

34 participants with an average age of 24 years were paid to participate in the experiment. Half the participants were male, half were female. Most participants were students of the University of Edinburgh. All were naive to the purpose of the experiment. Participants were asked about prior exposure to spoken dialog systems. Two (6%) mentioned that they had been in a car with a voice-based entertainment or navigation system.

3.3. Method and materials

For this experiment, we chose the flight booking domain, which involves complex trade-offs between different options. The human wizard translated users' requests into queries to an SQL database that contained flight information as provided by the travel web site Expedia (www.expedia.co.uk). The query output was converted into text using either the SR or the UMSR strategy. The text was then synthesized using Speechify (Nuance Communications, 2007), a highly intelligible commercial text-to-speech synthesis system. Since users prefer a voice that matches their gender (Nass and Brave, 2005), all participants heard a synthetic voice of their own gender. Participants were encouraged to speak naturally rather than merely responding to system prompts. The wizard used as few questions as possible to elicit reactions from participants. If users remained silent for more than five seconds after information had been provided by the system, they were re-prompted.

All participants sat facing a wall, in front of a desk equipped with a laptop computer, two microphones, and small speakers. The wizard sat on the opposite side of the room, hidden behind a partition that prevented participants from seeing or hearing the wizard during the experiment. The wizard's computer was connected to the speakers and the microphones on the participant's desk via cables running on the floor along the walls of the room in order to avoid attracting attention.

In order to enable reliable and rigorous comparisons, all participants were instructed to play the role of a business traveler for the flight booking task. In descending order of importance, the business traveler

1. preferred flying business class,
2. was equally concerned about arrival time, travel time, and number of stops, and
3. wanted to fly on KLM if possible.

The persona of the business traveller was further illustrated using a short story.

Participants were told that they were going to evaluate two systems. They were instructed to book two flights with each system. For each flight, the task description specified the departure city and the destination city. Having booked both flights, participants were asked to complete a four-item evaluation questionnaire modeled on Demberg and Moore (2006). Each item was rated on a seven-point Likert-type scale. Half the participants interacted with the UMSR system first, while the other half interacted with the SR system first.

To make the booking process more realistic, the flight booking tasks 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. All participants were debriefed after the experiment. Further details about the experimental setup can be found in Winterboer and Moore (2007).

Task success was determined by assessing how often participants had booked the optimal flight given the user model and the constraints of the scenario, efficiency was measured by the average number of dialog turns, and user satisfaction was assessed using the four items on the evaluation questionnaire.

3.4. Results and discussion

Results are summarized in Table 1. Two-tailed *t*-tests were used to assess whether differences between conditions were significant, except for task success, where a χ^2 test was used.

Our results confirm the preferences expressed by the judges in the original rating experiments (Demberg and Moore, 2006; Demberg et al., submitted for publication). When interacting with the UMSR system, users not only were significantly more likely to book the optimal flight, but they also needed substantially fewer turns to do so. The objective gain in efficiency also led to a significantly higher perceived efficiency. For the other three questionnaire items, our results differed from the outcome of Demberg and Moore's overhearer experiment. While users found the UMSR system easier to understand than the SR system, both systems gave an acceptable overview of the available options, and users did not have any strong views on the relevance of the options that had been presented.

4. Experiment 2: the effect of discourse markers on the comprehension of written UMSR summaries

4.1. Aims and design

A key aspect of the UMSR approach is the use of discourse markers that explicitly point out trade-offs among the presented options, rather than requiring users to compute the trade-offs themselves. In this experiment, we assessed to what extent the use of discourse markers facilitates recall and processing of the options presented in a UMSR summary. Users were presented with a series of UMSR summaries that are typical of a stage in the dialog when the relevant choices have been narrowed down to a small number of concrete options. For each summary, reading times were measured using an eye tracker, since previous studies suggest that discourse markers increase the speed of processing, leading to faster reading times (Haberlandt, 1982).

After each summary, users had to complete three tasks:

Verbatim recall: Users were asked to recall the name of one of the options, e.g., *Which restaurant's price is £34?*

Comparison: Users were asked to name the option that best fit a given criterion, e.g., *Which restaurant is the cheapest?*

Evaluation: Users were asked to choose between all options and give a reason for their choice, e.g., *Which restaurant would you like to go to and why?*

If discourse markers help readers build appropriate mental representations of a piece of text, we would expect users to perform better on the comparison and evaluation tasks if the summary contained discourse markers (McNamara and Kintsch, 1996; Degand and Sanders, 2002; Sanders et al., 2007). Conversely, we would not expect to find significant differences in verbatim recall between the two conditions, because recall mainly depends on memory, less on the quality of the underlying mental representation.

4.2. Materials

We created a total of 14 summaries. Each summary presented three options, using two to three sentences per option. Each option was characterised by three salient features in addition to cost. A sample summary is given in Fig. 2. Each summary was associated with a different domain. The 14 domains were rental cars, fridges, book bags, MP3 players, hotels, digital cameras, flights, mobile phone plans, restaurants, make-up palettes, laptops, monitors, microwaves, and cinemas. This guaranteed that almost every participant experienced both familiar and unfamiliar domains. For each summary, we constructed a version with and a version without discourse markers. Fig. 2 shows both versions of the restaurant summary. A full list of summaries can be found in Winterboer (2009).

