Scenario Design for Spoken Language Dialogue Systems Development

Laila Dybkjær, Niels Ole Bernsen and Hans Dybkjær

Centre for Cognitive Science, Roskilde University
PO Box 260, DK-4000 Roskilde, Denmark
E-mail: laila@cog.ruc.dk, nob@cog.ruc.dk, dybkjaer@cog.ruc.dk

ABSTRACT

Adequate data acquired through the Wizard of Oz experimental prototyping method are still crucial to the cost-effective development of advanced spoken language dialogue systems. One important source of data corruption is the unintended priming of subjects through the task scenario representations used in the experiments. The paper presents the three sets of development and test scenario representations which were used in the Danish Dialogue project. Based on the third set of scenarios an experiment was conducted to investigate the effects of a masking strategy which effectively avoids the possibility of priming the WOZ subjects. The experimental results are presented and discussed.

1. THE ROLE OF SCENARIOS IN SPOKEN LANGUAGE DIALOGUE SYSTEMS DESIGN

Scenarios are important tools in spoken language dialogue systems (SLDSs) development and testing. Nonetheless, the SLDS literature has little to say about scenario design and on the many problems to be aware of. This paper presents conclusions from the Danish Dialogue project as regards the construction, representation and use of scenarios in SLDS design. Over the last three years, the authors have designed and implemented the dialogue part of a realistic SLDS prototype, P2, which has been developed in collaboration with the Center for PersonKommunikation at Aalborg University and the Centre for Language Technology in Copenhagen. The domain of P2 is Danish domestic airline ticket reservation.

The P2 dialogue model was developed by means of the Wizard of Oz (WOZ) experimental prototyping method [3, 5, 6]. WOZ is an iterative process of testing and revising the dialogue model, which continues until the model is found acceptable for implementation. The implemented dialogue model is subjected to further testing. Each of these tests requires the use of predefined scenarios. The purpose of using scenarios is to develop and test the dialogue model on the basis of realistic situations of use of the SLDS under construction. Scenarios prescribe tasks embedded in realistic situations of use, which subjects, i.e. the persons acting as users, are asked to perform through spoken dialogue with the system. The scenario-based dialogues provide crucial data on user-system behaviour during dialogue, i.e. on user reactions to various aspects of the system's behaviour and vice versa, as well as on users' sublanguage vocabulary, utterance length, dialogue act types, number of turns per scenario, grammatical complexity, utterance ungrammaticality, task ordering preferences, problem-solving strategies, etc. An additional aim in using scenarios is to achieve some amount of systematicity in the testing process. There is, however, no known method for designing scenarios which are representative of all possible situations of use of the artefact being designed [7]. So the basic problem in scenario design is to capture, in a limited set of scenarios, as much as possible of the space of possible situations of use.

2. THE P2 DEVELOPMENT AND TEST SCENARIOS

Seven WOZ iterations were performed to design the P2 dialogue model which was then implemented and tested. Three different sets of scenarios were constructed in the process: one set for the first four WOZ iterations, a second set for the following three iterations, and a third set for the prototype user test.

The first set of scenarios was relatively small, comprising ten scenarios which were not designed to systematically represent as many situations of use as possible. The scenarios were simply considered as a set of cases for which the system should work and were mainly used for domain and task exploration and training by the two system designers acting as wizard and subject, respectively. The subject often revised a scenario and sometimes invented a new scenario on the fly which was never written down. The second set of scenarios was

designed on the basis of the dialogue structure that emerged from the fourth WOZ iteration. By then the scenarios could be designed in a more systematic way, as most of the domain and task structure had been uncovered.

The first two sets of scenarios conform to the notion of *development* scenarios, i.e. scenarios which are intended to more or less systematically cover the intended system functionality and are normally designed by the system designers [2]. Our third set of scenarios rather correspond to the notion of *evaluation and test* scenarios [2]. Based on the WOZ scenario experiences, we carefully considered what to test and why. We decided, i.a., not to do user testing on a number of possible but unlikely cases of communication failure. These have instead been tested in the black-box test. 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 [1, 4], it was possible to extract from the dialogue structure a set of sub-task components, such as number of travellers, age of traveller, and discount or normal fare, any combination of which should be handled by P2. The scenarios were generated through systematically combining these components.

3. MASKING THE SCENARIO REPRESENTATIONS

A *scenario representation* represents a task which subjects have to perform through dialogue with the system. A central problem addressed in our design of the test scenarios was the following. The sub-language vocabulary of P2 had been derived from the scenario-based WOZ dialogues. During the later WOZ experiments we discovered that subjects tended to repeat the date and hour of departure expressions used in the scenarios. This is a problem because a vocabulary defined on the basis of dialogues in which users model scenario phrases may not be sufficiently representative of realistic language use. On the other hand, scenarios clearly have to describe, to some necessary extent, the tasks to be performed by the subjects. It is not obvious, therefore, how one can avoid providing subjects with words or phrases which they will tend to repeat when answering the system's questions, rather than selecting their own forms of expression.

