# Spoken Language Dialogue Systems

**Report 9b, February 1996**

# Evaluation of Spoken Dialogues

# User Test with a Simulated Speech Recogniser

**Volume 1: Summarising report**

CPK - Center for PersonKommunikation, Aalborg University
CCS - Centre for Cognitive Science, Roskilde University
CST - Centre for Language Technology, Copenhagen

Printed at Roskilde University
February 1996
ISBN 87-7349-301-5

The present report is part of the documentation series from the strategic research programme *Spoken Language Dialogue Systems*. The programme is sponsored by the Danish Technical Research Council. The project partners are the Center for PersonKommunikation (CPK) (earlier Speech Technology Centre (STC)) (coordinating partner) at Aalborg University, the Centre for Cognitive Science (CCS) at Roskilde University, and the Centre for Language Technology (CST), Copenhagen.

# Authors:

**Laila Dybkjær, CCS**
**Niels Ole Bernsen, CCS**
**Hans Dybkjær, CCS**

The project partners can be contacted at:

**Center for PersonKommunikation (CPK):**

Paul Dalsgaard
Aalborg University
Frederik Bajers Vej 7
DK-9220 Aalborg Ø, Denmark
Phone: +45 98 15 85 22. Fax: +45 98 15 15 83
Email: pd@cpk.auc.dk

**Centre for Cognitive Science (CCS):**

Niels Ole Bernsen
Roskilde University
P.O.Box 260
DK-4000 Roskilde, Denmark
Phone: +45 46 75 77 11. Fax: +45 46 75 45 02
Email: nob@cog.ruc.dk

**Centre for Language Technology (CST):**

Bente Maegaard

Njalsgade 80
DK-2300 Copenhagen S, Denmark
Phone: +45 35 32 90 90. Fax: +45 35 32 90 89
Email: bente@cst.ku.dk

# Contents

**Volume 1. Summarising report**

**Volume 2. Appendices including Danish text**



**Volume 3. Appendices in English**

# 1 About this report

This report is report 9b in the documentation series from the research programme *Spoken Language Dialogue Systems*. Report 9b focuses on dialogue evaluation and is based on the user test of the Danish spoken language dialogue system prototype. The user test was carried out with a simulated speech recogniser. The system test of the prototype was presented in report 9a.

## 1.1 Structure of the report

The present report is divided into three volumes. The first volume contains a short summarising report in English. The second and the third volumes are each divided into a number of appendices that present the user test material, test data and analysis of the data. The second volume mainly contains Danish text whereas the third volume is in English.

*Volume 1: Summarising report*

The user test has been discussed in four published papers. The summarising report provides abstracts of the published papers and of two working papers on priming and vocabulary and on meta-communication, respectively. The full working papers are included among the appendices in Volume 3. Papers in preparation are briefly mentioned. Finally, the summarising report contains examples in English of the central parts of the data material. The full data material can be found in Volumes 2 and 3.

*Volume 2: Danish Appendices 1-7*

The appendices in Volume 2 include all the data material in Danish. Appendices 1-4 present the material given to subjects prior to their interaction with the system. Appendix 4 also contains subjects' answers to the questionnaire and to a brief telephone interview after the interaction. Appendix 5 presents data on age and background of subjects. Appendix 6 describes the creation and markup of the transcriptions. Finally, Appendix 7 presents the transcribed dialogues.

*Volume 3: English Appendices 8-16*

The appendices in Volume 3 include all data material in English together with two working papers. Appendices 8-14 present analysed data based on the transcribed dialogues. The aim of the analysis was to reveal problems in the dialogues and obtain an overview of their types, causes and remedies, if any. Appendices 15 and 16 present the short working papers of which abstracts are included in the main report.