Unlike the WOZ experiment (Experiment 1), where users were instructed to play the role of a user with a given set of preferences, the participants in this experiment were given no guidance about the options they were supposed to prefer. Instead, we used a generic user model with standard preferences for lower prices, higher food quality, higher service ratings, etc. If there were any differences in attribute values between an option and the previously mentioned options, this was signalled using appropriate discourse markers. All summaries and markers were hand-crafted. In Fig. 2, Raymond is better than Messina in terms of service, and décor. Fuji is worse than both Messina and Raymond for all three relevant attributes, but it is the cheapest of the three restaurants. The attributes themselves and the sequence

Sample Summaries, Restaurant Domain

Without Discourse Markers	*With Discourse Markers*
Messina's price is £22. It has very good food quality, attentive service, and decent décor.	Messina's price is £22. It has very good food quality, attentive service, and decent décor.
Raymond's price is £34. It has very good food quality, excellent service, and impressive décor.	Raymond's price is £34. It **also** has very good food quality, **but** excellent service, and **moreover** impressive décor.
Fuji's price is £16. It has good food quality, bad service, and plain décor.	Fuji's price is **only** £16. It has good food quality, **but** bad service, and **only** plain décor.

Sample Questions

Recall: What is the name of the second restaurant?
Comparison: What restaurant offers only plain décor?
Evaluation: Which restaurant would you prefer and why?

Fig. 2. Sample stimulus with and without discourse markers and corresponding tasks.

in which they were listed were kept constant. The only change between the versions with and without discourse markers was the insertion of markers to make trade-offs explicit.

Each participant saw seven summaries with discourse markers and seven summaries without discourse markers. The presentation of summaries in each domain was counterbalanced so that half the participants responded to a version with discourse markers and half the participant responded to the version without discourse markers.

The answers to the recall and comparison questions were rated correct if the participant selected the correct item. Correct answers to the evaluation question had to consist of one of the three items presented. The description of the item should contain at least two correct attributes. However, the phrasing of the evaluation question often led participants to select the cheapest option and give price as a reason, e.g., "The Fuji, because it was the cheapest.". As these were two attributes (name and price), answers of this type were considered correct. We attributed the occurrence of this type of answer to the application of a strategy to reduce memory load, as it requires only a comparison of the prices but none of the other attributes, of which there were always three or four, used to describe an item. This might have facilitated the task and in turn diminished the effect of the linguistic markers as they are considered to reduce processing costs associated with understanding and interpretation (Millis and Just, 1994; Sanders and Noordman, 2000; Ben-Anath, 2005).

4.3. Participants and method

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.

We used the SR Research Experiment Builder software and an EyeLink II eye-tracker[1] to implement the experiment and present the materials. In each trial, the summary was presented for 45 s or until the participant pressed "Enter".

After reading each summary, participants were presented sequentially with the three questions described above. The recall question was always first, followed by comprehension and evaluation. Summaries with and without discourse markers alternated. Three random orders for the domains were used.

4.4. Results

The experiment results are discussed in more detail in Winterboer et al. (2008) and Winterboer (2009). Here, we summarize the main outcomes that are relevant to our argument. Participants required an average of 37.93 s to read summaries with markers, and an average of 35.28 s to read summaries without markers. A detailed analysis of the

[1] http://www.eyelinkinfo.com/.

reading times of areas of interest showed that the longer reading times are mainly due to the additional words, not to increased sentence complexity.

Although we were unable to replicate the reading time results expected from the literature, task performance results were exactly as predicted. Results for the recall, comparison, and evaluation task are presented in Table 2 ("Lab" context). While users clearly benefit from the presence of discourse markers in the comparison, the explicit cues do not appear to aid verbatim recall or evaluation.

5. Experiment 3: the effect of discourse markers on the comprehension of spoken UMSR stimuli

5.1. Design

In the eye tracking study, we showed that users benefited from the presence of discourse markers in tasks that require them to compare different options. However, in a real-world SDS, users do not read system prompts, they hear them. Although the fundamental mechanisms for processing written and spoken language are the same (Just and Carpenter, 1984; Sinatra, 1990; Carpenter et al., 1995), there are many modality specific factors that also affect comprehension. For example, the decoding of written stimuli is affected both by properties of the reader such as visual acuity and literacy, and by properties of the stimulus, such as font size and contrast (Wickens et al., 2003). Similarly, the processing of spoken stimuli is affected both by properties of the listener such as hearing acuity (Roring et al., 2007; Wolters et al., 2007a) and by properties of the stimulus such as clarity of articulation and prosody (Duffy and Pisoni, 1992; Paris et al., 2000; Wolters et al., 2007b).

Therefore, we replicated the eye tracking experiment with spoken summaries as they would be presented to users in a real SDS. All summaries were produced using synthetic voices, since this is the type of voice that users of a real SDS would be exposed to. In addition to the presence of discourse markers, we varied two further aspects of the experiment, the presentation of the recall, comprehension and evaluation questions (written versus spoken) and the speech synthesis method used for generating the stimuli (unit selection versus statistical parametric synthesis). Presenting the three questions in spoken form ensures continuity, since participants hear the questions in the same voice that was used for generating the summaries. Written presentation, on the other hand, prevents problems due to misunderstandings.

The two speech synthesis methods chosen for this experiment, unit selection (Hunt and Black, 1996) and statistical parametric synthesis (Zen et al., 2009), represent the current state-of-the-art of synthetic speech. Both methods have their drawbacks. While good unit selection speech can sound highly natural due to the lack of post-processing, speech melody is difficult to control, utterances may be distorted by concatenation artefacts, and some sounds may be excessively shortened or lengthened (Black, 2002; Wolters et al., 2007b). Speech generated using statistical parametric synthesis tends to be more intelligible and of a more consistent quality, but has less natural variation and may sound more artificial (Zen et al., 2009). Therefore, it is good practice to expose participants to both speech synthesis methods.

The final experiment design was a $2 \times 2 \times 2$ full factorial design with presence of discourse marker as within-subject factor and synthesis method and modality of questions as between-subject factors. Since the design specifies $2 \times 2 = 4$ different between-subject conditions, we needed to recruit at least 4 times the number of participants of the eye tracking experiment, i.e. $24 \times 4 = 96$ people. Recruiting so many participants for a lab-based study is difficult unless one has access to large student participant pools. Therefore, we decided to conduct the experiment over a web-based platform, Amazon Mechanical Turk (AMT). Although web-based paradigms allow less control over the experimental environment, the results are potentially more realistic, since real users do not typically interact with SDS in distraction-free environments.

