A TYPOLOGY OF PROBLEMS OF COOPERATIVITY IN SPOKEN HUMAN-MACHINE DIALOGUE

Hans Dybkjær, Niels Ole Bernsen and Laila Dybkjær

Centre for Cognitive Science, Roskilde University
PO Box 260, DK-4000 Roskilde, Denmark
emails: dybkjaer@cog.ruc.dk, nob@cog.ruc.dk, laila@cog.ruc.dk
phone: +45 46 75 77 11 fax: +45 46 75 45 02

ABSTRACT

The paper presents a method for identifying problems of user-system interaction. Based on a consolidated set of 24 principles of cooperative spoken human-machine dialogue, the paper then proposes and illustrates a general typology of non-cooperative system dialogue behaviour for use in spoken language dialogue analysis and evaluation.

1. INTRODUCTION

Controlled user testing remains an important intermediate step in the evaluation of advanced spoken language dialogue systems (SLDSs). It enables in-depth assessments to be made of all parts of the system and their interaction, and provides a solid basis for judging field trial feasibility. One among several unsolved problems in performing in-depth evaluation of SLDSs is the establishment of a well-founded typology of inadequate system dialogue behaviours. Prior to the work to be described below, we had developed a set of principles for the design of cooperative system dialogue. The controlled user test of the Danish dialogue system demonstrated that these principles could be used to identify and classify the inadequate system dialogue behaviours that were identified following a rigorous methodology. The paper describes the methodology and proposes a general typology of non-cooperative system dialogue behaviour for use in SLDS dialogue analysis and evaluation.

2. PROBLEM DETECTION

The Danish SLDS prototype is an over-the-phone reservation system for domestic flights. The system is a walk-up-and-use application which runs in close to real time on a PC with a DSP board. The system understands speaker-independent continuous spoken Danish with a vocabulary of about 500 words. The dialogue model was developed by means of the Wizard of Oz technique and had to satisfy technological constraints on active vocabulary size and average and maximum user utterance length while being as natural as possible. The dialogue is mixed-initiative: domain communication is system-directed whereas users can initiate clarification and repair meta-communication through keywords.

A scenario-based user test was carried out on the implemented system (apart from the speech recogniser which was simulated). Twelve novice subjects, mostly professional secretaries, conducted a series of dialogues over the telephone in their normal work environments. The user test produced a corpus of 57 dialogues which were transcribed and analysed. The analysis aimed at detecting dialogue interaction problems and was done as follows. Based on the dialogue structure, a template was built which contained the system's questions. For each scenario, the key contents of normative user answers were filled into the template. The key contents of the actual user answers were then plotted into the template together with relevant key contents of system messages [4], cf. Figure 1. Finally, normative and actual user answers were compared which led to the identification of three major classes of interaction problems: (1) linguistic problems, (2) problems of dialogue interaction, and (3) other problems, such as system breakdown. (2) splits into (A) system cooperativity problems and (B) user errors. In the following we focus on describing and illustrating (A). An analysis of the user errors is presented in [3].

Figure 1 shows a template which revealed three different types of interaction problem. The user has expressed an interest in discount and wants a departure at 7:20. However, discount is not available on the departure at 7:20. This is a case of inconsistent user input. The system does not attempt to resolve the conflict, however. Without informing the user, the system always gives priority to discount over departure time and therefore claims that there is no departure at 7:20. In addition, it has only instructed users on how to change their immediately preceding answer. The user eventually works out how to repeatedly use the keyword 'change' to backtrack to the confirmative answer concerning discount which s/he wants to modify. The critical part of the dialogue is shown in Figure 2.

3. TOWARDS A GENERAL TYPOLOGY

Using the method presented in Section 2, a total of 119 system cooperativity problems were identified in the user test corpus. Each problem was analysed in detail. The analysis was based on a set of principles, or guidelines, for the design of cooperative spoken human-machine dialogue. These principles had been developed on the basis of our dialogue model development which was done using the

Wizard of Oz (WOZ) method. The principles had been subsequently refined through comparison with an already established set of maxims for cooperative human-human dialogue [1, 5].

