PRINCIPLES FOR THE DESIGN OF COOPERATIVE SPOKEN HUMAN-MACHINE DIALOGUE

Niels Ole Bernsen, Hans Dybkjær and Laila Dybkjær

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

ABSTRACT

This paper presents a consolidated set of 24 principles of cooperative spoken human-machine dialogue which are based on the development and controlled user testing of the dialogue component of the Danish dialogue system as well as on comparison with human-human dialogue theory. Potentially, the principles could be used as effective and systematic dialogue development and evaluation tools both during early design and in later phases of dialogue evaluation.

1. INTRODUCTION

Today's dialogue model design for spoken language dialogue systems (SLDSs) development is largely based on empirical techniques, such as the Wizard of Oz (WOZ) method and, for simple dialogues, implement-test-and-revise procedures based on emerging development platforms. These techniques mainly build on designers' common sense, experience and intuition, and on trial and error. WOZ supports the evaluation of quantitative and qualitative aspects of the dialogue model by producing data material on the interaction between a (fully or partially) simulated system and its users. WOZ is preferable to implement-test-and-revise when the costs of revising a seriously flawed implemented system are high. However, even WOZ does not tell how to design a habitable dialogue model, which evaluation metrics to use, nor whether the designers have overlooked important problems of user-system interaction. There is thus a strong need for improved tools to support habitable dialogue model design and reduce development cost and risk.

This paper presents a consolidated set of 24 principles of cooperative spoken human-machine dialogue. Potentially, the principles could be used as effective and systematic dialogue development and evaluation tools both during early design and in later phases of evaluation. This would significantly reduce the number of WOZ iterations needed to design habitable systems as well as reduce the risk of implement-test-and-revise methods. Dialogue cooperativity is crucial to habitable, task-oriented spoken human-machine dialogue. More or less tacitly, SLDSs designers have always relied on cooperative users. However, to ensure a habitable dialogue and support users in producing utterances which can be comprehended by the system, it is mandatory that the system's dialogue be cooperative as well. The presented principles state properties which should be controlled for to produce a cooperative dialogue model and support problem detection and diagnosis during evaluation. The principles were derived from a corpus of WOZ-simulated task-oriented spoken human-machine dialogue collected during the development of the dialogue component of the Danish dialogue system (Section 2). They were refined through comparison with an established body of maxims of cooperative human-human dialogue (Section 3). Including those maxims as a subset, the principles were then tested on the data from the user test of the implemented system (Section 4). The test showed that, with minor additions and revisions, the principles were capable of accounting for all the dialogue design problems encountered in the user test corpus. Examples are presented of how the principles are used in dialogue design evaluation. The concluding discussion (Section 5) addresses issues involved in developing the principles into a quasi-complete and practically useful set of design and evaluation guidelines.

2. CONSTRUCTING PRINCIPLES OF COOPERATIVE DIALOGUE

The Danish dialogue system may be briefly described as follows. The prototype addresses the domain of domestic airline ticket reservation. The system is a walk-up-and-use application which runs on a PC with a DSP board and is accessed over the telephone. The system understands speaker-independent continuous spoken Danish with a vocabulary of about 500 words and uses system-directed domain communication combined with keyword-based, user-initiated meta-communication. The prototype runs in close-to-real-time. The system is representative of advanced state-of-the-art systems. Comparable SLDSs are, e.g. [1,4].

The dialogue model for the Danish dialogue system was developed by the Wizard of Oz (WOZ) experimental prototyping method. Seven WOZ iterations involving a total of 24 users were performed to produce a dialogue model which satisfied the given design constraints [5]. The WOZ experiments produced a transcribed corpus of 125 scenario-based, task-oriented human-machine dialogues corresponding to approximately seven hours of spoken dialogue.

A major concern during WOZ was to detect problems of user-system interaction. Eventually, the following two approaches were used to systematically discover such problems: (i) prior to each WOZ iteration we matched the scenarios to be used against the current dialogue model in order to discover and remove po-

tential user problems. The dialogue model was represented as a graph structure with system phrases in the nodes and expected contents of user answers along the edges. If a deviation from the graph occurred during the matching process, this would indicate a potential dialogue design problem which should be removed, if possible. (ii) The recorded dialogues were plotted onto the graph representing the current dialogue model. As in (i), graph deviations indicated potential dialogue design problems. Deviations were marked and their causes analysed whereupon the dialogue model was revised, if necessary.