AMT is a micro-task platform that allows researchers and developers to put small tasks requiring human intelligence on the web. AMT provides a convenient online payment system, allows experimenters to place restrictions on the eligibility of participants, and makes it easy to recruit participants from outside the usual student population, because it attracts many visitors due to its affiliation with the well established Amazon website. Recently, AMT has been used for a number of linguistic annotation and rating tasks (e.g., Kaisser et al., 2008; Kittur et al., 2008; Snow et al., 2008). However, to the best of our knowledge, no experiments testing comprehension and recall have yet been performed on AMT. Therefore, we conducted a validation study to check whether the platform yielded comparable results to a lab-based version of our experimental paradigm.

5.2. Validating the use of the Amazon mechanical turk platform for information presentation experiments

5.2.1. Participants and methods

For our validation study, we used the same written materials and tasks as in the eye tracking experiment. Since we targeted a US-American audience, prices were changed from pounds sterling (£) to dollars ($). The resulting prices were still realistic. In order to eliminate confounders due to differences in user interface, we also replicated the look-and-feel of the eyetracker interface so that the overall appearance of the visual stimuli used in the web version of the experiment was as similar as possible to the appearance of the stimuli in the laboratory version.

Each summary contained three proper names, N1, N2, and N3, for referring to each of the three options. In order to control for name effects, a third of participants saw a version of the summary where N1 was the target stimulus for recall, a third saw a version where N2 was the target stimulus, and a third saw a version where N3 was the target stimulus. In the restaurant example given in Fig. 2, in one version, the cheapest restaurant would be Fuji, in the second, Messina, and in the third, Raymond. Thus, for each summary, we created a total of 2 (presence/absence of discourse marker) $\times$ 3 (permutation of proper names) versions. In our AMT experiments, each of the six versions was seen by nine participants, yielding a total of 6×9 participants.

We required all participants to be from the United States, since we assumed that visitors to the site from the US were sufficiently skilled English speakers. Participants were only allowed to take part in our experiment once to avoid contamination of results through learning effects. In order to enforce this restriction, we logged several pieces of information about the actual computers accessing the web site. We did not collect any demographic information since it was not clear to us whether a standard questionnaire covering age, gender, and literacy might violate Amazon's privacy policy. All experiments were only made available on the web-site during day-time hours in the US, 8 a.m. PST/11 a.m. EST to 6 p.m. PST/9 p.m. EST.

Each user was paid $2.50 for participation in a 30-min experiment, which is considerably higher than the remuneration offered for most other tasks available on AMT. We hoped that the financial incentive would attract more participants and motivate participants to complete the task more diligently. Participants were only paid if they answered more than 50% of all questions and took more than half the mean time to complete the experiment. This restriction reflects the fact that a minimal amount of time is needed to read and respond to the summaries. If participants need less time, this indicates that they may not have read the summaries.

The user interface was implemented using the *Adobe Flash* format (http://www.adobe.com/flashplatform/), which can be integrated easily into the AMT environment and has been used successfully for administering experiments (Reimers and Stewart, 2007). The technical requirements for running the experiment only excluded very few participants a priori, since Adobe Flash Player has a very high penetration (Adobe, 2008). We ensured that it was impossible for users to revisit a summary once it had been played or read.

Although we largely followed best practice in our implementation (Reips, 2002a,b), several important caveats apply. First of all, there is considerable sampling bias because participants were likely to be highly computer literate. Secondly, we did not gather demographic information because we were unsure whether this was permitted under Amazon's privacy policy. We were unable to ensure that our participants had no previous experience of speech technology, and it was not possible to ascertain that participants were native speakers of English. Finally, it was impossible to control the environment in which participants completed the task. We hypothesized that the potential presence of distractions might lead to lower performance. We were also unable to ensure that participants did not use memory aids, even though they had been explicitly instructed not to do so.

5.2.2. Results and discussion

Table 2 shows the mean and standard deviation of the number of correct answers for each combination of experimental context (Web/AMT versus Lab/Eye tracking) and stimulus condition (presence/absence of discourse marker). Differences in performance between stimulus conditions were assessed using the Wilcoxon test since the data are not normally distributed. There are no significant differences in variance between paradigms for any of the three tasks (Levene's Test).

Overall, performance on the verbatim recall task and the evaluation task is not affected by the presence of discourse markers. Users in the lab experiment find it significantly easier to compare options on a single attribute if discourse markers are present ($p < 0.02$). The same trend can be observed in the AMT data, but it is not significant ($p < 0.15$). While recall performance is highly similar between lab and web contexts, lab participants perform

slightly worse in the comparison task ($p < 0.08$) and significantly better in the evaluation task than web participants ($p < 0.04$).

In summary, results from the lab-based version and the web-based version of the reading experiment were relatively similar. There are no differences in verbatim recall performance, but the two user groups differ somewhat in their performance on the comparison and evaluation tasks. The absolute difference between mean scores in each condition is 2–8 points (out of a maximum score of 100), which corresponds to a relative difference of 3–10%. Further work is necessary to determine whether such differences are mainly due to the different participant pool accessed through AMT.

5.3. Materials

Since our validation study showed that it was feasible to conduct information presentation experiments on the AMT platform, we proceeded to implement the listening experiment on this platform. We used the modified materials described in Section 5.2.1. The stimuli were synthesized from the summaries and questions used in the lab-based experiment. Abbreviations were spelled out and pronunciations for proper names and words not in the pronunciation lexica were specified. The unit selection synthesiser used was a high-quality commercial system, CereVoice by CereProc (Aylett and Pidcock, 2007).[2] Variant tags were used to eliminate any major synthesis errors. For statistical parametric synthesis, we used the research system described in Yamagishi et al. (2008, 2009). The resulting audio-files were converted into MP3 format, because this format is easy to integrate into Flash-based web applications. After playing a spoken summary, there was a delay of 2 s until the first task, the recall task.