The set of principles included 13 generic principles and 9 specific principles. A generic principle may subsume one or more specific principles which specialise the generic principle to

Scenario: G-1-4-a User: 2 Date: 13 January 1995			
System questions	**Normative user answers**	**Actual user answers**	**Problems**
System already known	no / yes	yes	
Customer number	3	3	
Number of travellers	1	1	
ID-numbers	2	2	
Departure airport	Aalborg	Aalborg	
Arrival airport	Copenhagen	Copenhagen	
Return journey	yes	yes	
Interested in discount	no / yes	yes	
Day of departure (out)	January 16	Monday (January 16)	
Hour of departure (out)	7:20	7:20 (no departure) 7:20 (no departure) no, change [does not want one from list; change not caught by system] 7:20 (no departure) no [does not want one from list] 7:20 (no departure) no [does not want one from list] change [hour of departure] yes [out-day is January 16] 7:20 (no departure) change [hour of departure] change [day of departure] no [does not want discount] yes [out-day is January 16] yes [hour of departure is 7:20]	1, 13/10 1, 13/10 1, 13/10 1, 13/10 10/5 1, 13/10
Day of departure (home)	January 16	Same day (January 16)	
Hour of departure (home)	17:45	17:45	
Delivery	airport / send	airport	
More	no	no	

Figure 1. Key contents of the expected (normative) and actual user-system exchanges in the dialogue G14a. In the third column key contents of the system's replies are indicated in parentheses unless they can be derived from the explanatory comments in square brackets. The numbering in the fourth column refers to detected cooperativity problems, cf. Figure 3.

S: Monday January 16. At which time?
U: 7:20.
S: There is no departure at 7:20. The closest other red departure is at 10:50. Do you want this departure?
U: No thanks.
[The above four turns are repeated another three times with the only difference that the user adds 'change' after 'no thanks' in two cases.]
S: At which time?
U: Change.
S: The journey starts on Monday January 16. Is that correct?
U: Yes.
S: At which time?
U: 7:20.
S: There is no departure at 7:20. The closest other red departure is at 10:50. Do you want this departure?
U: Change.
S: The journey starts on Monday January 16. Is that correct?
U: Change.
S: Are you still interested in discount?
U: No.
S: The journey starts on Monday January 16. Is that correct?

U: Yes.
S: Departure at 7:20. Is that correct?
U: Yes.

Figure 2. The critical part of dialogue G14a. S means system and U means user.

certain classes of phenomena. Although subsumed by generic principles, we believe that specific principles are useful both as dialogue design guidelines and for the purpose of identifying and classifying problems of user-system interaction.

The user test served as a test of the principles and confirmed their broad coverage with respect to cooperative spoken user-system dialogue. Almost all of the 119 identified cooperativity problems could be ascribed to violations of one or other of the principles. Three additions had to be made, however. Two specific principles were added on meta-communication, cf. 13/10 and 13/11 in Figure 3. Since meta-communication had not been simulated during WOZ, as a result of which the WOZ corpus contained few examples of meta-communication, this came as no surprise. More interestingly, we had to add a modification to

