Corpus-based Approaches to Dialogue Modelling

Proceedings of the ninth
Twente Workshop on Language Technology

J.A. Andernach, S.P. van de Burgt & G.F. van der Hoeven (eds.)

CIP GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Andernach, J.A. , van de Burgt, S.P., van der Hoeven , G.F.

Corpus-based Approaches to Dialogue Modelling
Proceedings Twente Workshop on Language Technology 9 / J.A. Andernach, S.P. van de Burgt, G.F. van der Hoeven,
Enschede, Universiteit Twente, Faculteit Informatica

ISSN 0929-0672

trefw.: natural language processing, speech, dialogue, language engineering, empirical methods

Book orders:
University of Twente
Ch. Bijron
Dept. of Computer Science
PO Box 217
NL 7500 AE Enschede
fax: +31-53-315283
Email: bijron@cs.utwente.nl

Druk- en bindwerk: Reprografie U.T. Service Centrum, Enschede

PREFACE

TWLT is an acronym of Twente Workshop(s) on Language Technology. These workshops on natural language theory and technology are organised by Project Parlevink (sometimes with the help of others), a language theory and technology project conducted at the Department of Computer Science of the University of Twente, Enschede, The Netherlands. Each workshop has proceedings containing the papers that were presented. For the contents of these proceedings consult the last pages of this volume.

Previous workshops.
TWLT1, *Tomita's Algorithm: Extensions and Applications*. 22 March, 1991.
TWLT2, *Linguistic Engineering: Tools and Products*. 20 November, 1991.
TWLT3, *Connectionism and Natural Language Processing*. 12 and 13 May 1992.
TWLT4, *Pragmatics in Language Technology*. 23 September, 1992.
TWLT5, *Natural Language Interfaces*. 3 and 4 June, 1993.
TWLT6, *Natural Language Parsing*, 16 and 17 December, 1993.
TWLT7, *Computer Assisted Language Learning*, 16 and 17 June 1994.
TWLT8, *Speech and Language Engineering*, 1 and 2 December 1994.

TWLT9 was devoted to (spoken) natural language dialogues, the analysis of such dialogues, and the use of the results of the analysis in natural language dialogue systems. The workshop was sponsored by KPN Research, Leidschendam, and the NWO Prioriteitsprogramma Taal- en Spraaktechnologie. It took place in the Collegezalencomplex at the campus of the University of Twente in Enschede, The Netherlands. Just as with the previous workshop programs there were presentations by a select group of international researchers and other experts. Their contributions covered a wide variety of aspects of the theme of the workshop: there were examples of extensive analyses of particular aspects of the clauses in a corpus of dialogues, investigations into strategies that participants in a dialogue use to maintain coherence, presentations of methods to encode information on dialogue structure, studies into classification of dialogues and factors that influence their structure, and finally papers on systems that deal with (spoken) dialogue, and which are designed and evaluated on the basis of a corpus of dialogues.

A workshop is the concerted action of many people. It goes without saying that we are grateful to the authors and the organisations they represent for their efforts. But in addition we would like to mention here the people whose work has been less visible during the workshop proper, but whose contribution was evidently of crucial importance. Charlotte Bijron, Yvonne Sapulette and Alice Hoogvliet-Haverkate took care of the administrative tasks. Finally we also wish to thank the participants for being there and for contributing to the discussions.

We hope that TWLT10 on *Algebraic Methods in Language Processing*, which will be a joint event with the first AMAST workshop on language processing, in December 1995, will match the success of this workshop.

Toine Andernach
Stan van de Burgt
Gerrit van der Hoeven

June, 1995

CONTENTS

Workshop Papers:

Kinds of agents and types of dialogues 1
N. Dahlbäck (NLP Laboratory, Linköping, Sweden)

Clause-internal structure in spoken dialogue 13
J.H. Connolly, A.A. Clarke, S.W. Garner and H.K. Palmén (Loughborough University of Technology, UK)

The coding of dialogue structure in a corpus 25
J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon and A. Anderson (HCRC, Edinburgh, UK)

Designing the dialogue component in a speech translation system – a corpus-based approach 35
J. Alexandersson and N. Reithinger (DFKI, Saarbrücken, Germany)

Dialogue control in automatic inquiry systems 45
H. Aust and M. Oerder (Philips, Aachen, Germany)

Referring to topics – a corpus-based study 51
M. Rats (ITK, Tilburg, the Netherlands)

Design, formalization and evaluation of spoken language dialogue 67
H. Dybkjær, L. Dybkjær and N.O. Bernsen (Centre for Cognitive Science, Roskilde, Denmark)

Mutuality strategies for reference in task-oriented dialogue 83
D.G. Novick and B. Hansen (Oregon Graduate Institute of Science and Technology, Portland, USA)

Messy data, what can we learn from it? 95
N. Fraser (Vocalis Ltd, Cambridge, UK)

Predicting and interpreting speech acts in a theatre information and booking system 107
J.A. Andernach (University of Twente, Enschede, the Netherlands)

Sponsors and Support

We gratefully acknowledge help from:

KPN Research, Leidschendam

Prioriteitsprogramma Taal- en Spraaktechnologie, NWO

the European Chapter of the Association for Computational Linguistics

KINDS OF AGENTS AND TYPES OF DIALOGUES

Nils Dahlbäck*
Department of Computer and Information Science
Linköping University
S- 58183 Linköping, Sweden
nda@ida.liu.se

ABSTRACT

In recent years we have seen an increase in the work on the empirical foundations of computational theories of discourse, as well as in the increased use of empirical evaluation methods for natural language systems. In this paper I argue that presenting results from well conducted empirical studies of particular discourses or dialogues is necessary and important but not enough to foster the development of computational theories of discourse. Equally important is making clear to which other cases of agents and situations the results obtained apply. As far as the issue of agents is concerned, it is argued that present day computational theories of discourse can only be seen as theories of computer's processing of language, and not for all kinds of agents, and some consequences of this position are discussed. When, on the other hand, we come to the issue of dialogue situation I do not present any specific theoretical position. Instead a number of dimensions or parameters that seem to influence the language used, supporting the arguments with examples from our own studies of these issues, as well as results and observations from other workers in the field are described.

1 INTRODUCTION

In computational linguistics there is currently a notable trend towards a greater emphasis on the empirical base of the ongoing research. Examples of this are the Message Understanding Conferences (MUC), the recent AAAI Spring symposium on empirical methods in discourse; the large number of Natural Language Interface projects that make extensive use of Wizard of Oz-methods in the early development phases, e.g. SUNDIAL, PLUS, LinLin, to mention just a few of them on this side of the Atlantic.

In line with this, there has been a concomitant increase in the interest of different aspects of evaluation methods for natural language systems (e.g. Palmer and Finin, 1990, Chinchor, 1991, Neal and Walter, 1991, Chinchor, Hirschman and Lewis, 1993).

But presenting results from well conducted empirical studies of particular discourses or dialogues is not enough to foster the development of general computational theories of discourse. It is equally important to make clear which other cases the results obtained apply to.

There are two equally important aspects to any empirical investigation, whether it is concerned with evaluation of existing systems or development of new ones; finding the relevant metrics, and finding the relevant generalization domains for the results obtained. While most of the work cited above has been concerned with the former aspect, the present paper focuses on the latter.

My claim in this paper is that we need to be able to generalize to two different domains: one concerns for which kinds of *agents* the results of theories are taken to apply, the other to which kinds or classes of dialogue *situations* the results apply.

The present paper addresses both these issues, but treats them in a different manner. As far as the issue of agents is concerned, an argument is put forth below for the somewhat controversial position that present day computational theories of discourse can only be seen as theories of computer's processing of language, and not as general theories of discourse. When, on the other hand, we come to the issue of dialogue situation, I do not present any specific theoretical position. Instead I describe a number of dimensions or parameters that seem to influence the language used

*This research was financed by the Swedish National Board for Technical Development and the Swedish Council for Research in the Humanities and Social Sciences.

supporting the arguments with examples from our own studies of these issues, as well as with results and observations from other workers in the field. In both cases my hope is that the ideas put forth will contribute to the discussion on the theoretical status and practical relevance of computational theories of discourse as well as the empirical basis of these theories.

2 GENERALIZING THE RESULTS

There are two steps in the generalization of results, the statistical analysis and non-statistical. The statistical generalization is based on the use of methods of inferential statistics and is concerned with the generalization of results observed in a sample to the population from which it was drawn. The non-statistical generalizations are those that go beyond the population proper, for instance from college students to the population at large, or from the particular discourse to other discourses.

2.1 STATISTICAL GENERALIZATION

There are well known problems with the use of standard significance tests in language studies and varius solutions have been suggested to remedy this. Ever since H. Clark's (1973) influential paper 'The language-as-fixed-effect fallacy: A critique of language statistics in psychological research' it has been known that the standard analysis of variance method has its problems when used on linguistic data, since a standard ANOVA or other similar tests are concerned with the inference from the particular sample of subjects used to the rest of the subject population, but only for the particular specific linguistic materials used in the study. To be able to view the linguistic tokens as a sample from a population as well, Clark suggested the use of min F' instead of the common F. While this suggestion has not been uncontested by statisticians (e.g.Wike and Church, 1976, Cohen,1976, Smith, 1976 and Keppel, 1976), the use of F', or doing the statistical analysis both by materials and by subjects, has become something of the standard procedure in experimental psychological studies of language.

Another problem when making statistical analysis of linguistic materials is that the distribution of the variables is not known. Chinchor, Hirschman and Lewis (1993) therefore, in their analysis of the MUC-3 data, used the so-called approximate randomization method (Noreen, 1989).

2.2 NON-STATISTICAL GENERALIZATIONS

Whereas various problems with the use of inferential statistics in studies of language use have been addressed by among others the workers mentioned in the previous section, other problems when interpreting the results obtained still remain.

As is well known, true random sampling from a well defined population is the exception rather than the rule in much research in psychology and other related sciences. The use of inferential statistics can, however, still be defended in most cases, on the grounds that there are good reasons to believe that the population from which the sample is drawn does not in any important respects differ from the larger population of interest. When population characteristics are well known, this can be done with some certainty. But when this is not the case, the situation is problematic. And it is even more problematic in these cases to make non-statistical generalizations. If, for instance, a new pronoun resulution algorithm based on an empirical analysis of the pronoun-antecedent relationships in a language sample is published, how am I to know whether the system I am building will work on similar enough language to the one used in the development of this algorithm to make it possible for me to use it. I would claim that in many cases we do not even know what the relevant dimensions for judging similarity are.

My impression is that this is the situation we are confronted with today. It is for instance difficult to know to what kinds of discourse the results obtained on different aspects of discourse structure by Grosz and Hirschberg (1992), Swerts, Gelyukens, and Terken (1992) and others can be generalized.

3 FOR WHICH AGENTS?

Most work on dialogues in present-day computational linguistics does not make explicit to what extent the models and theories developed should be seen as theories about the processing of dialogue by computers or people or both. Though seldom explicitly stated, the underlying assump-

tion seems to be that the theories are to be taken as general theories of discourse for all kinds of agents and situations. There are, however, a number of reasons for assuming that the cognitive architecture of present day computers and people are sufficiently different to make it necessary to clarify to which extent a computational theory of discourse (or any other cognitive phenomenon, for that matter) is primarily to be seen as a psychological account or an account of computer's processing of discourse. It is important to note that this is true not only for those who, like Searle (1980, 1992), are critical to the computational theory of mind, but also for the outspoken defenders of that view (e.g. Pylyshyn 1984). It is thus, in a sense, an uncontroversial position. But perhaps less so are the consequences that I want to claim follow of necessity from it, one concerning the cognitive or procedural aspects, the other concerning the linguistic application domain.

As far as the internal or representational/algorithmic aspect is concerned, I want to claim that procedural computational accounts of the process of discourse using concepts from present day computer technology cannot be seen as a psychological account. To quote Pylyshyn (1984, p 91) "two programs can be thought of as strongly equivalent or as different realizations of the same algorithm or the same cognitive process if they can be represented by the same program in some theoretically specified virtual machine." A consequence of this is that "any notion of equivalence stronger than weak equivalence (i.e. the same input-output conditions) must presuppose an underlying functional architecture, or at least some aspects of such an architecture." (ibid., p 92) "Typical, commercial computers, however, are likely to have a far different functional architecture from that of the brain; hence, we would expect that, in constructing a computational model, the mental architecture must first be emulated (that is, itself modelled) before the mental algorithm can be implemented" (ibid., p 96).

Another way of formulating the same argument goes as follows. If you believe that the human mind is similar to a von Neuman computer in all respects important for the cognitive processes you are studying, i.e. accepts what Searle would call 'strong AI', then any *procedural*[1] theoretical account will obviously be applicable to both humans and machines. But if instead you believe that there are some important differences between men and machines, the obvious first step is to decide whether your account is about people or computers. If it is about computers there are no problems. But if you want to make a psychological account, you first need to specify this different 'machine', and implement it as the first level of your theory or program. First after having done so, you can specify your pronoun resolution algorithm or whatever. Well-known examples of such general cognitive psychological computational theories are ACT* (Anderson, 1983, 1990) and SOAR (Laird, Newell, and Rosenbloom, 1987, Newell 1990), for an overview and comparison of these see Newell, Rosenbloom, and Laird (1989). There are also other such symbolic theories. Furthermore, some connectionist work can be viewed as an attempt to implement a psychologically correct cognitive theory too, and in those cases its performance can be evaluated against what is known about humans performance in doing similar tasks.

There are, of course, other possible conclusions that can be drawn, and other possible theoretical positions can be taken than the one argued for here. One can, for instance, deny the validity of Pylyshyn's arguments. But since the points he makes seem rather uncontroversial taken one at a time, and the conclusion rather self-evident, the burden of the argument seems here to be on those who wish to argue against him.

Another possible stance is to believe that present day computers and people are similar enough in their basic architecture to make it possible to describe the details of a processing algorithm that will apply to both kinds of agents. But also this 'strong AI' position seems more controversial than the one drawn above.

A final comment, whichto some readers will seem self evident and superfluous, but which I believe needs to be made given the reactions I have sometimes encountered when presenting these arguments previously: Nothing stated in the paragraphs above implies or is meant to imply that it is impossible to simulate human cognitive processes on present day computers. The Church-Turing thesis states that any process that can be given a sufficiently precise decription can be simulated on a computer. And this includes human cognitive processes. So the argument is not

[1] Note that the argumet here only apllies to *procedural* and not *formal* but non-procedural theories. Formal descriptions of cognition, e.g. declaratively specified grammars, are not affected by it. For a theory to be labeled procedural in the sense used here, a specification of the algorithms and control structures to be used with the formal description must be defined too.

against the possibility of simulating cognitive processes, including linguistic ones. The argument is that either you present a computational theory of computers' processing of language or you present a computational theory of humans' processing of language. But you cannot do both at the same time. And you must do either.

4 SOME CONSEQUENCES

So much for the argument. But why then does it matter? One consequence if the arguments above are accepted is that most, if not all, present day theories and models in computational linguistic theories on discourse are about computer's processing of language and nothing else. Or, to phrase the same point somewhat differently, since there are no attempts to first emulate a theory of the human cognitive systems, it is difficult to regard them as anything but theories about computers. Another consequence is that psychological realism on the internal representational and procedural level is of no interest if your aim is to build useful systems. (This argument about 'representational agnosticism' is further elaborated and motivated in Dahlbäck 1989, 1991b.)

Another simple but important consequence of such a sub-language approach (Grishman & Kittredge, 1986) for those of us concerned with providing the empirical base for computational theories of discourse is that the language samples used for providing the empirical ground should come from relevant application domains for such software technology and from dialogues with computers and not between humans. In this context it can be noted that while workers in discourse have not been satisfied with theories based on "gedanken-data", but have strived to develop their theories through detailed analysis of empirical data of many diverse dialogue situations, the kinds of discourses studied do not always confirm to this requirement. In their review of the field, Grosz, Sidner, and Pollack (1989) mention work on task-oriented dialogues (Grosz 1978, Sidner, 1982), descriptions of complex objects (Linde, 1979), narratives (Polanyi 1985, Schiffrin 1982), informal arguments (Reichman-Adar, 1984), formal arguments, (Cohen, 1984), negotiations (Linde and Goguen, 1978), and explanations (Reichman-Adar, 1984). Note however that few of these dialogue situations resemble typical application domains for natural language interfaces, and the most prototypical situations for the technology such as information retrieval are lacking.

The reason for my wanting to stress this point is the well known fact that language use is situation dependent. Content and form differ depending on the situation in which they occur (e.g. Levinson, 1981, 1983), but also depending on the perceived qualities of the interlocutors; language directed to children is different from language directed to grown-ups (Phillips, 1973, Snow, 1972), as is the case with talking to foreigners, brain-injured peopl, and people that do not know who Jimi Hendrix was. The ability to modify the language to the perceived needs of the speaker seem to be present already at the age of four (Shatz & Gelman, 1973).

Since dialogue participants adapt to the qualities of their interlocutors, analysis of dialogues between people, or of people communicating with existing systems is not enough here. We have therefore based our work on the use of Wizard of Oz-studies (For early arguments see Dahlbäck & Jönsson, 1986; for a description of our present systems and methods see Dahlbäck, Jönsson & Ahrenberg, 1993).

The conclusion above is hardly controversial these days. But I would also claim that another important consequence of the position outlined above is therefore that goals of research on dialogue in computational linguistics such as "Getting computers to talk like you and me" (Reichman, 1985), or developing interfaces that will "allow the user to forget that he is questioning a machine" (Gal, 1988), are not only difficult to reach. They are misconceived. Since we always adapt to the qualities of our dialogue partner there is every reason to believe that NLI-users will adapt to the fact that they are interacting with a computer. An increasing body of research on the language used when communicating with computers seem to confirm this.

5 TOWARDS A DIALOGUE TAXONOMY

"That language varies according to the situation is a truism; however, the details and implications of that truism are far from obvious, whether your enterprise is theory formation or system construction" (Pattabhiraman, 1994). A necessary requirement for clarifying these consequences is the development of a descriptive classificatory scheme for dialogue situations. The aim of the present section is to provide the first steps towards the

development of such a taxonomy, with special emphasis on discourse aspects relevant for computational theories of discourse. While obviously influenced by the arguments put forth in the previous sections, I believe that the task of developing a descriptive taxonomy of discourses is important regardless of the position taken on that issue.

I make no claim that the dimensions described below are an exhaustive list, nor do I wish to claim that they are independent. It is much too early to make such conclusions. My goal is a more modest one; I hope to initiate a discussion on some of the aspects I consider important here, since I believe that the healthy development of the field requires clarification on these issues. Since my own work has been concerned with the development of natural language interfaces, there is a bias towards dialogue situations in the discussion below.

In Linköping we have recently been involved in a project aimed at comparing different kinds of computational discourse models empirically. We have not only used our own dialogue corpora from previous work, but have tried to gain access to other corpora as well. We found in the course of this work that different kinds of computational models seemed to be more adapted to some kinds of dialogues than to others. This led us to partly reformulate the aims of the project to also focus on a descriptive scheme of different kinds of dialogues. What I report below is hence a snapshot of work in progress. No claim of orginality is made here as far as the dimensions mentioned is concerned. As will be obvious to many readers, much of what is presented below is based on or influenced by work of others, probably even more so than is made evident in the references. But I have tried to enforce my argument that these factors need to be taken seriously by the computational discourse community by illustrating the possible ways in which the factors mentioned influences or might influence the computational treatment of discourse.

In one sense, the *type of agent* (person or computer) already discussed is one important dimension. But in this case the issue is not the internal architecture, but rather the influence of the agent on the language used. The few studies that I know of that have adressed this issue have also shown that it affects the dialogue on a number of dimensions. Guindon (1988) showed that the dialogue structure differed between dialogues with persons and with computers in similar situations. The work by Kennedy, Wilkes, Elder and Murray (1988) showed that the language used when communicating with a computer, as compared with a person in a similar situation, has the following characteristics: Utterances are shorter, the lexical variation is smaller and the use of pronouns is minimized. The results concerning the limited use of pronouns when communicating with has been established in a large number of studies (For a summary of a number of studies on this and other aspects of 'computerese', see Dahlbäck 1991 ch 9), but in most cases it is impossible to ascertain whether the differences found is caused by influences of the channel (typed vs spoken) or the perceived characteristics of the dialogue partner (human vs computer). I will return to the issue of channels below.

In a current project in Linköping we are comparing the language used when communicating with a computer or with a person in identical situations (typed information retrieval with or without the possibility of also ordering the commodities discussed). The only difference between the two situations is what the subjects are told they are interacting with, a person or a computer. In all other respects the situations are similar (and the 'wizards' are not told beforehand under which condition the specific subject is run). It is interesting to note that it is in this situation rather difficult to find any differences between the dialogues with humans and those with computers. If this result holds after a more thorough analysis, this indicates that *communication channel* and *kinds of tasks* influence the dialogue more than the perceived characteristics of the interlocutor. It is, however, still possible that there are differences between these dialogues in for example the dialogue structure, something which has not been analyzed as yet.

When talking about different dialogue types, a distinction is often made between spoken and written language. But the difference between the prototypical spoken and written language is really not one but many. Rubin (1980) suggests that the communicative medium, or what I here have called the communication channel, should be partitioned into the following seven dimensions: modality (written or spoken), interaction, involvement, spatial commonality, temporal commonality, concreteness of referents (are objects and events referred to visually present or not), separability of characters. Below I will present observations suggesting that at least some of these dimensions influence linguistic aspects of interest to computational linguists.

There is considerabel evidence suggesting that *type of medium* (spoken or written) influences the dialogue structure in human-computer dialogues. Cohen (1984) studied the effects of the communication channel on the language used in task oriented dialogues. When comparing spoken (telephone) and teletype conversations he noted that " keyboard interaction, with its emphasis on optimal packaging of information into the smallest linguistic "space", appears to be a mode that alters the normal organization of discourse". (Cohen, 1984, p 123) To take one example, the use of cue-words to introduce new discourse segments occurs in more than 90 % of the cases of spoken discourse, but in less than 45 % of the written dialogues. This seems to indicate that we should be careful when generalizing from spoken dialogues when constructing an NLI for keyboard interaction and vice versa.

The problem with this factor, as with those described above, is that even if we can assume that it affects the structure of the discourse, our current knowledge is not advanced enough to make it possible to predict with certainty how it will differ. But there is some evidence that it affects not only the use of cue-words and the other phenomena described by Cohen, but also the basic dialogue structure. An illustration of this is is found in the different kinds of basic dialogue structure proposed by us for typed dialogues (Dahlbäck 1991, Dahlbäck & Jönsson 1992, Jönsson 1993) and for Bilange (1991) for spoken dialogues. The dialogues involve in both cases information retrieval. The spoken dialogues seem to exhibit a three-move structure (called Negotiation, Reaction, Elaboration by Bilange), whereas in the typed a two-move structure (Initiative, Response) is sufficient. Before leaving this dimension I wish to suggest that one important difference between spoken and typed dialogues with computers affecting the discourse is that parts of the dialogue remain in front of the user when planning and executing the next move. We have for instance found that even with extremely long response times (due to a very slow simulation environment at the time), users make use of anaphoric expressions, including ellipsis in the dialogues.

The *interaction* dimension (dialogue versus monologue) seems to influence among other things the use of pronouns and the pattern of pronoun-antecedent relations. In typed human-computer dialogues pronouns are rarely used (Guindon, 1988, Dahlbäck & Jönsson, 1989, Kennedy et al, 1988). The anaphor-antecedent relations seem to be of a rather simple kind in these kinds of dialogue. To take one example, we found in an analysis of these patterns in one of our corpora of Wizard of Oz-dialogues that in those cases where the personal pronouns had an antecedent, the distance between pronoun and antecedent was very small. The analysis suggested that the antecedent could be found using a very simple algorithm which basically worked backwords from the pronoun and selected the first candidate that matched the pronoun on number and gender and which did not violate semantic selection restrictions (Dahlbäck, 1992). The algorithm described and evaluated on a number of computer manuals by Lappin and Leass (1994) is more complicated and uses among other things an intrasentential syntactic filter for ruling out anaphoric dependence of a pronoun on an NP on syntactic grounds. It is not clear that such a filter would improve the recognition of the antecedent in our dialogues, where instead the dialogue structure was needed to stop the search for antecedents to the pronouns when these were not found within the local structure unit. The reason for this rule was that in our corpus as many as 1/3 of the personal third person pronouns lacked an explicit antecedent, but instead made use of some kind of associative relation to the antecedent, or belonged to the class of pronouns called 'propositional' by Fraurud (1988).

Spatial and *temporal commonality* also seem to influence aspects of discourse. Not only is the use of deictic expressions made possible with a shared temporal/spatial context, but it is also possible that the use of other anaphoric devices is influenced. Guindon (1988) found, for instance, in her analysis of advisory dialogues for the use of a statistical computer package that pronouns either had their antecedent in the current subdialogue, or they referred to the statistical package that was present on the screen all through the dialogue. And as an aside, it is perhaps worth pointing out that the celebrated example from Grosz' dissertation (Grosz 1977, p 30), where the pronoun 'it' is used to refer to the pump just assembled, which has not been mentioned for 30 minutes and 60 utterances, could be seen as belonging to this categor,y too. But also in other kinds of discourse where there is no shared physical context, and where the interaction is minimized there sometimes occur privileged entities that can be referred to using a pronoun even if the antecedent in the strict sense has not been mentioned for a long time. These so-called 'primary referents' (Fraurud, 1988) are for instance the main actors in a novel.