The MP3 audio files containing the materials were around half a megabyte in size and each question file around 50 kilobyte. The amount of data the user had to download for the whole experiment amounted thus to around 8 megabyte. The data was downloaded file by file during the interaction of the user with the system, which meant that there was no waiting period at the beginning of the experiment. This strategy works well on fast connections, but may produce unacceptable delays during the experiment for participants who are on slow connections.

5.4. Participants

Participants were again recruited via the AMT website. As in the validation experiment, we prepared six variants per summary. In half of the variants, discourse markers were present, in half, they were absent. In a third of the summaries, the target stimulus was identified by the first of three possible names (N1), in a third, name N2 was used and in the final third, name N3 was used.

For each of the four between-subject conditions, data was collected separately. In each data collection, we initially made 50 tasks available on AMT. Responses were scored and discarded if participants answered at least 5 of all $14 \times 3 = 42$ questions correctly. This corresponds to a threshold of 10%. We assumed that participants who fell below this threshold were likely to be distracted, disinterested, or unable to understand the task. In the second phase, a further, smaller set of tasks was made available which was designed to ensure sufficient numbers for each summary variant. In order to obtain sufficient usable data, a total of $60 + 58 + 59 + 64 = 241$ participants was recruited. Data from 25 people was discarded. The remaining 216 users were distributed evenly across conditions, with nine participants per experiment and stimulus condition.

5.5. Results and discussion

Results were analyzed using the R package *lmer* (Bates and Maechler, 2009). For each of the three target variables, we constructed logistic mixed models with participant ID and stimulus ID as random effects and marker, question modality, and synthesis technology as fixed effects. We also added fixed effects representing the interaction between the presence of discourse markers and task modality, the interaction between discourse markers and speech synthesis technology, and the interaction between speech synthesis technology and task modality. Fig. 6 shows the coefficients of all fixed effects in addition to the intercept, while Table 3 illustrates the observed overall effect of the presence of

[2] http://www.cereproc.com.

Table 3
Effect of presence of marker (overall percent correct).

Marker	Recall	Comparison	Evaluation
Absent	44.18	46.10	52.05
Present	44.51	50.26	55.89

discourse markers on answers. The effect of synthesis technology and task presentation modality on performance is illustrated in Fig. 3 for recall, Fig. 4 for comparison and Fig. 5 for evaluation.

Fig. 6 shows estimates for the coefficients of each of the fixed effects (points) together with 95% confidence intervals for this estimate (lines). If the confidence interval includes 0, then it is possible that the real contribution of the fixed effect to the outcome is close to zero. If the confidence interval excludes 0, however, we can be sure that the effect is significant at least at the $p < 0.05$ level. Fig. 6 also illustrates the relative size of each effect. Highly influential factors will have associated confidence intervals that are far away from 0, whereas the confidence intervals associated with weak effects will be relatively close to the midline. In order to further assess the contribution of each of the fixed effects, we removed the corresponding term from the full mixed model and compared the models with and without the term using an analysis of deviance together with χ^2 significance tests. These more detailed significance levels will be reported throughout the text where appropriate.

The effect of the presence of discourse markers is consistent with the predictions found in the literature (Haberlandt, 1982; Degand and Sanders, 2002) and the results of the eye tracking study (Section 4). As predicted by Haberlandt (1982), verbatim recall performance is not affected by the presence of discourse markers ($p < 0.430$). This changes completely when participants complete the task that discourse markers are designed to facilitate, comparison of trade-offs between options. Now, the presence of discourse markers is the single strongest factor that affects participant performance ($p < 0.030$). Results for the evaluation task are equivocal. Although participants appear to benefit from discourse markers, the effect is not significant ($p < 0.084$).

There are no significant effects of synthesis technology or task modality on the recall task. For the evaluation task, we find that users benefit from seeing written task questions ($p < 0.008$), but the voice that is used for the experiment does not affect performance ($p < 0.564$). The first result is as expected, the second result suggests that the effect of speech synthesis technology is negligible, if high-quality synthesisers are used. On the comparison task, the voice

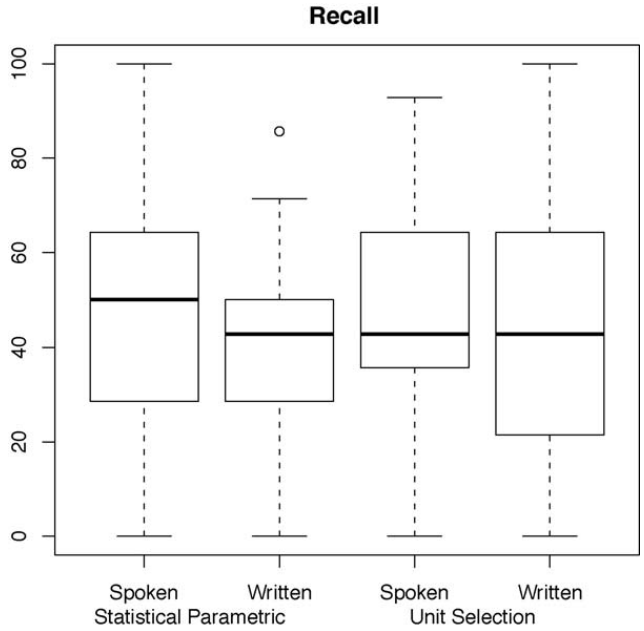

Fig. 3. Percentage of correct answers on the recall task by modality and synthesis technology. Scores are averages per participant. The mid lines indicate the mean, the length of the box corresponds to the range between the first and third quartile (interquartile range, IQR), and the whiskers extend to 1.5 * IQR. Data points outside 1.5 * IQR are represented by points.