We decided to investigate how to make it impossible for subjects to model the test scenario representations in unintended ways. We therefore had to consider which information to mask, and how. For each sub-task in the dialogue structure the type of question posed by the system was categorised. There were four types of question. One type invited a yes/no answer. A second type invited an answer containing an element chosen from an explicit list of alternatives, i.e. a multiple choice question. The third type invited the user to state a proper name or something similar to a proper name, such as an airport name or the user's own customer number. The fourth type were open questions about some topic, such as the date of departure. The interesting point is that in the first three cases, the key information can only be co-operatively expressed in one of several closely related ways, which means that it does not matter if users model the expressions of the scenario representation. It is only in the fourth case that co-operative user answers may express the key information in many different ways. It is exactly in these cases that it is desirable to know how users would normally express themselves and hence mandatory to prevent them from modelling the scenario representations. Questions of this type all concerned date and hour of departure. We therefore decided to concentrate on masking the scenario representations as regards date and hour of departure in order to avoid priming of the subjects.

In general, dates are either expressed in relative terms as being relative to, e.g., today, or in absolute terms as calendar dates. Hours are either expressed in quantitative terms, such as, e.g., 'ten fifteen a.m.' or 'between ten and twelve', or in qualitative terms, such as 'in the morning' or 'before the rush hour'. The masked scenario representations never contained re-usable expressions referring to dates or hours of departure. Relative dates were expressed using a list of the days from today onwards. Absolute dates were expressed as calendar indices such as might be used by a customer when booking a flight. Quantitative hours were expressed using the face of a clock. Qualitative hours were expressed using (travel) *goal state* temporal expressions rather than departure state temporal expressions, e.g. 'they want to arrive early in the evening'. This means that the user (subject), in order to determine when it would be desirable to depart, had to make an inference from the hour indicated in the scenario representation, thus excluding the possibility of priming.

4. THE EXPERIMENT

To test the effect on users' language of masking all temporal expressions in the scenario representations, subjects were divided into two groups, one serving as control group. Each test scenario was represented in two

Jens and Marie Hansen (ID-numbers 1 and 4) and Steen and Jane Sørensen (ID-numbers 6 and 7) live in Copenhagen. They will attend a meeting in Aarhus as shown in the calendar which starts with today in boldface and shows the day of departure as the next day in boldface. The meeting starts and ends as shown on the two clocks. The flight takes about 35 minutes. The

time to get from the airport to the meeting is about 45 minutes. The customer number is 4.

M T W T F S S M T W **T** F S S

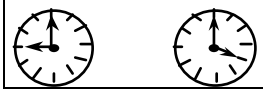

Figure 1. An analogue graphic scenario representation.

different ways. The masked version combines language and analogue graphics (cf. Fig. 1) whereas the control group version uses standard linguistic text (cf. Fig. 2) and roughly corresponds to the one used in the second set of WOZ scenarios.

Jens and Marie Hansen (ID-numbers 1 and 4) and Steen and Jane Sørensen (ID-numbers 6 and 7) live in Copenhagen. They will attend a meeting in Aarhus on Thursday next week. The meeting starts at 9 am and ends at 4 pm. The flight takes about 35 minutes. The time to get from the airport to the meeting is about 45 minutes. Therefore they want the departure at 7:20 and, for the return journey, the departure at 17:30. The customer number is 4.

Figure 2. A text scenario corresponding to the graphic scenario of Figure 1.

The user test involved a total of 12 subjects. Each subject received 4 scenarios. Subjects sometimes redid a scenario if they did not succeed the first time. Six of the subjects received text scenarios and the six other subjects received graphic scenarios. 32 dialogues based on text scenarios and 25 dialogues based on graphic scenarios were recorded, cf. Table 1.

Table 1. General data on the two scenario types. An * indicates that the figures only concern the dialogue parts on date and time.

	text scenarios	graphic scenarios
no. of subjects	6	6
no. of different scenarios	20	20
no. of dialogues	32	25
no. of user turns*	181	178
no. of user words*	705	451
no. of different user words*	85	63
average user utterance length*	3,9	2,5

5. COMPARISON OF RESULTS

Our hypotheses, as regards the parts concerning date and time, were that (1) there would be a massive priming effect from the text scenarios and none from the graphic scenarios, and (2) the dialogues based on graphic scenarios would contain a richer vocabulary material than those based on text scenarios in terms of (i) total number of different words and (ii) words out of vocabulary. The first hypothesis was confirmed, the second was not. In addition, we had an unexpected but interesting result which speaks in favour of using graphic scenarios.

5.1 Priming Effects

As expected, we found a large difference between the two scenario sets. The first row of Table 2 expresses the "cleaned" number of user turns for which priming from the scenarios was possible. "Cleaned" means that we have counted only the first occurrence of a user answer containing a date or a time in response to each of the four system questions concerning dates and times of out and home journey departures. In these cases there is no immediate priming from the expressions used by the system and figures are not influenced by repeated or changed user answers.

Table 2. Priming effect for WOZ7, text and graphic scenarios, respectively.