The other dimensions discussed by Rubin are probably also important when not only for human dialogues, but also for human-computer dialogues. They seem to be of use, for example, when discussing and comparing different kinds of multi-media or multi-modal interaction.

Rubin also discusses a number of message-related dimensions (without claiming them to be independent), especially topic, structure and function. I will here address two dimensions closely related to the ones mentioned by Rubin, namely task structure and kinds of shared knowledge.

That *task structure* influences the dialogue structure was an important aspect of Grosz' (1977) early work. But she also pointed out that for man-computer dialogues "there seems to be a continuum (...) from the totally unstructured table filing dialogues to the highly structured task dialogues (ibid, p 33). In the task oriented dialogues the structure of the task was shown to influence the structure of the dialogue and this result was the starting point for the use of the underlying task structure in the analysis of discourse. But it seems as if not only different tasks will influence the structure of the dialogue, but some of our observations seem to indicate that different kinds of task settings for the dialogues, and especially the *dialogue-task distance* influence the dialogue strucure with varying degree. Furthremore, this applies to the extent that different kinds of computational discourse structure models seem to be preferred depending on the value taken on this dimension.

Some of our observations suggest that while plan- or intention-based discourse models might be necessary for some kinds of human-computer dialogues, this is not true for all cases. There is a closer connection between task and dialogue in an advisory dialogue than in an information retrieval dialogue. I currently hypothesize that this difference makes different kinds of computational discourse models more or less applicable in the various cases. The closer the language-background task connection, the more apropriate become plan or intention based models. In these situations it is less difficult to infer the non-linguistic intentions behind a specific utterance from knowledge of the general task structure and from observations on the on-going dialogue. But with larger distance between the dialogue and the underlying task, as in the information retrieval case, the more difficult it becomes to infer the underlying intentions from the linguistic structure, and at the same time the need for this information in order to provide helpful answers diminishes.

As an example, to answer the question of when there are express trains to Stockholm within the next two hours, in most cases there seems to be no need to know why the questioner needs to know the answer. I am not denying that there are cases when the information provider can be more helpful when knowing this. But the prime case of this is probably when it is not possible to provide an answer, as for instance when in the case above, there are no trains of the requested kind within the specified time limit. In such cases humans often seem to ask for the information needed to provide additional help and presumably computer systems can do the same.

One observation from our on-going work that seems to support this position is that we have found that the coding of the underlying intentions in an information retrieval dialogue becomes really difficult if the coding is done move by move, i.e. when the move is classified without knowledge of what follows later in the dialogue. But this is, of course, the task a computer system will be in. A coding scheme based on more surface-oriented criteria seem to be in advantage in this situation.

It is not only the connectedness between the linguistic and the non-linguistic task that influences the complexity of the dialogue. The *number of different tasks* managed linguistically is another such factor. In our work we have compared cases of information retrieval dialogues with dialogues in the same domain (travel information) where the user also can order a ticket. In the latter case not surprisingly, a more complex topic management was required (Jönsson, 1993, Ahrenberg, Dahlbäck, Jönsson, forthcoming).

This dimension seems to us to point to an important difference between human dialogues and human-computer dialogues, since there are fewer different things that can function as topic in a dialogue with a computer system. (Not many of us chat with our couputer about the lousy weather while waiting for a manuscript to be printed, for example.)

The influence of different *kinds of shared knowledge* between dialogue participants on the use of referring expressions have been discussed by Clark and co-workers in a number of important papers (e.g. Clark & Marshall, 1981; Clark & Carlsson, 1981; for a summary see Clark, 1985). The basic point of this work is that a necessary pre-requisite for the successful use of a defi-

nite description is that speaker and listener share a common ground of mutual knowledge, beliefs, and assumptions, and furthermore that, were it not for a number of heuristics used by people, the acquisition of this mutual knowledge would require checking an infinite number of assumptions. The bewildered or sceptical reader of this claim is referred to the original sources. In this context I only want to use Clark's taxonomy of the basic classes of such heuristics for my present purposes. Clark's claim is that there are three basic such classes or kinds of information that can be used to infer the common ground between speaker and listener; shared perceptual, linguistic and cultural knowledge. Two of these have in different ways already been addressed previously. *Perceptual knowledge* is usable when the physical or visual context is shared; the shared *linguistic knowledge* is in this context another name for the shared knowledge of the previous text or dialogue. But what has not been discussed previously is the use of shared cultural knowledge, where 'cultural' here is used in its widest possible sense, including factual knowledge etc.

The basic idea here is that there are things that everybody in a community knows and which therefore can be used as common ground. The problem with this is, of course, to determine if my dialogue partner belongs to the same community as I do, or rather which cultural knowledge from different sub-communities that I can assume that we share. With my friends at the computer science department I can talk about 'a bug' meaning a malfunctioning part of a device or a scheme; with my friends at the department of biology I can't. The problem is for you, newcomer to our university, to know to which category the person I am talking to belongs when you walk up to us wanting to tell us about a programming bug that made you lose two hours of work. And the solution? Well, we all know what a hacker looks like, don't we? Joking aside, what this shows is simply the communicative value of stereotypes, including selecting your clothes to show which group or groups you belong to.

The point of this is that not only different knowledge content between the dialogue participants, but also different bases for inferring the necessary mutualness of this knowledge are involved. This seems to be an aspect worth considering not only when considering to what extent results obtained in one particular study can be used in another situation, but also when selecting tasks and domains for which an interactive computer system should be designed. Note that in many cases the computer is worse off than a human in the same situation not only since the computer's inferential abilities are less powerful than those of the person, but because it has a more impoverished empirical base to to build its deductions on. It cannot see its interlocutor and does not remember the person from previous encounters.

My suggestion here is that it will be difficult to develop dialogue systems for those kinds of applications where the common cultural ground needs to be acquired during the on-going dialogue. And a possible explanation for the successful information retrieval systems developed is that they operate in domains where it can be assumed that all users will have the same basic knowledge of the domain. Hence the need for clarification subdialouges is diminished or obsolete, as well as the need for user-modeling of a kind not yet achieved.

6 SUMMARY

In this paper I have adressed two interrelated issues for the empirical work on computational theories of discourse. I argued that given the basic difference between the architecture of humans and computers, procedural computational theories of discourse can only be seen as theories of computers' processing of discourse. I also argued that an important prerequisite for any empirical work on computational discourse theories is a clarification of which descriptive dimensions that classify different dialogue situations. As a first attempt I desribed a number of such dimensions that I believe influence one or more important parameters for any kind of computational theory of discourse and tried to illustrate their possible influence on different discourse phenomena. The list is by no means intended to be all-inclusive and final. There are in all probability other dimensions not mentioned here that are of equal or larger importance. Which these are is in the long run an empirical question. But the ones that we know of today are of sufficient importance to be taken seriously, if we are serious about our aim of placing computational linguistics on a firm empirical base.

7 ACKNOWLEDGEMENTS

The ideas presented in this paper have evolved during a number of years when I have involved in the developmenmt of a natural language interface

for Swedish at the Natural Language Processing Laboratory at the Department of Computer and Information Science, Linköping University. I greatfully acknowledge the inspiring discussions, help, and critique from the members of the group, and especially Lars Ahrenberg and Arne Jönsson. Thanks also to Ivan Rankin for correcting a number of linguistic errors, and to Tommy Persson for a last night LaTex-wizard performance.

REFERENCES

Ahrenberg, L., Dahlbäck, N., and Jönsson, A. (Forthcoming). Dialogue mangagement for natural language interfaces. *Manuscript in preparation.*

Anderson, J. R. (1983). *The Architecture of Cognition.* Harvard University Press.

Anderson, J. R. (1990). *The Adaptive Character of Thought.* Erlbaum.

Bilange, E. (1991). A task independent oral dialogue model. In *Fifth Conference of the European Chapter of the Association for Computational Linguistics (E-ACL'91).*

Chinchor, N. (1991). Evaluation metrics. In *Third Message Understanding Conference (MUC-3)*, San Mateo, CA. Morgan Kaufman.

Chinchor, N., Hirshman, L., and Lewis, D. D. (1993). Evaluating message understanding systems: An analysis of the third message understanding conference. *Computational Linguistics*, 19(3):409–449.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, 12:335–359.

Clark, H. H. (1985). Language use and language users. In Lindzey, G. and Aronson, E., editors, *The Handbook of Social Psychology (3rd edition).* Erlbaum.

Clark, H. H. and Carlson, T. (1981). Context for comprehension. In Long, J. and Baddeley, A., editors, *Attention and Performance IX.* Erlbaum.

Clark, H. H. and Marshall, C. (1981). Definite reference and mutual knowledge. In Joshi, A., Webber, B., and Sag, I., editors, *Elements of Disourse Understanding.* Cambridge University Press.

Cohen, J. (1976). Random means random. *Journal of Verbal Learning and Verbal Behavior*, 15:261–262.

Cohen, P. R. (1984a). The pragmatics of referring and the modality of communication. *Computational Linguistics*, 10:97–146.

Cohen, R. (1984b). A computational theory of the function of clue words in argument understanding. In *COLING'84*, Stanford, CA.

Dahlbäck, N. (1989). A symbol is not a symbol. In *Proceedings of the 11th Joint Conference on Artificial Intelligence (IJCAI'89).* Morgan Kaufmann.

Dahlbäck, N. (1991). *Representations of Discourse.* PhD thesis, Linköping University, Sweden.

Dahlbäck, N. (1992). Pronon usage in nli-dialogues: A wizard of oz study. In *Papers from the third Nordic Conference of Text Comprehension in Man and Machine.*

Dahlbäck, N. and Jönsson, A. (1986). A system for studying human-computer dialogues in natural language. Technical Report LiTH-IDA-R-86-42, Department of Computer and Information Science, Linköping University.

Dahlbäck, N. and Jönsson, A. (1992). An empirically based computationally tractable dialogue model. In *Proceedings of the 14th Annual Conference of the Cgnitive Science Society (CogSci'92).*

Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of studies - why and how. *Knowledge-Based Systems*, 6(4):258–266.

Fraurud, K. (1988). Pronoun resolution in unrestricted text. *Nordic Journal of Linguistics*, 11:47–68.

Gal, A. (1988). *Cooperative Responses in Deductive Databases.* PhD thesis, University of Maryland, College Park.

Grishman, R. and Kittredge, R. (1986). *Analyzing Language in Restricted Domains.* Erlbaum.

Grosz, B. and Hirschberg, J. (1992). Some intonational characteristics of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing*, Banff. ICSLP.

Grosz, B. J. (1977). *The Representation and Uses of Focus in Dialogue.* PhD thesis, University of California, Berkeley.

Grosz, B. J. (1978). Discourse analysis. chapter 9, pages 235–268. Elsevier North-Holland.

Grosz, B. J., Pollack, M. E., and Sidner, C. L. (1989). Discourse. In Posner, M. I., editor, *Foundations of Cognitive Science.* The MIT Press.

Guindon, R. (1988). A multidisciplinary perspective on dialogue structure in user-advisory dialgues. In Guindon, R., editor, *Cognitive Science and Its Application for Human-Computer Interaction.* Lawrence Erlbaum Publishers.

Jönsson, A. (1993). *Dialogue Management for Natural Language Interfaces.* PhD thesis, Linköping University.

Kennedy, A., Wilkes, A., Elder, L., and Murray, W. (1988). Dialogue with machines. *Cognition*, 30:73–105.

Keppel, G. (1976). Words as random variables. *Journal of Verbal Learning and Verbal Behavior*, 15:263–265.

Laird, J. E., Newell, A., and Rosenbloom, P. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, 33(1):1–64.

Lappin, S. and Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Levinson, S. C. (1981). Some pre-observations on the modelling of dialogue. *Discourse Processes*, 4:93–116.

Levinson, S. C. (1983). *Pragmatics.* Cambridge University Press.

Linde, C. (1979). Focus of attention and the choice of pronouns in discourse. In Givon, T., editor, *Syntax and Semantics*, pages 337–354. Academic Press, New York.

Linde, C. and Gougen, J. (1978). Structure of planning discourse. *J. Social Biol. Struct.*, 1:219–251.

Neal, J. G. and Walter, S. M. (1991). *Natural Language Processing Systems Evaluation Workshop.* Berkeley, CA.

Newell, a. (1990). *Unified Theories of Cognition.* Harvard University Press.

Newell, A., Rosenbloom, P. S., and Laird, J. E. (1989). Symbolic architectures for cognition. chapter 3, pages 93–131. The MIT Press.

Noreen, E. W. (1989). *Computer Intensive Methods for Testing Hypothesis: An Introduction.* John Wiley & Sons.

Palmer, M. and Finin, T. (1990). Workshop on the evaluation of natural language proseccing systems. *Computational Linguistics*, 16(3):175–181.

Pattabhiraman, T. (1994). Review of "user modeling in text generation" by ccile l. paris. *Computational Linguistics*, 20:318–321.

Philips, J. R. (1973). Syntax and vocabulary of mothers' speech to young children: Age and sex comparisons. *Child Development*, 44:182–185.

Polanyi, L. (1985). A theory of discourse structure and discourse coherence. In Eilfort, W., Kroerber, P., and Peterson, K., editors, *Proceedings of the 21st Regional Meeting of the Chicago Linguistics Society*, Chicago, Ill. University of Chicago Press.

Pylyshyn, Z. (1984). *Computation and Cognition.* Bradford Books/The MIT Press.

Reichman, R. (1985). *Getting Computers to Talk Like You and Me.* The MIT Press.

Reichman-Adar, R. (1984). Extended person-machine interface. *Artificial Intelligence*, 22(2):157–218.

Rubin, A. (1980). A theoretical taxonomy of the differences between oral and written language. In Spiro, R. J., Bruce, B. B., and Brewer, W. F., editors, *Theoretical Issues in Reading Comprehension.* Erlbaum.

Schiffrin, D. (1982). *Discourse Markers.* PhD thesis, University of Pennsylvania, Philadelphia PA.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–424.

Searle, J. r. (1992). *The Rediscovery of The Mind.* Bradford/MIT Press.

Shatz, M. and Gelman, R. *The development of communication skills: Modifications in the speech of young children ans a function of the listener*, volume 38 of *Monographs of the Society for the Research in Child Development.*

Sidner, C. L. (1982). Protocols of users manipulating visually presented information with natural language. Technical report, Bolt, Berenek and Newman.

Smith, J. E. K. (1976). The assuming-will-make-it-so fallacy. *Journal of Verbal Learning and Verbal Behavior*, 15:262–263.

Snow, C. (1972). Mothers' speech to children learning language. *Child Development*, 4:1–22.

Swerts, M., Gekyukens, R., and Terken, J. (1992). Prosodic correlates of discourse units in spontaneous speech. In *International Conference on Spoken Language Processing*, Banff.

Wike, E. L. and Church, J. D. (1976). Comments on clark's 'the language-as-fixed-effect fallacy'. *Journal of Verbal Learning and Verbal Behavior*, 15:249–255.

CLAUSE-INTERNAL STRUCTURE IN SPOKEN DIALOGUE

J.H. Connolly[1], A.A. Clarke[1], S.W. Garner[2] and H.K. Palmén[3]

1:Department of Computer Studies, Loughborough University of Technology, UK.

2: Department of Design and Technology, Loughborough University of Technology, UK.

3: Department of Psychology, University of Portsmouth, UK. Formerly at the Department of Computer Studies, Lougborough University of Technology, UK.

ABSTRACT

In this paper, some results are presented which derive from a recent research project relating to Computer-Supported Co-operative Work (CSCW), carried out by the authors in association with colleagues. The field of CSCW is concerned (inter alia) with the question of how computing systems may be designed to support groups of two or more individuals who are working in co-operation with one another to accomplish some common task. Clearly, most existing software is designed essentially for the use of single individuals, and is therefore not necessarily suitable as a basis for CSCW. Consequently, there is a need for the provision of systems engineered with the specific aim of supporting multi-user co-operative work. Such systems are often referred to under the heading of 'groupware'.

Of course, before high-quality groupware can be provided, it has first to be designed! Hence the question arises of what the design requirements of CSCW systems might be. The aim of our investigation was to address this issue in relation to one particular application domain, namely the design of industrial products through the co-operative activity of a pair of practitioners situated in separate geographical locations. Being physically remote form each other, these co-operating designers would be unable to communicate face-to-face, but would need to rely on electronic media to support their communicative interaction. This raised the problem of specifying what their exact communicational requirements would be – a question that would need to be answered in detail before an appropriate groupware-based system could be designed to support their work.

It is upon this specific question that our investigation was focused. The stated aim of the research was that of 'establishing the communicational requirements of IT systems that support humans co-operating remotely', and the investigation has thus become known as the ROCOCO (RemOteCOoperation and COmmunication) Project. It involved observing and recording various pairs of (student) product designers under different experimental conditions (e.g. with and without a video-link enabling them to see each other's faces) and then conducting a detailed analysis of their communicative activities in the context of their co-operative task.

A crucial part of the study consisted in a linguistic analysis of the dialogue between the designers in each pair. This analysis yielded a wide range of results, and it is the purpose of the present paper to present some of these.

It transpires that one of the salient linguistic characteristics of the dialogues in the ROCOCO corpus is that many of the clauses attested within it are not grammatically complete and well-formed, mainly as a result either of ellipsis or of being left unfinished by the speaker involved. In this paper, the structural details of these syntactically non-integral clauses will be outlined, and some relevant descriptive-statistical findings presented. In addition, an attempt will be made to draw some broader conclusions in respect of both language science and language technology.

1. INTRODUCTION

It is well known that natural speech does not always conform to the syntactic norms that are characteristic of the written language. This observation forms the starting point of the present paper, which is based on the analysis of a collection of data obtained during the course of a research

project in the field of Computer-Supported Co-operative Work (CSCW). Attention will be focused, in particular, upon the internal structure of main clauses in spoken dialogue, as evidenced by this corpus.

2. THE CORPUS

When two human beings are engaged in a co-operative task, it is possible for them to be situated either proximally (i.e. in the same place as each another, so that they may communicate face-to-face) or remotely (i.e. indifferent places, so that if they wish to communicate verbally, then they must do so by utilising some kind of technological artifact, such as a telephone line or a computer network). Computers may be used to support either proximal or remote co-operative work, but the stated aim of the particular research project which gave rise to the present corpus was that of 'establishing the communicational requirements of IT systems that support humans co-operating remotely'. The investigation was therefore named the ROCOCO (RemOte COoperation and COmmunication) Project; see Scrivener et al.(1993).

The co-operative task which formed the object of the investigation was that of industrial product design. Pairs of student product-designers were asked to work together to develop a design for some potential new product, such as a barbecue which would be suitable for taking on a picnic in the country and would derive its power-supply from a car battery. The designers were placed in separate locations, but each was provided with a workstation. In accordance with the purpose of the investigation, the two workstations were networked together. Both workstations were equipped with a keyboard, a graphics tablet and stylus, and two colour screens. In addition, a telephone-style link was installed, together with a headset at each workstation, enabling hands-free use. Furthermore, two video cameras were installed at each location. One of these was pointed at the person sitting at the workstation, and it was possible to display his/her image on one of the screens at the other workstation. The remaining cameras were mounted at an angle above the designers, and linked to a recording device, enabling the sessions to be taped for the benefit of the investigators.

A specially-developed item of software called the 'ROCOCO Sketch Pad' was mounted on each workstation. This displayed a shared (virtual) drawing surface in a window on the second screen at each workstation. By using the stylus and tablet, either designer could draw on the sketch pad, and an identical image of the drawings would be displayed on both screens simultaneously. Two distinguishable cursors were also displayed simultaneously on the screens, one controlled by each designer, enabling them to point at particular parts of the drawing for the benefit of the other participant.

In order to address the question of what the communication requirements might be for a system intended to support two individuals engaged in a co-operative task of this kind, an experiment was conducted which involved dividing the 20 designer-pairs into four equal-sized groups, and giving each group a different range of communication-channels. The first group was given the use of both the audio and the video links. This set-up will be referred to henceforth as the 'full configuration'. The second group was given the use of the audio link but not the video link, which was switched off in this case. This set-up will be termed the 'audio-only' condition. The third group was given use of the video link but not the audio link, while the fourth was given use of neither the audio nor the video link. However, all four groups were able to use the shared drawing surface and to communicate via this channel if they wished.

The intention was to compare the communicative activity in the four different experimental settings, and to observe how comfortably or otherwise the second, third and fourth groups coped with the various forms of communication-channel deprivation, in order to draw conclusions about the importance of providing the different types of channel. This work is still in progress, but the research to date has included the production of transcripts of the communicative interaction between the members of all twenty designer-pairs. A random ten-minute sample has been taken from each of the transcripts pertaining to the full configuration and to the audio-only conditions, and it is these ten samples that constitute the corpus on which the present paper is based.

3. CLAUSE STRUCTURE

3.1 CLAUSES WITHIN DIALOGUE

The dialogues that comprised the corpus have been analysed into conversational turns. It has been discovered that nearly every such turn conforms to the following structure:

(1) (a) Dialogue → Turn^{+}
(b) Turn → Sequence^{+}
(c) Sequence → (Discourse Particle +) (Connective^{+} +) Main Clause (+ Tag)

Here the superscript plus-sign means 'one or more'. Discourse particles included markers such as *well, anyway,* and so on, while connectives comprised syntactic links, including both conjunctions like *and* and conjunctive adverbials such as *therefore.* Tags are appended interrogative structures like *isn't it.* Optional items are placed between brackets.

What is presented in (1) is not, of course, intended to be interpreted as a serious attempt at a generative text-grammar. It simply offers a summary of the structure of the turn and shows how the clause fits into the latter.

3.2 CLAUSE-INTERNAL ANALYSIS

The principal focus of attention in the present paper, however, is upon the internal structure of the main clauses that occur within the corpus. For the purposes of the current investigation, the clauses concerned are subdivided into the following classes:

(2) (a) Major clauses, which are analysable in terms of the usual syntactic categories such as subject, verbal element object and so forth.
(b) Minor clauses (e.g. *Yes*) which are not thus analysable.

(This terminological usage is derived from Crystal, Garmanand Fletcher, 1989.)

The major clauses are further subdivided as follows:

(3) (a) Integral clauses, which are complete and well-formed in accordance with the accepted grammar of (written) English.
(b) Non-integral clauses, which may be categorised in terms of the following non-disjoint classes:
(i) Elliptical clauses.
(ii) False starts.
(iii) Otherwise non-integral clauses.

Within clauses, constituents are denoted in terms of the nomenclature of Quirk et al. (1985). They are also grouped under the following headings:

(4) (a) Major elements, i.e. subject, verbal element, object and complement.
(b) Minor elements, i.e. adverbials, vocative elements, connective elements and discourse markers.

3.3 CLAUSE STRUCTURE

As can be seen from Table 1, in all there are 2832 main clauses in the corpus. Of these, just under three quarters are major clauses, the remainder being minor. However, among the major clauses, over two fifths are non-integral Consequently, integral major clauses constitute only a minority (in fact, about three sevenths) of the total number of main clauses.

Table 1: Types of main clause

	Integral Major	*Non-integral Major*	*Total Major*	*Minor*	*Total Main Clauses*
Full	642	466	1108	353	1461
Config.	*57.9*	*42.1*			
	43.9	*31.9*	*75.8*	*24.2*	
Audio	565	410	975	396	1371
Only	*57.9*	*42.1*			
	41.2	*29.9*	*71.1*	*28.9*	
Total	1207	876	2083	749	2832
	57.9	*42.1*			
	42.6	*30.9*	*73.6*	*26.4*	

The figures in italics represent percentages, firstly of the totals of major clauses and secondly of the overall totals of main clauses.

Table 2: Classification of non-integral major main clauses[1]

	Elliptical	*False Start*	*Other*	*Combination*	*Total*
Full Config.	282	160	13	11	466
	60.5	*34.4*	*2.8*	*2.4*	
Audio Only	240	128	11	31	410
	58.5	*31.2*	*2.7*	*7.6*	
Total	522	288	24	42	876
	59.6	*32.9*	*2.7*	*4.8*	

In Table 2 the non-integral major main clauses are subclassified into four groups, namely those which are:

(5) (a) Elliptical.

(b) False starts.

(c) Non-integral for some other reason.

(d) Some combination of the above.

This table indicates that approximately three fifths of the non-integral major main clauses are elliptical, while approximately one third represent false starts. Those clauses which are non-integral for some other reason constitute less than 3% of the total, while those which fit simultaneously into more than one of our three non-integral categories amount to just under 5% overall.

If we absorb the latter into the categories identified in (5a), (5b) and (5c), then we arrive at Table 3, which represents a subdivision of non-integral major main clauses as follows:

(6) (a) Non-integral either solely through ellipsis or for a combination of reasons including ellipsis.

(b) Non-integral either solely as a result of being a false start or for a combination of reasons including being a false start.

(c) Non-integral solely for some other reason or else for a combination of reasons including, but not confined to, being elliptical and/or a false start.

