WIZARD-OF-OZ AND THE TRADE-OFF BETWEEN NATURALNESS AND RECOGNISER CONSTRAINTS

Hans Dybkjær, Niels Ole Bernsen and Laila Dybkjær

Centre for Cognitive Informatics (CCI), Roskilde University
PO Box 260, DK-4000 Roskilde, Denmark
emails: dybkjaer@ruc.dk, nob@ruc.dk, laila@ruc.dk

ABSTRACT

The Wizard-of-Oz simulation technique has been used in the development of the dialogue model for a spoken language dialogue system. The paper focuses on the trade-off between system naturalness and the technological constraints imposed by the speech recogniser. The constraints enforce a strongly system-directed dialogue. Phrases and subjects influence the trade-off whereas voice distortion apparently does not.

Keywords: *Spoken language dialogue systems, Wizard-of-Oz, dialogue model.*

1. INTRODUCTION

The context of this paper is the knowledge acquisition phase of a Danish national project on spoken language dialogue systems. The project aims at developing two prototype systems, P1 and a more advanced version P2, in the domain of ticket reservation for and information on Danish domestic flights [7]. The main system components are a speaker-independent continuous speech recogniser, a chart parser, a dialogue handling component linked to a database, and a synthesis module which uses pre-recorded speech. P1 has been implemented and will be tested during the summer of 1993.

This paper discusses the development of the dialogue model for P1 [3] by means of the Wizard-of-Oz (WOZ) simulation technique [4].

The dialogue model has to satisfy a number of technological constraints mainly imposed by the speech recogniser: a maximum user utterance length of 10 words, an average utterance length of 3-4 words, and at most 100 words active in memory at a time to allow real time performance which must be given high priority for the system to be usable in the chosen domain of application. Moreover, project resources limit the vocabulary to about 500 words.

At the same time the project aims to allow the use of natural forms of dialogue and language. This will contribute to making the system easy to use for both novices as experts but obviously conflicts with the constraints just mentioned. Given the recogniser constraints, naturalness therefore has to be traded for system feasibility. This trade-off process is, however, further limited by a number of basic usability constraints. In addition to real time performance, basic system usability requires sufficient domain and task coverage, sufficiency of task-related vocabulary, natural grammar, robustness, and that limitations on the naturalness of dialogue and language be principled and practicable by users [2].

It would seem to follow that it is the naturalness of the system's dialogue which has to be traded for system feasibility. We shall focus on four issues related to this problem: initiative, phrases, voice distortion and subjects. We begin with a survey of the WOZ method, the experimental set-up and the parameters which influence the evolving dialogue model.

2. WIZARD OF OZ

WOZ is a powerful empirical technique which is well-suited for the iterative development and evaluation of interface design when the input modality has a high computation /cognition ratio in the sense that it can be only partially decoded by computers but is easily understood by humans. Spoken language input belongs to this category. The WOZ method makes possible the testing of design ideas and the acquisition of knowledge of the system and its users and their interaction prior to system implementation. Design goals and constraints can be simulated and adjusted until an acceptable trade-off has been found.

Seven generations of experiments were performed to develop a dialogue model for P1. The set-up is shown in Figure 1.

The *graph* describes the dialogue structure including who has the initiative while the *predefined phrases* show the language used by the system. The graph structure and the phrases are the crucial variables involved in finding an appropriate trade-off between design constraints and naturalness. While the *interface medium* and the choice of *subjects* are not variables in this sense they can, however, be manipulated. This raises questions as to how they should be manipulated in order to optimise the usability of the final system.

Preferably subjects should believe that they are communicating with a real system in order that their behaviour approximates that of the intended end-users as much as possible.

