EXPLORING THE LIMITS OF SYSTEM-DIRECTED DIALOGUE

Dialogue Evaluation of the Danish Dialogue System

Niels Ole Bernsen, Hans Dybkjær and Laila Dybkjær
Centre for Cognitive Science, Roskilde University
PO Box 260, DK-4000 Roskilde, Denmark

ABSTRACT

Spoken language dialogue systems technologies are beginning to master the design and implementation of applied systems for complex well-structured tasks. Partly for this reason, there is a need for evaluation metrics which include general concepts of task and dialogue types. The paper reports on the scenario-based user test of the dialogue management of an airline ticket reservation system. The test data are compared to the data from the last Wizard of Oz iteration before the system was implemented. Detailed analysis of user dialogue behaviour reveals a series of principled limitations of system-directed dialogue for complex well-structured tasks. The discussion weighs those limitations against the demonstrated potential of system-directed dialogue for a broad class of tasks.

1. INTRODUCTION

Real mixed initiative dialogue has not yet been realised in spoken language dialogue systems (SLDSs) addressing complex tasks [9,12]. However, indications are that, as long as the task is highly structured, system-directed dialogue may be sufficient even for rather complex tasks [3]. The Danish SLDS uses system-directed dialogue for a complex task, and several other, similar SLDSs are currently moving from the laboratory environment into field trials [4,10,11]. Based on the evaluation of the Danish system, this paper addresses the problems involved in applying system-directed dialogue design to the management of complex tasks. Section 2 briefly describes the Danish system and the constraints imposed on its development. The main part of the paper reports on the evaluation of the system's dialogue management component. Section 3 describes the test setup. In Section 4, test results are compared with results from the last Wizard of Oz design phase. Section 5 analyses the task adequacy of system-directed dialogue, and Section 6 concludes the paper.

2. DESIGNING DIALOGUE TO CONSTRAINTS

The Danish prototype SLDS is a reservation system for Danish domestic flights. It has been developed in the Dialogue project by the Center for PersonKommunikation, Aalborg University, the Centre for Cognitive Science, Roskilde University, and the Centre for Language Technology, Copenhagen. The system runs on a PC and is accessed over the telephone. The prototype is a speaker-independent continuous speech understanding system which speaks and understands Danish. The speech recogniser uses HMMs to produce a 1-best string of words. The parser makes a syntactic analysis of the string and extracts the semantic contents which are represented in frame-like structures called semantic objects. The dialogue management module interprets the contents of the semantic objects and decides on the next system action which may be to send a query to the database, send output to the user, or wait for new input. In the latter case predictions on the next user input are sent to the recogniser and the parser. Output is produced by concatenating pre-recorded phrases.

The dialogue model was developed through seven Wizard of Oz (WOZ) iterations [5]. The primary design goals for this model were: sufficient task domain coverage, real time performance, robustness, natural forms of language and dialogue, and flexibility. These goals had to be traded off against the following resource- and technology-based constraints: a maximum vocabulary of about 500 words; enabling real-time performance through allowing at most 100 active words in memory at a time; and an average and a maximum user utterance length of 3-4 words and 10 words, respectively.

The WOZ-derived system-directed dialogue model that was implemented with a number of modifications (see below), satisfied the technological constraints except that subjects' vocabularies failed to show sufficient convergence towards the end of the WOZ phase [5]. This was no surprise, as related ATIS results from other languages suggest a domain vocabulary of 1000-1200 words [12]. As regards the dialogue design goals, real-time performance appeared feasible, task domain coverage was acceptable, and restrictions on user language and dialogue were principled so that the users were able to comply with them.

Dialogue robustness, however, remained essentially unknown. The dialogue model included the keywords 'correct' and 'repeat' through which users could initiate repair and clarification meta-communication, and the system could initiate meta-communication through the phrase "Sorry, I did not understand". However, errors which might force either subjects or the system to initiate meta-communication had not been simulated. It is, indeed, difficult to realistically simulate errors of recognition and understanding in a system which has not yet been implemented. However, this implied that, during implementation, meta-communication had to be elaborated from scratch. Much work went into defining the functionality of 'correct' and 'repeat'. Use of 'repeat' makes the implemented system repeat its most recent utterance, not including feedback. Use of 'correct' allows users to correct the latest piece of information given to the system. 'Correct' may be used recursively to correct information from an arbitrary earlier utterance. For natural and early error detection, the system provides feedback by echoing the key information in the latest user utterance. Furthermore, at the end of a reservation task, a summary is provided of the entire reservation made by the user.