## 1.2 The Danish dialogue system and the user test

The Danish dialogue system is a reservation system for Danish domestic flights. 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. It has five main components: 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 implemented system excluding the recogniser, was subjected to scenario-based testing with naive users in order to collect data on dialogue performance and sublanguage coverage. A wizard keyed in the users' answers into a simulated recogniser. This meant that 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 real recogniser's vocabulary and which conformed to the recogniser's grammar rules. Typos were automatically corrected. Twelve subjects took part in the test. Two of them were in-house staff and knew the wizard whereas the other ten subjects were external and did not know the wizard. However, all subjects believed that they had interacted with a real system. Each subject had to carry out four scenarios. All dialogues were recorded on tape and the internal system communication was logged.

This report focuses on dialogue performance.

## 1.3 Acknowledgements

We gratefully acknowledge the contributions of all those who participated as subjects in the user test of the Danish dialogue system. We particularly wish to thank Dansk Industri and the secretaries from Dansk Industry for the time and effort they gave to the project through acting as subjects, and N.O. Coops Olsen for establishing the contacts that made it happen.

# 2 Abstracts

This chapter provides abstracts of published papers on dialogue evaluation based on the user test. In addition, abstracts are provided of two short working papers on meta-communication and on statistical measurements of the user test data, respectively. Two papers in preparation are briefly mentioned. For each paper are provided (i) a reference to the full paper, (ii) an abstract, and (iii) a list of illustrations presented in the full paper.

## 2.1 Scenario design for spoken language dialogue systems development

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, Denmark, May 30 - June 2, 1995, 93-96.

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.

Illustrations:

An analogue graphic scenario representation.

A text scenario corresponding to the graphic scenario.

General data on the two scenario types.

Priming effects for WOZ7, text and graphic scenarios, respectively.

## 2.2 Design, formalisation and evaluation of spoken language dialogue

Dybkjær, H., Dybkjær, L. and Bernsen, N.O.: Design, Formalisation and Evaluation of Spoken Language Dialogue. *Proceedings of the TWLT9 Workshop*, Enschede, June 9, 1995, 67-82.

Dialogue model development is a major part of spoken language dialogue systems development. The dialogue model development process is a series of iterative interactions between design, formalisation and evaluation. This paper reports on the corpus-based development process of the dialogue model for the Danish dialogue system. The paper first describes dialogue model design through use of the Wizard of Oz method. Secondly, the continued formalisation of the dialogue model during the implementation phase is reported. The paper goes on to describe first results of the user test of the system, comparing these with the final results of the Wizard of Oz phase. Some issues for future work are raised in the conclusion.

Illustrations:

Overall architecture of the Danish dialogue system.

The introduction graph used in WOZ7.

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

Number of questions in per cent of total number of turns.

The communication structure of the recogniser, parser, player, dialogue handler and database with a detailed view of the dialogue handler and the database which represents domain knowledge.

An analogue graphic scenario representation.

A text scenario corresponding to the graphic scenario.

A translated example of a dialogue from the user test.

Comparison of results from WOZ7 and the user test.

Subjects' answers to the questionnaires from WOZ7 and the user test.

Data on the dialogues based on two different scenario types.

Priming effects for WOZ7, text and graphic scenarios, respectively.

## 2.3 Exploring the limits of system-directed dialogue. Dialogue evaluation of the Danish dialogue system

Bernsen, N.O., Dybkjær, H. and Dybkjær, L.: Exploring the Limits of System-Directed Dialogue. Dialogue Evaluation of the Danish Dialogue System. *Proceedings of Eurospeech '95*, Madrid, September 1995, 1457-60.

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.

Illustrations:

Comparison of results from WOZ7 and the user test.

Subjects' answers to the questionnaires from WOZ7 and the user test.

## 2.4 Designing co-operativity in spoken human-machine dialogue

Dybkjær, L., Bernsen, N.O. and Dybkjær, H.: Designing Co-operativity in Spoken Human-Machine Dialogue. Paper presented at the Second Workshop on Human Comfort and Security, Brussels, 26 October 1995. To be published in *Proceedings from the Second Workshop on Human Comfort and Security,* Springer Research Reports, 1996.