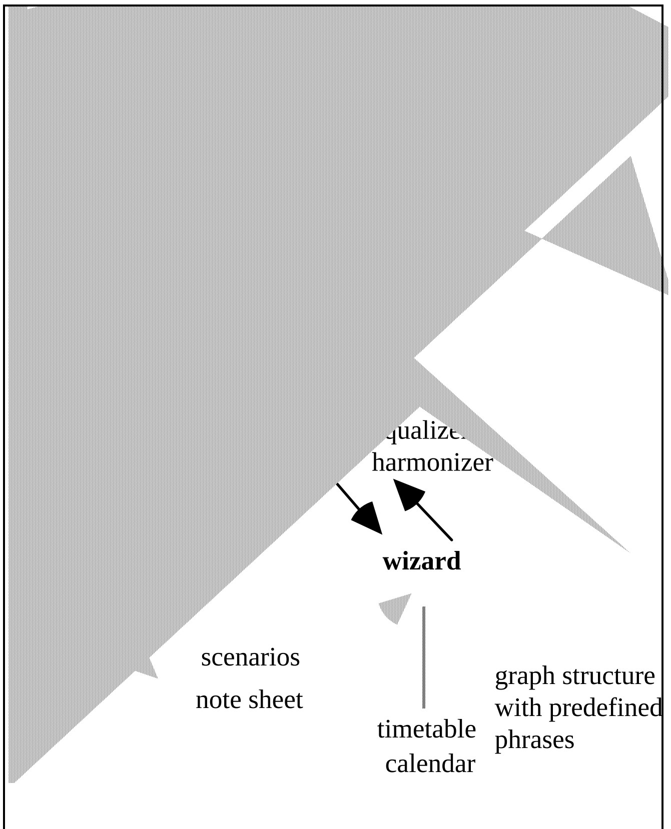


Figure 1: The set-up of the WOZ experiments.

Voice distortion is a mechanism which might help induce subjects into believing that they speak to a computer. Similarly, subjects' backgrounds might influence the way they interact with the system. It is therefore necessary to monitor how the choice of subjects affects the user dialogue behaviour observed in the experiments.

Subjects were asked to perform domain-relevant tasks described in written *scenarios*. Scenarios can be manipulated in many ways and how this is done may significantly influence the usability of the final system, particularly with respect to the sufficiency of its domain and task coverage but also with respect to language.

3. INITIATIVE

In the first two generations of experiments the dialogue structure was a loosely ordered set of predefined phrases. There were no constraints on which phrases could be used in which circumstances. The choice was fully left to the wizard who had great problems being consistent as a result. Subjects had as much of the dialogue initiative as they wanted to but the technological constraints were not met.

A more powerful tool was needed to obtain a consistent dialogue model which would eventually satisfy the technological constraints. A graph structure having predefined phrases in the nodes and predicted contents of user input along the edges was chosen as a tool for this purpose. The graph represented a more structured dialogue in which it was well-defined what information the system needed from the user in order to make, e.g., a reservation. The domain coverage was adjusted so that its limits became increasingly well-defined and the domain coverage more complete.

However, for a system such as P1 having limited vocabulary, at most 100 active words at a time and requiring limited user utterance length, user questions cause problems because of their unpredictability. The dialogue therefore had to be made increasingly system-directed. This can be done by converting user questions into system questions. Asking the questions itself allows the system to have well-defined expectations concerning user utterances in context. Users' answers are typically shorter and more predictable than their questions. As can be seen in Figures 2 and 3, users' average utterance length decreases while more and more of the dialogue initiative is left to the system which asks nearly all the questions in the last generation. The peak in the wizard's utterance length in Figure 2 reflects that more information was included in the wizard's utterances. This was done to have the system provide the information rather than having the user ask for it.

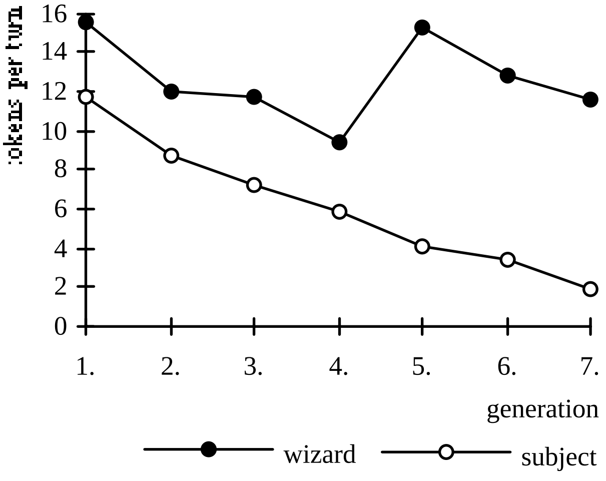


Figure 2: Average length of wizard and subject utterances in terms of tokens per turn.