Like robustness, flexibility represents a main problem in system-directed dialogue design. System-directed dialogue affords little dialogue flexibility. However, the task of reservation is well-structured, i.e., it has a certain number of sub-tasks most of which must be completed in order to achieve the reservation task, and for many of these sub-tasks there is a natural order in which to carry them out. For such tasks, system-directedness may allow acceptable dialogue, at least up to a certain level of dialogue complexity [6]. Adding flexibility to this structure was mainly obtained through a minimal user model which allows expert users to de-select (i) the

introductory instructions for novice users and (ii) the information on discount types.

The WOZ experiments also covered the tasks of changing reservation and obtaining travel information. However, as these tasks are not well-structured they would appear ill-suited for system-directed dialogue [3,6]. This was a main reason for not implementing these two tasks.

3. DIALOGUE MODEL TEST SETUP

The system, excluding the recogniser, was subjected to scenario-based testing with naive users. A wizard keyed in the users' answers into a simulated recogniser. The simulated recogniser ensured that typos were automatically corrected and that input to the parser corresponded to an input string which could have been recognised by our real speech recogniser. Recognition accuracy would be 100% as long as users remained within the vocabulary and grammars known to the system. Otherwise, the simulated recogniser would turn input into a string which only contained words and grammatical constructions that were within the recogniser's vocabulary and which conformed to the recogniser's grammar rules. In the second test phase which will not be reported here, the full system including the recogniser will be tested. These tests are not, of course, substitutes for field testing but constitute last steps before field testing can begin.

The 20 task scenarios used in test were systematically constructed to explore all aspects of the task structure. Since the flight ticket reservation task is a well-structured task in which a prescribed amount of information must be exchanged between user and system, it was possible to extract from the task structure a set of sub-task components, such as number of travellers, age of traveller, and discount vs. normal fare, any combination of which should be handled by the dialogue system. The scenarios were generated from systematically combining these components.

12 novice subjects, mostly professional secretaries, each received 4 scenarios and a brochure describing the system. After the experiment they received a telephone interview and filled in a questionnaire. In addition, the test included an experiment which addressed the following problem. The WOZ process had shown that subjects tend to copy important parts of the vocabulary used in their scenario description, thus jeopardising the sub-language acquisition goal of WOZ. To explore how to avoid scenario priming and hence to elicit a more realistic sublanguage, subjects were divided into two groups that received different versions of the scenario material. One group received standard task descriptions of the kind likely to be copied during dialogue, whereas the second group received a new version of the scenarios in which the copying effect had been effectively blocked [7].

4. TEST AND WOZ RESULTS COMPARED

The user test produced a corpus of 57 dialogues. Subjects sometimes repeated a scenario if they did not succeed the first time. The seventh and last WOZ corpus (WOZ7) and the test corpus are of similar size (see Figure 1). In the discussion of Figure 1, we focus on the data which show divergence from our dialogue design goals.

As predicted, the system's sub-language vocabulary is insufficient. The test corpus shows 51 out-of-vocabulary word types (excluding numbers as well as names of months, days of the week, airports, and false start items). Thus 28.2% or more than one fourth of the user word types were out of vocabulary.

The test results show that the average user utterance length is still well within the prescribed limits, cf. Figure 1 (average number of tokens per turn). However, the prescribed maximum user utterance length was exceeded in 17 cases. 10 of these utterances were produced by the same subject. Particularly in the first dialogue, this subject tended to repeat an utterance if the system did not answer immediately. However, the majority of long utterances, both for this subject and in general, was caused by user-initiated corrections which did not make use of the keyword 'correct' but were expressed in free style by users. Two long utterances were produced by subjects who took over the initiative when asked 'Do you want anything else?'. Finally, subjects sometimes provided more information than had been asked for, despite the fact that the system in its introduction had warned them that it would not be able to understand them unless they answered its questions briefly and one at a time. We shall return to this analysis in Section 5.

	WOZ7		User test	
No. of subjects	12		12	
No. of dialogues	47		57	
	User	Syst.	User	Syst.
No. of turns	881	905	998	998
Longest turn	12	92	23	87
Av. turns/dialogue	18.74	19.26	17.51	17.51
No. of tokens	1633	10495	2468	12185
Av. tokens/turn	1.85	11.59	2.47	12.20
No. of turns > 10 tokens	3	272	17	253
Turns > 10 tokens in % of all turns	0.34	30.06	1.70	25.35
Av. tokens/dialogue	34.74	223.30	43.30	213.77
No. of types	165	350	180	189
Av. types/ token	0.10	0.03	0.07	0.02
No. of questions	4	-	4	-
Questions in % of No. of turns	0.45	-	0.40	-

Figure 1. Comparison of results from WOZ7 and the user test. All system turns except for the closing phrase contained a question. Cardinals, ordinals, airport names, months, days of week, and false start items were counted as one group each, thus adding only six word types to the total number of types.