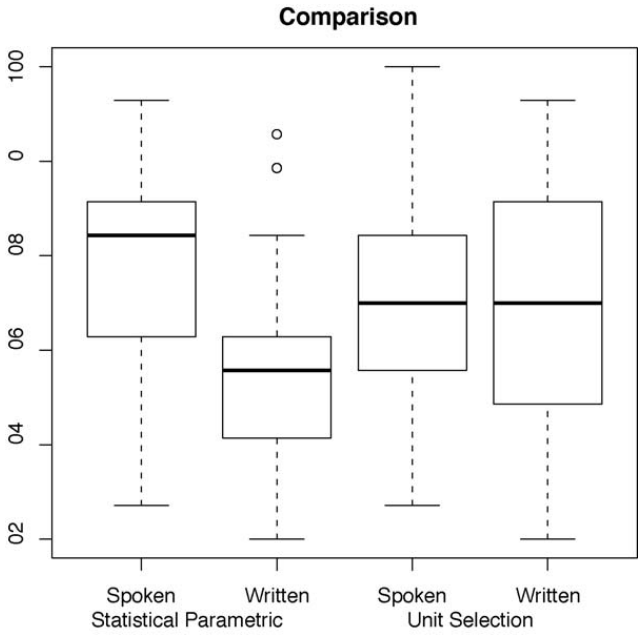

Fig. 4. Percentage of correct answers on the comparison task by modality and synthesis technology. Scores are averages per participant. The mid lines indicate the mean, the length of the box corresponds to the range between the first and third quartile (interquartile range, IQR), and the whiskers extend to 1.5 * IQR. Data points outside 1.5 * IQR are represented by points.

itself also does not significantly affect performance ($p < 0.299$), but there seems to be a problem with the written form of the tasks. Participants consistently perform more poorly when they read the questions instead of hearing them ($p < 0.000$). This is a particular problem for the statistical parametric synthesis voice, as Fig. 4 shows. However, if a summary contains discourse markers, performance is boosted back to normal levels. Without discourse markers, participants answer an average of 2.19 comparison questions correctly when the summary is read using statistical parametric synthesis; with discourse markers, participants get an average of 2.8 questions correct ($p < 0.032$).

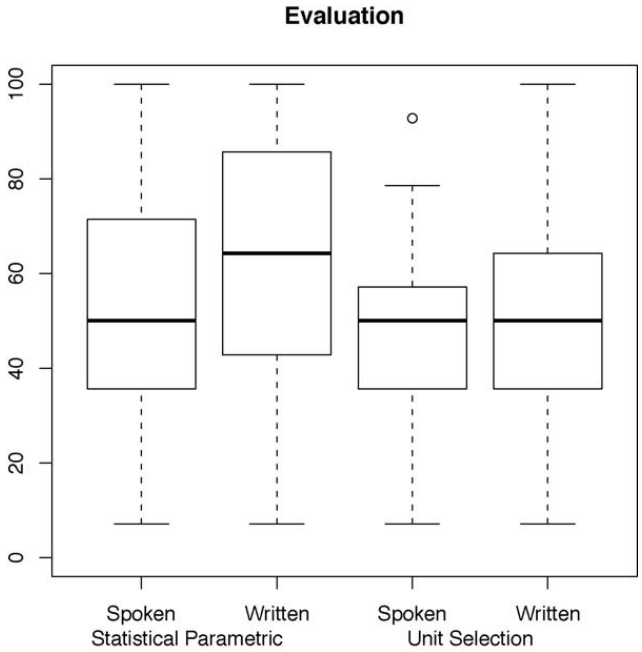

Fig. 5. Percentage of correct answers on the evaluation task by modality and synthesis technology. Scores are averages per participant. The mid lines indicate the mean, the length of the box corresponds to the range between the first and third quartile (interquartile range, IQR), and the whiskers extend to 1.5 * IQR. Data points outside 1.5 * IQR are represented by points.