COOPERATIVITY PROBLEM	CASES OBSERVED	N	TF	CAUSE/REPAIR
1: System provides less information than required.	Final question too open; withholding important information, requested or not.	19		Question design: 4. Response design: 15.
1/1: System is not fully explicit in communicating to users the commitments they have made.	Easy to ensure once it has been decided to follow 1/1.			
1/2: Missing system feedback on user information.	System misunderstandings only show up later in the dialogue.	2	1	Feedback response design.
2: System provides more information than required.	Difficult to test through identified cooperativity problems.			
3: System provides false information.	On departures.	2		Database design.
4: System provides information for which it lacks evidence.	Our system cannot do this. Problems 13/10 and 13/11 indirectly raise issues of this kind.			
5: System provides irrelevant information.	Irrelevant error message produced by grammar failure.	2	1	Speech recognition design.
6: Obscure system utterance.	Grammatically incorrect response; obscure departure information.	7		Response grammar design: 1. Response design: 6.
7: Ambiguous system utterance.	Question on point of departure.	2		Question design.
7/3: System does not provide same formulation of the same question to users everywhere in its dialogue turns.	Easy to provide once it has been decided to follow 7/3.			
8: Too lengthy expressions provided by system.	Difficult to test through identified cooperativity problems.			
9: System provides disorderly discourse.	Great care taken during dialogue design.			
10: System does not inform users of important non-normal characteristics which they should, and are able to, take into account to behave co-operatively in dialogue.	Users: provide indirect response; change through comments; ask questions; answer several questions at a time.	33		Reduce system demands on users.
10/4: Missing or unclear information on what the system can and cannot do.	System does not listen during its own dialogue turns.	33	1	Speech prompt design.
10/5: Missing or unclear instructions on how to interact with the system.	Undersupported user navigation: use of 'change'; round-trip reservations.	2	1	User instruction design.
11: System does not take users' relevant background knowledge into account.	Generic principle 11 was violated through specific principle 11/6.			
11/6: Lacking anticipation of domain misunderstanding by analogy.	User is unaware that discount is only possible on return fares.	3		User information design.
11/7: System does not separate when possible between the needs of novice and expert users.	Difficult to test through identified cooperativity problems.			
12: System does not consider legitimate user expectations as to its own background knowledge.	Generic principle 12 was violated through specific principle 12/8.			
12/8: Missing system domain knowledge and inference.	Temporal inference; inference from negated binary option.	4		Inference design.
13: System does not initiate repair or clarification meta-communication in case of communication failure.	Generic principle 13 was violated through specific principles 13/10 and 13/11.			

13/9: System does not initiate repair if it has failed to understand the user.	Easy to provide once it has been decided to follow 13/9.			
13/10: Missing clarification of inconsistent user input.	System jumps to wrong conclusion.	5		Clarification question design.
13/11: Missing clarification of ambiguous user input.	System jumps to wrong conclusion.	5	2	Clarification question design.

Figure 3. Typology of the 119 problems of cooperative dialogue design identified in the user test. In the left-most column, 1/1 refers to generic principle 1 which subsumes the stated specific principle 1. The number (N) of occurrences of each problem is shown as are the occurrences of transaction failures (TF) per problem type. The right-most column shows the cause(s) of the problems and hence what needs to be repaired to prevent those problems from occurring. Grey shading indicates that no cases were observed of a certain problem type in the user test. Suggested reasons why this was not the case have been added in the second column from the left.

generic principle 10, namely that it *should be feasible* for users to do what the system asks them to do. Despite the system's introduction which instructed users to answer the system's questions briefly and one at a time and use the keywords 'change' and 'repeat' for meta-communication purposes, a significant number of violations of those instructions occurred. We hypothesise that, for many users, it is not cognitively possible to modify their natural dialogue behaviour as much as required in the system's introduction to itself.

In analysing the results of the user test, it turned out that, suitably re-phrased, the 13 generic principles and 11 specific principles constituted a complete typology of the system cooperativity problems identified. For classification purposes the principles were rephrased to express the cooperativity problems they had been created to prevent. For instance, the principle "Make your contribution as informative as is required (for the current purposes of the exchange)" was re-phrased as "The system provides less information than required", cf. Cooperativity Problem 1 in Figure 3. Of the 24 types of cooperativity problem, 16 were found in the user test as shown in Figure 3. The figure also provides a brief characterisation of the actual cases of each cooperativity problem type. We also analysed the likely reasons why 8 problem types were absent from the user test corpus (see Figure 3).

Based on the typology each problem identified in the user test was described in terms of its symptom (**S**), a diagnosis (**D**) was made and a cure (**C**) proposed. Figures 4-6 illustrate the analysis with respect to the three different problem cases presented in Figure 1. In the figures, U means user and S means system.

