The influence of written task descriptions in Wizard of Oz experiments

Heidi Brøseth
Department of Language
and Communication Studies
Norwegian University of
Science and Technology
NO-7491 Trondheim
broseth@hf.ntnu.no

Abstract

In this paper we investigate an assertion made by Richards and Underwood (1985), who claim that people interacting with a spoken information retrieval system, structure their information in such a uniform manner that this regularity can be used to enhance the performance of the dialog system. We put forward the possibility that this uniform ordering of information might be due to the design of the written task descriptions used in Wizard of Oz experiments.

1 Introduction

As noted in for instance Richards and Underwood (1985), Jönsson and Dahlbäck (1988) and Wooffitt et al. (1997), people tend to behave differently when interacting with a machine as opposed to a human being. These differences are found in various aspects of the dialog, such as the type of request formulation, the frequency of response token, the use of greetings and the organization of the opening and closing sequences, to mention a few. As a consequence of these findings, so-called Wizard of Oz experiments (abbrev. WOZ) are widely used for collecting data about how people interact with computer systems. In a Wizard of Oz experiment the subjects are led to believe that they are interacting with a computer when they are in fact interacting with a human being (a wizard). The wizard can act as a speech synthesizer, speech recognizer, and/or perform various tasks which will eventually be performed by the future system. It is vital that the subjects really think they are communicating with an implemented system in order to obtain reliable data concerning human-computer interaction. The findings in WOZ experiments can serve as an important guide in further development and design of the system (Dahlbäck et al., 1993).

In this paper we investigate the methodology used in WOZ experiments to see how various factors can influence the results. We will start by introducing the experiment done by Richards and Underwood (1985). Then a similar experiment performed in Trondheim will be presented. The results from this experiment will serve as our stance for questioning the claim made in Richards and Underwood (1985). We will also use data material from human-human dialogs to support our claim.

2 Richards and Underwood (1985): significant regularity in the user utterances

The domain for Richards and Underwood's (1985) experiment was train tables. 48 subjects were asked to carry out 6 inquiry tasks concerning these train tables via the phone. Richards and Underwood (1985) claim that the participants in their WOZ experiment rendered

S. Werner (ed.), *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, Ling@JoY 1, 2006, pp. 30–37
ISBN 952-458-771-8, ISSN 1796-1114. http://ling.joensuu.fi/lingjoy/

the information to the system in the following order: 1) place of departure, 2) place of arrival, 3) day and 4) approximate time of travel. They also investigated how the participants responded to different introductory messages. The conclusion was: "*In all cases the finding was sufficiently well established to provide for a potentially useful means of improving recognition accuracy by allowing recognition probabilities to be appropriately weighted*" (217:1985). This conclusion seems very promising, but there is an important aspect of this experiment to be accounted for. Richards and Underwood (1985) mention that the information necessary to perform the request in the WOZ experiment was included in a written task description given to the participants. Unfortunately, they do not give any examples of such a description. The possible influence that these could yield on the results was neither investigated nor discussed in the paper. It is reasonable to ask whether the regularity in information structure reported by Richards and Underwood (1985) is really caused by a spontaneous and natural way of asking about traveling, or perhaps this uniform ordering of information could be caused by the written task description given to the participants. In order to shed some light on this question, we will present some interesting findings in the Trondheim WOZ experiment, as well as some results from the recorded human-human dialogs.

3 The Trondheim WOZ experi-ment

The Trondheim WOZ experiment (abbrev. TWOZ) is within the domain of bus information, and it was conducted in 2003/2004 (Johnsen et al., 2003). The TWOZ is part of the BRAGE[1]-project, and the aim is to develop a mixed-initiative spoken dialog system. The system was built on an existing written query system called BussTUC (Amble, 2000), i.e. a speech interface and a dialog manager were added. The wizard's task was to act as a "perfect speech recognizer". All other tasks were performed by the BussTUC system, the dialog manager and speech synthesis. 64 participants made 3 phone calls asking about bus information resulting in 192 inquiries. The data material consists of 455 user turns (4063 tokens). The participants were either students or staff at Norwegian University of Science and Technology (NTNU).