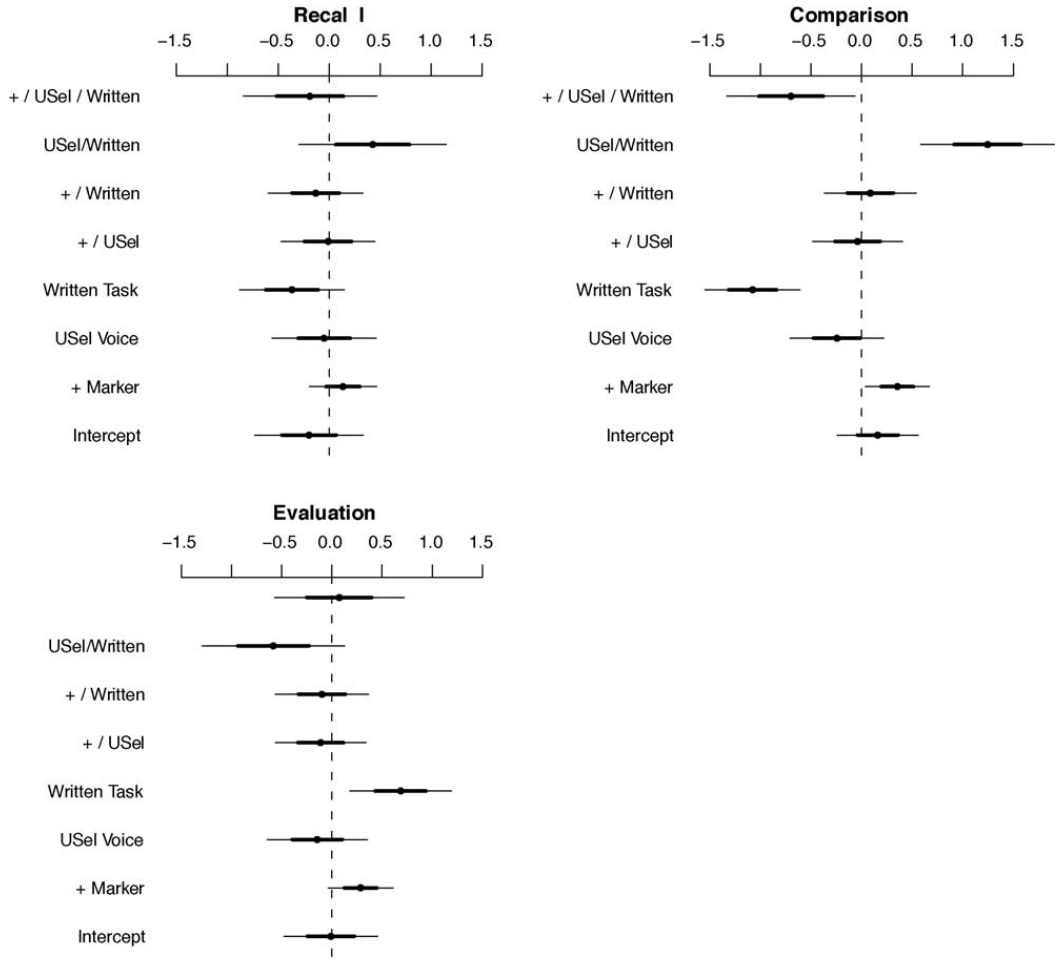

Fig. 6. Results for fixed effects of presence of discourse markers, synthesis method and modality of questions. The dots are the actual coefficient estimates, while the lines correspond to 95% confidence intervals. If the intervals do not include zero, then the corresponding fixed effect can be considered significant at the $p < 0.05$ level.

5.6. Discussion

Our results show that discourse markers do not aid verbatim recall, but they help users compare different options. There is no clear evidence that they may also help users select an optimal choice, although the trends are encouraging. In conditions where discourse markers show a significant effect, overall performance typically improves by 10–15%, if we express the improvement as a percentage of the baseline. This is the outcome we would expect based on the studies reviewed in Section 2.2. Verbatim recall questions can be answered correctly without understanding the relations between the units of the discourse that is presented. Comparison and evaluation questions, on the other hand, rely on a detailed mental representation that allows users to compute the trade-offs between the options that have been presented. It is not surprising that the effect of discourse markers is largest for comparison tasks, since the markers have been chosen to signal contrast and comparison. However, as detailed statistical analysis shows, seemingly innocuous design decisions such as using spoken instead of written questions can have stronger effects on performance than the linguistic organisation of the summary itself, as Fig. 6 illustrates.

Since most of the studies on the effects of linguistic markers on language processing have used written materials (Ben-Anath, 2005), the results of our experiment make an important contribution to the literature on discourse markers. We show that markers of comparison and contrast are not only effective in a traditional psycholinguistic eye-tracking experiment, but also in a listening study where both the voice in which the summaries were presented and the modality used for presenting the tasks varied considerably.

Our results also show that the benefit of discourse markers persists even when we do not assume a specific user model. In Experiments 2 (eye-tracking) and 3 (web-based listening experiment), users were free to judge the three

options presented in each summary according to their own priorities and explain these priorities in their response to the evaluation task. We conclude that even when discourse cues are based on a generic user model that may be at odds with actual user preferences, discourse cues make it easier for users to compare different options.

It is beyond the scope of the present paper to describe how discourse markers are generated depending on the user model, and how values are sorted into desirable and undesirable values. Sample algorithms are described in detail in Moore et al. (2004), Carenini and Moore (2006), and Demberg et al. (submitted for publication). In spoken dialog systems, discourse markers need to be accompanied by appropriate intonation, since the position of pitch accents affects the meaning of some of the discourse markers used in this study, such as "only" (Beaver and Clark, 2003). In White et al. (2010), we discuss how the effect of discourse markers can be enhanced by using appropriate prosodic markers.

6. Conclusions and future work

In this paper, we showed that the user-model based summarize and refine (UMSR) approach clearly outperforms the summarize and refine (SR) approach in terms of task success and dialog duration (Experiment 1). User ratings on our four user satisfaction criteria demonstrate a consistent trend favoring presentations based on the UMSR approach. This is in line with results from previous experiments (Hu et al., 2007; Winterboer et al., 2007).

We then examined one possible reason for the effectiveness of the UMSR approach, explicit indication of trade-offs using discourse markers (Experiments 2 and 3). We found that both for written and spoken summaries, the presence of discourse markers helped users compare options. Importantly, we saw this beneficial effect even though the underlying user model is based on generic potential user preferences instead of specific user priorities. Therefore, when designing information presentation strategies for SDS, we recommend that trade-offs should be made explicit to help users select from a set of available options.

Although UMSR has been successfully implemented in an end-to-end spoken dialog system (Paksima et al., 2009), it has not been compared to an end-to-end SR system. It remains to be seen to what extent the relative benefits of UMSR remain in this context.

The precise algorithm for choosing appropriate discourse markers depends not only on the natural language generation system available, but also on the specificity of the user model. To date, we have used hand-crafted (Winterboer et al., 2008; Winterboer, 2009) or template-based generation (Winterboer et al., 2007; Hu et al., 2007; Winterboer and Moore, 2007). Work on a full natural language generation system using jSPaRKy (Stent and Molina, 2009) and SimpleNLG (Gatt and Reiter, 2009) is in progress.

Discourse markers are not the only way of signalling comparison and contrast relations. This information can also be made explicit using comparative forms of adjectives (e.g., "Raymond is more expensive than Messina. Its price is $32.") or predicative constructions (e.g., "The main advantages of Raymond are its impressive décor and its excellent service."). Although discourse markers are often less explicit than these constructions, they are more concise. It remains to be seen whether more explicit linguistic choices are more beneficial than well-chosen discourse markers.