	WOZ7	text	graphic
first date and time answers	74	106	84
primed answers	59	59	1
primed out date	91%	45%	-
primed home date	83%	23%	-
primed out hour	68%	78%	-
primed home hour	73%	71%	-

Each date or time expression in the users' answers was compared to the scenario text. Complete matches and matches where *optional* parts of the date or time expression had been left out or added were counted at primed cases. If *non-optional* parts of the date or time expression had been changed, however, the case was counted as non-primed. For example, if the scenario said 'Friday the 2nd of January' then 'the 2nd of January' and 'Friday the 2nd' would count as primed but not 'the 2nd of first' which is a common Danish calendar expression.

In the text scenario dialogues, priming was not equally distributed across date and time. This may have the following explanation. The time expressions used in the

scenarios were similar to the feedback expressions used by the system and chosen from among the most common time expressions in Danish. A broader variety of date expressions was used in the text scenarios although most frequently of the form 'the 2nd of January'. Furthermore, there are several frequent date expression formats. The system's feedback was of the form 'the 2nd of first'. The decrease from 45% to 23% partly seems to be due to users changing from modelling the scenario text to modelling the system feedback when it came to answering the question about home date, and partially to the use of relative dates such as 'the same day'.

Throughout the WOZ scenarios the date format 'Friday the 2nd of January' was used, which was in accordance with the system's feedback. This and the general frequency of the expression may explain the high date priming percentage from WOZ7.

5.2 Vocabulary Effects

The use of graphic scenarios did not result in a much richer vocabulary than using the text scenarios, nor in more new words. On the contrary, dialogues based on graphic scenarios contained fewer different words, cf. Table 1. The scenario sets generated no out-of-vocabulary dates and only 9 new words for times.

Graphic scenario users massively replaced relative dates with absolute ones. This may be because people generally tend to do that on reservation tasks, or because people tend to do that in dialogue with machines which they know are inferior in language understanding. Whichever hypothesis is true, the effect is that subjects tended to standardise their date vocabulary by using exact dates rather than using their diverse relative dates vocabulary.

Similarly, graphic scenario users tended to replace qualitative time with quantitative time, although less strongly so than when replacing relative dates by absolute dates. Again, the tendency is towards exactitude at the expense of using the language of qualitative time. The effect is another limitation on the vocabulary used.

We see three implications of these findings:

(i) The introduction, in SLDSs development, of graphic scenarios is not a means of doing away with good task scenario designs which may efficiently test task domain, user language and user task performance. Good scenario design, however represented in the scenarios, is still essential to good dialogue design.

(ii) Given the fact that neither text nor graphic scenarios are able to elicit the full diversity of potential user language vis-á-vis the system, field trials of SLDSs developed by means of scenarios are still essential to the design of workable real-life systems.

(iii) The good news is that, in the graphic scenarios, subjects demonstrated a clear tendency towards expressing themselves in exact terms for dates and times.

5.3 An Unexpected Result

We found a significant difference in tokens (words) per turn between dialogues based on text and graphic scenarios, respectively, cf. Table 1. Apart from the scenario representations, all subjects received identical material. They were asked the same questions, and they all believed that they communicated with a machine. Task contents were identical in the two sets of scenarios. There are no significant differences between the two user populations. The most plausible explanation seems to be that the observed difference is produced by the different scenario representations themselves. In the text-based dialogues, subjects read aloud from their scenario representation. *They produce, in effect, spoken language which is not spontaneous, or which is not spoken discourse but read-aloud text.*

In the graphic-based dialogues, subjects cannot read aloud from their scenario representation because it does not contain textual expressions for date and time. To communicate the task contents of the graphic scenarios, subjects *have to* produce spontaneous spoken language.

When developing realistic SLDS applications, we need to copy or imitate realistic situations of use to the extent possible. Use of read-aloud text in communicating with the system is hardly close to realistic situations of use of most SLDSs. This would imply that textual development scenarios which afford read-aloud solutions to communications with the system are unsuited for SLDS development. Other means of solution should be found in order to ensure that subjects do produce spontaneous spoken language in communicating with the system. One solution is to use analogue graphic representation of scenario sub-tasks when necessary. We have shown that this is possible, and that it works, for the representation of temporal scenario information.

7. REFERENCES

[1] 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 Conference*, Yokohama, Japan, September 1994, 875-78.

[2] Campbell, R.L.: Will the Real Scenario Please Stand Up? *SIGCHI Bulletin* 24, 2, 1992, 6-8.

[3] 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, 947-50.

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

[5] Dybkjær, L. and Dybkjær, H.: Wizard of Oz Experiments in the Development of a Dialogue Model for P1. *Report 3 from the Danish Project in Spoken Language Dialogue Systems.* Roskilde University, February 1993.

[6] Fraser, N.M. and Gilbert, G.N.: Simulating Speech Systems. *Computer Speech and Language* 5, 1991, 81-99.

[7] Klausen, T. and Bernsen, N.O.: CO-SITUE: Towards a methodology for constructing scenarios. In Hollnagel, E. and Lind, M. (Eds.): *Proceedings of the Fourth European Meeting on Cognitive Science Approaches to Process Control (CSAPC '93): Designing for Simplicity*. Copenhagen, August 1993, 1-16.