In terms of this classification, in the corpus as a whole, 59.8% of the non-integral major main clauses are elliptical and 34.4% represent false starts, while the clauses which are non-integral for some other reason amount to 5.8% of the total.

As is also clear from Tables 1, 2 and 3, there is very little difference between the statistics relating to the two distinct experimental conditions.

Table 3: Classification of non-integral major main clauses[1]

	Elliptical	*False Start*	*Other*	*Total*
Full Config.	290	167	20	477
	60.8	*35.0*	*4.2*	
Audio Only	259	149	33	441
	58.7	*33.8*	*7.5*	
Total	549	316	53	918
	59.8	*34.4*	*5.8*	

[1] The figures in italics represent percentages of the totals in the rightmost column.

3.4 PATTERNS OF ELLIPSIS

Let us now look in greater detail at the structural characteristics of the elliptical clauses in the corpus. As shown in Table 4a, approximately 70% of these clauses completely lack a verbal element. (This compares with only 37% in the case of false starts, as can be seen from Table 4b.) On the other hand, the verbal element is present but incomplete in just over 10% of elliptical clauses (very similar, this time, to false starts) and present in its entirety in fewer than 20% of the total (compared with over 53% in false starts).

Table 4: Elliptical major main clauses and false starts with the verbal element complete, incomplete or absent[1]

	4a: major main clauses				4b: false starts			
	Complete	*Incomplete*	*Absent*	*Total*	*Complete*	*Incomplete*	*Absent*	*Total*
Full Config.	61	26	203	290	96	14	57	167
	21.0	*9.0*	*70.0*		*57.5*	*8.4*	*34.1*	
Audio Only	43	32	184	259	72	17	60	149
	16.6	*12.4*	*71.0*		*48.3*	*11.4*	*40.3*	
Total	104	58	387	549	168	31	117	316
	18.9	*10.6*	*70.5*		*53.2*	*9.8*	*37.0*	

Table 5: Omission of constituents other than the verbal element[1]

	5a: from major main clauses containing a complete verbal element				5b: from false starts containing a complete verbal element			
	Preverbal Only	*Postverbal Only*	*Other*	*Total*	*Preverbal Only*	*Postverbal Only*	*Other*	*Total*
Full Config.	33	23	5	61	7	84	5	96
	54.1	*37.7*	*8.2*		*7.3*	*87.5*	*5.2*	
Audio Only	26	9	8	43	2	59	11	72
	60.5	*20.9*	*18.6*		*2.8*	*81.9*	*15.3*	
Total	59	32	13	104	9	143	16	168
	56.7	*30.8*	*12.5*		*5.4*	*85.1*	*9.5*	

Table 6a: Omission of constituents other than the verbal element from elliptical major main clauses containing an incomplete verbal element[1]

	Preverbal Only	*Postverbal Only*	*Nothing Else Incomplete*	*Other*	*Total*
Full Config.	15	9	0	2	26
	57.7	*34.6*	*0.0*	*7.7*	
Audio Only	21	6	2	3	32
	65.6	*18.8*	*6.3*	*9.4*	
Total	36	15	2	5	58
	62.2	*25.9*	*3.4*	*8.6*	

[1] The figures in italics represent percentages of the totals in the rightmost column.

Table 6b: Omission of constituents other than the verbal element from false starts containing an incomplete verbal element[1]

	Preverbal Only	*Postverbal Only*	*Nothing Else Incomplete*	*Other*	*Total*
Full Config.	0	11	1	2	14
	0.0	*78.6*	*7.1*	*14.3*	
Audio Only	0	16	0	1	17
	0.0	*94.1*	*0.0*	*5.9*	
Total	0	27	2	2	31
	0.0	*87.1*	*6.5*	*6.5*	

Among elliptical major main clauses containing a complete verbal element (Table 5a), in over half the cases ellipsis is confined to preverbal elements only, whereas it is limited to postverbal elements in approximately three tenths of the total number of instances. Where the verbal element is incomplete, a fairly similar distribution is found, but with a slightly greater difference emerging between the preverbal and postverbalf requencies-of-occurrence. In fact, as we see from Table 6a, ellipsis affects just the preverbal and verbal elements in over three fifths of the clauses in question, while it is confined to the verbal and preverbal elements in about a quarter of the total number of cases.

As is clear from Tables 5b and 6b, there is a considerable difference here between elliptical clauses and false starts. Whereas the omission of material is, as we have just seen, concentrated in the earlier part of elliptical clauses, not surprisingly the reverse is true of false starts. Here, in cases where the verbal element is intact, over 85% have material omitted from the postverbal region only, while among those false starts where the verbal element is present but incomplete, omission of material is confined to the verbal and postverbal parts of the clause in over 87% of the total.

Returning to elliptical clauses, let us look more closely first of all at those in which the verbal element is present and complete but preverbal material is omitted. Here we find (Table 7) that in two thirds of instances the whole of the preverbal part of the clause is missing, and in almost a further quarter of the

Table 7: Omission of preverbal constituents from elliptical major main clauses containing a complete verbal element[1]

	All Preverbal Constituents Absent	*At least One but not All Preverbal Constituents Absent*	*Part of One Preverbal Constituents Absent*	*Total*
Full Config.	22	10	1	33
	66.7	*30.3*	*3.0*	
Audio Only	20	4	2	26
	76.9	*15.4*	*7.7*	
Total	42	14	3	59
	71.2	*23.7*	*5.1*	

[1] The figures in italics represent percentages of the totals in the rightmost column.

Table 8: Omission of postverbal constituents from elliptical major main clauses containing a complete verbal element[1]

	All Postverbal Constituents Absent	*At least One but not All Postverbal Constituents Absent*	*Part of One Postverbal Constituent Absent*	*Total*
Full Config.	16	3	4	23
	69.6	*13.0*	*17.4*	
Audio Only	6	2	1	9
	66.7	*22.2*	*11.1*	
Total	22	5	5	32
	68.8	*15.6*	*15.6*	

total, while there is some preverbal material present, nevertheless at least one whole element is omitted. In instances where the verbal element is present and complete but there is postverbal material missing (Table 8), in over two thirds of cases the omission extends to the whole postverbal part of the clause. The remaining third are divided equally between those instances where at least one whole postverbal constituent is absent and those where only part of one such constituent is left out.

Table 9: Omission of preverbal constituents from elliptical major main clauses containing an incomplete verbal element[1]

	All Preverbal Constituents Absent	*At least One but not All Preverbal Constituents Absent*	*Part of One Preverbal Constituent Absent*	*Total*
Full Config.	14	1	0	15
	93.3	*6.7*	*0.0*	
Audio Only	20	1	0	21
	95.2	*4.8*	*0.0*	
Total	34	2	0	36
	94.4	*5.6*	*0.0*	

Table 10: Omission of postverbal constituents from elliptical major main clauses containing an incomplete verbal element[1]

	All Postverbal Constituents Absent	*At least One but not All Postverbal Constituents Absent*	*Part of One Postverbal Constituent Absent*	*Total*
Full Config.	8	1	0	9
	88.9	*11.1*	*0.0*	
Audio Only	5	1	0	6
	83.3	*16.7*	*0.0*	
Total	13	2	0	15
	86.7	*13.3*	*0.0*	

[1] The figures in italics represent percentages of the totals in the rightmost column.

In Tables 9 and 10, the corresponding statistics are given for those clauses in which the verbal element is present but incomplete. Among those cases where the omission affects preverbal material, almost 95% show a total absence of the elements in question. As for clauses exhibiting omission of postverbal material, again a very high proportion (nearly 87%) of instances show a complete rather than a partial absence of the material in question.

Combining the totals in Tables 7 and 9 and in Tables 8 and 10, we find (Table 11a) that if we take all the elliptical clauses in which the omission of material outside of the verbal element is either confined to the preverbal region or limited to the postverbal region, then the exclusively preverbal ellipsis predominates over the exclusively postverbal ellipsis by a ratio of almost exactly 2:1. This contrasts markedly with the corresponding ratio for false starts, which is 20:1 in favour of exclusively postverbal ellipsis (Table 11b).

In elliptical major main clauses from which the verbal element is completely omitted (Table 12), we find that the most common pattern is for all but one major constituent also to be absent. This pattern

Table 11: Omission of preverbal and postverbal constituents from elliptical major main clauses containing a complete or incomplete verbal element[1]

	11a: from elliptical major main clauses containing a complete or incomplete verbal element			11b: from false starts containing a complete or incomplete verbal element		
	Preverbal Material Omitted	*Postverbal Material Omitted*	*Total*	*Preverbal Material Omitted*	*Postverbal Material Omitted*	*Total*
Full Config.	48	32	80	7	95	102
	60.0	*40.0*		*6.9*	*93.1*	
Audio Only	47	15	62	2	75	77
	75.8	*24.2*		*2.6*	*97.4*	
Total	95	47	142	9	170	179
	66.9	*33.1*		*5.0*	*95.0*	

Table 12: Omission of constituents from elliptical major main clauses containing no verbal element[1]

	All Major Constituents Abent	*All but Part of One Major Constituent Absent*	*All but One Major Constituent Absent*	*Other*	*Total*
Full Config.	58	13	124	8	203
	28.6	*6.4*	*61.1*	*3.9*	
Audio Only	49	17	111	7	184
	26.6	*9.2*	*60.3*	*3.8*	
Total	107	30	235	15	387
	27.6	*7.8*	*60.7*	*3.9*	

[1] The figures in italics represent percentages of the totals in the rightmost column.

amounts to just over two thirds of the total. However, in over a quarter of cases, all the major constituents are omitted, leaving only minor elements overtly expressed.

Ellipsis affects the full range of functionally-defined element (subject, verbal element, object, complement and adverbial). Similarly, it extends to more-or-less all types of structurally-defined unit (main and subordinate clauses, as well as nominal, verbal, adjectival and prepositional phrases).

3.5 INSERTIONS AND REPETITIONS

Although ellipsis is the most usual reason why clauses in our corpus may be classed as non-integral, it is not the only such ground, since the insertion or repetition of material can also be implicated. In major main clauses classed as elliptical (Table 13a), false starts (Table 13b) or otherwise non-integral(Table 13c), material is sometimes inserted between constituents and sometimes into the interior of constituents, with the former pattern predominating overall (ranging from 62.5% to 75.8%). Repetitions also occur in all three types of clause just listed(Tables 14a, 14b and 14c). These may consist in the reiteration either of entire constituents or of material within constituents. The former pattern predominates in elliptical clauses (100%) and in false starts (100%), while the latter is slightly more common among the remaining non-integral clauses (57.1%).

Table 13: Insertion of material *between* and *within* constituents[1]

	13a: into elliptical major main clauses			13b: into false starts			13c: into otherwise non-integral major main clauses		
	Between	*Within*	*Total*	*Between*	*Within*	*Total*	*Between*	*Within*	*Total*
Full	1	2	3	1	3	4	8	6	14
Config.	*33.3*	*66.7*		*25.0*	*75.0*		*57.1*	*42.9*	
Audio	7	1	8	9	3	12	17	2	19
Only	*87.5*	*12.5*		*75.0*	*25.0*		*89.4*	*10.5*	
Total	8	3	11	10	6	16	25	8	33
	72.7	*27.3*		*62.5*	*37.5*		*75.8*	*24.2*	

Table 14: Repetitions *of* and *within* constituents[1]

	14a: within elliptical major main clauses			14b: within false starts			14c: within otherwise non-integral major main clauses		
	Of	*Within*	*Total*	*Of*	*Within*	*Total*	*Of*	*Within*	*Total*
Full	2	0	2	3	1	4	5	5	10
Config.	*100.0*	*0.0*		*75.0*	*25.0*		*50.0*	*50.0*	
Audio	0	0	0	0	1	1	1	3	4
Only	*0.0*	*0.0*		*0.0*	*100.0*		*25.0*	*75.0*	
Total	2	0	2	3	2	5	6	8	14
	100.0	*0.0*		*60.0*	*40.0*		*42.9*	*57.1*	

[1] The figures in italics represent percentages of the totals in the rightmost column.

4. CONCLUSIONS

Let us now step back from the detailed empirical findings and draw out some general implications. The fact that such a large proportion of the sample consisted of clauses other than the integral major types is not without significance both for language science and language technology.

We have seen from the empirical findings presented in this paper that structurally integral major clauses are in the minority within the corpus of spoken dialogue examined here. Moreover, we have found ellipsis to be ubiquitous, affecting all kinds of syntactic unit, whether phrasal or clausal, and all kinds of functionally-defined element However, in the linguistic literature generally, the properties of clauses other than the integral major type, although they are not entirely neglected, do not, perhaps, receive as much attention as the results presented in this paper (at least, if these are reasonably typical) suggest they deserve. Of course, we have to acknowledge that linguistic description involves a certain amount of idealisation, and that phenomena like false starts were sidelined long ago by Chomsky (1965) as constituting performance errors. However, elliptical major clauses and minor clauses cannot be dismissed so easily. In the case of elliptical clauses, it may(or may not) be possible to reconstruct the corresponding fully-formed clause, but the ellipsis is not, in general, due to any kind of production error. It is intentional, it is governed by certain constraints (see Quirk et al.: ch. 12), and often it is stylistically preferable to an otiose, if syntactically complete, alternative. As for minor clauses, in general these simply cannot be treated as reduced versions of other, fully-fledged structures, but require recognition in their own right. The extension of existing descriptive linguistic frameworks in order to enable them to account properly for minor and elliptical clauses is a task that needs to be addressed more fully than has happened hitherto.

A related issue is the extent to which the same kind of grammatical description should be used for both the written and the spoken forms of a language which, like English, has a long written tradition. Studies such as Crystal (1980)have raised this question in the past, but it remains in need of further investigation.

Another important point is that, despite the frequent occurrence of non-integral clauses, the designers within each pair manage to communicate effectively enough to enable them to make progress in their co-operative task. This fact bears witness to the functional resilience of natural language, whereby it can be used successfully as a vehicle of communication even when the discourse shows a certain lack of grammatical integrity.

As far as language technology is concerned, the implications of findings such as those presented here are that if we want to have systems capable of handling spoken language in real time, then the problem of dealing with non-integral material cannot be avoided. At this point the obvious question is whether such systems would indeed be desirable, The answer is surely that they would. For example, systems which automatically kept a written record of discussions, or which translated between one language and another and thus made possible a live (and not severely restricted) dialogue between people with no common tongue, would be a great boon if they could be achieved. (For further discussion see Connolly, 1994).

The problem of processing non-integral input is well known within the field of language technology, and various approaches have been advanced in order to deal with it. (See for example Volume 9, Parts 3 and 4, of the *American Journal of Computational Linguistics*, and also Frederking, 1988.) However, it has certainly not been solved, and requires a good deal of further research. This is one of a number of reasons why it is unrealistic to expect the early appearance of broad-coverage systems capable of processing natural speech in real time. (Another is the sheer amount of time needed by a computer to process data of the complexity of natural language.)

On the other hand, to end on a positive note, the problem of describing and processing natural speech is undoubtedly a source of much potential intellectual interest, and the development of improved technology in this area would surely have commercial value as well. The findings presented in this paper may serve, then, as a reminder of the significance of the problem of handling the full breadth of linguistic data in the context of both the science and the technology of natural language.

ACKNOWLEDGEMENTS

The ROCOCO Project was funded by S.E.R.C. grant GR/F35814. We are pleased to acknowledge the part played in this project by Stephen A.R. Scrivener (principal investigator), Sean M. Clark, André Schappo and Michael G. Smyth. Our thanks are also due to those who acted as our experimental subjects.

REFERENCES

Chomsky, N, (1965). *Aspects of the Theory of Syntax*. Cambridge, Mass.: MIT Press.

Connolly, J.H. (1994). Artificial Intelligence and Computer-Supported Co-operative Working in International Contexts. In Connolly, J.H. and Edmonds, E.A. (eds.), *CSCW and Artificial Intelligence*. London: Springer-Verlag. 141-159.

Crystal, D. (1980). Neglected grammatical factors in conversational English. In Greenbaum, S., Leech, G.N. and Svartvik, J. (eds.), *Studies in English Linguistics for Randolph Quirk*. London: Longman. 153-166.

Crystal, D., Garman, M. and Fletcher, P. (1989). *Grammatical Analysis of Language Disability*. 2nd edition. London: Cole and Whurr.

Frederking, R.E. (1988). *Integrated Natural Language Dialogue: a Computational Model*. Boston: Kluwer.

Quirk, R., Greenbaum. S., Leech, G.N. and Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Scrivener, S.A.R., Clark, S.M., Clarke, A.A., Connolly, J.H., Garner, S.W., Palmén, H.K., Smyth, M.G., and Schappo, A. (1993). Real-time communication between dispersed work groups via speech and drawing. *Wirtschaftsinformatik* 35, 149-156.

THE CODING OF DIALOGUE STRUCTURE IN A CORPUS

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline Kowtko,
Gwyneth Doherty-Sneddon, and Anne Anderson
Human Communication Research Centre
Universities of Edinburgh and Glasgow
2 Buccleuch Place, Edinburgh EH8 9LW, U.K.
56 Hillhead Street, Glasgow G12 9YR, U.K.

ABSTRACT

Currently, many researchers in discourse and dialogue are subjectively coding discourse and dialogue phenomena on collected corpora in order to correlate discourse structure with other aspects of conversation and dialogue. The HCRC Dialogue Group, working with a corpus of spontaneous task-oriented spoken dialogues, has developed coding schemes for conversational moves (defined by function), conversational game structure, and higher level conversational transaction structure. These schemes are proving very useful for investigating the relationship between intonation and utterance function, the differences between face-to-face, audio, and video-mediated communication, and the conversational characteristics of aphasics. In this paper we describe the coding schemes.

1. INTRODUCTION

Work on dialogue analysis has traditionally been done using isolated examples, either constructed or real. Now many researchers are beginning to try to code large dialogue corpora for higher level dialogue structure in the hope of establishing their findings statistically. One of the major issues for such work is how to make the coding scheme replicable, so that other researchers will be able to use it and so that results based on the coding are believable. In this paper we introduce the dialogue coding distinctions which are central to our work. The HCRC Dialogue Group's overall method is to use a vertical analysis of a corpus of spoken task-oriented dialogues in order to test hypotheses which arise from our theories about dialogue. That is, we independently code the linguistic structure of the dialogues at a number of levels ranging from speech characteristics to high level dialogue structure, and test hypotheses relating the codings at different levels or relating coding to non-linguistic aspects of interactions. By making public both our dialogue coding distinctions and the method by which we operate, we hope to allow coders using similar schemes to benefit from our experience.

2. DIALOGUE STRUCTURE IN THE HCRC MAP TASK

Our coding centres on the HCRC Map Task Corpus [Anderson et al. 1991], which is a collection of 128 task-oriented dialogues involving approximately fifteen hours of speech, available on compact disk. In the dialogues, two participants have slightly different versions of a simple map with approximately fifteen landmarks on it. One participant has a route drawn on the map; the task is for the other participant to duplicate the route. An example route giver map is given in Figure 1 (at the end of this paper). The trials balance the familiarity of the speakers — whether they were acquainted before the experiment — and whether eye contact was possible between the speakers or blocked by a screen. There is also variation in matching between landmarks on the participants' maps, opportunities for contrastive stress, and phonological characteristics of landmark names. Some trials were video taped as well as being tape-recorded.

Many of the distinctions which are useful for empirical dialogue work require some kind of subjective coding to be performed. The HCRC coding schemes for three different levels of structure in task-oriented dialogue are similar

to the three middle levels of structure in Sinclair and Coulthard's [Sinclair and Coulthard 1975] analysis of classroom discourse. At the highest level, dialogues are divided into *transactions*, which are subdialogues that accomplish one major step in the participants' plan for achieving the task. The size and shape of transactions is largely dependent on the task. In the map task, route givers typically divide the task up by dividing the route into manageable segments; a typical transaction is a subdialogue which gets the route follower to draw one route segment on the map. Transactions are made up of *conversational games*, which are often also called dialogue games [Carlson 1983, Levin and Moore 1977, Power 1979], interactions [Houghton 1986], or exchanges [Sinclair and Coulthard 1975]. All forms of conversational games encode the idea that, by and large, questions are followed by answers, statements by acceptance or denial, and so on. Game analysis makes use of this regularity to differentiate between "initiations" which set up an expectation about what will follow, and "responses" which fulfill those expectations. In addition, games are differentiated by the kind of purpose which they have (e.g., getting information from the partner vs. providing information). A conversational game is a set of utterances starting with an initiation and encompassing all utterances up until the purpose of the game has been either fulfilled (e.g., the requested information has been transferred) or abandoned. Games can nest within each other if one game is initiated to serve the larger goal of a game which has already been initiated (for instance, if a question is on the floor but the hearer needs to ask a clarificatory question before answering). Games are themselves made up of *conversational moves*, which are simply different kinds of initiations and responses classified according to their purposes.

Researchers at the HCRC have been using subjective coding of dialogue structure to study a variety of issues. One issue is the relationship between intonational contour and conversational move function. Previous work on this subject tends to follow one of two opposite approaches. It either describes very general discourse functions (e.g. connecting, continuing and segmenting; [McLemore 1991]) or it identifies very specific discourse contexts (e.g. anaphor distribution and turn-taking; [Hockey 1992]). In order to make progress in this area, these two approaches need to be combined. This in turn requires an independent description of dialogue context as the basis for a robust account of intonational function. Such an independent description is the analysis of conversational moves and games outlined in this paper. It is used in a study of intonation function in single-word utterances taken from spontaneous and read-aloud dialogue [Kowtko 1995]. Other issues currently studied include the differences between face-to-face and computer-mediated conversation [Doherty-Sneddon et al. 1995, Newlands et al. 1995], and the conversational characteristics of aphasics [Merrison et al. 1994]. Dialogue structure coding has also been used in some investigations outside of the HCRC (e.g., [Alexandersson et al. 1995, Condon and Cech 1995]).

3. THE MOVE CODING SCHEME

Our move coding analysis is the most substantial of our coding schemes. We developed it by extending the moves which make up Houghton's [Houghton 1986] interaction frames to fit the kinds of interactions which we found in the map task dialogues. The following parameters underly our move classes:

A. **Game role:** whether the move is of a type that initiates games, responds to an initiation, acknowledges a partner's contribution, or prepares the conversation for the initiation of a new game.

B. **Force:** (for initiating moves only) whether the move initiates commanding, informing, or questioning type games.

C. **Question type:** (for questions only) whether the move initiates a yes-no question or an open-ended wh-question.

D. **Scriptedness:** (for responses only) whether the move gives simply the information requested by the question or amplifies by giving some additional information.

E. **Reply type:** (for scripted responses only) whether the move essentially answers "yes" or "no", or gives information requested by a wh-question.

F. Information novelty: (for questions only) whether the move asks for completely new information or information which could have been remembered or inferred from previous conversation.

G. Information type: (for questions only) whether the move asks for domain information or plan-related information ("meta" information about the state of the dialogue participants' shared plan).

The first four of these parameters are fairly standard for move classification systems. The others are possibly peculiar to our coding, since one of the immediate purposes of developing the coding was to help study the relationship between intonation and move function. Putting together these distinctions gives us twelve different move classes. The conversational moves are defined as follows.

3.1. INITIATING MOVES

The coding scheme distinguishes the following move types, all of which set up the expectation of a response. Initiating moves often occur at the beginning of a game, where they introduce a new discourse purpose into the dialogue.

Instruct: An instruction commands the partner to carry out any action other than the one implicit in queries (i.e., "tell me the answer to this question"). The instruction can be quite indirect, as in (4) below, as long as it is obvious that there is a specific action which the instructor intends to convey (in this case, putting the pen down at the start). In the map task, this usually involves the route giver telling the route follower how to navigate part of the route. Participants can also give other instructions, such as telling the partner to go through something again but more slowly. In these and later examples, "G" denotes the instruction giver, the participant who knows the route, and "F", the instruction follower, the one who is being told the route. Editorial comments which help to establish the dialogue context are given in square brackets.

1. G: Go right round, ehm, until you get to just above them.
2. G: If you come in a wee bit so that you're about an inch away from both edges.
3. G: And I want you to go towards the left-hand side of the page.
4. G: We're going to start above th... directly above the telephone kiosk.
5. F: Say it... start again.
6. F: Go. [as first move of dialogue]

Explain: An explanation states information which has not been elicited by the partner. (If the information were elicited, the move would be a response, such as a reply to a question.) The information can be some fact about either the domain or the state of the plan or task.

1. G: Where the dead tree is on the other side of the stream there's farmed land.
2. G: I've got a great viewpoint away up in the top left-hand corner.
3. F: I have to jump a stream.
4. F: I'm in between the remote village and the pyramid.
5. G: I do hope that's better than the last one was.
6. F: Yeah, that's what I thought you were talking about.

Check: A checking move requests the partner to confirm information that the checker has some reason to believe, but is not entirely sure about. Typically the information to be confirmed is something which the partner has tried to convey explicitly or something which the checker believes was meant to be inferred from what the partner has said. In principle, checks could cover past dialogue events (e.g., "I told you about the land mine, didn't I?") or any other information that the partner is in a position to confirm. The only use we have noted of checking anything other than what the checker has been told is one where a participant is explaining a route for the second time to a different route follower, and checks for a feature on the partner's map that has not yet been mentioned in the current dialogue.