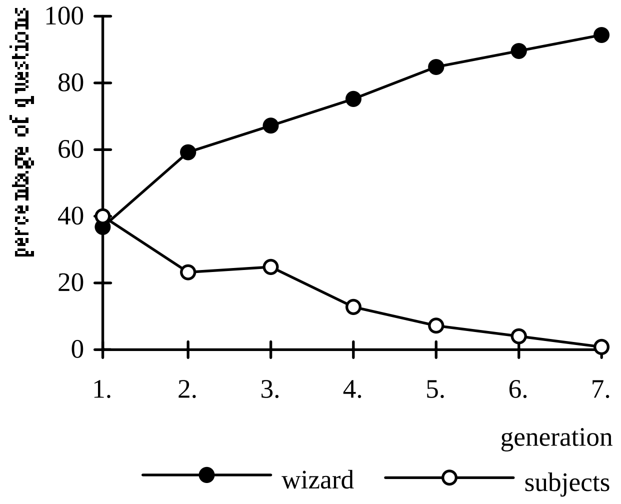


Figure 3: Number of questions in per cent of total number of turns.

Interestingly, system-directed dialogue seems acceptable in some tasks. Recordings of dialogues from a travel agency showed that when the customer has expressed a goal and a few constraints then the travel agent typically takes over and asks questions. This is particularly clear in reservation which is a well-structured task whereas customers typically ask more questions when performing information tasks. The difference between reservation and information tasks is that reservation tasks are closed in the sense that the goal is known and it is known how to reach it whereas information tasks are open: users' goals are much more diverse and satisfying them all is therefore no simple matter.

The field recordings also showed that the average number of words per system/user exchange as well as per task is at the same level in the seventh WOZ generation as in similar human-human dialogues. This may be taken to indicate that a natural level of information exchange has been reached.

4. PHRASES

The phrases used by the simulated system are a powerful tool for manipulating users' language. There is evidence that users model the system's language [10]. So concise, consistent and yet informative system phrases are important.

Especially from the fifth generation of experiments onwards focus was on elaborating the language used in the predefined phrases (cf. Figure 2). The idea was that this might contribute to decreasing the size of the vocabulary used by subjects since they often reused the system's formulations. Care was taken that the same formulations were used in similar contexts. It was particularly clear that subjects model the phrases used by the system when offered a choice among a number of possibilities (closed questions), but similar behaviour could be observed in other situations as well.

However, subjects do not only model the system's phrases. They also model the formulations of the written scenarios given to them [6]. This is a problem since one cannot know if subjects would have used different words in the corresponding real life situations. The solution seems to be to use less explicitly and more diversely formulated scenarios.

Consistency of formulation is one way of influencing subjects through system phrases. Another important point is that there must be phrases enough in the dialogue structure and that each phrase must be sufficiently precise and well-delimited to elicit a predictable answer. The number of ad hoc generated system phrases in a generation of experiments provides an estimate of how well the predefined phrases cover the task domain needs. Walk-throughs of the recorded dialogues were made after each of the later generations of experiments. They gave a good indication not only of phrases missing but also of jumps in the dialogue structure and inadequate formulations which were too open and caused too lengthy or different answers or which confused the subjects. Such phrases were reformulated and made more specific. Sometimes intonation was used to make the meaning clearer.

5. VOICE DISTORTION

Several authors emphasize the use of voice distortion in simulations in order to maintain the illusion of dialogue with a computer [1, 4, 5, 8, 9]. This might seem odd since voice response systems which are used by fairly many people and which involve communication with computers simply use normal pre-recorded human voice. However, people are not used to talking to computers and, moreover, their expectations as to how computers speak may be based on what they have read in science fiction novels or seen on television or in the cinema. There, computers are usually equipped with a somewhat metallic and monotonous voice which is markedly different from a human voice.