Apart from user-initiated meta-communication through keywords, the dialogue is designed to be entirely system-directed. In particular, great care has been taken to prevent users from asking any kind of clarification or repair meta-communication question of the system. Based on post-hoc analysis of the WOZ process, a set of SLDS design guidelines were defined for this purpose [1,2]. This aspect of the dialogue design has been successful. In the user test, only 4 out of 998 user utterances were questions. One question was asked because the subject had misread the scenario text. The three remaining user questions all concerned available departure times. This is not surprising since departure times constitute a type of information which users often do not have in advance but expect to be able to obtain from the system. We shall return to this analysis in Section 5.

Figure 2 shows how the subjects evaluated qualitative aspects of the dialogue system they had interacted with in WOZ7 and in the user test, respectively. In interpreting the results, it must

be kept in mind that the system was tested without the real recogniser.

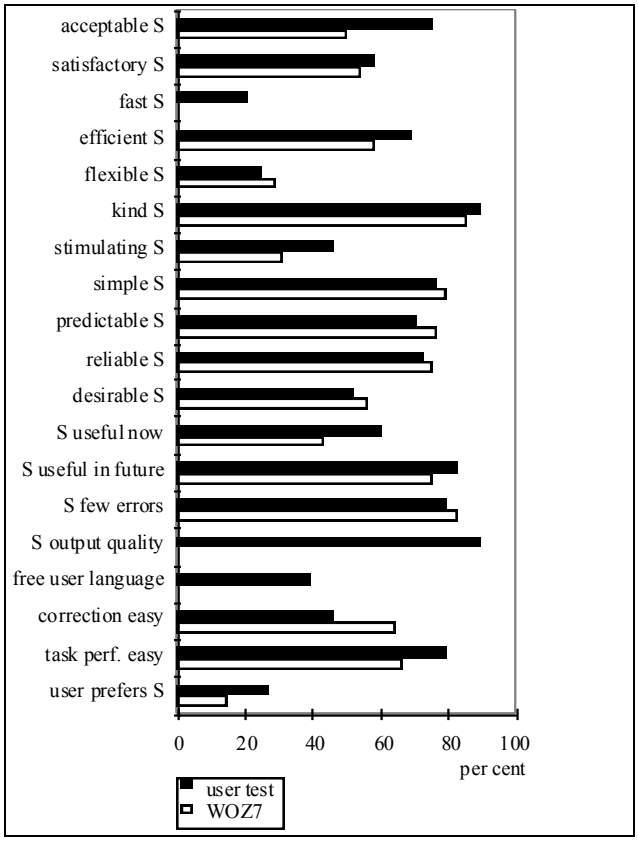

Figure 2. Subjects' answers to the questionnaires from WOZ7 and the user test in per cent of maximum positive score. A score of less than 50 per cent indicates a negative opinion of the system. 'S' in the left-hand column refers to the system.

In many cases there is no real difference between the two sets of answers. This is true of the properties of satisfactoriness, kindness, simplicity of use, predictability, reliability, desirability, future usefulness and lack of errors, all of which were evaluated positively in the sense that positive evaluations range from 50% upwards. The positive evaluation of robustness (few errors) is encouraging (about 80%). Positive inprovements from WOZ7 to the tested system can be seen on acceptability (to 75%), efficiency (to 70%), usefulness now (to 60%), and ease of task performance (to 80%). There are also improvements in the evaluation of stimulatingness and preference of the system over a human travel agent but both are still low (45% and 25%, respectively). The main reasons probably are the rigid dialogue structure and, in particular for the latter percentage, the correct impression that such a system has limited capabilities and cannot cope with non-routine matters.

There are drops in the positive evaluation on two important parameters, namely on flexibility (to 23%) and ease of making corrections (to 45%). The low evaluation on flexibility is probably due to the rigid, system-directed dialogue structure and the restriction to keywords for meta-communication. The negative development with respect to ease of making corrections is probably due to the fact that misunderstandings were not simulated in WOZ7. This meant that hardly any user-initiated meta-communication was required. In the user test, the simulated recogniser sometimes misunderstood what the user said. In addition, the use of keywords for making corrections does not form part of the natural human linguistic skills.