1. G: ... you go up to the top left-hand corner of the stile, but you're only, say about a centimetre from the edge, so that's your line.
   F: OK, up to the top of the stile?
2. G: Ehm, curve round slightly to your right.
   F: To my right?
   G: Yes.
   F: As I look at it?
3. F: I'm in between the remote village and the pyramid.
   G: Are you?
4. G: Right, em, go to your right towards the carpenter's house.
   F: Alright well I'll need to go below, I've got a blacksmith marked.
   G: Right, well you do that.
   F: Do you want it to go below the carpenter? [*]
   G: No, I want you to go up the left hand side of it towards green bay and make it a slightly diagonal line, towards, em sloping to the right.
   F: So you want me to go above the carpenter? [**]
   G: Uh-huh.
   F: Right.

Note that in example 4, the move marked * is not a check because it asks for new information, but the move marked ** is a check because the information was meant to be inferred from G's prior contributions.

Align: An align move checks the attention or agreement of the partner, or his readiness for the next move. The most common type of align requests the partner to confirm that the goal of some open game has now been achieved and that they are ready to move on. Participants often initiate this sort of align game when they do not know of any outstanding problems with the goal, but the partner has not clearly indicated that they are ready to move on. Some participants ask for this kind of confirmation immediately after they have issued an instruction, probably in order to force more explicit responses to what they say. Aligns also occur at top level checking that "everything is OK" (i.e., that the partner is ready to move on) without asking about anything in particular. Aligns can also be used to check that the partner understands what is being referred to by some description of a place on the map, "aligning" the partners' positions.

1. G: OK? [after an instruction and an acknowledgement]
2. G: You should be skipping the edge of the page by about half an inch, OK?
3. G: ... and then straight up so that you're... see where your farmer's gate is?
4. G: Then move that point up half an inch so you've got a kind of diagonal line again.
   F: Right.
   G: This is the left-hand edge of the page, yeah?
   F: Yeah, okay.

Query-yn: A query-yn asks the partner any question which takes a "yes" or "no" answer and does not count as a check or an align. In the map task, these questions are most often about what the partner has on the map.

1. G: Do you have a stone circle at the bottom?
2. G: I've mucked this up completely have I?
3. F: I've got Dutch Elm.
   G: Dutch Elm. Is it written underneath the tree?
4. G: Have you got a haystack on your map?
   F: Yeah
   G: Right just move straight down from there, then,
   F: Past the blacksmith? [with no previous mention of blacksmith]

Query-w: A query-w is any query which is not covered by the other categories. This includes polar questions where the polar distinction is not "yes" or "no".

1. G: Towards the chapel and then you've
   F: Towards what?

3.2. RESPONSE MOVES

The following moves are used within games after an initiation, and serve to fulfill the expectations set up within the game.

Acknowledgement: An acknowledgement is a verbal response which minimally shows that the speaker has heard the move to which it responds, and often also demonstrates that the move was understood and accepted and that the partner may continue (e.g. giving instruction). Verbal acknowledgements do not have to appear even after substantial explanations and instructions, since acknowledgement can be given non-verbally, especially in dialogue modalities with eye contact, and because the partner may not wait for one to occur.

1. G: Ehm, if you... you're heading southwards.
   F: Mmhmm.
2. G: Do you have a stone circle at the bottom?
   F: No.
   G: No, you don't.

Reply-y: A reply-y is any reply to any query with a yes-no surface form which means "yes", however that is expressed. Since reply-y's are elicited responses, they normally only appear after query-yn's, aligns, and checks.

1. G: See the third seagull along?
   F: Yeah.
2. G: Do you have seven beeches?
   F: I do.
3. F: Green Bay?
   G: Uh-huh.
4. G: Do you want me to run by that one again?
   F: Yeah, if you could.

Reply-n: Similar to reply-y's, a reply to a a query with a yes-no surface form which means "no" is a reply-n.

1. G: Do you have the west lake, down to your left?
   F: No.
2. G: So you're at a point that's probably two or three inches away from both the top edge, and the left-hand side edge. Is that correct?
   F: No, no at the moment.

One caveat about the meaning of the difference between reply-y and reply-n: there is a rare class of queries which include negation (e.g., "You don't have a swamp?"; "You're not anywhere near the coast?"). Replies to these questions are coded as reply-y or reply-n depending on the surface form of the answer, even though in this case "yes" and "no" can mean the same thing.

Reply-w: A reply-w is any reply to any type of query which does not simply mean "yes" or "no".

1. G: And then below that, what've you got?
   F: A forest stream.
2. G: No, but right, first, before you come to the bakery do another wee lump
   F: Why?
   G: Because I say.
3. F: Is this before or after the backward *s*?
   G: This is before it.

Clarify: A clarification is a reply to some kind of question in which the speaker tells the partner something over and above what was strictly asked. If the new information is substantial enough, then the utterance is coded as two moves, a reply followed by an explain, but in many cases, the information added is insubstantial enough that we would not want to code it as a separate move. Route givers tend to give clarifications when the route follower seems unsure of what to do, but there is not a specific problem on the agenda (such as a landmark now known not to be shared).

1. G: And then, have you got the pirate ship?
   F: Mmhmm.
   G: Just curve from the point, go right ... go down and curve into the right til you reach the tip of the pirate ship
   F: So across the bay?
   G: Yeah, through the water.

F: So I just go straight down?
G: Straight down, and curve to the right, til you're in line with the pirate ship.

2. [... instructions which keep them on land...]
F: So I'm going over the bay?
G: Mm, no, you're still on land.
F: Oh, the left, the left, sorry, yes. The left.
G: The left.

3.3. THE READY MOVE

In addition to the initiation and response moves, we identify "ready" moves as transitional moves between games. They occur after the close of a conversational game and prepare the dialogue for a new game to be initiated. Speakers often use utterances such as "OK" and "right" to serve this purpose. It is a moot point whether ready moves should form a distinct move class or should be treated as discourse markers attached to the subsequent moves, but the distinction is not a critical one, since either interpretation can be placed on the coding. We usually choose to consider ready moves as distinct, complete moves to emphasise the comparison with acknowledgements, which are often just as short and may even contain the same words as ready moves.

1. G: Okay. Now go straight down.

2. G: Now I have banana tree instead.

3. G: Right, if you move up very slightly to the right along to the right.

4. THE GAME CODING SCHEME

Moves are the building blocks for conversational game structure, which reflects the goal structure of the dialogue. In our move coding, we differentiated a set of initiating moves, all of which signal some kind of purpose in the dialogue. For instance, instructions signal that the speaker intends the hearer to follow the command, queries signal that the speaker intends to acquire the information requested, and statements signal that the speaker intends the hearer to acquire the information given. A conversational game is a sequence of moves starting with an initiation and encompassing all moves up until that initiation's purpose is either fulfilled or abandoned.

There are two important components of any game coding scheme. The first is an identification of the game's purpose; we identify the purpose simply by the name of the game's initiating move (e.g. an explaining game). The second is some explanation of how games are related to each other. The simplest, paradigmatic relationships are implemented in computer-computer dialogue simulations, such as those of Power [Power 1979], Houghton [Houghton 1986], and Guinn [Guinn 1994]. In these simulations, once a game has been opened, the participants work on the goal of the game until they both believe that it has been achieved or that it should be abandoned. This may involve embedding new games with subservient purposes to the top level one being played (for instance, clarification subdialogues about crucial missing information), but the embedding structure is always clear and mutually understood. Although some natural dialogue is this orderly, much of it is not. Participants are free to initiate new games at any time (even while the partner is speaking), and these new games can introduce new purposes rather than serving some purpose which is already present in the dialogue. In addition, natural dialogue participants often fail to make clear to their partners what their goals are and why they are saying what they are. This makes it very difficult to develop a replicable coding scheme for complete game structure.

Our game coding scheme simplifies these issues to the aspects of embedded structure which concern us the most. First, we code where new games begin, naming the game's purpose according to the game's initiating move. Although all games begin with an initiating move (possibly with a ready move prepended to it), not all initiating moves begin games, since some of our initiating moves serve to continue existing games or remind the partner of the main purpose of the current game again. Second, we code where games end or are abandoned. Finally, we mark games as either occurring at top level or being embedded (at some unspecified depth) in the game structure, and thus being subservient to some top level purpose, and we mark them as either eventually

being completed or being abandoned. The goal of these definitions is to give us enough information to study relationships between game structure and other aspects of dialogue whilst keeping those relationships simple enough to code.

5. THE TRANSACTION CODING SCHEME

Transaction coding gives the subdialogue structure of complete task-oriented dialogues, with each transaction being built up of several conversational games and corresponding to one step of the task. In most map task dialogues, the participants break the route into manageable segments and deal with them one by one. Because transaction structure for map task dialogues is so closely linked to what the participants do with the maps, we include the maps in the analysis. Our coding system has two components: we code both (1) how route givers divide conveying the route into subtasks and what parts of the dialogue serve each of the subtasks, and (2) what actions the route follower takes and when.

Our basic route giver coding identifies the start and end of each segment and the subdialogue which conveys that route segment. However, map task participants do not always proceed along the route in an orderly fashion; as confusions arise, they often have to return to parts of the route which were previously discussed and which at least one of them thought had been successfully completed. In addition, participants occasionally overview an upcoming segment in order to provide a basic context for their partners, without the expectation that their partners will be able to act upon their descriptions, as in the following transaction:

G: And what we're basically going to be... where we're basically going to be going is towards.. I'll t-say this sort of globally
F: Mmm-hmmm.
G: then I'll do it more precisely. What we're basically doing is going, erm, south-east and then, erm, north-east, so you can imagine... a bit like a diamond shape if you like.
F: Mmm.
G: Southeast then northeast
F: Mmm.
G: and then northwest and then north, but the line's a lot more wavy than that. I'm just trying to give you some kind of overall picture.
F: Mmm.
G: It may not be very useful but.

They also sometimes engage in subdialogues which are not relevant to any segment of the route, sometimes about the experimental setup but often nothing at all to do with the task. Other types of subdialogues are possible (such as checking the placement of all map landmarks before describing any of the route, or concluding the dialogue by reviewing the entire route), but were prejudged to occur infrequently enough that we do not include them in our coding scheme. This gives us four transaction types: 'normal', 'review', 'overview', and 'irrelevant'.

Coding involves marking where in the dialogue transcripts a transaction starts and which of the four types it is, and for all but 'irrelevant' transactions, indicating the start and end point of the relevant route section using numbered crosses on a copy of the route giver's map. We do not explicitly code endings of transactions because, generally speaking, transactions are large enough that they do not appear to nest; if a transaction is interrupted to, for instance, review a previous route segment, participants by and large restart the goal of the interrupted transaction afterwards rather than picking up where they left off. It is possible for several transactions (even of the same type) to have the same starting point on the route.

Our basic route follower coding identifies whether the follower action was drawing a segment of the route or crossing out a previously drawn segment, the start and end points of the relevant segment, indexed using numbered crosses on a copy of the route follower's map.

6. REPLICABILITY OF CODING SCHEMES

It is important to show that subjective coding distinctions can be understood and applied by people other than the coding developers, both to make the coding credible in its own right and to establish that it is suitable for testing empirical hypotheses. We have established successful replicability of our coding system through ex-

periments which we describe in another paper [Carletta et al. in preparation].

7. CONCLUSIONS

We have described subjective coding for three different levels of task-oriented dialogue structure, which we call conversational moves, games, and transactions. Our move coding divides the dialogue up into segments corresponding to the different discourse goals of the participants and classifies the segments into one of twelve different categories, some of which initiate a discourse expectation and some of which respond to an existing expectation. Our game coding shows how moves are related to each other by placing into one game all moves which contribute to the same discourse goal, including the possibility of embedded games, such as those corresponding to clarification questions. Our transaction coding divides the entire dialogue into subdialogues which correspond to major steps in the participants' plan for completing the task. We have found that dialogue structure coding in the Map Task Corpus is largely reproducible by other coders [Carletta et al. in preparation]. The coders were able to reproduce the most important aspects of the coding reliably, such as move segmentation, classifying moves as initiations or responses, and subclassifying initiation and response types.

AUTHOR NOTES

This work was completed within the Dialogue Group of the Human Communication Research Centre, funded by an Interdisciplinary Research centre Grant form the Economic and Social Research Council (U.K.) to the Universities of Edinburgh and Glasgow and grant number G9111013 of the Joint Councils Initiative.

REFERENCES

References

[Alexandersson et al. 1995] Alexandersson, J., Maier, E., and Reithinger, N. (1995). A robust and efficient three-layered dialogue component for a speech-to-speech translation system. In *Proceedings of the Seventh European Meeting of the ACL*, pages 188–193.

[Anderson et al. 1991] Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., and Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366.

[Cahn 1992] Cahn, J. (1992). An investigation into the correlation of cue phrase, unfilled pauses, and the structuring of spoken discourse. In *Proceedings of the IRCS Workshop on Prosody in Natural Speech (IRCS Report 92-37).*

[Carlson 1983] Carlson, L. (1983). *Dialogue Games: An Approach to Discourse Analysis.* D. Reidel.

[Carletta et al. in preparation] Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., Anderson, A. (in preparation). Coding dialogue structure in the map task Ms. in preparation for journal submission.

[Condon and Cech 1995] Condon, S. L. and Cech, C. G. (1995). *Functional comparison of face-to-face and computer-mediated decision-making interactions.* John Benjamins.

[Doherty-Sneddon et al. 1995] Doherty-Sneddon, G., Anderson, A. H., O'Malley, C., Langton, S., Garrod, S., and Bruce, V. (1995). Face-to-face and video mediated communication: A comparison of dialogue structure and task performance. Submitted for publication.

[Greene and Cappella 1986] Greene, J. O. and Cappella, J. N. (1986). Cognition and talk: The relationship of semantic units to temporal patterns of fluency in spontaneous speech. *Language and Speech*, 29(2):141–157.

[Guinn 1994] Guinn, C. (1994). *Meta-Dialogue Behaviors: Improving the Efficiency of Human-Machine Dialogue — A Computational Model of Variable Initiative and Negotiation in Collaborative Problem-Solving.* PhD thesis, Duke University.

[Hockey 1992] Hockey, B. A. (1992). Prosody and the interpretation of cue phrases In *Proceedings of the IRCS Workshop on Prosody in Natural Speech*, University of Pennsylvania, IRCS Report No.: 92–37, 71-77.

[Houghton 1986] Houghton, G. (1986). *The Production of Language in Dialogue: A Computational Model.* PhD thesis, University of Sussex.

[Kowtko 1995] Kowtko, J. (1995). The function of intonation in spontaneous and read dialogue To appear in the Proceedings of the XIIIth Congress of Phonetic Sciences, Stockholm.

[Levin and Moore 1977] Levin, J. A. and Moore, J. A. (1977). Dialogue games: Metacommunication structures for natural language interaction. *Cognitive Science*, 1(4):395–420.

[Litman and Hirschberg 1990] Litman, D. and Hirschberg, J. (1990). Disambiguating cue phrases in text and speech. In *Proceedings of the Thirteenth International Conference on Computational Linguistics (COLING-90)*, volume 2, pages 251–256.

[McLemore 1991] McLemore, C. A. (1991). *The Pragmatic Interpretation of English Intonation: Sorority Speech* Ph.D. dissertation, University of Texas at Austin.

[Merrison et al. 1994] Merrison, A., Anderson, A., and Doherty-Sneddon, G. (1994). An investigation into the communicative abilities of aphasic subjects in task-oriented dialogue Technical report, HCRC RP-50, June.

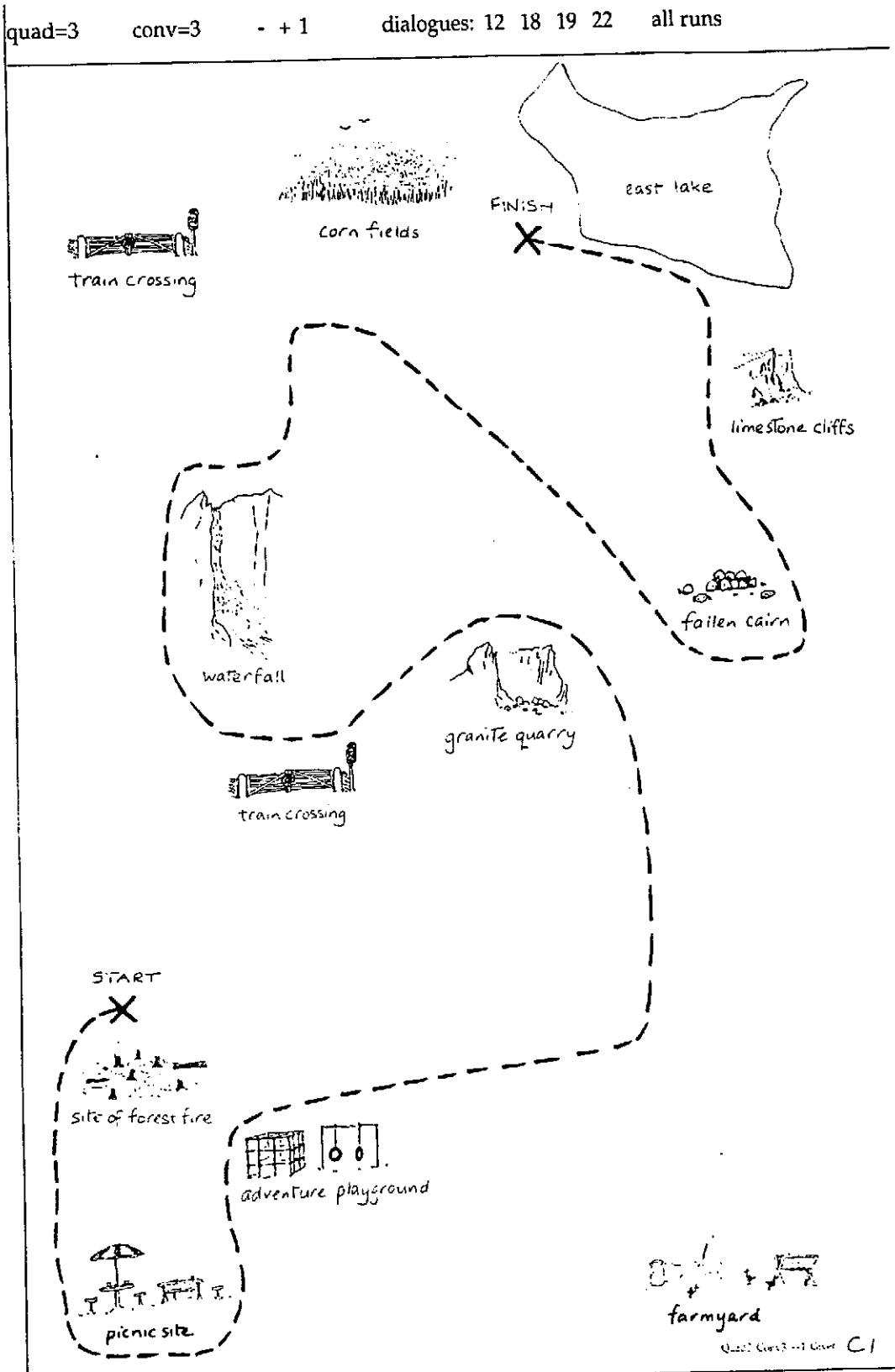
[Newlands et al. 1995] Newlands, A., Anderson, A. H., and Mullin, J. ((forthcoming) 1995). *Dialogue Structure and Co-operative Task Performance in two CSCW Environments.* Springer-Verlag.

[Passonneau and Litman 1993] Passonneau, R. J. and Litman, D. J. (1993). Intention-based segmentation: human reliability and correlation with linguistic cues. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 148–155.

[Power 1979] Power, R. J. D. (1979). The organisation of purposeful dialogues. *Linguistics*, 17:107–152.

[Sinclair and Coulthard 1975] Sinclair, J. M. and Coulthard, R. M. (1975). *Towards an Analysis of Discourse: The English used by teachers and pupils.* Oxford University Press.

Figure 1: An Example Route Giver Map



DESIGNING THE DIALOGUE COMPONENT IN A SPEECH TRANSLATION SYSTEM

—

A CORPUS BASED APPROACH

Jan Alexandersson and Norbert Reithinger *
DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbrücken
{alexandersson,reithinger}@dfki.uni-sb.de

ABSTRACT

New and challenging requirements arise for the dialogue processing component in the speech-to-speech translation system VERBMOBIL. It has to cope with both unexpected and vague input as well as gaps in the input. The design is based on a large corpus of transliterated dialogues. A careful analysis of this corpus and of the requirements from other components of VERBMOBIL resulted in a hybrid approach consisting of both knowledge based as well as statistic based processing. In this paper, we present the design process and the resulting architecture. Using the corpus, we made various experiments to evaluate the first design of the component.

1 INTRODUCTION

The role of the dialogue processing component in a speech-to-speech translation system like VERBMOBIL [20, 7] differs in various respects from other natural language systems with typed or speech input. One important point is that the translation system is not an active dialogue participant, except in cases where clarification dialogues between the system and the user are necessary. The users of the system interact in English and activate VERBMOBIL only if the owner of VERBMOBIL lacks knowledge of English and demands the translation of utterances in her mother tongue.

In contrast, a system like SUNDIAL [1, 14] – where the user requests travel information – is a dialogue participant that has the ability to control the ongoing dialogue to fulfill its task. It does not only monitor the dialog, it also actively engages in the interaction. In such a system the dialogue component plays an important role in controlling the overall system and the dialogue.

The application context of VERBMOBIL sets up demanding requirements for a dialogue component. In this paper we present the design process and the resulting architecture for the component and show first results for the fully implemented system. We conclude the paper with a discussion and a suggestion for further topics.

2 DESIGNING THE SYSTEM

The first application scenario for VERBMOBIL is an appointment scheduling dialogue between two business persons, one of them German, where both are non-native speakers of English. If the German partner is not able to express himself adequately he can switch to his mother tongue indicating the need for translation by pressing a button. VERBMOBIL is then expected to translate the utterances into English.

The Corpus

The empirical basis for the development of the dialogue component was a corpus of speech data. For different purposes in the development of VERBMOBIL, e.g. training the speech recognizers,

*This work was funded by the German Federal Ministry for Research and Technology (BMBF) in the framework of the VERBMOBIL Project under Grant 01IV101K/1. The responsibility for the contents of this study lies with the authors.

a large number of German-German scheduling dialogues has been collected and transliterated [8]. Like previous approaches for modeling task-oriented dialogues, we assume that a dialogue can be modeled by means of a limited but open set of dialogue acts (see e.g. [3], [10] for speech processing and [17] for the use for machine translation). We examined this corpus for the occurrence of dialogue acts as proposed by e.g. [2, 18] and for the necessity to introduce new, sometimes problem-oriented dialogue acts.

In a first step, we defined 17 dialogue acts together with semi-formal rules for their assignment to utterances [9]. Following the assignment rules, which also served as starting point for the automatic determination of dialogue acts within the semantic evaluation component, we hand-annotated over 200 dialogues with dialogue act information to make this information available for training and test purposes. After one year of experience with these acts, the users of dialogue acts in VERBMOBIL selected them as the domain independent "upper" concepts within a more elaborate hierarchy (see figure 1) that becomes more and more propositional and domain dependent towards its leaves [5]. Such a hierarchy is useful e.g. for translation purposes.

From the analysis of the annotated corpus we derived a standard model of admissible dialogue act sequences. Figure 2 shows our dialogue model which consists of a network representation of admissible sequences of speech acts. The model for the usual sequence of dialogue acts is described in the left network; digressions that can occur everywhere in the dialogue are displayed at the right side of the main net.

This dialogue model and the acts defined therein are the basic units for the processing in the dialogue component. Main input from the other modules is based on dialogue acts for an utterance, either determined during deep processing or while spotting the English parts of the dialogue. Also information for the other components is based on the dialogue acts.

Requirements from the other Components

In a system like VERBMOBIL that combines deep analysis for translation, shallow dialogue tracking by a keyword spotter, and speech as input, the tasks of a dialogue component are manifold. The three main requirements are

1. to provide and to store contextual information which is used by the linguistic modules of VERBMOBIL e.g. the transfer component
2. to provide top down expectations about what dialogue steps are most likely to follow. This information is used to support the analysis components for narrowing down the search space which is extremely important for speech processing systems (see e.g. [14]).
3. to integrate both modes of processing within a unified approach to get a hold on the overall flow of the dialogue

In contrast to the abovementioned systems like SUNDIAL, VERBMOBIL does not control the dialogue. Therefore the dialogue component cannot take over the part it plays in these other systems, namely guiding the user so that the information he gives can finally be used (e.g. to return a scheduling information). In the VERBMOBIL scenario the dialogue is between the two humans, and VERBMOBIL is only a tool for one of them.

The Internal Structure

To store the contextual information, we follow the approach of [4] for modeling the context, and describe it by three interconnected knowledge sources

- an intentional structure, a tree-like structure which contains information about the intentions for parts of the dialogue. This information is used amongst others to determine the dialogue act of the actual utterance.
- a thematic structure which represents local and global focus and the development of the different topics mentioned in the dialogue. It is for instance used by the transfer component.
- a referential structure which links the conceptual and language-related information for the objects mentioned. One example of the application of this knowledge is the generation of noun phrases in the target language.

To build and maintain these structures and to provide predictions, we had to select the appropriate processing mechanisms.