As already mentioned, the experiment performed by Richards and Underwood (1985) concerned train tables, hence the domains in the two experiments are comparable to a great extent. Both WOZ experiments were conducted via telephone, and written task descriptions containing different scenarios were given to the participants beforehand. In the TWOZ experiment the descriptions also informed that there were no restrictions regarding the actual spoken formulation of the inquiries.

3.1 Scenario groups

The scenarios were divided into five main groups which gave slightly different instructions to the participants.

Group A
The participants should include all information given in the scenario in one utterance, like in a query system.
Example of scenario from which the user should formulate a request to the system:
Place of departure: Munkvoll
Place of arrival: Kalvskinnet
Time: On Monday. You want to arrive at 16:00.

Group B
The participants should divide their inquiry in several utterances.
Example of scenario from which the user should formulate a request to the system:
Place of departure: Hospitalkirka
Place of arrival: Fagerheim
Time: You want to leave after 13 o'clock.

Group C
The scenario consisted of a short narrative which should be the basis for the inquiry.
Example of scenario:
You are working at Dragvoll and wish to go to a football match at Lerkendal. The game starts at 21:00 but you want to be there in due time to meet some old friends and enjoy the supporter band before the match.

[1] BRAGE is an acronym for Brukergrensesnitt for naturlig tale (User interface for natural speech).

Group D
The participant should alter parts of their original inquiry after receiving an answer from the dialog system.
Example of scenario:
Formulate an inquiry that contains the following information:
Place of departure: Lade
Place of arrival: Saupstad
Time: Tomorrow, after 14:30
You are not satisfied with the answer and ask a question about a later bus.

Group E
The participants were allowed to ask freely.
You should freely formulate an inquiry about bus schedules in Trondheim. Beforehand, consider what information you seek. Remember that you can ask questions about only one bus schedule. You don't have to reveal all the information in your inquiry at once.

3.2 Scenario information ordering

An investigation of the written task descriptions in the TWOZ experiment showed that except from Group E, and two occurrences in group C, all the information that the participants should use in their inquiry, were presented to the participants in the following order:

1) place of departure 2) place of arrival 3) time of travel.

3.3 Information categories

The language data obtained in the TWOZ experiment was then divided into four main information categories, based on semantic content that was regarded as vital information to the future dialog system.

a. Main category PLACE contains subcategories DEPARTURE and ARRIVAL.
b. Main category TIME contains subcategories EXACT TIME, PERIOD, INDICATION OF TIME.
c. Main category DAY contains subcategories D(AY)-SPECIFIC, D(AY)- RELATIVE and DAYS.
d. Main category BUS contains subcategories X(BUS), BUS NUMBER, X(BUS NUMBER)[2].

Category (a) contains references to places, usually names of bus stops[3]. Example of the two subcategories in (a) are given in the following.

Jeg skal fra Fiolsvingen til Ugla.
"I am going [from Fiolsvingen] [to Ugla]."
[DEPARTURE] [ARRIVAL]

Category (b) includes phrases referring to the exact time of the day. These are labeled EXACT TIME. References to parts of the day, such as morning, evening, early or late are assembled in PERIOD. Temporal expressions like *as soon as possible* and *now* are gathered in subcategory INDICATION OF TIME. Category (c) contains phrases referring to days like *today* and *tomorrow*, labeled DAY RELATIVE. All seven proper names like *Monday, Tuesday,* etc. are in the subcategory DAY SPECIFIC. The use of the phrases w*eekends* and *weekdays* are assembled in subcategory DAYS. Examples of the subcategories in (b) and (c) are given in the following.

Jeg skal være på Lade halv ti på lørdag.
"I ought to be at Lade [half ten] [on Saturday]."
[EXACT TIME][D-SPECIFIC].

Når går første buss fra Ila i morgen tidlig?
"When does the first bus leave from Ila [tomorrow] [early]?"
[D-RELATIVE][PERIOD].

Jeg vil til Være så fort som mulig.
"I want to go to Være [as soon as possible]."