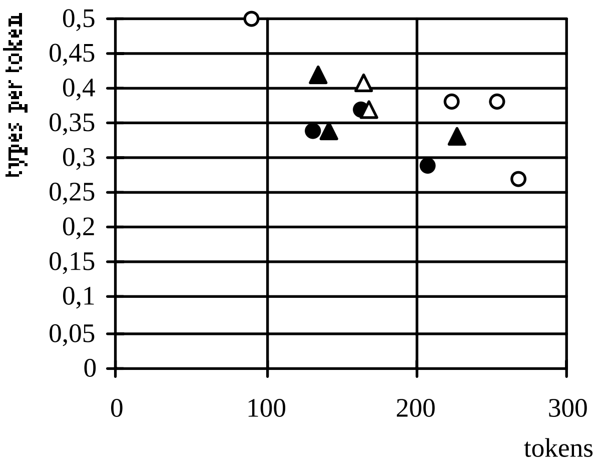
It was therefore decided to use voice distortion in the hope that this would support subjects' illusion of speaking to a computer. In the seventh generation of experiments an equalizer and a harmonizer were used to distort the wizard's voice during the dialogues with about half of the subjects. The hardware gave the wizard's voice a slightly metallic sound with a distant echo-effect. However, this did not seem to have any effect on the subjects. There were about as many subjects who thought they had been dealing with a real computer among those who had heard the distorted voice as among those who had heard the wizard's normal voice. When such parameters as number of turns, types (i.e. new words), tokens (i.e. all words), and tokens per turn are compared, there is a small difference. But this contrasts with other investigations since the subjects who heard the distorted voice on the average used more turns, types, tokens, and tokens per turn even excluding results from two colleagues who had tried the system before and knew that it was simulated. However, the difference is not significant. Probably voice distortion had no effect because other parameters —primarily system directedness— already had caused the effect that voice distortion may have.

6. SUBJECTS

Subjects' backgrounds seem to be important to their interaction with the system. The majority of subjects in the sixth and seventh generations came from outside the lab. The rest were colleagues.

Figure 4 shows that people with a linguistic background tended to use many words (tokens) and many different words (types). They experimented with the system and tested which grammatical constructions and words it understood. Secretaries, on the other hand, were much more cooperative and focused on reaching the goal as easily and quickly as possible. One group of computer scientists were very cooperative and focused on the goal and apparently took care of expressing themselves briefly as they were asked to (engineering behaviour). The second group experimented with the system, not, like the linguists, with grammatical constructions but rather with the system's semantics (academic behaviour).

The sixth and seventh generations of experiments seem to demonstrate the importance of choosing the right kind of subjects. Subjects' backgrounds do seem to affect their performance with the system. It is therefore important to choose subjects having a background corresponding to that of the users of the final system (in our case mostly secretaries) in order to obtain data which are as reliable as possible. Probably a certain type of background, experience and training cannot be simulated. It is therefore not sufficient just to ask a person to behave as s/he believes that, e.g., a secretary would do. The parameters which have been in focus in the experiments such as vocabulary and utterance length cannot just be simulated. What is a natural way of communication for a secretary is not necessarily quite as natural for a linguist.



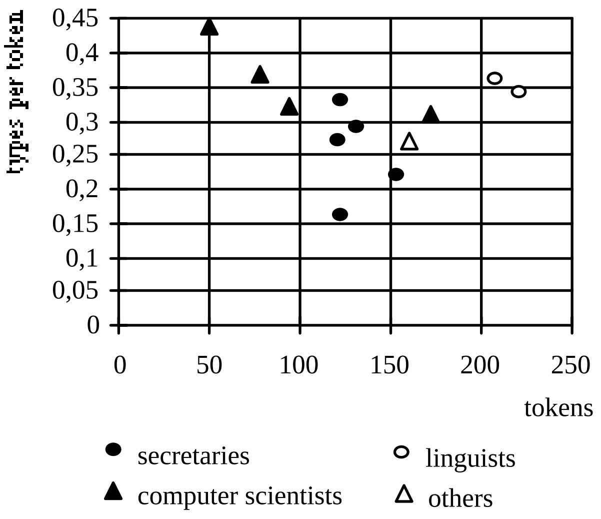


Figure 4: Average number of types per token in relation to number of tokens used by each subject in the sixth (top) and seventh generations of experiments.