The first and obvious step for the implementation is to put the dialogue model into software, i.e. to implement it as an automaton. This is a technique frequently to be found in speech processing

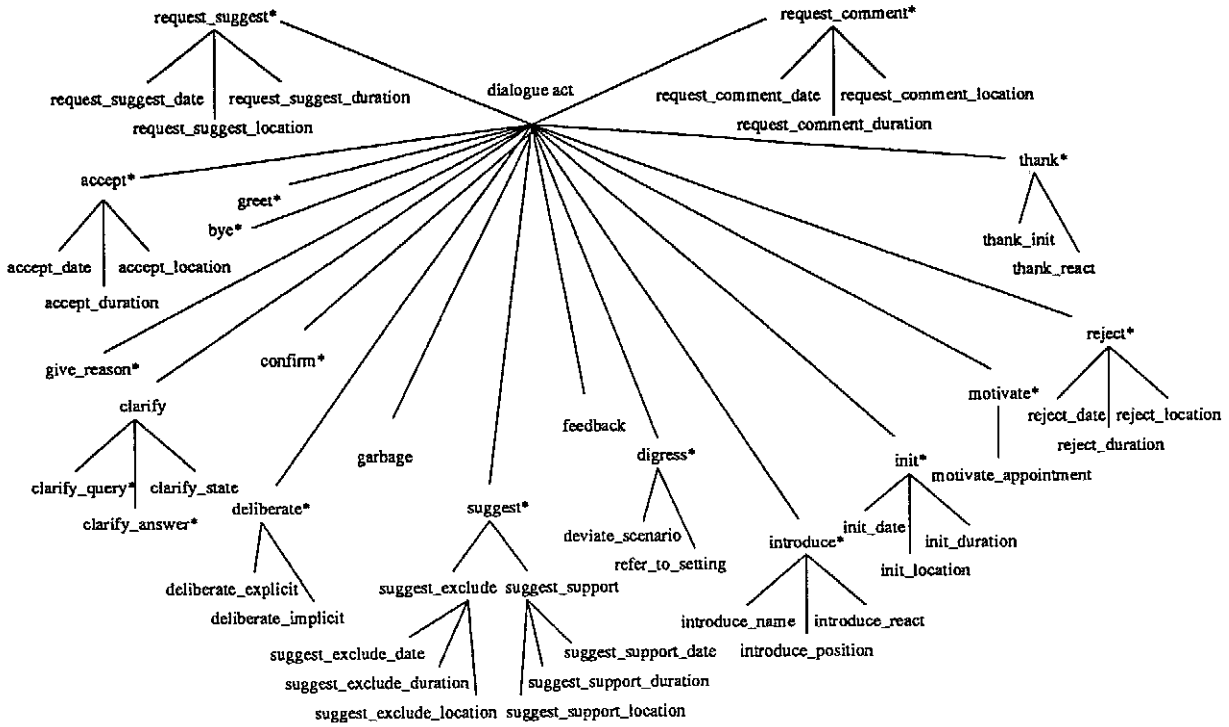


Figure 1: The taxonomy of dialogue acts

systems. It is a simple way to follow and control the flow of discourse. However, as VERBMOBIL does not actively participate in the dialogue, it has no control over the dialogue steps, and cannot rely on a reasonable sequence of dialogue acts, as it is e.g. the case in travel information systems. Also, the first two of the abovementioned requirements are hardly met with such a simple model.

We therefore divided processing up into two parts. One is responsible for building up the intentional and thematic structure, and one for the prediction process.

The development of the knowledge structures is the task of a plan recognizer. Its input consists of the dialogue acts and the propositional content expressed by the utterance when deep processing takes place. The leaves of the intentional structure are the acts, while the intermediate nodes represent subsections of the dialogue like the greeting or negotiating phase. This tree is built up incrementally, as the dialogue acts are provided by the other components of VERBMOBIL.

The source of the dialogue acts can either be deep linguistic processing in cases where one dialogue partner presses the activation button and demands a translation, or either the keyword spotter which tracks the English parts superficially. This spotter is one component provided with predictions for follow-up dialogue acts.

Structural knowledge sources are usually useless for prediction purposes since they provide too many, unscored predictions. To compute weighted dialogue act predictions we evaluated two methods: The first method is to attribute probabilities to the arcs of our network by training it with annotated dialogues from our corpus. The second method adopts information theoretic methods from speech recognition. We implemented and tested both methods and currently favor the second one because it is insensitive to digressions from the dialogue structure as described by the dialogue model and generally yields better prediction rates (see below).

In the next section, we describe, how these two processing approaches can be combined to form a synergetic processing environment.

3 ARCHITECTURE

Figure 3 shows an overview of the internal structure of the dialogue component. In the middle the three processing modules *plan recognizer*, *finite state machine (FSM)* and *statistics* are given.

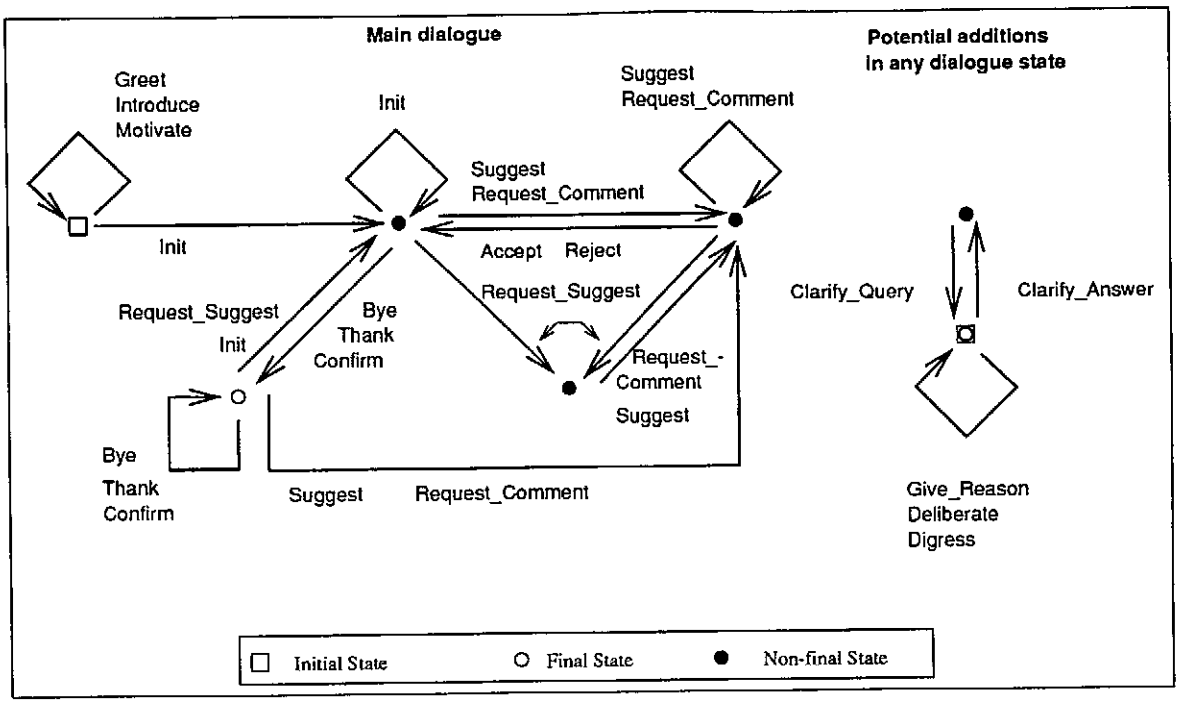


Figure 2: The Dialogue Model

On the left and the right hand side the different components we interact with and the type of interaction we perform are sketched. Below we first present the implementation of the three different components, and in the next section we present their evaluation.

The Finite State Machine

The finite state machine uses the dialogue model (see figure 2) as knowledge base. It parses the incoming dialogue acts and checks whether the ongoing dialogue is in accordance with the dialogue structure as defined in the networks. If the incoming dialogue act is not within the language generated by the model, the FSM uses a fall-back strategy, based on the information in the statistics component, to find the most probable follow-up state.

We have also tried to use the FSM to predict the next dialogue act to come but as we will see in section 4 the results are quite poor.

Currently, we use this module only to drive the graphical user interface as one means to visualize the ongoing dialogue.

The Statistics Component

To provide weighted predictions for the most probable follow-up speech acts, we use a statistical approach. A knowledge source like the network model cannot be used for this task since the average number of predictions in any state of the main network is 10 when including the speech acts that can occur everywhere in the dialogue. Given a total number of 17 dialogue acts this is not sufficiently restrictive, especially when there is no ranking information attached.

The statistical method is based on n-gram speech act probabilities, a method adapted from speech recognition (see [6, 14, 13], and [15] for an evaluation of the method used in VERBMOBIL).

The task of this module is to compute those dialogue acts that will most probably follow a given dialogue history, i.e. to compute

$$s_n := \max_s P(s|s_{n-1}, s_{n-2}, s_{n-3}, \cdots)$$

To approximate the conditional probability $P(.|.)$ the standard smoothing technique known as deleted interpolation is used [6] with

$$P(s_n|s_{n-1}, s_{n-2}) = q_1 f(s_n) + q_2 f(s_n|s_{n-1}) + q_3 f(s_n|s_{n-1}, s_{n-2})$$

where f are the relative frequencies of dialogue

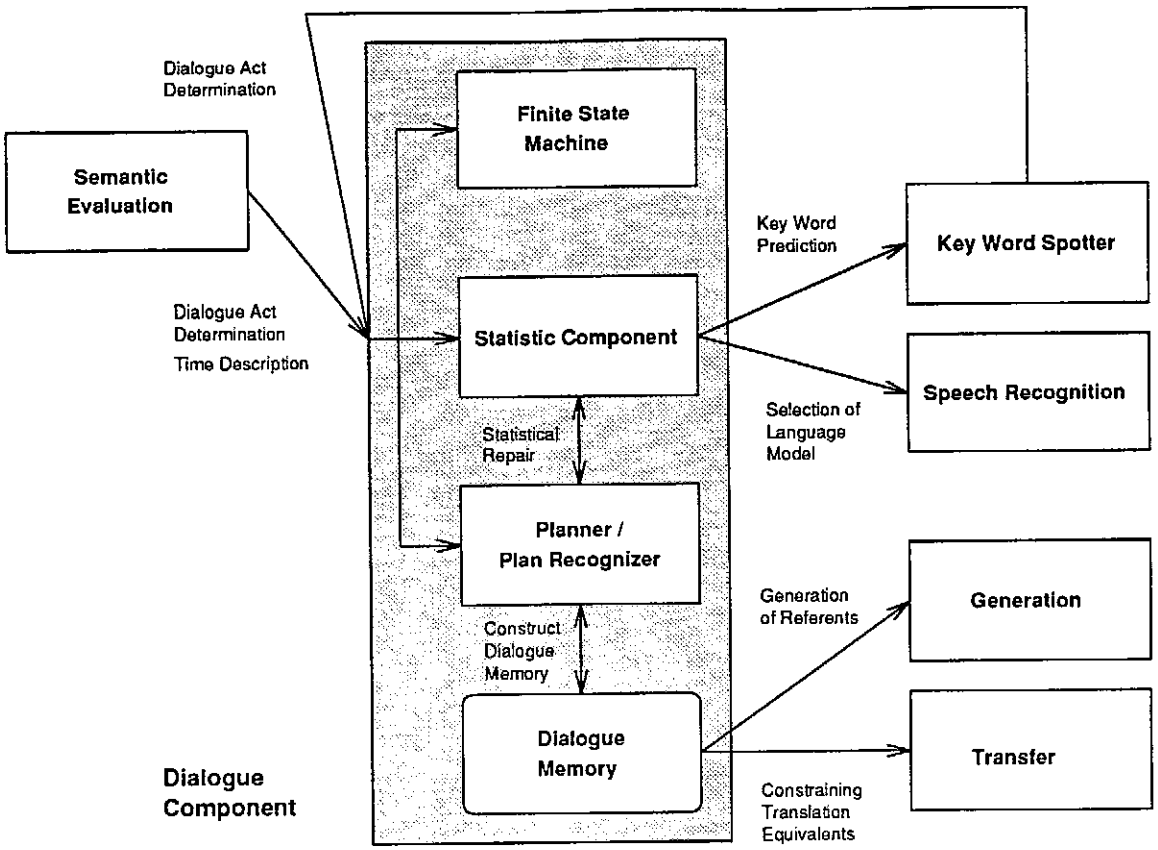


Figure 3: The Architecture of the Dialogue Module

act uni-, bi-, and trigrams computed from a training corpus and $\sum q_i = 1$. The weights q_i are determined with the HMM-based algorithm described in [6].

In [15] we show that a training corpus size of about 50 dialogues is sufficient to get an approximation that is based mainly on bigrams. I.e. q_2 is large compared to the uni- and trigramm coefficients. The annotated corpus is currently too small to allow for stable frequencies of higher n-grams.

The Plan Recognizer

The plan recognizer explores the connection between plan recognition and parsing as pointed out by Vilain [19]. In his paper he shows how a plan hierarchy can be compiled into a context free grammar. This approach is convenient since parsing is a well understood technique with a lot of both fast and robust approaches. Our first version is basically a simple top down parser with backtracking where the plan operators are processed strictly left-to-right (mimicking the behaviour of a prolog interpreter).

The plan recognizer operates on a set of *plan operators*. In our perspective, however, they are rules forming a grammar. The rules are used to encode both the dialogue model and methods for recovery from erroneous dialogue states. The latter is especially important: even when the dialogue partners deviate from a well-formed dialogue as defined in the dialogue model, the planner has to continue to construe the intentional structure of the dialogue.

Each rule represents a specific `goal` which it is able to fulfill in case specific `constraints` hold. These constraints mostly address the context, but they can also be used to check pragmatic features, like e.g. whether the dialogue participants know each other. Also, every plan operator can trigger follow-up `actions`. To be able to fulfill a goal a plan operator can define `subgoals` which have to be achieved in a pre-specified order (see for instance [11, 12] for comparable approaches).

One of the big problems with parsers is that

they are recognizers – they either accept or reject the input. We cannot allow the recognizer to fail since it would cause the whole module to fail. To prevent this we use two techniques when our input deviates from what the grammar allows for. The first method relies on the information of the statistical component allowing for reinterpreting our input, while the second uses a set of so-called *repair operators* for "repairing" the parse tree.

Below we first present the methods and then we give two examples taken from our corpus of annotated dialogues. The English translation, however, is not produced by VERBMOBIL.

Statistical Repair

This method for error recovery is based on the hypothesis that the attribution of only one speech act to a given utterance is insufficient and that an utterance has more than one speech act reading.

...

mhw3_1_07: `ja Montag der vierzehnte Juni paßt mir ausgezeichnet` (ACCEPT)
(*Yes Monday the 14th suits me well*)

`wir können jede Zeit nehmen die Ihnen gefällt` (SUGGEST)
(*We could pick any time you want*)

mps1_1_08: `ja morgens um halb elf hab' ich Zeit` (SUGGEST)
(*I have time at half past ten in the morning*)

mhw3_1_09: `also gut treffen wir uns am Montag den vierzehnten Juni um zehn Uhr dreißig zu unserem Termin` (CONFIRM)
(*Ok let us meet on Monday the 14th of June at half past ten*)

...

Figure 4: Example turns – statistical repair

If a dialogue act not compatible with our dialogue model is encountered, the statistical component is looked up in order to find out whether any statistically relevant dialogue acts exists which are able to bridge the previous and the current (incompatible) dialogue act. If such a speech act can be found and if the insertion of this speech act renders the dialogue compatible, a multiple reading is proposed for either the current, or one of the former turns.

Plan Based Repair

The second mechanism uses a set of special *repair operators* which are used when the plan recognizer does not succeed in parsing the next token using the normal plan operators. The simplest case covers the dialogue acts in the subnet in figure 2. The problem with these acts is that they can appear anywhere in the dialogues. One could handle this by adding these dialogue acts to each state in the dialogue model. However, this method is costly in performance and grammar size. We instead process these dialogue acts using the repair operators.

Also, when an input is not admissible by the grammar, and our statistical repair technique has not been able to adjust the input, we repair the tree with this technique.

An Example of Statistical Repair

In this example (see fig. 4) a confirmation (CONFIRM) follows a suggestion (SUGGEST) – a sequence not admissible for the plan recognizer. The trace in fig 5 shows how the recognizer discovers that it can not process the sequence. It consults the statistical component for suggestions to bridge the two dialogue acts. The only suggestion from the statistical component in this example (ACCEPT[1]) is then checked with the surrounding dialogue acts to see which reading to modify. Here the CONFIRM gets an additional reading of ACCEPT.

An Example of Plan Based Repair

We here show a typical example of a clarification dialogue and how the recognizer inserts a clarification dialogue using the repair technique. In the example (see figure 6) the ongoing sub-dialogue is interrupted by a clarification dialogue between the dialogue participants.

Figure 7 shows a screen snapshot from the plan recognizer. It is taken from a system version with the "old" 17 core dialogue acts named in German. Also, the names of the plan operators are truncated to 20 characters[2]. In the figure we see the difference between how the plan recognizer construes the intentional structure for a normal sub-negotiation dialogue SUGGEST (`vorschlag`) – REJECT (`ablehnung`) to the left, and the repaired SUGGEST (`vorschlag`) – CLARIFICATION_QUESTION (`klaerungsfrage`) – CLARIFICATION_ANSWER (`klaerungsantwort`) – REJECT (`ablehnung`) to the right.

[1] The score is the product of the transition probabilities times 1000 between the previous dialogue act, the potential insertion and the current dialogue act.

[2] ...when possible

```
...
Planner: -- Processing ACCEPT
Planner: -- Processing SUGGEST
Planner: -- Processing SUGGEST
Planner: -- Processing CONFIRM
Warning -- Repairing...

Trying to find a dialogue act to
bridge SUGGEST and CONFIRM ...

Possible insertion(s) and
its (their) score(s):
((ACCEPT 98256))

Testing ACCEPT for compatibility
with surrounding dialogue acts...

The current dialogue act CONFIRM has
an additional reading of ACCEPT:

CONFIRM -> ACCEPT CONFIRM !

 Planner: -- Processing
 ...
```

Figure 5: Trace of statistical repair

The repair operator `repair-operator` is inserted and allows for the insertion of the clarification sub-dialogue.

4 TESTING THE MODULE WITH THE CORPUS

In this section we describe the results on testing our component on the corpus. For evaluating the dialogue model and plan recognizer we used 177 hand-annotated dialogues containing 7469 dialogue acts.

Evaluating the dialogue model

To test the coverage of the dialogue model we parsed the above mentioned dialogues with the FSM. In 6633 (91.1 %) cases admissible state changes were encountered. In 836 (8.9 %) cases a non valid sequence of dialogue acts was encountered. However, when trying to use the model to predict the next dialogue act to come, the results is not as good as when using the statistical method (see below).

`turn_2_speaker_b_MW1001': wie wär's denn am Dienstag den dreizehnten April vormittags` (SUGGEST)
(*How about Tuesday the 13th of April in the morning*)

`turn_3_speaker_a_PS1002: tut mir leid ,am dreizehnten April bin ich noch im Urlaub . genauso wie am zwölften April Montag` (REJECT)
(*I'm sorry, but I'm on vacation the 13:th. The same with Monday the 12:th*)

`turn_3_speaker_a_PS1002: ich habe erst wieder ab dem vierzehnten April Zeit.` (SUGGEST)
(*I'm free from the 14th of April*)

`turn_4_speaker_b_MW1003: der vierzehnte is' ein Mittwoch , richtig` (CLARIFICATION_QUESTION)
(*The 14th is a Wednesday, isn't it*)

`turn_5_speaker_a_PS1004: ja genau` (CLARIFICATION_ANSWER)
(*Yes, exactly*)

`turn_5_speaker_a_PS1004: allerdings hab' ich da von neun bis zehn Uhr schon einen Arzt-Termin` (REJECT)
(*I have to see my doctor at ten*)

`turn_5_speaker_a_PS1004: deshalb würde ich vielleicht den Donnerstag vorschlagen` (SUGGEST)
(*Maybe I could propose Thursday*)

Figure 6: Example turns – plan based repair

Pred.	Statistics	FSM
3	2127 (57.25 %)	2069 (55.69 %)
2	1687 (56.53 %)	1427 (38.41 %)
1	1082 (38.79 %)	970 (30.07 %)

Table 1: Predictions without update

Evaluating the plan recognizer

We also tested the plan recognizer with the same 177 dialogues. We got 1249 repairs. 795 of them (63.65 %) are concerned with digressions. Of the remaining 454, 95 could be repaired using statistics, i.e. 7.61 % of all repairs and 20.92 % of the repairs without digressions. It shows, that the repair mechanism plays an important role in the plan processing module. The role of statistical repair covering one fifth of all "real" repairs is important but has to be investigated further.

Evaluating The Prediction Process

To evaluate the prediction process, we took 52 dialogues with 2538 dialogue acts and trained both the FSM and the statistical component. We then

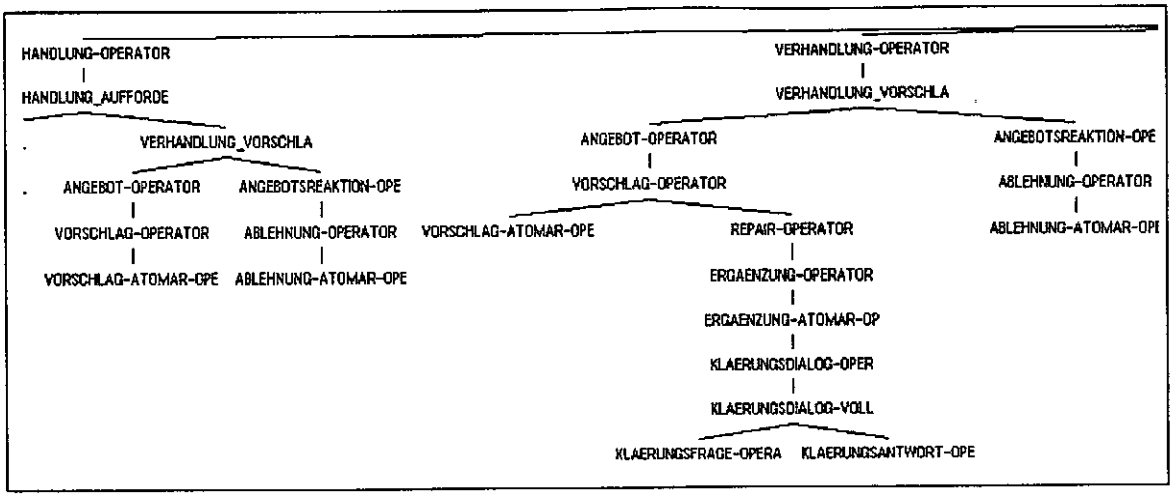


Figure 7: Screen dump – Plan Recognizer

Pred.	Statistics	FSM
3	2572 (69.23 %)	2073 (55.80 %)
2	2100 (56.53 %)	1764 (47.48 %)
1	1441 (38.79 %)	1642 (30.07 %)

Table 2: Predictions with update

took 81 dialogues containing 3715 dialogue acts and tried to predict the next dialogue act to come. Two methods where used. In the first, the predictions where made without update, and in the second with update, i.e. processed dialogue acts are added to the training data. The results are shown in tables 1 and 2. We tested the hit rates for one to three predictions. It can bee seen that the n-gram based statistical method performs better than the FSM, because it is trained on real data and not hand-crafted, and because it is possible to integrate longer histories of dialogue acts by using trigrams. The difference is even more obvious when the two components are allowed to adjust for the new dialogues. For more information about the prediction method and its evaluation, see [15].

5 DISCUSSION AND FUTURE WORK

In this paper we gave an overview of the design process and the inner structure of the dialogue component of VERBMOBIL. One point we want to stress is the importance of a careful analysis of the application environment. It was not possible to simply take the approaches of dialogue processing as used in earlier speech processing systems. Due to the passive, non-controlling character of VERBMOBIL in the scenario, the dialogue structures to be processed can vary unforeseeable. Yet, they have to be processed by the dialogue component.

Our design process for the dialogue component of VERBMOBIL consists of the following steps

1. annotate a corpus
2. extract a "standard" dialogue model from the annotations
3. check the requirements from the other components in the system and identify information needed from the dialogue component
4. select appropriate processing methods, in our case a plan based and a statistical approach which are combined for robustness reasons
5. evaluate the system with real data
6. tune the system, again using real data

Evaluation shows certain deficits in e.g. prediction. We are currently in the process of replacing the prediction module with a re-implementation that delivers up to five percent better prediction results. Still, the prediction process is far from optimal. Since the structure of the dialogue varies a lot [16], we are now testing whether dialogues with similar dialogue structure can be automatically clustered together in different training sets. The idea is to switch between the training data to find the best one for a dialogue.

Our current dialogue model is hand-crafted, which may explain its poor results in the prediction process. To automate the extraction of a dialogue model given an annotated corpus is also topic for further research.

For the plan recognizer we have two main challenging tasks to work on. As mentioned above the input of the dialogue component contains gaps. Extending the plan recognizer to cope with this is a big challenge. The current version also processes the plan operators strictly left to right. In future versions the plan operators will be selected on basis of statistical information collected from a corpus.

