The effect of linguistic devices in information presentation messages on comprehension and recall

Martin I. Tietze and Andi Winterboer and Johanna D. Moore
University of Edinburgh, Edinburgh, United Kingdom
mtietze@inf.ed.ac.uk, A.Winterboer@ed.ac.uk, J.Moore@ed.ac.uk

Abstract

In this paper we examine the effect of linguistic devices on recall and comprehension in information presentation using both recall and eye-tracking data. In addition, the results were validated via an experiment using Amazon's *Mechanical Turk* micro-task environment.

1 Introduction

In this paper, we present two experiments designed to examine the impact of linguistic devices, such as discourse cues and connectives, on comprehension and recall in information presentation for natural language generation (NLG) as used in spoken dialogue systems (SDS).

Spoken dialogue systems have traditionally used simple templates to present options (e.g., flights, restaurants) and their attributes to users (Walker et al., 2004). Recently, however, researchers have proposed approaches to information presentation that use linguistic devices (e.g., 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, 2006) and contrasts (Winterboer and Moore, 2007). Previous research indicates that linguistic devices such as connectives facilitate comprehension (see Ben-Anath, 2005, for a review). However, to our knowledge, no empirical validation has been performed to test whether using linguistic devices has an effect on comprehension and recall of the information presentated.

2 Experiment 1: Recall of written materials

In order to test whether there are differences in recall, we performed a within-participants reading experiment comparing recall for experiment material presented with or without linguistic devices[1] 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 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 Table 1 and Table 2 for examples. The texts are the type of presentation typically produced by spoken dialogue systems designed to help users select an entity from a set of available options. Participants' eye-movements during reading were recorded as described in section 3.

> 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 1: *Experiment material without discourse cues*

> Messina's price is £22. It has very good food quality, attentive service, and decent décor.
> Ray's price is £34. It has **also** 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 2: *Experiment material with discourse cues*

There were two types of messages, one containing linguistic devices to point out similari-

[1] This experiment has been presented as an one-page abstract, (Winterboer et al., 2008)

Proceedings of the 12th European Workshop on Natural Language Generation, pages 114–117,
Athens, Greece, 30 – 31 March 2009.

ties and differences among the options, and one without these linguistic 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 took particular care to add discourse devices without modifying the propositions in any other way. After each message, the participant had to answer three questions testing different levels of recall. Examples of each type of question are given in figure 3.

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

Figure 3: *The three types of evaluation questions with examples*

2.1 Experimental procedure

In each trial, participants read a text presented for up to 45 seconds on the screen. Users could press *Enter* on the keyboard when they were finished reading. They were then presented with the 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.

2.2 Results

Overall, we found a consistent numerical trend indicating that items in messages containing linguistic devices could be recalled more easily (see Table 2.2). In particular, answers to comparison questions were correctly recalled significantly more often when linguistic markers were present.

	Verb. Q.	Comp. Q.	Eval. Q.
w/o cues	0.79	0.68*	0.73
with cues	0.82	0.79*	0.81

Figure 4: *Average recall on a scale from 0 to 1 for the 3 questions. t-test, "*" indicates a significant difference with $p < 0.5$.*

3 Comprehension of written materials

In this experiment we used an eye-tracker in order to measure reading times, because reading times are considered to be sensitive to people's ongoing discourse processing/comprehension (Haviland and Clark, 1974). We found that reading the presentation messages containing linguistic devices took generally slightly longer, with participants reading messages containing discourse cues taking 37.93 seconds per message on average, and messages without discourse cues taking 35.28 seconds on average to read. The question, however, was whether this difference could be attributed exclusively to the number of additional words or whether readers also spent more time to build a mental representation of the presentation's content by reading the parts marked by discourse cues more carefully. Alternatively, sentence complexity might also increase with the introduction of linguistic cues, which in turn increases reading times. In order to answer this question, we compared the reading times of interest areas (IA) located directly (one word) after the (potential) location of the discourse marker. In total, we determined 46 IAs within the 14 messages, each one consisting of two words or around nine characters on average.

3.1 Results

The results of the different reading time measures, established with linear-mixed effects model (LME) analyses in R[2] (see Table 1), do not reveal any significant differences between the two conditions, although, surprisingly, IAs had a numerically shorter reading time when linguistic 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 linguistic devices. We compared first pass and remaining pass reading times per IA, the total number of passes, and regressions in and out of the IA.