Dialogue model design for spoken language dialogue systems (SLDSs) is still based mainly on common sense, experience and intuition, and trial and error, rather than on established design principles. Co-operativity in dialogue is crucial to habitable human-machine spoken dialogue.

The paper presents a set of principles of co-operative user-system dialogue which have been derived from a corpus of task-oriented spoken human-machine dialogue. The set of principles is shown to include as a sub-set an established body of principles of co-operative human-human dialogue. Analysis of results from a user test of an implemented SLDS prototype shows the set of principles to be adequate to account for the dialogue problems identified in the test corpus. Both empirical and theoretical grounds thus indicate that the principles presented in the paper may constitute a comprehensive set of guidelines for the design of co-operative human-machine dialogue.

Illustrations:

A plotted END sub-graph from WOZ6.

Part of the HOUR sub-graph from WOZ5.

Part of the HOUR sub-graph from WOZ6.

The co-operative dialogue design principles for spoken language dialogue systems and their justifications.

Grice's maxims and comments.

Generic principles of co-operative spoken dialogue.

Specific principles of co-operative spoken dialogue.

An example of system questions, expected key contents of user answers, actual key contents of user answers, and problems identified in a dialogue from the user test.

An example of a dialogue design problem: inconsistent user input.

Dialogue design problem types identified during the user test.

## 2.5 Priming and vocabulary

Dybkjær, H., Bernsen, N.O. and Dybkjær, L.: Priming and Vocabulary. Appendix 15, 1995.

Priming effects and variations in expressions of time in human-computer conversation are investigated, mainly from a quantitative/statistical point of view. Material from a user test of the Danish spoken language dialogue system is presented.

We show that subjects in this user test are primed by the written task scenarios they carry out, cf. Section 3.1.1, and that they are primed by the system. Together, the two results indicate that people are primed by the most recent, relevant linguistic expression received.

The test included a control group of subjects using scenarios in which expressions of time (dates and hours) were masked through the use of analogue graphic representations of the faces of clocks, calendar pages and the like, cf. Section 3.1.2. These subjects were expected to use more spontaneous (i.e. less primed) phrases exhibiting greater linguistic variation. Unexpectedly, it turned out that the linguistic variation in the dialogues in the control group was less than for dialogues from the other group. The explanation may be that the text group used read-aloud speech in addition to their own spontaneous phrases. However, the statistical data associated with this result have some as yet unexplained implications.

Illustrations:

Priming: definition by example.

Priming effects for text and graphics dialogues.

Priming criteria applied to graphics dialogues and text scenarios.

Priming after a system question with one or more hour expressions.

Relation between outdate and homedate priming.

Primed date and hour expression types.
Date and hour expressions accumulated.
Comparison of results from WOZ7 and the user test.
Type figures.
Basic type statistics.
Test statistics of types.
Statistics of out-of-vocabulary types and tokens.
Out-of-vocabulary types.
Statistics of tokens and turns, graphics group.
Statistics of tokens and turns, text group.
Test statistics for tokens per utterance.
Types per log(tokens).

## 2.6 Meta-communication

Dybkjær, L., Bernsen, N.O. and Dybkjær, H.: Meta-communication. Appendix 16, 1995.

The dialogue model for the Danish dialogue system was designed to be as co-operative as possible in order to avoid the need for unrestricted repair and clarification dialogue which the system would not be able to handle. However, given the quality of current speech recognisers, the need for making corrections remains no matter how co-operative the dialogue has become. In the Danish dialogue system, the user is offered the possibility of initiating repair meta-communication in order to correct misunderstandings by using the keyword 'change'. Users are also allowed to initiate clarification meta-communication dialogue through the use of the keyword 'repeat'. The system's possibilities of initiating meta-communication are limited to informing the user that it did not understand the input and, if the user does not answer for a long time, to ask if s/he is still present. The total percentage of turns spent on meta-communication in the test corpus was 3.61. Meta-communication was mainly caused by insufficient system sub-language coverage.