As we have seen in Section 2, UMSR not only helps users identify trade-offs among options by choosing an appropriate linguistic realization, but also favours a path through the set of possible options that is strongly weighted towards options that will be highly rated by the user. It is not clear which aspect of UMSR has the bigger impact on usability, the optimised path through the set of options, the focus on relevant attributes, or the explicit indication of trade-offs. In this paper, we concentrated on the effect of explicitly indicating trade-offs. The benefits of focusing on relevant attributes and pursuing a more efficient path through the set of options need to be investigated in an experiment similar to our Experiment 1, where users are guided through a number of steps until arriving at the final set of options.

Clearly, UMSR can only be effective if the original user model is sufficiently accurate, and there is a lot of work in the user modeling community on acquiring user models. It is the responsibility of application designers to define relevant user models and ensure that they are appropriate for the intended audience. Given these user models, the spoken dialog interface then needs to provide opportunities for users to state relevant preferences, and then use this information for guiding them through the set of possible choices. This is exactly the kind of navigation that UMSR facilitates. As we move from highly restrictive WOZ scenarios to more realistic implementations of UMSR systems, we will need to work closely with experts in user modeling to ensure that relevant aspects of the user model can be determined efficiently and accurately. In particular, we hope to investigate the effect of mismatches between the system's user model and the user's actual preferences.

References

Adobe, 2008. Flash player penetration. http://www.adobe.com/products/player_census/flashplayer/ (accessed 30.10.08).

Aylett, M., Pidcock, C., 2007. The CereVoice characterful speech synthesiser SDK. In: The 2007 Annual Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB), pp. 174–178.

Bates, D., Maechler, M., 2009. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-31. URL http://CRAN.R-project.org/package=lme4.

Beaver, D., Clark, B.Z., 2003. Always and only: why not all focus sensitive operators are alike. Natural Language Semantics 11 (4), 323–362.

Becker, T., Blaylock, N., Gerstenberger, C., Kruijff-Korbayová, I., Korthauer, A., Pinkal, M., Pitz, M., Poller, P., Schehl, J., 2006. Natural and intuitive multimodal dialogue for in-car applications: The sammie system. ECAI, 612–616.

Ben-Anath, D., 2005. The role of connectives in text comprehension. Working Papers in TESOL and Applied Linguistics 5 (2), 1–27.

Black, A., 2002. Perfect synthesis for all of the people all of the time. In: Proceedings of 2002 IEEE Workshop on Speech Synthesis, Sept, pp. 167–170.

Britton, B.K., Glynn, S.M., Mayer, B.J.F., Penland, M.J., 1982. Effects of text structure on use of cognitive capacity during reading. Journal of Educational Psychology 74, 51–61.

Carenini, G., Moore, J.D., 2006. Generating and evaluating evaluative arguments. Artificial Intelligence 170 (11), 925–952.

Carpenter, P., Miyake, A., Just, M., 1995. Language comprehension: sentence and discourse processing. Annual Review of Psychology 46 (1), 91–120.

Chung, G., 2004. Developing a exible spoken dialog system using simulation. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics Morristown, NJ, USA.

Dahlbáck, N., Jónsson, A., Ahrenberg, L., 1993. Wizard of Oz studies—why and how. Knowledge-Based Systems 6, 258–266.

Degand, L., Sanders, T., 2002. The impact of relational markers on expository text comprehension in L1 and L2. Reading and Writing 15, 739–757. URL http://www.ingentaconnect.com/content/klu/read/2002/00000015/F0020007/05086433.

Demberg, V., Moore, J., 2006. Information presentation in spoken dialogue systems. In: Proceedings of the EACL 2006, pp. 65–72.

Demberg, V., Winterboer, A. K., Moore, J. D., submitted. A strategy for information presentation in spoken dialogue systems. Computational Linguistics.

Duffy, S., Pisoni, D., 1992. Comprehension of synthetic speech produced by rule: a review and iheoretical interpretation. Language and Speech 35, 351–389.

Gatt, A., Reiter, E., 2009. Simplenlg: a realisation engine for practical applications. In: ENLG'09: Proceedings of the 12th European Workshop on Natural Language Generation, Association for Computational Linguistics, Morristown, NJ, USA, pp. 90–93.

Haberlandt, K., 1982. Reader expectations in text comprehension. In: Leny, J., Kintsch, W. (Eds.), Language and Comprehension. North-Holland, Amsterdam, pp. 239–249.

Hu, J., Winterboer, A., Nass, C., Moore, J., Illowsky, R., 2007. Context and usability testing: usermodeled information presentation in easy and diffcult driving conditions. In: Proceedings of the 25th SIGCHI Conference on Human Factors in Computing Systems, ACM Press, New York, NY, USA, pp. 1343–1346.

Hunt, A., Black, A.W., 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In: Proceedings of ICASSP. Vol. 1, Atlanta, Georgia, pp. 373–376.

ISO, 1998. ISO Ergonomic Requirements for Office Work With Visual Display Terminals (VDTs) Part 11: Guidance on Usability.

Just, M., Carpenter, P., 1984. Reading skills and skilled reading in the comprehension of text. In: Mandl, H., Stein, N., Trabasso, T. (Eds.), Learning and comprehension of text. Erlbaum, Hillsdale, NJ.

Kaisser, M., Hearst, M., Lowe, J., 2008. Improving search result quality by customizing summary lengths. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL/HLT).

Kamalski, J., Lentz, L., Sanders, T., Zwaan, R.A., 2008. The forewarning effect of coherence markers in persuasive discourse: evidence from persuasion and processing. Discourse Processes 45, 546–579.

Kittur, A., Chi, E., Suh, B., 2008. Crowdsourcing user studies with Mechanical Turk. In: Proceeding of the Twenty-sixth Annual SIGCHI Conference on Human Factors in Computing Systems.