7. CONCLUSION

The dialogue model from the seventh generation of experiments satisfies the constraints on average and maximum utterance length and at most 100 active words at a time. The dialogues from the sixth and seventh generations were used as a basis for defining a sublanguage having a 500 word vocabulary. Experiments with the P1 system will show whether the vocabulary has sufficient coverage.

To satisfy the constraints dialogue naturalness was traded for system feasibility. The dialogue was made increasingly system-directed by converting user questions into system questions. Dialogue naturalness suffered differently from the conversion depending on the nature of the task. Naturalness of language use has not been constrained except that users are asked to use short phrases to be understood by the system. P2 should allow a less system-directed dialogue.

The phrases used by the simulated system may be used to limit the user's vocabulary partly by using the same formulation in the same situation which the user in many cases will model, and partly by making user utterances precise and easy to understand to avoid long and confused user utterances.

Voice distortion apparently had no effect on users' language, perhaps because other parameters had already caused the desired effect. The choice of subjects seems important.

ACKNOWLEDGEMENTS

The project is carried out in collaboration with the Speech Technology Centre at Aalborg University (STC) and the Centre for Language Technology at the University of Copenhagen (CST). We gratefully acknowledge the support of the project by the Danish Government's Informatics Programme.

REFERENCES

[1] Amalberti, R., Carbonell, N., and Falzon, P.: User representations of computer systems in human-computer speech interaction, to appear in Int. Jour. Man-Machine Studies, 1992.

[2] Bernsen, N.O.: Design of a Spoken Language Dialogue System. A Study of the Initial Specification Phase. Working Papers in Cognitive Science WPCS-92-5. Centre for Cognitive Science, Roskilde University, 1992.

[3] Dybkjær, L. and Dybkjær, H.: Wizard of Oz Experiments in the Development of the Dialogue Model for P1. Report 3, Spoken Language Dialogue Systems, STC Aalborg University, CCI Roskilde University, CST University of Copenhagen, 1993.

[4] Fraser, N. and Gilbert, N.: Simulating Speech Systems. Computer Speech and Language, no. 5, 1991.

[5] Guyomard, M. and Siroux, J.: Experimentation in the Specification of an Oral Dialogue. In: H. Niemann, M. Lang, and G. Sagerer (eds.): Recent Advances in Speech Understanding and Dialog Systems. NATO ASI Series F, vol. 46, pp. 497-502, 1988.

[6] Klausen, T.: Talking to a Wizard: Report from the design of a natural speech understanding system. Working Papers in Cognitive Science WPCS-93-7. Centre for Cognitive Informatics, Roskilde University, 1993.

[7] Larsen, L.B., Brøndsted, T., Dybkjær, H., Dybkjær, L., and Music, B.: Overall Specification and Architecture of P1, Report 2, Spoken Language Dialogue Systems, STC Aalborg University, CCI Roskilde University, CST University of Copenhagen, 1993.

[8] Luzzati, D. and Neel, F.: Dialogue Behaviour Induced by the Machine. Eurospeech ‘89, Paris, pp. 601-604, 1989.

[9] Richards, M. A. and Underwood, K.: Talking to Machines. How are People Naturally Inclined to Speak. In E. D. Megaw (ed.): Contemporary Ergonomics, London, Taylor and Francis, 1984.

[10] Zoltan-Ford, E.: How to get people to say and type what computers can understand. In: International Journal on Man-Machine Studies, vol 34, pages 527-547, 1991.