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

[2] `www.r-project.org`

	RT	FPRT	NoP	RegrIn	RegrOut
with cues	473.83	1055.56	3.639	0.430	0.322
w/o cues	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 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 discourse cues*

4 Experiment 2: Web-based recall of written materials

We carried out a web-based user study on Amazon's *Mechanical Turk*[3] (MT) platform both in order to verify the results obtained in the previous recall experiment 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. We recruited native English speakers online to carry out the same experiment previously conducted in the lab. MT is a web-based micro-task platform that allows researchers and developers to put small tasks requiring human intelligence on the web. Deploying MT 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 usual 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, MT has recently been used in a number of language experiments (e.g., Kaisser et al., 2008; Kittur et al., 2008).

4.1 Participants

We had 60 participants reading the same materials that were used in experiment 1. MT does allow to place restrictions on participant location (only users from the US were allowed to participate to ensure English language skills), for instance, or the number of trials (each participant was only allowed to participate once). However, one cannot balance gender of participants 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. We hoped that the higher reward would encourage participants to take the task more seriously.

[3] `https://www.mturk.com/mturk/`

4.2 Experimental setup and procedure

In order to resemble the interface that was used in the previous 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 MT environment easily and allows for tighter user control 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 HMTL users would have been able to use their browser's back button to do just that. The experiment was then made available to the users on Amazon's MT website. The procedure was otherwise exactly the same as in experiment 1.

4.3 Results

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 MT 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 lab-based experiment. Average recall rate was nearly identical with 0.76 (web-based) and 0.77 (lab-based). In addition, the average answer time was also almost identical 23 minutes (web-based) and 26 minutes (lab-based) per participant. However, the results from three of the 60 participants had to be excluded from the analysis (and payment withheld), as they answered less than 50% of the questions while performing the task in less than

half of the average time.

We did not find an effect on the comparison questions. 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 linguistic markers facilitates recall of information.

	Verb. Q.	Comp. Q.	Eval. Q.
w/o cues	0.83	0.62	0.83*
with cues	0.80	0.65	0.88*

Figure 5: *Average recall on a scale from 0 to 1 for the 3 questions in the web-based experiment. t-test, "*" indicates a significant difference with* $p < 0.5$.

5 Discussion and outlook

Taken together, we found a small but significant effect of discourse cues on recall. The combination of eye-tracking and recall data seems to provide a relatively clear picture: Although sentences with linguistic devices took more time to read, this is exclusively due to the additional words and not caused by a differences in the construction of the internal representation. While these findings are in line with results from psycholinguistics which demonstrated that linguistic devices may improve comprehension and recall (Ben-Anath, 2005), given the small effect, it does not fully explain the improvements in terms of task effectiveness found in information presentation for SDS (Winterboer and Moore, 2007).

We additionally validated the results using participants recruited online. The similar results show that this 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 SDSs 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 (Sinatra, 1990). However, to confirm this, we are repeating the experiments in order to examine whether linguistic devices also facilitate recall and comprehension in auditorily presented messages, using stimuli created with a speech synthesiser. We plan to use the auditory moving window paradigm (Ferreira et al., 1996) to assess the impact of lingustic devices in this modality in more detail.

References

D. Ben-Anath. 2005. The Role of Connectives in Text Comprehension. *Working Papers in TESOL and Applied Linguistics*, 5(2):1–27.

F. Ferreira, JM Henderson, MD Anes, PA Weeks, and DK McFarlane. 1996. Effects of Lexical Frequency and Syntactic complexity in Spoken-Language Comprehension: Evidence From the Auditory Moving-Window Technique. *Journal of experimental psychology. Learning, memory, and cognition*, 22(2):324–335.

S.E. Haviland and H.H. Clark. 1974. What's new? acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behaviour*, 13:512–521.

Michael Kaisser, Marti Hearst, and John Lowe. 2008. Improving Search Result Quality by Customizing Summary Lengths. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*.

J. Polifroni and M. Walker. 2006. Learning database content for spoken dialogue system design. In *5th International Conference on Language Resources and Evaluation (LREC)*.

G.M. Sinatra. 1990. Convergence of listening and reading processing. *Reading Research Quarterly*, 25:115–130.

Marilyn A. Walker, Steve Whittaker, Amanda Stent, Preetam Maloor, Johanna D. Moore, Michael Johnston, and Gunaranjan Vasireddy. 2004. Generation and evaluation of user tailored responses in multimodal dialogue. *Cognitive Science*, 28:811–840.

Andi Winterboer and Johanna D. Moore. 2007. Evaluating information presentation strategies for spoken recommendations. In *Proceedings of the ACM conference on Recommender Systems (RecSys '07)*.

Andi Winterboer, Johanna D. Moore, and Fernanda Ferreira. 2008. Do discourse cues facilitate recall in information presentation messages? In *Proceedings of the 9th International Conference on Spoken Language Processing*.