Illustration:

Data concerning meta-communication in the user test dialogues.

## 2.7 Papers in preparation

Two papers are in preparation based on the user test material. The first paper will present a detailed analysis of the types of dialogue design problem revealed in the test corpus, cf. Section 3.4.1. The second paper will focus on the types of user error identified in the test corpus, cf. Section 3.4.2.

# 3 Examples from Appendices 3, 7, 9 and 11

## 3.1 Test scenarios (Appendix 3)

### 3.1.1 An example of a text scenario

The following figure shows a text scenario representation (translated from Danish).

Scenario T-5-4-a

Lene Offersgaard (ID-number 4, customer number 2) is going from Copenhagen to Karup on Tuesday, February 21. She must arrive no later than nine o'clock. The flight plus subsequent transport takes about one hour and 15 minutes. Early next morning she drives to Esbjerg to attend a meeting. From there she wants to go home later that day. She wants to arrive in Copenhagen no later than at seven o'clock. The flight takes about 50 minutes.

### 3.1.2 An example of a graphic scenario

The figure below shows an analogue graphic scenario representation corresponding to the text scenario above (translated from Danish).

Scenario G-5-4-a

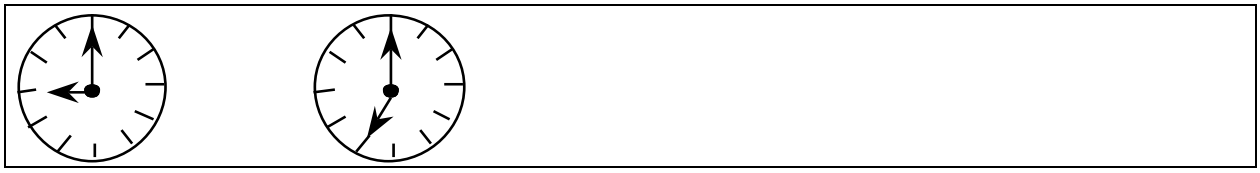
Lene Offersgaard (ID-number 4, customer number 2) is going from Copenhagen to Karup as shown in the calendar. She must arrive no later than the hour shown on the first clock. The flight plus subsequent transport takes about one hour and 15 minutes. Early next morning she drives to Esbjerg to attend a meeting. From here she wants to go home as shown in the calendar. She wants to arrive in Copenhagen no later than the hour shown on the second clock. The flight takes about 50 minutes.

**JANUARY 1995**

| M | **T** | W | T | F | S | S | **WEEK** |
|---|---|---|---|---|---|---|---|
| | | | | | | 1 | 52 |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 | 1 |
| 9 | 10 | 11 | 12 | 13 | 14 | 15 | 2 |
| 16 | 17 | 18 | 19 | 20 | 21 | 22 | 3 |
| 23 | 24 | 25 | 26 | 27 | 28 | 29 | 4 |
| 30 | **31** | | | | | | **5** |

**FEBRUARY 1995**

| M | T | **W** | T | F | S | S | **WEEK** |
|---|---|---|---|---|---|---|---|
| | | **1** | 2 | 3 | 4 | 5 | **5** |
| 6 | 7 | 8 | 9 | 10 | 11 | 12 | 6 |
| 13 | 14 | 15 | 16 | 17 | 18 | 19 | 7 |
| 20 | 21 | 22 | 23 | 24 | 25 | 26 | 8 |
| 27 | 28 | | | | | | 9 |



## 3.2 Transcriptions of recorded dialogues (Appendix 7)

An example of a dialogue from the user test is presented below. It has been translated from Danish and is shown in an abbreviated form compared to the transcriptions in Volume 2, Appendix 7. Only system output and user input is shown below whereas the internal communication between system modules has been left out. The dialogue is based on the graphic scenario shown in Section 3.1.2. S means system and U means user. The subject number is 10 as indicated after S and U. The number after the hyphen indicates the turn number. The first turn in the dialogue below is 47 since the dialogue is based on the fourth scenario given to the user and turns are counted consecutively from the first to the last dialogue.