[2] The categories Bus number and X(Bus number) were not found in the TWOZ material, hence the examples are from the human-human dialogs.
[3] In this paper, the terms DEPARTURE and ARRIVAL are exclusively used for referring to localities.

[INDICATION OF TIME]

Når går siste buss til Lade i helgene?
"When does the last bus to Lade leave [in weekends]?"
[DAYS]

Category (d) includes references to buses, i.e. phrases like *first/last/next bus* or *bus number four*, as in the following examples.

Når går neste buss til Studenterhytta?
"When does the [next bus] to Studenterhytta leave?"
[X(BUS)]

Jeg vil vite når 46eren går fra Tiller.
"I want know when [the 46] leaves from Tiller."
[BUS NUMBER]

Jeg vil vite når neste 24 går fra Tunga.
"I want know when [next 24] leaves from Tunga."
[X(BUS NUMBER)]

The categories described above were then plotted according to where they were located in the opening sequence relative to the other categories. The opening sequence equals the subjects' first turn in the dialog.

We will not go into details about all the subcategories in all the main groups since this is not relevant for this paper. The investigation was limited to the categories (a) and (d), namely DEPARTURE, ARRIVAL and BUS. This selection was based on Richards and Underwood's claim that the distribution of place of departure as the first piece of information and place of arrival as the second was so significant that is could be used to enhance the performance of the system. The category BUS is selected due to some interesting patterns emerging when comparing the TWOZ material with a corpus containing 106 human-human dialogs (abbrev. H-H).

4 The results

4.1 Dispersion of Departure-Arrival in the TWOZ experiment

The categories DEPARTURE and ARRIVAL are the overall most frequently used categories in the opening sequence, cf. Figure 1.

Figure 1: Dispersion of departure-arrival in the opening sequence in the TWOZ material.

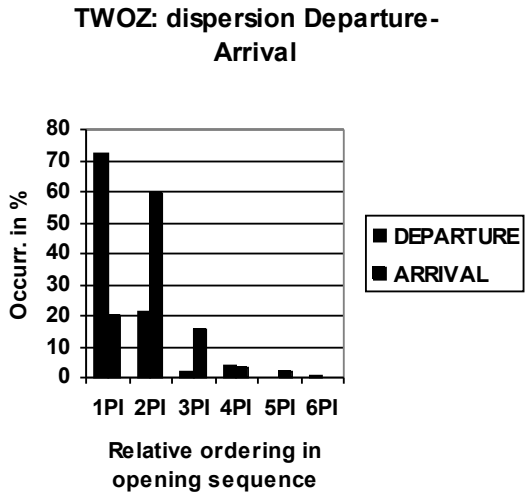


Figure 1 clearly shows that DEPARTURE is most frequently used as the first piece of information (abbrev. 1PI) with 72% of the occurrences of DEPARTURE emerging in this position. Only 21% of the occurrences of category DEPARTURE occur as the second piece of information (abbrev. 2PI). (The remaining 7% are divided amongst the remaining positions in the utterance.) As we can see, the difference between DEPARTURE as the 1PI and the 2PI is clearly significant.

Occurrences of the ARRIVAL, on the other hand, emerge more frequently as the 2PI with 59% in contrast with 21% as the 1PI. (15% of ARRIVAL is found as the 3PI.)

The picture rising from the TWOZ experiment largely coincides with the findings in Richards and Underwood (1985), and could be taken as supportive evidence to their claim. However, there is one important issue to be noted. As mentioned in section 3.2, there is a uniform ordering of the written task descriptions given to the participants, and the ordering of DEPARTURE and ARRIVAL illustrated in

Figure 1 coincides with the ordering of information categories found in the descriptions. In other words, the strong tendency for subcategory DEPARTURE to occur as the 1PI in the opening sequence may be due to influence from the written task description. The occurrences of ARRIVAL crowding together in the 2PI do also follow the pattern from the task description. Thus, the ordering of DEPARTURE and ARRIVAL might not be a result of human beings spontaneously presenting their inquiries about travels in a particular order, but might be due to influence from the written task descriptions.