McNamara, D., Kintsch, W., 1996. Learning from texts: effects of prior knowledge and text coherence. Discourse Processes 22, 247–288.

Millis, K.K., Just, M.A., 1994. The influence of connectives on sentence comprehension. Journal of Memory and Language 33 (1), 128–147.

Moore, J.D., Foster, M.-E., Lemon, O., White, M., 2004. Generating tailored, comparative descriptions in spoken dialogue. In: Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference, pp. 917–922.

Nass, C., Brave, S., 2005. Wired for Speech: How Voice Activates and Advances the Human–Computer Relationship. The MIT Press.

Nuance Communications, 2007. Speechify.

Paksima, T., Georgila, K., Moore, J., September 2009. Evaluating the effectiveness of information presentation in a full end-to-end dialogue system. In: Proceedings of the SIGDIAL 2009 Conference. Association for Computational Linguistics, London, UK, 1–10. URL http://www.aclweb.org/anthology/W09-3901.

Paris, C.R., Thomas, M.H., Gilson, R.D., Kincaid, J.P., 2000. Linguistic cues and memory for synthetic and natural speech. Human Factors 42, 421–431.

Polifroni, J., Chung, G., Seneff, S., 2003. Towards the automatic generation of mixed-initiative dialogue systems from web content. In: Proceedings of the Eighth European Conference on Speech Communication and Technology. ISCA.

Polifroni, J., Walker, M., June 2008. Intensional summaries as cooperative responses in dialogue: automation and evaluation. In: Proceedings of ACL-08: HLT. Association for Computational Linguistics, Columbus, OH, 479–487. URL http://www.aclweb.org/anthology/P/P08/P08-1055.

Polifroni, J. H., 2007. Enabling browsing in interactive systems. Ph.D. thesis, University of Sheffield.

Reimers, S., Stewart, N., 2007. Adobe Flash as a medium for online experimentation: a test of reaction time measurement capabilities. Behavior Research Methods 39 (3), 365–370.
Reips, U.D., 2002a. Internet-based psychological experimenting - Five dos and five don'ts. Social Science Computer Review 20, 241–249.
Reips, U.D., 2002b. Standards for Internet-based experimenting. Experimental Psychology 49, 243–256.
Roring, R.W., Hines, F.G., Charness, N., 2007. Age differences in identifying words in synthetic speech. Human Factors 49, 25–31.
Sanders, T., Land, J., Mulder, G., 2007. Linguistic markers of coherence improve text comprehension in functional contexts. Information Design Journal 15 (3), 219–235.
Sanders, T., Noordman, L., 2000. The role of coherence relations and their linguistic markers in text processing. Discourse Processes 29 (1), 37–60.
Sinatra, G., 1990. Convergence of listening and reading processing. Reading Research Quarterly 25, 115–130.
Snow, R., O'Connor, B., Jurafsky, D., Ng, A., 2008. Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing.
Stent, A., Molina, M., September 2009. Evaluating automatic extraction of rules for sentence plan construction. In: Proceedings of the SIGDIAL 2009 Conference. Association for Computational Linguistics, London, UK, 290–297. URL http://www.aclweb.org/anthology/W09-3941.
Walker, M., Whittaker, S., Stent, A., Maloor, P., Moore, J., Johnston, M., Vasireddy, G., 2004. Generation and evaluation of user tailored responses in multimodal dialogue. Cognitive Science 28, 811–840.
Weng, F., Varges, S., Raghunathan, B., Ratiu, F., Pon-Barry, H., Lathrop, B., Zhang, Q., Bratt, H., Scheideck, T., Xu, K., 2006. CHAT: a conversational helper for automotive tasks. In: Ninth International Conference on Spoken Language Processing.
White, M., Clark, R.A.J., Moore, J.D., 2010. Generating tailored, comparative descriptionswith contextually appropriate intonation. Computational Linguistics 36, 159–201.
Wickens, C., Lee, J., Liu, Y.S., Gordon-Becker, S., 2003. Introduction to Human Factors Engineering, 2nd edition. Prentice Hall.
Winterboer, A., Hu, J., Moore, J., Nass, C., 2007. The influence of user tailoring and cognitive load on user performance in spoken dialogue systems. In: Proceedings of Interspeech.
Winterboer, A., Moore, J., 2007. Evaluating information presentation strategies for spoken recommendations. In: Proceedings of the 2007 ACM Conference on Recommender Systems, ACM Press, New York, NY, USA, pp. 157–160.
Winterboer, A., Moore, J.D., Ferreira, F., 2008. Do discourse cues facilitate recall in information presentation messages ? In: Proceedings of Interspeech.
Winterboer, A.K., 2009. Evaluating information presentation strategies for spoken dialogue systems. Ph.D. thesis, University of Edinburgh.
Wolters, M., Campbell, P., DePlacido, C., Liddell, A., Owens, D., 2007a. The effect of hearing loss on the intelligibility of synthetic speech. In: Proceedings of the 2007 International Congress of Phonetic Sciences, Saarbrücken, Germany, Aug.
Wolters, M., Campbell, P., DePlacido, C., Liddell, A., Owens, D., 2007b. Making synthetic speech accessible to older people. In: Proceedings of the 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, pp. 288–293.
Yamagishi, J., Ling, Z., King, S., 2008. Robustness of HMM-based speech synthesis. In: Proceedings of Interspeech 2008.
Yamagishi, J., Nose, T., Zen, H., Ling, Z., Toda, T., Tokuda, K., King, S., Renals, S., 2009. A robust speaker-adaptive HMM-based text-to-speech synthesis. IEEE Audio Speech and Language Processing 17 (7), 1–15.
Zen, H., Tokuda, K., Black, A.W., 2009. Statistical parametric speech synthesis. Speech Communication 51 (11), 1039–1064.