Finally, a number of parameters were only evaluated after the user test. In view of the fact that the test used a bionic wizard system, it is no surprise that subjects did not find the system fast (20%) and thus did not perceive real-time performance. Output quality was rated high (87%). Not surprisingly in view the requirement to use keywords in initiating meta-communication and the missing sub-vocabulary parts, subjects did not find that they could use free natural language (40%).

No methodology exists for synthesising the results of user evaluations of SLDSs [8]. The discussion in the next section provides a perspective from which to view the user evaluation results.

5. LIMITATIONS OF SYSTEM-DIRECTED DIALOGUE

The user test has confirmed that, with the exception of its vocabulary limitations and some minor problems which can be corrected relatively easily, the system is functionally adequate. The minor problems include missing capability of making certain inferences based on user input and an obscure system response**.** Apart from theses limitations, all intended reservation tasks can be performed using brief sentences and unrestricted syntax. The low percentage of lengthy user utterances (1.7%, cf. Table 1) indicates that the user utterance length aimed at has been achieved. The low percentage of user questions (0.4%, cf. Table 1) indicates success for the dialogue design guidelines which were applied to prevent user-initiated repair and clarification meta-communication [2]. The low number of cases in which users failed to answer the system's questions briefly and one at a time shows how the combination of appropriate instructions to users and strict system-directedness may serve to keep utterance lengths sufficiently low for the needs of the speech recogniser.

On three conditions, therefore, it might be argued that the system could be made commercially available. The conditions are that (a) the sub-language vocabulary be made adequate, (b) the system exhibit close-to-real-time performance when coupled with the real speech recogniser, and, not least, (c) that the speech recogniser do not damage overall system performance. Solving problem (a) is primarily a matter of resources. Preliminary tests on the full system indicate that problem (b) has been solved. The seriousness of problem (c) will become clear in the user test of the full system.

A functionally adequate system is not necessarily an optimally usable one relative to the intended user group(s). We believe that the user test has demonstrated several deeper usability problems with the system, problems which may be solved only through the introduction of mixed-initiative dialogue. Problems of this nature cannot be identified from the quantitative or qualitative test data reported in Figures 1 and 2 above, but appear through analysis of human-machine dialogues which fail on some performance measure or other. The analysis in Section 4 showed two such problems.

Non-preventable user questions, given the task: Analysis of the user questions identified three user questions concerning available departure times. We want to argue that questions of this nature cannot be prevented from occurring during reservation tasks, no matter how successful a strategy one adopts to prevent users from asking questions of the system. The reason is that reservation or ordering tasks are inherently *informed* reservation or ordering tasks. In other words, it is a natural part of ordering, or reserving, something, to request information in order to decide what to order or reserve. The implication is *not* that mixed-initiative SLDSs are necessary

for complex reservation tasks. System-directed dialogue systems may still work well enough, at shown above. But mixed-initiative dialogue is the ideal dialogue model for complex ordering and reservation tasks.

Non-naturalness of keyword-based user meta-communication initiative: We have seen (Section 4) that the majority of long utterances was caused by user-initiated repairs which did not make use of the keyword 'correct' but were expressed in free style. In general, this violation occurred in 17 cases and the system only correctly understood the user's intention in one of these. Several transaction failures (see below) were due to the fact that it did not ocur to novice users to use the unlimited backtracting facility offered by 'correct' to back out of some problem. As we have seen, however, despite its unnaturalness, the keyword-based user meta-communication solution may work as long as keywords are few, non-ambiguous and well-explained. But it is equally clear that this solution is inferior to the ideal solution which is to allow free-style user-initiated meta-communication and hence mixed-initiative meta-communication dialogue.

In computing the transaction success [13] we first split the test dialogues into three groups. In the first group, users achieved the scenario task to the extent possible. Thus, for instance, if some scenario-prescribed departure was full, the user made a reasonable alternative choice. In the second group, users achieved what they asked for, but what they asked for was not exactly what their scenarios prescribed. In the third group, users failed to achieve what they asked for. We count only the latter case as a transaction failure. Transaction successes and failures are counted relative to the prescribed task rather than to a single dialogue. Sometimes subjects hung up on the system in the middle of a task and completed the task in a second dialogue. On these principles, the test produced 44 transaction successes (86.3%) and 7 failures 13.7%).