4.2 Dispersion of Departure-Arrival in the human-human dialogs

A collection of H-H dialogs were recorded at the manual bus information service in Trondheim in 1996. This service is public available and the H-H dialogs consist of a randomly chosen sample of 106 phone calls. The callers did not get any instructions or information before their inquiries.

The observation described in 4.1 does not prove the hypothesis that written task descriptions influence the participants in a WOZ experiment, since the congruent tendency of ordering information in both the TWOZ experiment and Richards and Underwood's experiment (1985) can also be interpreted as a natural inclination for humans to structure these kinds of inquiries in a specific way. In order to test the above hypothesis, we compared the TWOZ material with the H-H dialogs to see if the same regularity in the information pattern was found here. If humans prefer to order their inquiry in a specific way when asking about traveling, we should expect the same pattern to emerge here as in the TWOZ experiment. Figure 2 shows the occurrences of DEPARTURE and ARRIVAL in the H-H dialogs.

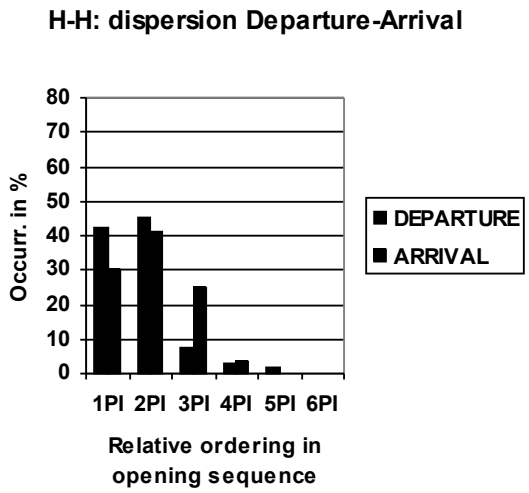


Figure 2: Distribution of Departure-Arrival in the opening sequence in the H-H dialogs.

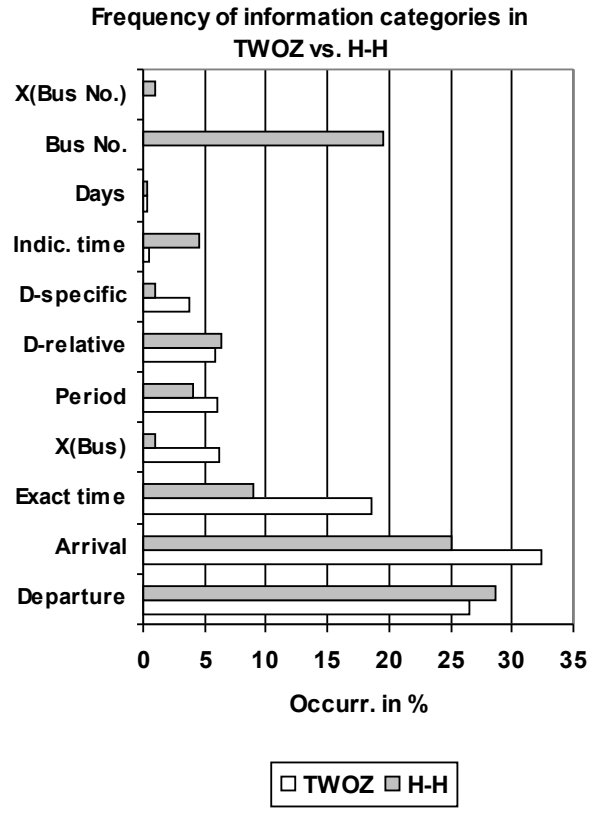
The categories DEPARTURE and ARRIVAL are still the overall most frequently used categories, but the data material shows a rather different distribution. DEPARTURE is now slightly more frequent as the 2PI (42% vs. 45%), and do not follow the pattern discovered in Figure 1 (72% vs. 21%). DEPARTURE is now actually more common than ARRIVAL both as the 1PI and the 2PI. ARRIVAL is still most frequently found as the 2PI compared to its occurrences in the 1PI, but the difference which was 38% in the TWOZ is now decreased to 11%. Both the TWOZ material and the H-H dialogs display a predominance of ARRIVAL as the 3PI. Figure 2 weakens the hypothesis that humans tend to follow a "pre-established" ordering when inquiring about time tables since the regularity in the H-H dialogs is not the same as shown in Figure 1. The findings in Figure 2 can also support the hypothesis that written task descriptions influence the inquiries made in a WOZ experiment. The callers in the H-H dialogs did not have any user instructions on how to conduct an inquiry, and this might have caused a greater diversity in the dispersion of the ordering of the information in these dialogs.