At the end of the WOZ design phase, all problems of interaction uncovered during WOZ were analysed and represented as violations of principles of cooperative dialogue. Each problem was considered a case in which the system, in addressing the user, had violated a principle of cooperative dialogue. The principles were made explicit, based on the problems analysis. The WOZ corpus analysis led to the identification of 14 principles of cooperative spoken human-machine dialogue based on analysis of 120 examples of user-system interaction problems [2]. If the principles were observed in the design of the system's dialogue behaviour, we assumed, this would serve to reduce the occurrence of user dialogue behaviour that the system had not been designed to handle.

3. COMPARISON WITH GRICE'S THEORY

The 14 principles of cooperative spoken human-machine dialogue were refined and achieved their present formulation as shown in Figure 1 through comparison with Grice's Cooperative Principle and maxims for cooperative human-human dialogue [6]. Grice's Cooperative Principle is a general principle which says that, to act cooperatively in conversation, one should make one's "conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which one is engaged" [6].

Dialogue Aspect	GP No.	SP No.	Generic or Specific Principle
Group 1: **Informativeness**	GP1		*Make your contribution as informative as is required (for the current purposes of the exchange).
		SP1	Be fully explicit in communicating to users the commitments they have made.
		SP2	Provide feedback on each piece of information provided by the user.
	GP2		*Do not make your contribution more informative than is required.
Group 2: **Truth and evidence**	GP3		*Do not say what you believe to be false.
	GP4		*Do not say that for which you lack adequate evidence.
Group 3: **Relevance**	GP5		*Be relevant, i.e. be appropriate to the immediate needs at each stage of the transaction.
Group 4: **Manner**	GP6		*Avoid obscurity of expression.
	GP7		*Avoid ambiguity.
		SP3	Provide same formulation of the same question (or address) to users everywhere in the system's dialogue turns.
	GP8		*Be brief (avoid unnecessary prolixity).
	GP9		*Be orderly.
Group 5: **Partner asymmetry**	GP10		Inform the dialogue partners of important non-normal characteristics which they should take into account in order to behave cooperatively in dialogue. Ensure the feasibility of what is required of them.
		SP4	Provide clear and comprehensible communication of what the system can and cannot do.
		SP5	Provide clear and sufficient instructions to users on how to interact with the system.
Group 6: **Background knowledge**	GP11		Take partners' relevant background knowledge into account.
		SP6	Take into account possible (and possibly erroneous) user inferences by analogy from related task domains.
		SP7	Separate whenever possible between the needs of novice and expert users (user-adaptive dialogue).
	GP12		Take into account legitimate partner expectations as to your own background knowledge.
		SP8	Provide sufficient task domain knowledge and inference.
Group 7: **Repair and clarification**	GP13		Initiate repair or clarification meta-communication in case of communication failure.
		SP9	Provide ability to initiate repair if system understanding has failed.
		SP10	Initiate clarification meta-communication in case of inconsistent user input.
		SP11	Initiate clarification meta-communication in case of ambiguous user input.

Figure 1. Principles of cooperative system dialogue. GP means generic principle. SP means specific principle. The principles that were found violated in the user test are indicated in dark shading. Grice's maxims are marked with an asterisk.

Grice proposed that the CP can be further explicated in terms of four groups of simple maxims which are neither claimed to be jointly exhaustive nor to be mutually exclusive. A detailed discussion of, and comparison with, Grice's work is presented elsewhere [2]. The comparison between our principles and Grice's maxims yielded a clear-cut result. It turned out that the principles include the maxims as a subset (Figure 1). In addition, the principles manifest aspects and principles of cooperative task-oriented dialogue which were not addressed by Grice. The distinction between *principle* and *aspect* (Figure 1) is useful because an aspect represents the property of dialogue addressed by a particular maxim or principle. Finally, the comparison made us aware of the distinction between *generic* and *specific* principles. Grice's maxims are all generic. However, a generic principle may subsume one or more specific principles which specialise the generic principle to certain classes of phenomena. Although subsumed by generic principles, we believe that specific principles are useful to SLDS dialogue design (see Section 5).