Points of maximum complexity in dialogues on well-structured tasks: space does not permit a full analysis of the transaction failures. However, analysis of the transaction failures demonstrated a third type of problem in using system-directed dialogue for complex well-structured tasks. Mixed-initiative dialogue would seem required in SLDSs for complex unstructured tasks [3,6] However, one may hypothesise that many complex well-structured tasks have points of maximum complexity at which system-directed dialogue comes close to its limits. In the case of our reservation task, such points occur when, e.g., 4 persons want to fly out together and only 2 persons want to fly back together, or when a person wants to fly out to airport X and back from airport Y. Functionally speaking, such cases are simply dealt with through reservation of one-way tickets. To users, however, this is a counter-intuitive and overly complex way of doing things. The other side to the dilemma is that if such problems are to be solved within the system-directed dialogue paradigm and its fixed task scheme, as they well may be, dialogue with the system becomes overly cumbersome even for the completion of simple reservation tasks. It is not clear that there are other ways out of this dilemma than through adopting mixed-initiative dialogue.

6. CONCLUSION

Tests with the full system (including the recogniser) and the following field tests will undoubtedly show more functional and usability problems with the implemented system. SLDSs which understand long utterances and conduct mixed-initiative dialogue are obviously preferable to system-directed SLDSs. In this paper, we have pointed out three principled and general reasons why human-machine dialogue on complex well-structured ordering tasks ideally requires mixed-initiative. For such tasks, however, mixed-initiative SLDSs still present major scientific and technological problems. In the meantime, the test results reported in this paper give reasons for believing that system-directed SLDSs are a viable alternative for a broad class of complex well-structured tasks.

REFERENCES

[1] Bernsen, N.O.: Types of User Problems in Design. A Study of Knowledge Acquisition Using the Wizard of Oz. Esprit Basic Research project *AMODEUS II Working Paper RP2-UM-WP 14*, 1993. In Deliverable D2: Extending the User Modelling Techniques. June 1993.

[2] Bernsen, N.O., Dybkjær, H. and Dybkjær, L.: Co-operativity in Human-Machine and Human-Human Spoken Dialogue. To appear in *Discourse Processes*.

[3] Bernsen, N.O., Dybkjær, L. and Dybkjær, H.: A Dedicated Task-Oriented Dialogue Theory in Support of Spoken Language Dialogue Systems Design. *Proceedings of the ICSLP '94,* Yokohama, September 1994, 875-878.

[4] Cole, R., Novick, D.G., Fanty, M., Vermeulen, P., Sutton, S., Burnett, D. and Schalkwyk, J.: A Prototype Voice-Response Questionnaire for the US Census. *Proceedings of the ICSLP '94,* Yokohama, September 1994, 683-686.

[5] Dybkjær, H., Bernsen, N.O. and Dybkjær, L.: Wizard-of-Oz and the Trade-off between Naturalness and Recogniser Constraints. *Proceedings of EUROSPEECH '93*, Berlin, September 1993, 947-950.

[6] Dybkjær, L., Bernsen, N.O. and Dybkjær, H.: Different Spoken Language Dialogues for Different Tasks. A Task-Oriented Dialogue Theory. *Human Comfort and Security,* Springer Research Report 1995 (in press).

[7] Dybkjær, L., Bernsen, N.O. and Dybkjær, H.: Scenario Design for Spoken Language Dialogue Systems Development. *Proceedings of the ESCA Workshop on Spoken Dialogue Systems,* Vigsø, 30 May to 2 June, 1995, 93-96.

[8] Fraser, N.M..: Quality Standards for Spoken Language Dialogue Systems: A Report on Progress in EAGLES. *Proceedings of the ESCA Workshop on Spoken Dialogue Systems,* Vigsø, 30 May to 2 June, 1995, 157-160.

[9] Grau, B., Sabah, G. and Vilnat, A.: Control in Man-Machine Dialogue. *THINK*, Vol. 3, Tilburg, The Netherlands, May 1994, 32-55.

[10] Mazor, B., Braun, J., Ziegler, B., Lerner, S., Feng, M.-W. and Zhou, H.: OASIS - a Speech Recognition System for Telephone Service Orders. *Proceedings of the ICSLP '94,* Yokohama, September 1994, 679-682.

[11] Oerder, M. and Aust, H.: A Realtime Prototype of an Automatic Inquiry System. *Proceedings of the ICSLP '94,* Yokohama, September 1994, 703-706.

[12] Peckham, J.: A New Generation of Spoken Dialogue Systems: Results and Lessons from the SUNDIAL Project. *Proceedings of EUROSPEECH '93*, Berlin, September 1993, 33-40.

[13] Simpson, A. and Fraser, N.M.: Black Box and Glass Box Evaluation of the SUNDIAL System. *Proceedings of EUROSPEECH '93*, Berlin, September 1993, 1423-1426.