4.3 Frequency of information categories

In addition to investigate the dispersion of the categories DEPARTURE and ARRIVAL in the TWOZ and the H-H dialogs, we looked at the overall frequency of the various categories in the opening sequence. This investigation yielded some interesting insights particularly eye-

catching with respect to the category BUS. Figure 3 shows the overall frequency in opening sequence in the TWOZ material and H-H material.

Figure 3: The overall frequency of information categories in the opening sequence in the TWOZ and the H-H dialogs.



There is a noticeable difference in the use of the category EXACT TIME which is twice as common in the TWOZ material as in the H-H material. Questions about X(BUS) show a similar tendency. This category amounts to only 0.9% of the information categories in the H-H, while it is the forth most frequent category in the TWOZ material with 6.2%. DAY SPECIFIC is also more frequently used in the TWOZ material than in the H-H material. The opposite tendency is found in the categories INDICATION OF TIME, and BUS NUMBER/X(BUS NUMBER) which is more frequent in the H-H. The most eye-catching difference is use of the category BUS NUMBERS which is the third most frequent category in the H-H dialogs (almost as frequent as ARRIVAL) but non-existing in the TWOZ material. This striking difference is extremely odd considering that the domain of both corpora is the very same bus schedules in Trondheim. If the use of BUS NUMBER is the third most frequently used category in the H-H dialogs, why are there zero occurrences in the TWOZ material? The difference can be explained if we again look at the written task description given to the participants. The possibility of using bus number as a strategy to obtain information was not described neither exemplified in any of the scenarios, and it is reasonable to claim that this is why the bus numbers are absent in the TWOZ dialogs[4].

The fact that occurrences of the category BUS NUMBER was so frequent in the H-H dialogs while never used in the opening sequence in the TWOZ dialogs sustains our claim that written task descriptions do influence participants in a WOZ experiment.

5 Rigid structure restricted to computer-oriented talk?

As mentioned in the introduction, people behave differently when interacting with a computer. Further investigation might show that people actually order their inquiries in a much more rigid and uniform manner when the interlocutor is a computer, but as long as we do not have any human-computer dialogs unaffected by written tasks descriptions, we cannot pin down the actual cause for the rigid order of information categories. In order to support or dismiss the hypothesis that written task descriptions can yield misleading data, one must perform another experiment without the ready-made scenarios given to the participants in forehand. Unfortunately, the results from the scenario in group E in the TWOZ experiment would not give us any pointer with regard to this matter because the participants first performed two inquiries that

[4] If we extend the investigation of the categories beyond the opening sequence, there are three occurrences of bus numbers in TWOZ material. They occur when the participants are unsatisfied with the answer from the dialog system, and utter for instance: "But what about bus number four? I know that this bus also arrives at…" This indicates that the possibility of using bus numbers is known to at least some of the participants. Still they do not use this strategy in the opening sequence.

were based on structured schemas. Not until their third call could the participants ask freely. Consequently, the participants conducted two inquiries based on Group A-D before carrying out the scenario described in Group E. As noted by Thomson (1980), users interacting with a computer tend to follow what she calls a "success strategy". This means that the users repeat the same type of request if they have experienced that this particular way of interacting with the computer functions well. It is plausible that the participants in the TWOZ experiment had already learned how to successfully communicate with the system when they performed their last and unrestricted inquiry. If so, the unrestricted dialogs would pattern with the restricted one due to the influence from the previously executed inquiries. As a matter of fact, the dialogs based on the unrestricted inquiry show exactly the same ordering of information, namely departure, arrival and time.