When the user calls, the system introduces itself by the following phrase:

S: Hello, this is the DanLuft reservation service for domestic flights. Do you know how to use this system?

When the user has finished a reservation task the system will ask if s/he wants anything else, thereby offering the possibility of continuing with a new reservation. The user accepts this offer in the following dialogue.

S10-47: Do you want anything else?

U10-47: Yes.

S10-48: Customer number 2. Is that correct?

U10-48: Yes

S10-49: How many people will travel?

U10-49: One person.

S10-50: One person. Please state the id-number of the person.

U10-50: Four.

S10-51: Id-number four, Lene Christensen. Where does the journey start?

U10-51: From Copenhagen to Karup.

S10-52: From Copenhagen to Karup. Do you want a return ticket?

U10-52: Yes.

S10-53: Are you particularly interested in discount?

U10-53: No.

S10-54: On which date will the journey start?

U10-54: The 31st of January.

S10-55: Tuesday the 31st of January. At which time of day?

U10-55: At 16:50.

S10-56: At 16:50. On which date will the return journey start?

U10-56: The 1st of February.

S10-57: Wednesday the 1st of February. At which time of day?

U10-57: At 18:10.

S10-58: There is no departure at 18:10. The closest other departures are at 18:00 and at 22:00. Do you want one of these departures?

U10-58: No.

S10-59: At which time of day?

U10-59: 15:45 from Esbjerg.

S10-60: There is no departure at 15:45. The closest other departures are at 13:00 and at 16:20. Do you want one of these departures?

U10-60: Yes.

S10-61: Which?

U10-61: 16:20.

S10-62: At 16:20. You have now booked one ticket for Lene Christensen from Copenhagen to Karup on Tuesday the 31st of January at 16:50, returning on Wednesday the 1st of February at 16:20. Your reference number is 52. The price is 1360 kroner. Is the ticket to be mailed or will the traveller pick it up at the airport?

U10-62: Mailed.

S10-63: The ticket will be mailed in a couple of days. Do you want anything else?

U10-63: No.

S10-64: Goodbye.

U10-64: Goodbye.

## 3.3 Expected and actual user answers to each scenario: locating dialogue problems (Appendix 9)

The table below shows the system's questions, expected key contents of user answers, actual key contents of user answers, and problems identified in a subject's completion of Scenario G-5-4-a-1. Contents in parentheses (third column) indicate key contents of the system's next utterance. Comments in brackets are explanatory. One system problem type (undersupported user navigation) and one user error type (answering several questions at a time) were identified.

| Scenario: G-5-4-a-1 User: 10 Date: January 25 1995 | | | |
|---|---|---|---|
| **System questions** | **Normative user answers** | **Actual user answers** | **Problems** |
| System already known | no / yes / - | - | |
| Customer number | 2 | yes (2) | |
| Number of travellers | 1 | 1 | |
| ID-numbers | 4 | 4 | |
| Departure airport | Copenhagen | Copenhagen | |
| Arrival airport | Karup | Karup | |
| Return journey | no | yes | Under-supported user navigation |
| Interested in discount | - | no | |
| Day of departure (out) | January 31 | January 31 | |
| Hour of departure (out) | around 7:30 / around 19:30 | 16:50 | |
| Day of departure (home) | - | February 1 | |
| Hour of departure (home) | - | 18:10 (no departure)<br>no [does not want one from list]<br>15:45 from Esbjerg (no departure [from Karup])<br>yes [wants one from list]<br>16:20 | Answering several questions at a time |
| Delivery | airport / send | send | |
| More | yes | no | |

# 3.4 Categorisation and analysis of each identified dialogue problem (Appendix 11)

## 3.4.1 An example of a system problem

The following example illustrates the dialogue design problem: undersupported user navigation. The system did not provide clear and sufficient instructions to users on how to interact with the system in the present case. **S** (boldface) means symptom, **D** means diagnosis and **C** means cure. S (normal) means system and U means user. The problem shown is the first one revealed in the example in Section 3.3.