References

[1] Francois Andry. Static and dynamic predictions: a method to improve speech understanding in cooperative dialogues. In *Proceedings of the International Conference on Spoken Language Processing*, pages 639–642, Banff, October 1992.

[2] John Austin. *How to do things with words*. Oxford: Clarendon Press, 1962.

[3] Eric Bilange. A task independent oral dialogue model. In *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics (EACL-91)*, pages 83–88, Berlin, Germany, April 1991.

[4] Barbara J. Grosz and Candace L. Sidner. Attention, intentions and the structure of discourse. *Journal of Computational Linguistics*, 12(3), 1986.

[5] Susanne Jekat, Alexandra Klein, Elisabeth Maier, Ilona Maleck, Marion Mast, and J. Joachim Quantz. Dialogue acts in verbmobil. Verbmobil Report 65, Universität Hamburg, DFKI Saarbrücken, Universität Erlangen, TU Berlin, 1995.

[6] Fred Jelinek. Self-Organized Language Modeling for Speech Recognition. In A. Waibel and K.-F. Lee, editors, *Readings in Speech Recognition*, pages 450–506. Morgan Kaufmann, 1990.

[7] Martin Kay, Jean Mark Gawron, and Peter Norvig. *Verbmobil. A Translation System for Face-to-Face Dialog*. Chicago University Press, 1994. CSLI Lecture Notes, Vol. 33.

[8] Klaus Kohler, Gloria Lex, Matthias Pätzold, Michael Scheffers, Adrian Simpson, and Werner Thon. Handbuch zur Datenerhebung und Transliteration in TP14 von Verbmobil - 3.0. Verbmobil Technisches Dokument 11, Universit"at Kiel, 1884.

[9] Elisabeth Maier. Dialogmodellierung in VERBMOBIL – Festlegung der Sprechhandlungen für den Demonstrator. Technical Report Verbmobil–Memo 31, DFKI Saarbrücken, Juli 1994.

[10] M. Mast, R. Kompe, F. Kummert, H. Niemann, and E. Nöth. The dialog module of the speech recognition and dialog system EVAR. In *Proceedings of the International Conference on Spoken Language Processing, Banff, Canada*, pages 1573–1576, 1992.

[11] Mark T. Maybury. *Planning Multisentential English Text Using Communicative Acts*. PhD thesis, University of Cambridge, Cambdridge, GB, 1991.

[12] Johanna Moore. *Participating in Explanatory Dialogues*. The MIT Press, 1994.

[13] Masaaki Nagata and Tsuyoshi Morimoto. An experimental statistical dialogue model to predict the Speech Act Type of the next utterance. In *Proceedings of the International Symposium on Spoken Dialogue (ISSD-93)*, pages 83–86, Waseda University, Tokyo, Japan, November 1993.

[14] Gerhard Th. Niedermair. Linguistic Modelling in the Context of Oral Dialogue. In *Proceedings of International Conference on Spoken Language Processing (ICSLP'92)*, volume 1, pages 635–638, Banff, Canada, 1992.

[15] Norbert Reithinger. Some Experiments in Speech Act Prediction. In *AAAI 95 Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995.

[16] Norbert Reithinger and Elisabeth Maier. Utilizing Statistical Speech Act Processing in VERBMOBIL. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95)*, Cambridge, MA, June 1995. forthcoming.

[17] Bärbel Ripplinger and Folker Caroli. Konzeptbasierte Übersetzung in Verbmobil. Technical report, IAI Saarbrücken, May 1994.

[18] John R. Searle. *Speech Acts*. Cambridge/GB: University Press, 1969.

[19] Marc Vilain. Getting Serious about Parsing Plans: a Grammatical Analysis of Plan Recognition. In *Proceedings of AAAI-90*, pages 190–197, 1990.

[20] Wolfgang Wahlster. Verbmobil–Translation of Face-to-Face Dialogs. Technical report, German Research Centre for Artificial Intelligence (DFKI), 1993. to appear in Proceedings of MT Summit IV, Kobe, Japan, July 1993.

DIALOGUE CONTROL IN AUTOMATIC INQUIRY SYSTEMS

Harald Aust, Martin Oerder

Philips GmbH Forschungslaboratorien
P.O. Box 1980, D-52021 Aachen, Germany
E-mail aust@pfa.philips.de

ABSTRACT

In a flexible, natural-language inquiry system, the dialogue between a caller and the machine can be complex and extensive. Therefore, traditional means for modeling the dialogue flow explicitly would result in very complicated and lengthy descriptions that would not be manageable anymore.

To overcome these difficulties, we have developed a description language specifically for the kind of dialogues that are encountered in automatic inquiry systems. Our idea is to identify those constituents of these dialogues that are common to all of them, independent of the particular application. Then, we represent them in highly separated sections and use an interpreter for controlling the dialogue flow based upon these parts.

This approach, which has been successfully tested in thousands of real-world dialogues, has the additional advantage that only the mere dialogue description must be provided for a new application while the interpreter can remain unchanged.

1. INTRODUCTION

We have developed a system that people can call in order to obtain information on the schedule of the German railway. This system, which is described in more detail elsewhere [1, 7], provides accurate connections between roughly 1200 German cities over the telephone. Callers can talk to it in unrestricted, natural, and fluent speech, very much like they would communicate with a human operator.

The timetable information is stored in a database, and it is the task of the system to extract that information from the users' utterances that it needs for generating an appropriate query. In other words, values for *slots* in the query pattern, like origin, destination, date etc., must be determined from the incoming speech signal. To this end, the system contains a *speech recognizer* and a *speech understanding component* [2], as well as a *speech output module*.

Since it cannot be reasonably expected that all callers always provide all necessary information, and nothing else, in their first sentence, a *dialogue control module* is also needed. It has to decide upon, and construct, appropriate questions as long as the system does not yet have all the information for the database query. The problem is then to find a way to model the ensuing dialogue between the system and the caller, i.e. to give the system the ability to decide whether to access the timetable database or to come up with another question — and, of course, to determine its exact form.

2. DIALOGUE IN AUTOMATIC INQUIRY SYSTEMS

In our opinion, people should be able to use an automatic inquiry system without having to be experts in this field, or having to listen to lengthy explanations first. This means that the dialogue flow must be habitable and natural, and it has to resemble everyday conversation between humans. In particular, we do not want to apply a rigid question-answer scheme to obtain the slot values; instead, the caller should be allowed, at any given moment, to use whatever phrases he prefers. As a consequence, the system must be able to adapt to the user, and not vice versa. For example, if a caller supplies more information in his answer to a question than was explicitly asked for, the system should make use of it; if the question "Where would you like to go to?" is answered by "To Hamburg, tomorrow at nine", the additional information about the desired travel time

must not be discarded just because it did not directly refer to the question asked. Especially, the following question should not be something like "When would you like to go?".

But there are more things to be considered for a system to have a natural and user-friendly appearance. To avoid confusion, no more questions than necessary should be asked, and no more confirmation than needed should be given, either. Also, the sequence of questions must be in an order that appears to be logical and comprehensible for the caller. For example, the dialogue should stay focused on a single subject, and should not jump back and forth between them.

3. GENERAL DIALOGUE FLOW

Our principal approach to satisfy the above requirements is as follows: every incoming sentence is processed by basically the same recognition and understanding components which may, however, be modified in order to account for specific dialogue situations. Such modifications may consist of temporary supplements to the system's inventory of understandable phrases (like expressions for "yes" and "no", where applicable), but also in changes of meaning (the response "Sunday" after the question "When would you like to go?" would normally be processed differently from the same word after "On what day exactly would you like to go at Easter?").

We can distinguish between three different classes of questions:

- *Disambiguation questions* are asked if there is already a candidate value for a slot, but further information is needed for the system to decide upon its exact contents. The most common situations for these questions are apparent ambiguities and contradictions ("Did you say you want to go to Hamburg or to Homburg?"), either caused by the form of the caller's utterance or by recognition or understanding errors, and not sufficiently specified values ("Which Neustadt would you like to go to?").

- *Extension questions* are asked if there are slots that do not yet have a value at all: "Where would you like to go to?"

- *Follow-up questions* are used for the continuation of the dialogue after the results from the database query were read to the caller: "Is there anything else I can do for you?"

After an initial greeting of some kind, the dialogue control module repeatedly checks these three groups for applicable questions according to the following scheme:

- If there is a disambiguation question, ask it.
- Else, if there is an extension question, ask it.
- Else, if there is a follow-up question, ask it.
- Else, the dialogue is finished.

Note that with this procedure, we automatically achieve a question sequence that most people consider logical since individual subjects are dealt with one after the other, and the questions are asked in a hierarchical, increasingly more specific way.

4. DIALOGUE DESCRIPTION WITH A GRAPH

Now that we have defined the desired behavior of our system, the next question to be addressed is how it can actually be made to act in the described way. A straightforward approach, of course, would be to simply write an appropriate program. This would mean, however, that the resulting system would only be a suitable solution for a single problem. In particular, the program would have to be completely rewritten whenever a new application had to be developed. Therefore, we would like to have a more general, and reusable, dialogue description formalism.

A common method for specifying dialogues employs a directed graph whose nodes are labeled with the questions the system can ask, as well as the actions to be taken, and whose edges are marked with transition conditions that indicate which questions can follow each other, and when. Every path through the graph corresponds to a possible sequence of system questions.

While this is a uniform and general approach that has been successfully employed in many systems, e.g. as described in [5] or [6], it is not well-suited for a problem like ours. For a flexible and natural dialogue flow — including, for example, a different formulation whenever a question must be repeated — a very large number of possible questions might be necessary: in our case, more

than 1000. Worse yet, if we want the system to adapt flexibly to the caller's behavior, almost any question can follow almost any other, leading to tens of thousands of transitions. It is obvious that in such a situation the graph representation is not appropriate: the necessary effort for its construction, as well as the inherent potential for errors, would be unacceptably high.

5. DECLARATIVE DIALOGUE DESCRIPTION

We get a first clue on possible alternatives when we think about what the dialogue description graph would look like if it were actually constructed. Let us consider, for example, a node labeled with the question "Where would you like to go to?". There will be many transitions ending at this node, with many different conditions attached. But they all have one thing in common: every condition contains an expression for "the destination is unknown" — otherwise it would not make sense to pose this particular question. So instead of explicitly modeling the transitions, we can simply write down all the questions, together with appropriate preconditions, and leave it to an *interpreter* to decide upon the question to be asked next, depending on the dialogue history and the current situation. With this declarative representation, the complexity of a dialogue description can be greatly reduced: not a huge quantity of transitions, but only the much smaller number of questions and preconditions must be specified.

6. PREDEFINED DIALOGUE STRUCTURE

Of course, the sequence of questions is not the only thing that must be addressed in a dialogue specification. It is just as important — though sometimes neglected in the literature — to define how the caller's answers should be processed and made use of. This is particularly true if, as in our case, verifications and corrections, possibly accompanied by new information in the same utterance, can make for rather complicated responses. Besides, we would like to simplify the dialogue description even further by delegating more tasks to the interpreter than just the selection of the next question.

The approach now is not to strive to model any dialogue that is possible among humans or between a human and a machine, but to limit ourselves to that subset that typically occurs in the context of database inquiries.

The final goal of such a dialogue is, of course, to create a database query. As mentioned earlier, this means that certain values for the slots of the query pattern must be determined. To this end, questions must be asked, answers must be processed, ambiguities must be clarified, results should be verified and so on. This general process is the same for all inquiry dialogues, with only very little variation, and it is completely independent of the subject or the particular application. Furthermore, the order in which these tasks are performed is always the same, too, which is obvious because one has to ask a question before the answer can be processed, and it is only thereafter that ambiguities or other problems can be detected and resolved.

The idea that follows is then to identify the parts a "generic" inquiry dialogue consists of, and to define a separate formalism for each of them, as well as a general strategy that determines how to process, combine, and use these parts. It will be executed by an interpreter.

With this approach, a dialogue need not be fully specified. Only the "holes" in the general pattern must be filled. This results in considerably less programming effort when developing a new application. On the other hand, because of the given structure, the flexibility is limited. In particular, the chosen framework may be inadequate for certain applications. The goal of such a definition would therefore be to provide as much structure as possible, and to maintain as much flexibility as necessary.

7. A DIALOGUE DESCRIPTION LANGUAGE

Based on the ideas presented above, we have developed a special programming language for inquiry dialogues. A program in this language consists of several *sections* in which all those aspects of the dialogue are specified that the interpreter needs to know. This is done in a way as declarative as reasonably possible. The most important of these sections are the following:

- Although the *grammar rules* of the speech understanding module are not really part of dia-

logue control, they affect, of course, the overall behavior of the system. Besides, they are a purely declarative description of an important system component and therefore fit nicely into our scheme.

- Database query *slot definitions* contain rules for the combination of two different slot values (" 'Sunday' and 'the 4th of June' can refer to the same day"), the disambiguation questions, and all other information necessary to describe the properties of a slot. This is done in a declarative way, too.
- The already introduced extension questions make up a third section which is also declarative.
- The *general framework* of the dialogue includes, among other things, initial greeting, database access, and output of the results, as well as the follow-up questions. Note that in this context the dialogue flow actually can be modeled by a graph or a similar procedural means: these questions will typically be asked in a more or less linear way, one after the other, and they will often be answered by just "yes" or "no".

In addition, we defined two *patterns* that can be used in any of the above sections so that the interpreter can automatically handle repeating tasks:

- A pattern for *questions* computes the actual output, for example by inserting slot values into a carrier phrase. It starts the speech recognition, reads and processes the caller's answer, and automatically handles reformulations in case of question repetitions. Furthermore, it can modify the behavior of the speech understanding component in order to account for the dialogue state that led to the question.
- A pattern for *verifications* determines the slots whose values are not yet verified, modifies questions accordingly or creates new ones, and modifies the answer processing with respect to confirmations or corrections.

The block structure of the entire system is depicted in Fig. 1. A dialogue always begins with the procedural definitions of the general framework. After the first question, control is passed to the interpreter automatically, which subsequently selects, creates, poses, and works with all following questions based upon the information in the program's grammar rules, slot definitions, and extension questions, as well as the system status. The latter represents the current state of the dialogue and contains, in particular, the information gathered by the system from the caller so far. When the interpreter does not find any applicable questions anymore, it passes control back to the procedural program part, which creates the database query and continues the dialogue according to its contents.

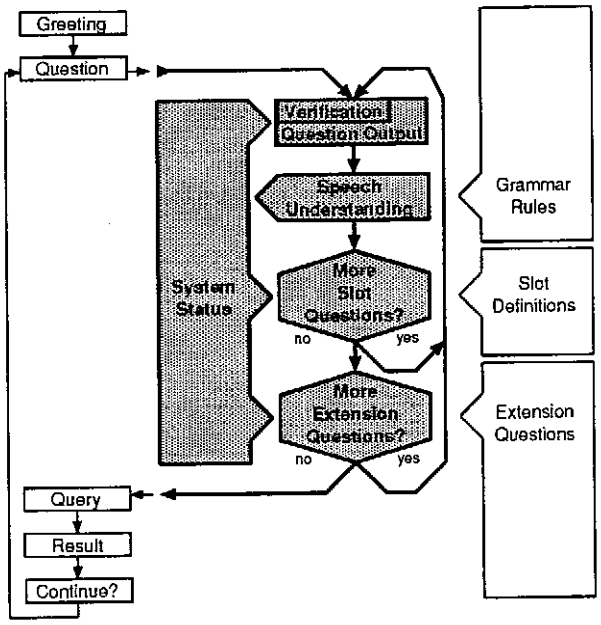


Figure 1: *Block diagram of the dialogue control component of our automatic inquiry system. For new applications, only the constituents of the dialogue description program (white) must be updated, while the central interpreter (shaded) remains unchanged.*

8. CONCLUSION

The methods introduced in this paper are successfully employed in our train timetable information system where the validity of this approach has proven in several thousand successfully completed telephone calls. The dialogue description language itself is still a research subject and therefore not yet in its final form; however, it already allows the easy and straightforward creation of completely new applications like automatic telephone switchboard operators. We will continue our research in order to further improve both the system and the underlying technology.

9. REFERENCES

[1] H.Aust et al.: *Experience with the Philips Automatic Train Timetable Information System.* In *Proceedings of the 2nd Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA94)* pp. 67–72, Kyoto, Japan, 1994.

[2] H.Aust, M.Oerder: *Database Query Generation from Spoken Sentences.* In *Proceedings of the 2nd Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA94)* pp. 141–144, Kyoto, Japan, 1994.

[3] E.Bilange: *A Task Independent Oral Dialogue Model.* In *Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics (EACL 91)* pp. 83–88, Berlin, Germany, 1991.

[4] W.Boogers: *Dialogue Construction by Compilation.* In *Proceedings of the 2nd European Conference on Speech, Communication, and Technology (EUROSPEECH 91)* pp. 853–856, Genova, Italy, 1991.

[5] C. Müller, F.Runge: *Dialogue Design Principles — Key for Usability of Voice Processing.* In *Proceedings of the 3rd European Conference on Speech, Communication, and Technology (EUROSPEECH 93)* pp. 943–946, Berlin, Germany, 1993.

[6] P.B.Nielsen, A.Baekgaard: *Experience with a Dialogue Description Formalism for Realistic Applications.* In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 92)* pp. 719–722, Banff, Canada, 1992.

[7] M.Oerder, H.Aust: *A Realtime Prototype of an Automatic Inquiry System.* In *Proceedings of the International Conference on Spoken Language Processing (ICSLP 94)* pp. 703–706, Yokohama, Japan, 1994.

[8] S.J.Young, C.E.Proctor: *The Design and Implementation of Dialogue Control in Voice Operated Database Inquiry Systems.* In *Computer Speech and Language* 3: pp. 329–353, 1989.

REFERRING TO TOPICS
-a corpus-based study-

Mieke Rats
Institute of Language Technology and Artificial Intelligence
Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

E-mail: M.M.M.Rats@kub.nl

ABSTRACT

An earlier study (Rats [1994]; Rats [1995]) of a corpus of 111 telephone conversations recorded at the information service of Schiphol Airport made clear that one aspect of the way in which dialogue participants control their information exchange is topic management. The result of topic management is that information is exchanged in such a way that the dialogue contains one or more topical chains, sequences of utterances that all communicate information about the same topic. This paper reports a study of the referential expressions that are used to introduce or continue topics in a topical chain. Not only their types, also their informational content is considered. The study ends with some informal guidelines for the use of referential expressions in a topical chain.

INTRODUCTION

Commonly, speakers aim at a coherent and understandable organization of the information exchange of their conversation, which means that they will organize it around one or more topics. This organization is done by a form of dialogue control that is called *topic management* (Bunt [1994]; Rats [1995]). The function of topic management is to regulate the introductions, continuations, and shifts of topics in the dialogue.

Locally at the level of the formulation of the utterance, topic management is part of the process by which the speaker tries to adjust the presentation of the informational content of his utterance to his assumption of the informational state of his dialogue partner. The term *information packaging* (Chafe [1978]; Chafe [1987]; Vallduví [1993]; Vallduví [1994a]; Vallduvi [1994b]) is used to describe the phenomenon at issue here.

Generally, people follow a common strategy while *packing* the information content of their utterances. It is descibed by (Chafe [1987]) p.36, while he was describing spontaneous spoken monologue:

> *The usual technique for presenting information is to choose some concept, typically some referent, as a starting point and then add information about it. As a speaker proceeds to verbalize one focus of consciousness after another, each added piece of information is attached to some other piece that is in some sense already present.*

As the following example illustrates, the same kind of thing happens when two speakers are involved:

**2063

1	I:	Inlichting Schiphol
2	S:	Ja,
3		u spreekt met de Wijl
4		Vlucht KL 550,
5		hoe laat is die gepland?
6	I:	Die wordt nu definitief verwacht om vijf voor twaalf.
7	S:	Vijf voor twaalf?
8	I:	Ja hoor.
9	S:	Oke,
10		bedankt.
11	I:	Tot uw dienst.
12	S:	Dag.
13	I:	Dag.

1	I:	Schiphol Information.
2	S:	Yes,
3		you are speaking with de Wijl.
4		Flight KL 550,
5		for what time is it scheduled?
6	I:	It is now definively expected at five to twelve.
7	S:	Five to twelve?
8	I:	Yes.
9	S:	Okay,
10		thank you.
11	I:	You're welcome.
12	S:	Goodbye.
13	I:	Goodbye.

At the point where the information exchange starts, utterance 4, the information seeker intro-

duces a common starting point *Flight KL 550.* This entity serves as the point of attachment for the question that follows: it is the entity about which *for what time is it scheduled* is asked. The answer of the information service takes up this same entity to provide the asked information, as do the check and the confirmation that follow.

The introduction of a common entity and the continuation of that entity, which constitutes the linkage part of *information packaging* (Chafe [1978]; Chafe [1987]; Vallduví [1993]; Vallduví [1994a]; Vallduví [1994b]) is the work of topic management. It regulates that information is exchanged in such an orderly and understandable way that it is clear for both speaker and hearer what piece of new information is exchanged about what discourse entity. Topic management regulates that the new information in an individual utterance will be connected with an entity introduced in a preceding utterance. If there is no preceding discourse or a new connected discourse fragment has to be opened, it regulates that a new entity will be introduced.

Part of topic management is the choice of referential expressions that will be used to introduce, continue and shift topics. This choice does not only concern the type of expression -the question if f.i. a definite NP or a pronoun should be used-, but also the informational content of the expression. In the choice of the referential expression, the speaker is confronted with the same considerations as with the formulation of a whole utterance: he will have to *pack* it in such a way that the listener is able to evoke the right referent.

This paper will elaborate on this part of topic management. What referential expressions are used to introduce, continue and change topic? The study, which is based on a corpus of 111 telephone conversations recorded at the information service of Schiphol Airport, will lead to some guidelines for the use of referential expressions that refer to topics in naturally occurring information dialogues.

Before we really start with the referential expressions, a summary will be given of the results of a research of the same corpus to the linkage part of topic management (Rats [1994]; Rats [1995]). This will be the framework for the work about referential expressions.

1 TOPICS AS LINKS

In (Rats [1994]; Rats [1995]), the method for finding the principles behind the linkage part of topic management was bottom up. First, a syntax-based rule was found for determining the topic of an individual utterance. Then, the topic-comment structures of the individual utterances were combined to find the topical structure of a whole dialogue.

The guidelines behind the rule for finding the topic of an utterance were the following intuitive definitions of topic and comment (Compare (Gundel [1985]) p.86):

> "An entity, T, is the topic of an utterance, U, iff U is intended to increase the addressee's knowledge about, request information about or otherwise get the addressee to act with respect to T."
>
> An information unit, C, is the comment of an utterance, U, if C is "...what is actually communicated, i.e., asserted, questioned with respect to the topic."

How these definitions should work is illustrated by applying them to the information exchange of example dialogue mentioned on page 1.

In utterance 4 the topic is introduced: Flight KL 550. In utterance 5 information is asked about it:

> *hoe laat is die gepland?*
> *(for what time is it scheduled?)*

The topic of this utterance, the entity about which the information is asked is represented by *it.* The rest of the utterance *for what time is _ scheduled?* represents the information that asked about it, the comment.

Utterance 6 provides the asked information:

> *Die wordt nu definitief verwacht om vijf voor twaalf.*
> *(It is now definively expected at five before twelve.)*

The topic of this utterance, the entity about which the information is provided is again represented by *it.* The comment, the information that asked about it is represented by the rest of the utterance *_ is now definively expected at five before twelve..*

Utterance 7 checks the information simply by repeating part of utterance 6.

Vijf voor twaalf?
(Five before twelve?)

The topic and also a large part of the comment is deleted. Only the newest information of the comment of the preceeding utterance is expressed. Nevertheless, the topic the information is checked about is still the same: Flight KL 550.

The result of this topic-comment analysis for the individual utterances leads to the following topical analysis for the whole information exchange:

**2063

	Topic	
hoe laat is	Vlucht KL 550, die Die	gepland? wordt nu definitief verwacht om vijf voor twaalf. Vijf voor twaalf?
Ja hoor.		
for what time is	Flight KL 550, it It	scheduled? is now definitively expected at five before twelve. Five before twelve?
Yes.		

A close look at the syntactic function of the NPs that represent the topics tells us that the topic of utterances with a normal word order is generally represented by the subject. Application of this rule to the clauses with normal word order in the whole corpus confirms this conclusion. The same rule is suggested for English, which is a language of the same family as Dutch (Gundel [1985]; Reinhart [1981]): subject position is topic position in the unmarked case.

This rule for simple sentences can be extended to complex sentences in the following way: split the sentences in their simple clauses and attribute to each clause a topic according to the rule for simple utterances (Rats [1995]).

Also complex NP's will be split. Modifications as relative clauses, appositions, the second PP after another one, will be seen as an extra comment on the topic. The reason for this is that in the corpus, very complex NP's are used to introduce a topic into the dialogue. These NP's cannot be seen as a part of one simple topic comment structure, because each extra modifying phrase really intends to give extra information about the topical referent. It is intuively more plausible to take these extra informative phrases as extra comments on the topic. See the example below:

**4258

	Topic	
Weet	u	ook
of	het toestel dat eh.. dat	van de AL Italia is vertrokken uit Rome om tien over tien
of	dat	al binnen is?
Do	you	know
if	the plane that uh.. that	of the AL Italia has departed from Rome at ten past ten
if	that	has already arrived?