4. TEST OF THE PRINCIPLES

Having consolidated the principles of cooperative system dialogue through comparison with Gricean theory, we tested the revised principles in the user test of the implemented Danish dialogue system. It should be noted that the principles had not been systematically used as guidelines in designing the implemented system. The user test thus cannot be considered a test of the principles in the crucial sense of indicating to which extent successful design of cooperative (system) dialogue behaviour may result from using the principles as design guidelines. Rather, the user test worked as a test of the *scope* of the principles. It provided an indication of whether the present set of principles converge on a complete set. If analysis of user-system interaction in the user test would produce a significant number of novel principles, then the present set of principles are still far from complete. On the other hand, if the analysis failed to produce more principles, then this would at least suggest that convergence on completeness is well under way.

The system was tested with 12 users from the intended user population. The resulting 57 scenario-based reservation dialogues covered the full functionality of the system. Figure 2 shows an excerpt from one of the dialogues. The transcribed dialogues were analysed to detect problems of dialogue interaction, as follows. Based on the dialogue structure, a template was built which contained the system's questions. For each scenario, key contents of normative system and user answers were filled into the template. The key contents of the actual dialogues were then plotted into the template, cf. Figure 3. Comparison between normative and actual system and user utterances led to the identification of more than 100 dialogue interaction problems, splitting into (a) dialogue design problems and (b) user errors. This section focuses on (a). For a full account of the user errors see [3]. An example of a user error is shown in Figure 2 (U3) and identified as E2 in Figure 3. The user ignores part of the system's feedback (February 10) and does not change the date that was misunderstood by the system.

S1: Are you particularly interested in discount?
U1: Yes.
S2: On which date will the journey start?
U2: The first weekend in February.
S3: Friday February 10th. At which time?
U3: It must be on Saturday at 7:20.
S4: There is no departure at 7:20. The closest other red departure is at 10:50. Do you want this departure?
U4: No.
S5: At which time?

Figure 2. Excerpt of the reservation dialogue T32a translated from the Danish. S means system and U means user.

System questions	**Normative user answers**	**Actual user answers**	**Problems**
Interested in discount	no/yes	yes	
Day of departure (out)	February 4	first weekend in February (Friday February 10)	
Hour of departure (out)	7:20	Saturday at 7:20 [attempt to change Friday] (no departure) no [does not want one from list]	E2, GP10, SP10, GP1

Figure 3. Key contents of the expected (normative) and actual user-system exchanges in the excerpt from dialogue T32a in Figure 2. In the third column the system's understanding of user input is indicated in parentheses when different from what was expected. Explanatory comments are in square brackets. GP means generic principle, SP means specific principle and E means user error.

S: S: At which time? U: It must be on Saturday at 7:20. S: There is no departure at 7:20. The closest other red departure is at 10:50.
D: The system withholds important 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. Violation of principle GP1: Make your contribution as informative as is required (for the current purposes of the exchange).

S: S: Are you particularly interested in discount? U: Yes. ... S: At which time? U: It must be on Saturday at 7:20. S: There is no departure at 7:20. The closest other red departure is at 10:50.
D: There is a departure at 7:20, but without discount. S gives priority to discount over time without proper reason.
C: S should ask U about priority: 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 5. Violation of principle SP10: Initiate clarification meta-communication in case of inconsistent user input.

S: S: Friday February 10. At which time? U: It must be on Saturday at 7:20.
D: The user is too occupied with the present problem to remember to use the keyword 'change'.
C: 'Change' is not natural. Allow natural sentences in meta-communication.

Figure 6. Violation of principle SP10: Inform the dialogue partners of important non-normal characteristics which they should take into account in order to behave cooperatively in dialogue. Ensure the feasibility of what is required of them.