G-5-4-a

| |
|---|
| **S:** When asked about home departure hour for the second time, the user tries and fails to input a departure airport for the home journey that is different from the arrival airport for the out journey. |
| **D:** Roundtrip tickets can only be handled as two separate single reservations in the Danish dialogue system. The system does not provide information on how to handle roundtrips. |
| **C:** The system should inform users on its limitations with respect to roundtrip reservations. |

## 3.4.2 An example of a user error

The example below illustrates the user error: answering several questions at a time. **ER** means error, **EX** means explanation and **PM** means preventive measure. S means system and U means user. The problem shown is the second one revealed in the example in Section 3.3. SP5 refers to the design principle "Provide clear and sufficient instructions to users on how to interact with the system" whose violation was shown in Section 3.4.1.

G-5-4-a

| |
|---|
| **ER:** S: At what time? U: 15:45 from Esbjerg. |
| **EX:** The user may be influenced by SP5. |
| **PM:** None, except for SP5. |

# Project Reports

The following is a list of project reports from the research programme *Spoken Language Dialogue Systems.* The final versions of reports 9a and 10 may have slightly different titles and authors.

1. Larsen, L.B., Brøndsted, T., Dybkjær, H., Dybkjær, L., Music, B. and Povlsen, C: State-of-the-art of Spoken Language Systems—A Survey. September 1992.
2. Larsen, L.B., Brøndsted, T., Dybkjær, H., Dybkjær, L. and Music, B.: Overall Specification and Architecture of P1. February 1993.
3. Dybkjær, L. and Dybkjær, H.: Wizard of Oz Experiments in the Development of a Dialogue Model for P1. February 1993.
4. Povlsen, C.: Sublanguage Definition and Specification. April 1994.
5. Brøndsted, T. and Larsen, L.B.: Representation of Acoustic and Linguistic Knowledge in Continuous Speech Recognition. January 1994.

5a. Larsen, L.B. and Steingrimsson, P.: Representation of Acoustic and Linguistic Knowledge in Continuous Speech Recognition. Documentation of Training and Test Databases. To appear 1996.

5b. Brøndsted, T. and Larsen, L.B.: Representation of Acoustic and Linguistic Knowledge in Continuous Speech Recognition. Program Descriptions. May 1994.

6a. Bernsen, N.O., Dybkjær, L. and Dybkjær, H.: Task-Oriented Spoken Human-Computer Dialogue. February 1994.

6b Dybkjær, H. and Dybkjær, L.: Representation and Implementation of Spoken Dialogues. May 1994.

7. Music, B. and Offersgaard, L.: The NLP Module. April 1994.
8. Lindberg, B. and Kristiansen, J.: Real-time Continuous Speech Recognition within Dialogue Systems. December 1995.

9a. Dybkjær, L., Bernsen, N.O., Brøndsted, T., Bækgaard, A., Dybkjær, H., Larsen, L.B., Lindberg, B., Povlsen, C.: Test of the Danish Spoken Language Dialogue System. January 1996.

9b. Dybkjær, L., Bernsen, N.O. and Dybkjær, H.: Evaluation of Spoken Dialogues. User Test with a Simulated Speech Recogniser. February 1996.

9c. Povlsen, C.: Adequacy Evaluation of the Linguistic Module. To appear 1996.

10. Bækgaard, A.: The Generic Dialogue System. To appear 1996.