Utterance 4 of the dialogue fragment about flight KL 550 is an example of a special syntactic structure that has a great impact on the topical structure of the dialogue: a left-dislocation. The dialogue fragment illustrates the effect of a left-dislocation construction. The left-dislocated NP introduces a discourse-new topic into the dialogue, that stays the topic of the utterances that follow at least until the end of the information exchange.

The same kind of effect can be expected from right-dislocation and topic topicalization. A speaker that uses one of these *marked* constructions shows explicitly what he intends to communicate about (the topic) and what he is communicating, i.e. asserted, requested and so on, about it (comment) (Gundel [1985]; Gundel [1988]). These explicit topic-comment structures have a very radical influence on the topical structure of a dialogue. They are always used at those places in the dialogue where the fronted topic isn't the current topic. Either they introduce a discourse-new topic into the dialogue or they bring about a topic shift to an entity that wasn't topic in the utterance(s) before (Rats [1994]; Rats [1995]).

In a dialogue, the topic-comment structures of the individual utterances are connected via topics to form a topic-comment structure of a whole dialogue. The result is generally a topical structure with one or more topical chains. On the next page a topic-comment analysis is given for dialogue of page 1. The analysis ilustrates how topics form a connecting thread in a dialogue.

Of course, not all dialogues have such a clear and linear topical structure. A main topical line can be interupted by temporary and permanent topic shifts to entities, that were non-topical in the preceeding utterance (Rats [1994]; Rats [1995]). Nevertheless, in all dialogues coherence

4		Vlucht KL 550,	T_1 ↓		
5		hoe laat is die gepland?	T_1 ↓	-	C_1
6	I:	Die wordt nu definitief verwacht om vijf voor twaalf.	T_1 ↓	-	C_2
7	S:	Vijf voor twaalf?	T_1	-	C_2
8	I:	Ja hoor.			
4		Flight KL 550,	T_1 ↓		
5		for what time is it scheduled?	T_1 ↓	-	C_1
6	I:	It is now definively expected at five to twelve.	T_1 ↓	-	C_2
7	S:	Five to twelve?	T_1	-	C_2
8	I:	Yes.			

is preserved by reducing too many topic fading topic shifts and by a general preference for topic continuity (Rats [1995]).

Most dialogues in the corpus are relatively short, so that all utterances are locally connected in the above mentioned way. For longer dialogues however, it is important to define *topic breaks*. Topic breaks are shifts to entities that cannot be directly related to the preceeding utterance. At a topic break, a connected chain ends and a new one is started (Rats [1995]).

As a result, topical chains have three kinds of origins: either they are started at the beginning of the information exchange as the topic of the first utterance, or they are started farther in the dialogue at a non-topic place of an utterance, or they start at the beginning of a topic break. In what follows, we will investigate the referential expressions that are used to start and continue all topical chains that appear in the corpus.

2 TOPICAL CHAINS

Table 1 in the appendix of this paper gives a general overview of the types of referential expressions that are used to introduce and continue topics in the corpus. It shows by which referential expressions topics are referred to at what place in the referential chain. The crossing of the first row and the first column for example shows how often a definite NP is used for the first reference to a topic.

It is generally assumed that discourse entities that are introduced for the first time into the discourse are introduced by full expressions, i.e., expressions that contain, according to the speaker's estimation, the information needed to be able to identify the referent. Continuing reference is usually done by anaphoric expressions which presuppose the information given in the first reference and therefore contain less information.

We see that the referential chains in the corpus confirm this general view. The first references are done by referential expressions that contain full information about the referent. The definite NP is most popular. For the second reference the use of full referential expressions decreases considerably, and pronouns and ellipsis become more popular. The demonstrative pronoun is most popular. After the second reference the decrease of full referential expressions continues and ellipsis becomes the most important reference form.

We see that reference by full referential expressions happens mostly at the beginning. Then demonstrative pronouns, personal pronouns and ellipsis follow, whereby demonstratives are used during the shortest and ellipsis during the longest time. The use of personal pronoun lies in between.

The following table gives us information about the length of the referential chains:

Table 2: Length of the referential chains for topics in information dialogues:

Length	Number of chains
1	32
2	121
3	72
4	51
5	23
6	13
7	13
8	11
9	8
10	5
11	2
12	2
13	1
14	1
15	0
16	0
17	0
18	0
19	0
20	1

We see that most referential chains are two references long, but a substantial number of chains reaches lengthes up to four or five references.

The interesting point is now to get to know something about the informational content of the expressions used. What information do they contain? How do speakers present a topic that is in some sense considered to be already given for the dialogue partner? And what do speakers do if the referent is considered to be completely new?

The answer to these questions will start with an explanation of how people express given and new information in utterances. Then the focus will be directed on the informational status of the topics themselves.

3 TOPICS AND THE GIVEN-NEW CONTRACT

Generally, people construct their message in such a way that they start with some context setting information and end with the information they really want to communicate. In linguistic literature (Chafe [1978]; Chafe [1987]; Clark & Haviland [1977]; Halliday [1985]), this way of information packaging is related with the notions given and new information.

(Clark & Haviland [1977]) argue that the construction of an utterance is done according to an implicit agreement between language users, which they call the *given-new contract.* It is described in the following way (Clark & Haviland [1977]) p.4:

> *The speaker tries, to the best of his ability, to make the structure of his utterances congruent with his knowledge of the listener's mental world. He agrees to convey information he thinks the listener already knows as given information and to convey information he thinks the listener doesn't yet know as new information. The listener, for his part, agrees to interpret all utterances in the same light. The result is what we have called the given-new contract...*

It is commonly assumed that the given-new distinction in utterances is generally expressed by word order a.o. (Halliday [1985]; Quirk et al. [1985]). In the most idealized form, an utterance starts with a given element and ends with a new element. This idealized form is not always present. There are two reasons for that. One is that discourse has to start somewhere, so there can be discourse-initiating utterances consisting of new elements only. The other is that by its very nature the given can be left out in cases where it refers to something already present in the verbal or non-verbal context. This means that structurally, an utterance consists of an obligatory new element plus an optional given.

Linguistic literature is not so clear about the exact meaning of the notions *given* and *new.* Most linguists agree on the global description of *given.* Given information is information that the speaker takes for granted as part of the background of the conversation. Other ways to formulate the same thing is: given information is taken by the speaker to be in the *common ground,* the *common knowledge* or the *mutual knowledge* of the participants in the conversation. But when it has to come to a more strict definition, linguists assume several ways in which a sentence element can become part of the *common ground.* These are (Clark & Marshall [1981]; Prince [1981]):

1. the information is perceived in the utterance situation, or
2. the information was introduced in the preceding discourse, or
3. the information belongs to the shared background knowledge of the discourse participants, or
4. the information is inferrable from or associated with information that is available in one of the first three ways.

But when applied to concrete examples, it is clear that these different kinds of giveness do not have the same weight. See for example the next dialogue fragment:

**5322

5 S: Ik wou van u een inlichting
over de KL 775 van 26 februari.

5 S: I would like from you information
about the KL 775 from 26 february.

The starting point of this utterance is not problematic. It is the speaker himself, who is given by the utterance situation. Problematic is the end of the utterance. Here an entity is introduced by a definite NP, which suggests that the speaker assumes that the listener has previous knowledge about it. Nevertheless, as it occurs at the end of the utterance, it is presented as new information.

A solution could be found in making a strict distinction between information that is available in the speaker's mind and information that is available in the utterance situation or via the discourse. In fact, the last two sources provide information that is directly accessible for both discourse participants, which is not true for the first one. So if it is assumed that the participants aim at a common basis for integration of new information, it is plausible that a speaker would rather start with information that is inevitably shared

than with information of which that is less certain.

Crucial in understanding of what given information is, is then the point that the information is recoverable through the actual presence of that information in the context at the time of the utterance. By context we mean the complete discourse record of the discourse at a given point, including both linguistic and situational information. So if the information seeker introduces a new entity by saying *the KL 775 from 26 februari*, he doesn't assume that the information service has no previous knowledge of that flight. The speaker considered *the KL 775 from 26 februari* as new information because it wasn't introduced earlier in the dialogue.

It seems natural to associate the topic-comment distinction with the given-new distinction, to associate the topic with given and the comment with new information. However, these two distinctions do not denote the same thing. It is true that the topic is very often (part of) the given information of an utterance. Nevertheles, there are utterances in which it is not. These utterances introduce a discourse-new topic into the discourse. See for example the dialogue fragment below:

**3126

6 S: gisteravond om negen uur
is een vliegtuig vertrokken naar
de eh.. Antillen.
7 Kunt u me zeggen
8 waar die nu zit?

6 S: Yesterday at nine o'clock
a plane departed to
the uh.. Antillen.
7 Can you tell me
8 where it is now?

Utterance 6 of this fragment contains only new information. It introduces a new topic into the dialogue and at the same time it predicates something about that topic. The same holds for utterance 4 of the next example:

**5193

3 S: Kunt u mij zeggen
4 of het vliegtuig uit eh Los Angeles
op tijd is?
5 I: Ja hoor,
6 die komt om tien over drie binnen.

3 S: Can you tell me
4 if the plane from uh Los Angeles
is in time ?
5 I: yes,
6 it will arrive at ten past three.

It introduces a new topic into the dialogue and at the same time it asks some information about that topic. We already saw that speakers also split the two acts, by first expressing the topic introducing act and then the informative act. The function of the first is simply to introduce the topic of the utterance(s) that follow(s). See again utterance 4 of the next dialogue fragment:

**2063

4 S: Vlucht KL 550,
5 hoe laat is die gepland?
6 I: Die wordt nu definitief verwacht
om vijf voor twaalf.

4 S: Flight KL 550,
5 for what time is it scheduled?
6 I: It is definitively expected
at five past twelve.

We see in all these examples with a topic introducing utterance that the topic of the utterance that follows is part of the given information of that utterance. Topics become part of the given information in an utterance when that utterance continues the topic that was introduced in a preceding utterance. Topic introducing utterances appear at those place in a dialogue where a discouse-new topic is introduced: at the beginning of the dialogue exchange and the start of a topic break.

4 TOPIC INTRODUCTION

We see that in the examples in the preceding section, topics are introduced by different kinds of referential expressions: *a plane*, *the plane from Los Angeles*, and *Flight KL 550*. Each of these expressions, present the topic in a different way and as such require different strategies for evoking the right referent. Depending on the information contained in the NP that is used to introduce it, three kinds of newness will be distinguished (Prince(1981), p.235-236):

1. *Unused new* , in which case the NP is a proper name or an uniquely identifying description. The speaker assumes that the hearer has previous knowledge about the referent. By introducing it, he expects that the hearer will recollect it from his background knowledge with the result that it will become part of the dialogue context. An example of such an NP is *Flight KL 550*.

2. *Brand-new* , which means that on the basis of the information contained in the NP the

hearer cannot find the unique entity it refers to. The speaker has to learn about it by the information exchange itself. There are two ways in which an entity can be brand-new:

(a) *Anchored brand-new*, if the NP representing it is linked by means of another NP or "anchor" contained in it to another entity, that is unused new or situationally evoked. An example of such a NP is *the plane from Los Angeles.*

(b) *Unanchored brand-new*, if there is no anchor, for example *a plane.*

In the context of the telephone conversations at Schiphol Airport, the NPs that at the first sight seem to introduce an anchored brand-new entity into the discourse, are most of the time not as new as they look like. Usually, the information service replies such that it is clear that she knows what referent was meant by the information seeker. This means that the dialogue context generally contains enough extra information to find the unique referent. So in the context of the Schiphol corpus anchored brand-new entities have normally the same kind of impact as unused new entities.

Another speciality of the dialogues in the corpus is that there is a large preference to introduce the topic with rather complex NP's. Even when speakers use a proper name, they still add one or more modifications. It seems as if they want to add more modifications to the head noun to be sure that they achieve a unique reference. The complex NP's are constructed by attaching one or more PP's, one or more relative clauses, one or more appositions, by forming noun-noun combinations, or by using a combination of these possible modifications. See for example the example below:

**5201

3 S: Ik wilde vragen
4 hoe laat toestel eh...
van Aria,
5 vluchtnummer RJ 261,
6 aankomt.

3 S: I wanted to ask
4 at what time the plane uh...
of Aria,
5 flightnumber RJ 261,
6 will arrive.

Here the head noun is modified by a PP and an apposition.

Now we have a categorization of newness for discourse-new topics we can relate the newness of the topic to the place in the utterance where it is introduced. The examples in the preceding section made already clear that topics can be introduced both near the beginning and towards the end of the utterance. If introduced near the beginning, they take a subject place or are fronted. If introduced towards the end they are part of the comment. This means that topics are introduced both at the giveness and the newness side of the utterance. We may expect now that the referential expressions used to introduce a topic near the beginning of utterances refer to unused new or anchored brand new entities, while NP's used towards the end of the utterance refer to brand new entities. And as definite NP's are generally considered to refer to known entities and indefinite NP's to unknown entities, it is plausible to assume that the NP's used in the beginning will be definite NP's and those used towards the end will be indefinites.

To the assumptions we already had we can also add the assumption that probably more complex NP's will be used at the beginning of an utterance and fewer towards the end.

The next table shows what NP's are used when they are introduced in the beginning of an utterance:

Table 3: Reference forms for topic introductions at the beginning of an utterance:

Defnp	Indefnp	Demnp	Qnp
75	11	1	4

We see that definite NP's are preferred. So this confirms our expectation. The table below shows the giveness status of the entities the utterance-initial NP's represent:

Table 4: Giveness status of the NP's used to introduce topics at the beginning of the utterance[1]:

	UN	ABN	UBN
Defnp	42	28	5
Indefnp	-	4	7
Demnp	-	1	-
Qnp	-	3	1
Total	42	36	13

Unused new and anchored brand new entities are preferred. This also confirms the expectation.

The table below shows the complexity of the NP's.

Table 5: Complexity of the NP's used to introduce topics

[1] UN stands for *Unused new*, ABN stands for *Anchored brand new*, and UBN stands for *Unanchored brand new*

at the beginning of the utterance[2]:

Simple	Complex
35	56

		Total	SE	UN
	PP	29	-	29
	RC	3	1	2
	NN	6	3	3
	A	3	-	3
	C	15	-	15

We see that complex NP's are preferred over simple NP's and that within the group of complex NP's the NP's linked to unused new entities are preferred. One side of the hypothesis is confirmed. We now have to look for evidence for the other side.

Table 6 below shows the full referential expressions that introduce a topic towards the end of an utterance:

Table 6: Reference forms for topics introduced towards the end of an utterance:

Defnp	Indefnp	Demnp	Qnp
31	35	1	2

We see that although indefinite NP's are preferred, there is not a very big difference with definite NP's. Nevertheless, compared with the topics introduced at the beginning of an utterance there are far more indefinite NP's here.

The table below shows the giveness status of the NP's used to introduce the topic in the comment part of an utterance:

Table 7: Giveness status of topics introduced towards the end of an utterance:

	UN	ABN	UBN
Defnp	16	10	5
Indefnp	-	5	30
Demnp	-	1	-
Qnp	-	-	2
Total	16	16	37

We see that unachored brand new entities are preferred over unused new and anchored brand new entities, which is in agreement with the assumption made in the beginning of this section. But if we count the unused new and the anchored brand new entities together the difference is not that large. Nevertheless as compared with the topics introduced at the beginning of the utterance there are more unanchored brand-new entities now.

The table below presents the complexity of the NP's used to introduce a topic towards the end of an utterance:

[2]SE stands for *Situationally evoked entity*, PP stands for *Prepositional phrase*, RC stands for *Relative clause*, NN stands for *Noun-noun combination*, A stands for *Apposition*, and C stands for *Combination* .

Table 8: Complexity of the NP's used to introduce topics towards the end of the utterance:

Simple	Complex
13	19

		Total	SE	UN
	PP	10	1	9
	RC	3	2	1
	NN	2	1	1
	C	4	-	4

Again we see that there are more complex NP's that simple ones, although their difference is not as large as for utterance initial topic introductions. Compared with the topics introduced at the beginning of the utterance there are more simple NP's towards the end of an utterance.

We can conclude that a topic that is introduced near the beginning of an utterance is preferrable introduced by a complex definite NP. It introduces preferrable an unused new entity or an entity that is connected with a unused new entity. When a topic is introduced towards the end of an utterance indefinite NPs, that introduce unanchored brand-new entities are slightly more preferred, although complex definite NPs, that introduce an unused-new entity or an entity connected to an unused-new entity do not really lag behind.

5 TOPIC CONTINUITY

The preceding sections made clear that topics that are introduced for the first time into the discourse are introduced by full expressions, i.e., expressions that contain, according to the speaker's estimation, the information needed to be able to evoke the intended referent. It is clear that after the establishment of the referent less informational expressions will do. In fact, continuing reference is usually done by anaphoric expressions, - i.e. expressions which presuppose the information given in the first reference and therefore contain less information- and even ellipsis.

Table 9 below, gives an impression of the distribution of ellipsis and the various kinds of anaphors in the corpus.

Table 9: Topic continuations

Total	661	100%
Identity anaphora	295	45%
Subsectional anaphora	27	4%
Relational anaphora	27	4%
Ellipsis	312	47%

It shows that more than half (53%) of the

topic continuations are lexicalized and almost half (47%) are ellided. The greatest part (85%) of the anaphors are identity anaphors. It is clear that ellipsis and identity anaphora are the most important means to continue a topic.

The notion of anaphora is taken very broad: it comprises identity anaphora, subsectional anaphora and relational anaphora (Deemter [1991]). In the subsections that follow, I will elaborate on each of these.

5.1 IDENTITY ANAPHORS

Identity anaphora are the most well known anaphora: they refer to the same entity as their antecedent. Because of that, the meaning or reference of an identity anaphor is supposed to be already established in the preceeding discourse, so generally only a small reference is needed to refer to the same entity again. This gives us a reason to suppose that more pronouns than full expressions are used to continue the topic. In fact, table 10 below confirms that:

Table 10: Identity anaphors used in topic continuations:

Pronoun	243	82%		
Rel. p			8	3%
Pers. p			90	37%
Dem. p			145	60%
Full exp.	52	18%		
Defnp			38	73%
Indefnp			3	6%
Demnp			4	8%
Qnp			2	4%
Prop			5	9%

82% of the identity anaphors are pronouns and only 18% are full expressions. We see that demonstrative pronouns are the most popular pronouns used; 60% of the pronouns are demonstrative pronouns.

Another question we could ask ourselves is whether it is possible to say something more about the way in which people proceed in more than one topic continuation. Is there a certain preferred order in which the different expressions are used? The next table is an attempt to give an answer to such a question. Each row shows the number of times the expression leading the row followed each of the expressions in the columns.

Table 11: Order of referential expressions in a topical chain:

	Defnp	Indefnp	Demnp	Qnp	Prop
Defnp	42	3	2	1	-
Indefnp	1	-	-	2	-
Demnp	4	4	2	1	-
Qnp	-	-	-	2	-
Prop	-	-	-	-	5
Rel.p	7	8	1	1	-
Dem.p	78	20	7	11	78
Pers.p	33	7	4	-	7
Ellipsis	19	8	4	9	10

	Rel.p	Dem.p	Pers.p	Ellipsis
Defnp	-	9	6	7
Indefnp	-	2	1	-
Demnp	-	2	1	4
Qnp	-	3	1	1
Prop	-	5	-	3
Rel.p	-	-	-	-
Dem.p	7	58	13	35
Pers. p	3	23	47	27
Ellipsis	-	83	39	135

We see that after each kind of full expression, a demonstrative pronoun is preferred. This changes however when the pronouns come into play: after a demonstrative pronoun ellipsis is preferred, after a personal pronoun a personal pronoun, and after ellipsis ellipsis.

We see that although demonstrative pronouns are in the majority, also a lot of definite NP's are followed by a definite NP. There are even more definite NP's than personal pronouns. Table 1 in the appendix reveals that most of them are used in the first references to the topic, which supports the assumption that they play a role in the identification phase, the phase in which the dialogue participants establish the topic.

These anaphors can have the following kind of informational content:

1. the anaphoric NP is a complete or partial lexical copy of the antecedent NP, e.g. *the KL 364 from Madrid* and *364*,

2. the anaphoric NP contains information that was not included in the antecedent. There are three possibilities:

   (a) the anaphoric NP and the antecedent NP belong to the same frame of reference, e.g. *The plane from London* and *the KL 128*,

   (b) the anaphoric NP partially reproduces the antecedent NP and adds restrictive information, e.g. *a flight from London* and *the flight of ten past one.*

   (c) the anaphoric NP simply contains different information about the antecedent, e.g. *the EL AL* and *the LY 337.*

3. the anaphoric NP is a synonym of the antecedent NP, e.g. *that man* and *that mister.* Also included in this category will be the case

in which the anaphoric NP defines a superordinate category of which the antecedent NP is a member, e.g. *The VA 587* and *that flight.*

In general, by uttering a complete or partial copy of the topic of the preceeding utterances, speakers show or express linguistic feedback. They show that they are or have been processing the expression, like in utterance 5 of the dialogue fragment below:

**4999

4	S:	Klopt het dat vandaag vanuit Thailand vlucht BR 722 aankomt?
5	I:	BR 722?
6		Een ogenblikje.
7	S:	Ja
8	I:	Ja,
9		die is al geland,
10		kwart voor een

4	S:	Is it true that today from Thailand flight BR 722 will arrive?
5	I:	BR 722?
6		One moment.
7	S:	Yes
8	I:	Yes,
9		that has already landed,
10		a quarter before one

or they check if they perceived or understood the previous expression well, as in utterance 6 of the dialogue below:

**4093

5		wanneer de KL 602 aankomt?
6	I:	De KL 602?
7	S:	Ja.
8	I:	Ogenblik graag.
9	S:	Ja
10	I:	Even kijken hoor,
11		die wordt om ... tien over drie op tijd verwacht.

5		when will the KL 602 arrive?
6	I:	The KL 602?
7	S:	Yes.
8	I:	Moment please.
9	S:	Yes
10	I:	Let's see,
11		it is expexted on time at ... ten past three.

NP's that contain extra information also have a special function in the dialogue. They are information providing, information requesting or information correcting, depending on the communicative function of the utterance. See for example the dialogue fragment below:

**4065

3	S:	Het vliegtuig S 4 20
4		dat naar Singapore gaat,
5		dat zou om drie uur vertrekken,
6		heeft dat soms nog vertraging?
7	I:	De SQ vieren... eh 23?
8	S:	20,
9		de SQ 20.
10	I:	Naar Singapore he?
11	S:	Naar Singapore.
12	I:	En die zou om drie uur vertrekken, dacht u?

3	S:	The plane S 4 20
4		that will go to Singapore,
5		that would depart at three o'clock,
6		is that still delayed?
7	I:	The SQ four... uh 23?
8	S:	20,
9		the SQ 20.
10	I:	To Singapore?
11	S:	To Singapore.
12	I:	And that would depart at three o'clock?

In utterance 7, the NP contains other information than the NP that is supposed to be its antecedent (in utterance 3). The function of the NP is to check the information given in that antecedent. The NP's in utterances 8 and 9 correct the information content of the NP given in utterance 7. We see that the NP's do not only refer, they also communicate information.

Synonyms and more common terms than those that were used to introduce the entity into the discourse do not really have an extra function in the information exchange because their use rests on general linguistic knowledge everyone has about meanings of referential expressions. Common terms can even be compared with pronouns because they contain in fact less specific information about the referent.

Table 12 below demonstrates the distribution of the different anaphorically definite NP's in the corpus:

Table 12: Informational content of the full referential expressions used in topic continuations:

	Extra information	Synonym	Lexical copy
Defnp	20	3	15
Indefnp	-	1	2
Demnp	-	1	3
Qnp	-	-	2
Prop	2	-	3
Total	22	5	25

We see that lexical copies and NP's that contain extra information are preferred, which means that most anaphorically definite NP's play a role in the identification of the topic.

If we want to formulate a general guideline for the use of referential expressions for topic continuity we could state it as follows:

(Full expression(s)) → Demonstrative pronoun → Ellipsis → Ellipsis

In the phase where the topic isn't established yet, one or more full expressions are used, which either repeat (part) of the information of the introducing expression or contain extra information. After the establishment of the topic the demonstrative pronoun is preferred for the next reference. For the references that follow after the use of the demonstrative pronoun, ellipsis is preferred.