In the analysis of dialogue design problems, each identified problem was (a) characterised with respect to its *symptoms*, (b) a *diagnosis* was made, and (c) a *cure* was proposed. The analysis of the dialogue design problems identified in Figure 3 is shown in Figures 4 to 6. In these figures, **S** means symptom, **D** means diagnosis and **C** means cure. S means system and U user.

The user test confirmed the broad coverage of the principles with respect to cooperative spoken user-system dialogue. Almost all of the 119 individual dialogue design problems identified could be ascribed to violations of the cooperative principles. Only three additions had to be made to the principles established during WOZ. Two specific principles of meta-communication were added, i.e. SP10 and SP11 in Figure 1. Since meta-communication had not been simulated during WOZ and the WOZ corpus therefore contained few examples of meta-communication, this came as no surprise.

More interestingly, we had to add a modification to GP10, namely that it *should be feasible* for users to do what they are asked to do. For instance, in its introduction the system asks users to use the keywords 'change' and 'repeat' for meta-communication purposes and to answer the system's questions briefly and one at a time. Despite the introduction, a significant number of violations of those instructions occurred in the user test. For instance, users attempted to make changes through full-sentence expressions rather than by saying 'change' (Figure 6). Almost all of these cases led to misunderstanding or non-understanding. These violations of clear system instructions were initially categorised as user errors. However, upon closer analysis they were re-categorised as dialogue design problems. Although the system has clearly stated that it has non-normal characteristics due to which users should modify their natural dialogue behaviour, this is not cognitively possible for many users.

5. CONCLUDING DISCUSSION

We have described the development of a set of principles of cooperative spoken human-machine dialogue. The principles were shown to include as a sub-set a well-established body of maxims of cooperative human-human dialogue. The principles have a considerably wider scope than the maxims and split into generic and specific principles. At the generic level, the principles address three aspects of cooperative dialogue which are not addressed by the maxims. The specific principles have no counterparts among the maxims. Yet these principles appear useful to SLDS design. What we need in order to discover dialogue problems at an early stage, is to know what to look for in the emerging dialogue structure. The specific principles extend the generic principles by further specifying their import. Analysis of the corpus that was produced from the user test of the implemented system shows that the generic principles, including an addition to GP10 (Section 4), are able to subsume all the identified dialogue problems. The user test corpus analysis increased the number of specific principles by two which both address dialogue issues that were not prominent in the original corpus of simulated human-machine dialogue. Jointly, these results suggest that the principles of cooperative system dialogue represent a step towards a more or less complete and practically applicable set of design guidelines for cooperative SLDS dialogue.

Two further lines of investigation must be pursued in order to test and improve the completeness and practical utility of the principles. First, it cannot be excluded at this stage that the principles are somehow tied to the task domain and dialogue complexity of our particular SLDS. Analysis of dialogue problems caused by systems that address different task domains or have lower or higher dialogue complexity than our system may thus reveal additional specific or even generic principles. Secondly, principles of cooperative dialogue are not necessarily the same as practically applicable design guidelines. An SLDS designer who simply receives the principles as represented in Figure 1, may not quite know what to do with them in practice. We believe that a representation of the principles which includes their justification as well as an extensive set of example violations might be of help. Current work aims to provide the necessary support for the principles to become of maximum benefit to dialogue design practice, thereby reducing the cost of producing habitable dialogue for SLDSs.

REFERENCES

1. Aust, H., Control in Automatic Inquiry Systems," *Proceedings of the ESCA Workshop on Spoken Dialogue Systems*, Vigsø, 121-124, 1995.

2. Bernsen, N.O., Dybkjær, H., and Dybkjær, L. "Cooperativity in Human-Machine and Human-Human Spoken Dialogue", *Discourse Processes*, Vol. 21, No. 2, 1996, 213-236.

3. Bernsen, N.O., Dybkjær, L., and Dybkjær, H. "User Errors in Spoken Human-Machine Dialogue", *Proceedings of the ECAI '96 Workshop on Dialogue Processing in Spoken Language Systems*, Budapest, 1996 (to appear).

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, 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, 1993, 947-50.

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