S: S: Are you particularly interested in discount?. U: Yes please. ... S: At which time? U: 7:20. S: There is no departure at 7:20. The closest other red departure is at 10:50.
D: The system provides insufficient information. It does not tell that there is a blue departure at 7:20.
C: The system should provide sufficient information, e.g. by telling that there is no red departure but that there is a blue departure at the chosen hour.

Figure 4. Cooperativity problem 1 in dialogue G14a. The system provides less information than required, thereby misleading the user.

S: The user wants to de-select discount and enters a loop. Eventually the user figures out that 'change' can be used repeatedly to backtrack more than one step, cf. Figure 2.
D: The user has not been told about repeated use of 'change'.
C: Provide sufficient instruction to users on how to make repeated use of 'change'.

Figure 5. Cooperativity problem 10/5 in dialogue G14a. The instructions on how to interact with the system are unclear or missing.

S: S: Are you particularly interested in discount? U: Yes. S: At which date will the journey start? U: Monday. S: Monday January 16. At which time? U: 7:20. S: There is no departure at 7:20. The closest other red departure is at 10:50. Do you want this departure?.
D: There is, in fact, a departure at 7:20 but it does not offer discount. The system implicitly assigns priority to discount over departure time without sufficient reason.
C: S should ask U about priority, e.g.: "7:20 is not a discount departure. Red discount can be obtained on the departures at x, y and z. Which departure do you want?"

Figure 6. Cooperativity problem 13/10 in dialogue G14a. Clarification of inconsistent user input is missing.

4. CONCLUSION

The paper has described and illustrated a method for detecting problems of spoken dialogue interaction by comparing expected and actual user-system exchanges. Based on a consolidated set of principles for cooperative spoken human-machine dialogue, a typology of non-cooperative system dialogue behaviour was proposed. The typology was applied to the problems of system dialogue cooperativity detected in a controlled user test of an SLDS. The typology was found useful and sufficient for classifying the problems and providing clues to repair. This suggests that the typology might be

useful more generally for classifying and repairing system dialogue cooperativity problems.

A secondary finding was that the template which was used in describing expected and actual dialogue interaction (Figure 1), might gain in usability if more explicitly representing the actual user and system utterances. Our cooperativity problems analysis often required use of the full transcriptions and sometimes also of the logged transactions between the system modules.

Ideally, however, all non-cooperative system dialogue behaviours should be *prevented* through good dialogue design rather than being identified, classified and repaired at the post-implementation stage. We believe that the principles for cooperative spoken dialogue design underlying the presented typology might serve the purpose of problem prevention if used as design guidelines [2].

REFERENCES

1. Bernsen, N.O., Dybkjær, H., and Dybkjær, L. "Cooperativity in human-machine and human-human spoken dialogue," *Discourse Processes*, 21: 2, 213-236, 1996.

2. Bernsen, N.O., Dybkjær, H., and Dybkjær, L. "Principles for the design of cooperative spoken human-machine dialogue," *Proc. of ICSLP '96* Philadelphia, Oct. 1996 (to appear).

3. Bernsen, N.O., Dybkjær, L., and Dybkjær, H. "User errors in spoken human-machine dialogue," *Proc. of ECAI '96 Workshop on Dialogue Processing in Spoken Language Systems*, Budapest, 1996 (to appear).

4. Dybkjær, L., Bernsen, N.O., and Dybkjær, H. "Evaluation of spoken dialogues. User test with a simulated speech recogniser," *Report 9b from the Danish Project in Spoken Language Dialogue Systems*, Roskilde University, 1996.

5. Grice, P. "Logic and conversation." In P. Cole and J.L. Morgan (Eds.), *Syntax and Semantics,* Vol. 3, *Speech Acts,* Academic Press, New York, 41-58, 1975. Reprinted in P. Grice, *Studies in the Way of Words,* Harvard University Press, Cambridge MA, 1989.