6 Conclusion

Our investigation of the Trondheim Wizard of Oz- experiment (TWOZ) agrees with the findings in Richards and Underwood (1985) in that people render their information request in a uniform and strict manner. However, we have questioned whether the ordering is actually a natural way of asking about time tables, or rather follows from written task descriptions given to the participants before the experiment. Both Richards and Underwood and the TWOZ experiment made use of such written task descriptions. An investigation of these task descriptions in the TWOZ experiment showed that the categories were presented in the same order in practically every scenario, and that this order was congruent with the order of the categories found in the opening sequence in the TWOZ dialogs.

A comparison with human-human dialogs that were not influenced by any written task descriptions, did not display the same striking distribution difference with regard to DEPARTURE as the first piece of information (1PI) and ARRIVAL as the second piece of information (2PI). This supports our suspicion that written task descriptions may influence the result in a WOZ experiment.

The eye-catching difference in the use of bus numbers is also an argument for sustaining that written task descriptions do influence the results in a Wizard of Oz experiment. We found that occurrences of the category BUS NUMBER was the third most frequently used category in the human-human dialogs, while non-existing in the TWOZ material. An investigation of the written task description in the TWOZ experiment showed that the category BUS NUMBER were not presented as a possible strategy to obtain information about bus schedules.

Unfortunately, our data contains no examples of human-computer dialogs not influenced by any written task description. Further investigation is necessary to see whether the differences between the findings in the TWOZ and the H-H dialogs nevertheless can be explained by a tendency to perform the inquiry in a more structured way when interacting with a computer. To completely dismiss or confirm this hypothesis another experiment must be performed, but without the ready-made scenarios.

Based on our findings, it is questionable whether the pattern of information structure found in the TWOZ experiment, or other WOZ experiments based on similar written task descriptions, can be used as a reliable source for improvement of performance in a dialog system. A possible improvement of the written task descriptions avoiding some of the problems addressed in this paper would be to present the information in a more haphazard way.

Even though our findings are not devastating to Richards and Underwood's (1985) claim about a statistically significant regularity in the ordering of information, it makes it less plausible that this ordering will be prominent enough to be of considerable aid in enhancing the performance of a future dialog system.

In addition, our study raises some methodological issues in the use of WOZ experiments in dialog system development. If the participants are as influenced by the written task description as it might seem from our findings, the design of these written task

descriptions should be given much more consideration than previously acknowledged

References

Tore Amble. 2000. BussTUC – a natural language bus route oracle. *ANLP-NAACL 2000- 6th Applied natural language conference*, Seattle. PDF-file: http://acl.ldc.upenn.edu/A/A00/A00-1001.pdf

Nils Dahlbäck, Arne Jönsson and Lars Ahrenberg. 1993. Wizard of Oz studies – why and how. *Proceedings of the 1st international conference on intelligent user interfaces*, 193-200. Orlando.

Magne H. Johnsen, Tore Amble and Erik Harborg. 2003. A Norwegian Spoken Dialog System for Bus Travel Information – Alternative Dialog Structures and Evaluation of a System Driven Version. *Teletronikk* 2:126-31.

Arne Jönsson and Nils Dahlbäck. 1988. Talking to a computer is not like talking to your best friend. *Proceedings in the First Scandinavian Conference on Artificial Intelligence*, 53-68, Tromsø.

M.A. Richards and K.M. Underwood. 1984. Talking to machines: how are people naturally inclined to speak. *Contemporary Ergonomics*, 62-7, (ed.) Megaw. Taylor&Francis, London.

M.A. Richards and K.M. Underwood. 1985. How should people and computers speak to each other?. *INTERACT'84*, 215-18, Elsevier Science Publisher, Amsterdam.

Bozena Henisz Thomson. 1980. Linguistic analysis of natural language communication with computers. *Proceedings of the 3rd International Conference of Computational Linguistics*, Tokyo.

Robin Wooffitt, Norman M. Fraser, Nigel Gilbert and Scott McGlashan. 1997. *Humans, Computers and Wizards. Analysing human (simulated) computer interaction*. Routledge, London.