5.2 SUBSECTIONAL ANAPHORS

A subsectional anaphor refers to a subsumption (a subset of a set, a part of a quantity, a substructure of a given structure) of its antecedent (Deemter [1991]). In the following dialogue fragment *the first flight* is a subsectional anaphor of *some flights to Paris.*

**7011

5	S:	Ik heb hier een meneer
6		die wil weten
7		*enkele vluchten naar Parijs*
8		voor morgenochtend,
9		hoe laat die vertrekken.
10	I:	Morgenochtend,
11		momentje graag.
12	S:	Ja
13	I:	Nou,
14		*de eerste vlucht* vertrekt om
		zeven uur dertig...
15	S:	Ja
16	I:	en *de tweede vlucht* om
		acht uur vijfenveertig...
17	S:	Ja
18	I:	en dan...
19		even kijken...
20		om negen uur dertig gaat er *een*...
21	S:	Ja
22	I:	dan ∅ dertien uur vijf,

5	S:	Here is a mister
6		who wants to know
7		*some flights to Paris*
8		for tomorrow morning,
9		at what time they will leave.
10	I:	Tomorrow morning,
11		moment please.
12	S:	Yes
13	I:	Well,
14		*the first flight* will leave at
		seven thirty...
15	S:	Ja
16	I:	and *the second flight* at
		eight forty-five...
17	S:	Yes
18	I:	and then...
19		let's see...
20		at nine thirty there is *one*...
21	S:	Yes
22	I:	then ∅ five past one

Also elements of the same set will be understood as subsectional anaphors, e.g. *the first flight* and *the second flight.* Strictly speaking, these kinds of anaphors involve a shift to another entity. However, the entity the anaphor refers to cannot be found on the basis of its linguistic form alone. We need the antecedent to find the right reference. The referent of *the first flight* for example can only be found by looking for it within the reference of *some flights to Paris.* From this we can argue that although a subsectional anaphor introduces a new (sub)topic, it is so related to the previous one that it continues the main topic and as such continues the topical chain in the dialogue (Compare (Grosz [1977]; Grosz [1981]; Sidner [1979]; Sidner [1981])).

This special relation between subtopics and their heads leads to a special behavior of the referential expressions that are used to refer to them. One aspect of that behavior is that the general rule that entities, introduced into the discourse for the first time will be introduced by full referential expressions doesn't really hold for subsectional anaphors. As the dialogue fragment about some flights to Paris shows, a sequence of shifts from one entity to another exhibits a decrease of informational content from one anaphor to the other. It is a decrease from two full NP's, via a partly elliptical quantified NP to ellipsis. Table 13 below shows that more subsectional entities are "introduced" by ellipsis.

Table 13: Referential expressions as subsectional anaphor

Defnp	Indefnp	Demnp	Qnp	Prop
7	2	-	8	7

Dem.p	Pers.p	Rel.p	Ellipsis
-	-	-	3

Also the view that reintroduced entities are introduced by full referential expressions doesn't hold for subsectional anaphors. The dialogue fragment below illustrates that a sequence of shifts from one entity to another is interrupted by a reference to the orgininal entity (utterance 16). The reference to the original entity *Flights that depart between twelve and three for Frankfurt* is even ellided, which means that it is so highly accessible, that it doesn't have to be introduced by whatever referential expression.

**2004

1	S:	Kunt u mij misschien ook vertellen
2		welke *vliegtuigen* er tussen
		twaalf uur en drie uur naar eh...
		Frankfurt vertrekken?
3	I:	Een ogenblikje graag.
4	I:	Nou
5		twaalf uur vijfenvijftig gaat *de KLM*...
6	S:	Ja
7	I:	*De KL 243*...
8		en om 13 uur de gaat *de Garuda*...
9	S:	Ja
10	I:	en *die* maakt zijn eerste tussenstop in
		Frankfurt,
11	S:	Ja
12	I:	*de GA 895*,
13	S:	*De GA 895*,
14		ja.
15	I:	en...,
16		∅ tussen twaalf en drie zei u he?
17	S:	Ja.
18	I:	Ja,
19		en er gaat er nog *een* dertien uur
		dertig van de Turkish Airlines.
20	S:	∅ Turkish Airlines?
21	I:	Ja.
22	S:	O.
23	I:	En er gaat er nog precies om drie
		uur *eentje* van de Phillipan Airlines
24	S:	∅ Phillipan Airlines?
25	I:	Ja.

1	S:	Can you tell me
2		which *planes* will leave for Frankfurt
		between twelve and three o'clock?
3	I:	A moment, please.
4	I:	Well
5		twelve fifty-five *the KLM* will leave...
6	S:	Yes
7	I:	*The KL 243*...
8		and at one am en *the Garuda* will leave...
9	S:	Yes
10	I:	and *it* will make its first intermediate
		stop in Frankfurt,
11	S:	Yes
12	I:	*the GA 895*,
13	S:	*The GA 895*,
14		yes.
15	I:	and...,
16		∅ between twelve and three you said,
		didn't you?
17	S:	Yes.
18	I:	Yes,
19		and there is another *one* at thirteen
		thirty of the Turkish Airlines.
20	S:	∅ Turkish Airlines?
21	I:	Yes.
22	S:	O.
23	I:	And there is *one* more at exactly three
		o'clock of the Phillipan Airlines.
24	S:	∅ Phillipan Airlines?
25	I:	Yes.

The dialogue illustrates too that a sequence of shifts from one entity to another can easily be interrupted by a sequence of identity anaphors that refer to one of the sub-entities (utterances 7, 10, 12, 13, 20 and 24). It doesn't influence the decrease in information content of the subsectional anaphors. This special behavior confirms the assumption that subsectional anaphors continue the topic of the dialogue.

A general guideline for the use of subsectional anaphors is difficult to make because we do not have enough instances of subsectional anaphors (see table 9 at the beginning of this section) in the corpus. Nevertheless, on the basis of the facts shown here we can imagine the form it could take: Use in the first instances a full expression that contains just enough information to distinguish the referent from the rest of the group. Then change to less informational anaphors like *one* and finish with ellipsis. Of course this guideline should be tested on a corpus that contains more subsectional anaphors.

5.3 RELATIONAL ANAPHORS

A relational anaphor is cognitively associated with its antecedent. In the dialogue below *the flight number* and *company* are relational anaphors of *the flight from Malaga*.

**6244

4	S:	ik wilde vragen
5		hoe laat *het vliegtuig uit Malaga*,
6		*dat* vanmiddag om half twee zou
		vertrekken,
7		aankomt.
8		*Het vluchtnummer* is niet precies
		bekend..
9		of althans..
10	I:	Welke *maatschappij*?
11	S:	Wat?
12	I:	Welke *maatschappij*?
13	S:	Ja
14		dat weet ik ook niet.
15		Ik weet alleen
16		dat *ie* uit Malaga zou vertrekken
		om half twee.
17	I:	Ja,
18		maar eh.. is *het* een chartervlucht,
		een lijnvlucht?

4	S:	I wanted to ask
5		at what time *the plane from Malaga*,
6		*that* would leave this afternoon
		at half past two,
7		will arrive.
8		*The flightnumber* I don't know exactly..
9		although at least..
10	I:	Which *company*?
11	S:	What?
12	I:	Which *company*?
13	S:	Yes
14		that I don't know too.
15		I only know
16		that *it* would leave from Malaga
		at half past two.

17 I: Yes,
18 but uh.. is *it* a charter flight,
a scheduled flight?

Also for these kinds of anaphors it holds that they strictly speaking involve a shift to another entity, that can be picked up again by an identity anaphor. But again, the referential expression used to refer to them is in such a way related to the previous expression that we have to speak about an anaphoric relation. Also these kinds of anaphors are considered to continue the topic of the dialogue because the entity the anaphor refers to cannot be found on the basis of its linguistic form alone. The referents of *the flightnumber* and *which company* for example can only be found if they are linked with the referent of *the plane from Malaga.*

The relation between the relational anaphor and its antecedent is based on our knowledge about inextricable relations of entities with other entities. The idea is that entities that are introduced into the dialogue bring with them entities that we immediately associate with the original entity. So flights bring with them a scheduled arrival time, a real arrival time, a flight number, a place of departure, a place of arrival, a company etc. While talking about one entity the other entities become immediately available for communication.

Also the relationship between relational anaphors and their antecedent leads to a special behavior concerning the use of referential expressions.

The relationship between these entities is so strong that the introduction of one entity can be done by taking one of its characteristics as referring term. See the following dialogue fragment:

**4458

3 S: eh.. Kunt u mij ook zeggen
4 of de KL 312 soms vertraagd is
uit Zurich ?

3 S: uh.. Can you tell me
4 if the KL 312 happens to be delayed
from Zurich ?

A flight is introduced by taking its flightnumber and placing the definite article in front of it. The dialogue fragment below shows another effect of the special relationship between associated entities: the topic is introduced as a group of associated entities.

**2172

7 S: Ik wilde graag vragen
8 of de geplande aankomsttijd
van de vlucht KL 602,
9 vanuit Los Angeles,
10 om vijftien uur tien,
11 of dat ook de juiste tijd wordt.
12 I: Ja,
13 die wordt op het ogenblik nog
steeds om vijftien uur tien verwacht.

7 S: I would like to ask
8 if the scheduled arrival time
of the flight KL 602,
9 from Los Angeles,
10 at fifteen hours ten,
11 if that will be the right time.
12 I: Yes,
13 it is at the moment still expected
at fifteen houres ten.

The topic is introduced by a very complex NP containing a reference to a scheduled arrivaltime, its flight, the flightnumber, the place of departure, and the expected arrivaltime.

Utterance 13 shows one effect more: a switch is easily made from one entity to the other. The flight is picked up by a pronoun, without being the topic of the preceeding utterance (which was the scheduled arrivaltime). The dialogue fragment about the plane from Malaga illustrates the same kind of phenomenon. The orginal entity is easily picked up by a pronoun after two shifts from one associated entity to another.

In the sequence of relational anaphors the information content of the used anaphors does not really decrease so considerably as in the case of subsectional anaphors. This is because the words used in the antecedent can generally not be repeated in the words that have to refer to the associated entity.

If we want to state a guideline about the use of referential expressions for relational anaphors we could say that the information contained in the relational NP need not be complete. It is enough that it only contains the information that is needed to distinguish the intended referent from the cluster it is part of. But as is clear from the beginning of this whole section, there are relatively few relational anaphors in the corpus. To give the rule more universality it should be tested once more on a corpus with more relational anaphors.

SUMMARY

We can now summerize the results of the study in the following way:

Topical chains are generally started by complex NPs, that refer to unused new entities or to entities that are anchored by unused new entities. If

simple NPs are used that refer to an unanchored brand-new entity, they are usually expressed in the comment part of the utterance.

After introduction, the topic is not always immediately settled. Sometimes, one or more full NPs, that either repeat or contain extra information are needed to establish the referent.

If the topic is clear for both dialogue participants, it will generally be continued first by a demonstrative pronoun, then by ellipsis.

A guideline for subsectional and relational anaphors can not really be based on the corpus studied here, because there are not enough instances. A qualitative analysis however suggests that these anaphors generally contain only that information that is needed to distinguish the intended entity from the cluster of entities it belongs to. A series of subsectional anaphors seem to exhibit the same kind of decrease of informational content as identity anaphors.

REFERENCES

H.C. Bunt [1994], "Context and Dialogue Control," *Think* 3, 19–31 .

W.L. Chafe [1978], "Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of view ," in *Subject and Topic*, C.N. Li, ed., Academic Press, New York , 25–55 .

W.L. Chafe [1987], "Cognitive Constraints on Information Flow," in *Coherence and Grounding in Discourse*, R.S. Toulmin, ed., John Benjamins, Amsterdam, 21–51 .

H.H. Clark & S.E. Haviland [1977], "Comprehension and the given-new contract," in *Discourse production and comprehension discourse processes: Advances in research and theory*, R.O. Freedle, ed. #1, Norwood (NY) Ablex.

H.H. Clark & C.R Marshall [1981], "Definite reference and mutual knowledge," in *Elements of discourse understanding*, A. Joshi, B. Webber & I. Sag, eds., Cambridge University Press.

K. van Deemter [1991], in *On the Composition of Meaning*.

B.J. Grosz [1977], "The Representation and Use of Focus in Dialogue Understanding," SRI International, Technical Note 151.

B.J. Grosz [1981], "Focusing and Description in Natural Language Dialogues," in *Elements of Discourse Understanding*, A.K. Joshi, B.L. Webber & I.A. Sag, eds., Cambridge University Press, Cambridge (etc.).

J.K. Gundel [1985], "'Shared Knowledge' and Topicality' ," *Journal of Pragmatics* 9, 83–107.

J.K. Gundel [1988], "Universals of topic-comment structure," in *Studies in syntactic Typology*, M. Hammond, E. Moravcsik & J. Wirth, eds., John Benjamins Publishing Company.

M.A.K. Halliday [1985], in *An Introduction to Functional Grammar*, Edward Arnold , London .

E.F. Prince [1981], "Toward a Taxonomy of Given-New Information," in *Radical Pragmatics*, P. Cole, ed., Academic Press.

R. Quirk, S. Greenbaum, G. Leech & J. Svartvik [1985], in *A Comprehensive Grammar of the English Language*, Longman Group, London, New York .

M.M.M. Rats [1994], "Topic-Comment Structures in Information Dialogues," in *Focus and Natural Language Processing. Proceedings of a conference in celebration of the 10th anniversary of the Journal of Semantics*, P. Bosch & R. van der Sandt, eds., Workings paper of the IBM Institute for Logic and Linguistics, Heidelberg, Germany, 591–599.

M.M.M. Rats [1995], "Topic Management in Information Dialogues," in *Proceedings of the International Conference on Cooperative Multimodal Communication CMC/95*, H. C. Bunt, R. J. Beun & T. Borghuis, eds., Eindhoven, The Netherlands, .

T. Reinhart [1981], "Pragmatics and Linguistics: An Analysis of Sentence Topics," *Philosophica* 27, 53–94.

C.L. Sidner [1979], "Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse,," MIT Technical Report AI-TR-537. .

C.L. Sidner [1981], "Focusing for Interpretation of Pronouns," *American Journal of Computational Linguistics* 7.

E. Vallduví [1993], "Information Packaging: Survey," Human Communication Research Centre, Research Paper HCRC/RP-44, University of Edinburgh.

E. Vallduví [1994a], "Detachment in Catalan and Information Packaging," *Journal of Pragmatics* 22, 573–601.

E. Vallduví [1994b], "Updates, Files, and Focus-Ground," in *Focus and Natural Language Processing. Proceedings of a conference in celebration of the 10th anniversary of the Journal of Semantics*, P. Bosch & R. van der Sandt, eds., Workings paper of the IBM Institute for Logic and Linguistics, 649–658.

Appendix

Table 1: Referential chains for topics in Information Dialogues:

Number of reference:	Referential expression that is used to refer to the topic:								
	Defnp	Indefnp	Demnp	Qnp	Prop	Dem.p	Pers.p	Rel.p	Ellipsis
1	126	34	8	18	67	0	0	0	0
2	26	1	6	0	3	149	42	15	57
3	24	0	2	2	6	60	33	1	75
4	4	1	1	1	0	32	17	0	73
5	3	0	2	1	0	15	18	1	44
6	4	0	0	0	2	12	13	0	32
7	2	0	3	1	1	7	7	0	25
8	0	0	0	0	0	11	3	0	16
9	0	0	0	1	0	3	2	0	11
10	0	0	1	0	0	2	3	0	8
11	0	0	0	0	0	0	2	0	7
12	0	1	0	0	0	0	3	0	3
13	0	0	1	0	0	1	2	0	4
14	0	0	0	0	0	0	2	0	2
15	0	0	0	0	0	0	2	0	1
16	0	0	0	0	0	0	1	0	1
17	0	0	0	0	0	0	1	0	1
18	0	0	0	0	0	0	1	0	1
19	0	0	0	0	0	0	0	0	1
20	0	0	0	0	0	0	0	0	1
21	0	0	0	0	0	0	0	0	0

DESIGN, FORMALISATION AND EVALUATION OF SPOKEN LANGUAGE DIALOGUE

Hans Dybkjær, Laila Dybkjær and Niels Ole Bernsen
Centre for Cognitive Science, Roskilde University
PO Box 260, DK-4000 Roskilde, Denmark
Phone: (+45) 46 75 77 11 Fax: (+45) 46 75 45 02
Email: dybkjaer@cog.ruc.dk, laila@cog.ruc.dk, nob@cog.ruc.dk

ABSTRACT

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.

1 INTRODUCTION

Dialogue model development is a major part of the development of spoken language dialogue systems (SLDSs). The entire development process, from the design of a first dialogue model through to the final user tests of the implemented system, may be viewed as a series of iterations, each iteration encompassing interacting aspects of design, formalisation and evaluation.

This paper describes how we addressed these interacting aspects when developing the dialogue model for the Danish prototype SLDS for domestic flight reservation.

The prototype SLDS, often simply termed the Danish dialogue system, has been developed in the Danish dialogue project which involves an effort of 30 man/years by the Center for Person-Kommunikation, Aalborg University, the Centre for Cognitive Science, Roskilde University, and the Centre for Language Technology, Copenhagen [Baekgaard et al. 1995].

The system runs on a PC and is accessed over the telephone. It understands continuous spoken Danish with a vocabulary of about 500 words and uses system-directed dialogue. The prototype runs in close-to-real-time. It consists of the main components shown in Figure 1. When a user calls the system, this will be detected by the *telephone line interface*. The *speech recogniser* then receives the user's speech signals. The speech recogniser is speaker-independent and 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 consists of the *ICM* and the *dialogue description*. 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. The *database* contains information on timetables, flights, reservations and customers and rules for managing the information and queries it receives. System output is produced by concatenating pre-recorded phrases. The phrases are selected by the dialogue management module and replayed by a separate *reproductive speech module*. The *text recogniser* is only used when the speech recogniser is disabled, as has been desirable during debugging and test, cf. Sections 3 and 4. The *DDL-tool* is not part of the running system but is a tool used to create the dialogue description, i.e. the implemented dialogue model. The *Dialogue Communication Manager* is a data bus which transfers messages between all other modules.

The dialogue model for the system was iteratively designed by means of the Wizard of Oz method. The model resulting from the last WOZ iteration was implemented and debugged and the implemented system was tested with naive users.

The WOZ experiments produced a corpus of transcribed dialogues, user questionnaires, and inter-

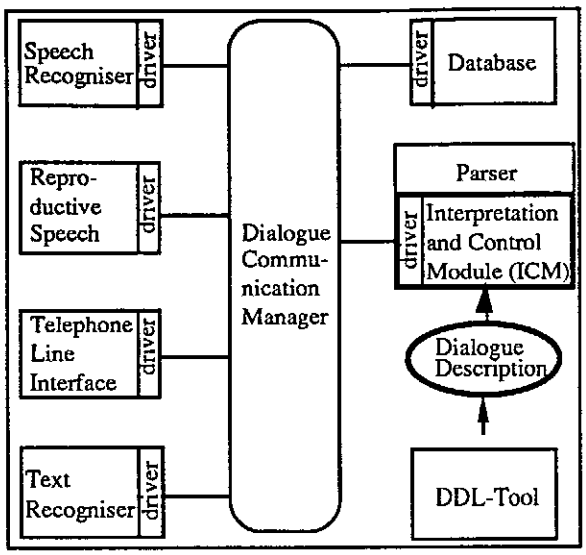


Figure 1. The overall architecture of the Danish dialogue system.

views; the implementation and debugging phase produced logfiles; and the user test produced logfiles and a corpus of transcribed dialogues, user questionnaires and interviews. Throughout the development process, these sources have served as a basis for evaluating the dialogue model by identifying user problems and revealing unsatisfied design goals and constraints.

The outcome of each evaluation cycle in the development process has served partly as a basis for improving the dialogue model and partly as input to the development of an applied theory of task-oriented dialogue. The evolving, formalised expression of the theory in its turn interacted with the dialogue design process. In addition, the dialogue design process as a whole has generated a consolidated series of guidelines for the design of usable SLDSs.

The remainder of this paper describes the dialogue development process for the Danish dialogue system in terms of iterative interaction between design, formalisation and evaluation based on corpora. Section 2 presents the WOZ experiments and the resulting corpus. Section 3 describes implementation and debugging. Section 4 reports on the user tests and their results. Section 5 summarises and concludes the paper.

2 DIALOGUE MODEL DEVELOPMENT

The Wizard of Oz (WOZ) experimental prototyping method is an iterative simulation technique which is well suited to the testing of dialogue models and the adjustment of design goals and design constraints prior to implementation. During each iteration a human (the 'wizard') simulates the system in dialogue with users who should preferably believe that they are speaking to a real system [Fraser and Gilbert 1991]. The dialogues are recorded, transcribed and analysed and results are used to improve the dialogue model. This iterative process continues until an acceptable dialogue model has been achieved.

2.1 The first dialogue model

The initial dialogue model was based on a number of different sources, including literature, field interviews with human travel agents and a standard timetable for Danish domestic flights which, in addition to departure and arrival times, contained information on i.a. fares and travel conditions. Two other important and intertwined sources were the technological constraints which were primarily imposed by the speech recogniser, and the goals to be achieved as regards usability [Dybkjær et al. 1993, Dybkjær et al. 1995a].

Since the application is based on access over the telephone, real-time performance was considered a constraint which had to be satisfied in order to obtain a usable system. However, this constraint, together with the chosen hardware, gave rise to new compulsory constraints caused by the speech recogniser:

- At most 100 words can be active in memory at a time for real time performance to be possible.
- The average user utterance length should not exceed 3-4 words.
- The maximum user utterance length should not exceed 10 words.

The two last-mentioned constraints were also meant to maintain the recogniser error rate at an acceptable level.

Furthermore, because of limited project resources the system vocabulary size was set to about 500 words.

The main usability constraints, apart from real-time performance, were sufficient task domain coverage, robustness, natural forms of language and dialogue, and flexibility. These goals had to be traded off against the above resource and technological constraints. This was done during seven iterations of WOZ experiments.

2.2 The WOZ experiments

The first five WOZ iterations mainly served to train the wizard and adjust the dialogue model so that major shortcomings were repaired. Each WOZ iteration produced only a few dialogues. The dialogue model was initially represented as a loosely ordered set of predefined phrases. This made it difficult for the wizard to maintain consistency and quickly find an appropriate phrase. In addition, as the domain coverage was not yet complete, sometimes the needed phrase would not even be present in the dialogue model. To solve the wizard's problems we decided instead to use a graph structure for representing the dialogue model, cf. Figure 2. The graph has predefined system phrases in the nodes and expected contents of user input along the edges and turned out to significantly facilitate the wizard's job. Domain coverage was gradually made more complete. Users (subjects) were during this period exclusively system designers and colleagues.

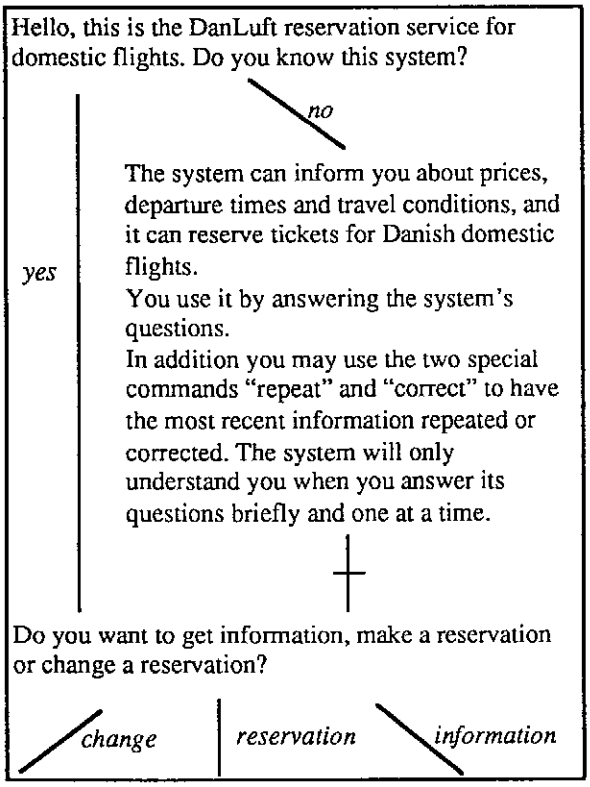


Figure 2. The introduction graph used in WOZ7 (translated from the Danish).

Throughout the experiments, interaction with the system was based on scenarios, i.e. domain-relevant tasks which the subject performed over the phone through dialogue with the system. The first four WOZ iterations were based on a set of ten scenarios which were simply considered a set of cases for which the system should work and which were mainly used for domain and task exploration and training of the wizard. Most decisions on precise reservation details such as date of departure were left to the subjects. Subjects often revised a scenario or invented a new scenario on the fly which was never written down.

In the last three WOZ iterations a new set of scenarios was used. This second set included a total of 28 scenarios. Only some of them were used in WOZ5 whereas all were used in WOZ6 and WOZ7. The scenarios were 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 scenarios from the second set contained more details than those in the first set and left few or no decisions to the subject. This facilitated the wizard's job because he would approximately know what a user would answer at a certain point during dialogue. However, the use of such detailed scenarios also had a negative effect in terms of users modelling the scenario phrases. This will be further discussed in Section 4 which also presents example scenarios.

The last two WOZ iterations were larger than the five first ones and were aimed directly at forming a basis for the dialogue model to be implemented and for the sub-language to be defined. Each of these two iterations involved 12 subjects. The majority of the subjects were external (non-in-house) and the rest were colleagues. Apart from three colleagues none of the subjects in the last two iterations had tried the system prior to the WOZ experiment. External subjects were selected so that half of them had a background as secretaries and the other half were computer scientists. The expected end-user group is mainly secretaries. The computer scientists were included in order to study the reactions of people who had general system knowledge.

Having agreed to participate, each subject in the sixth and seventh iterations received an envelope containing (i) a letter which briefly introduced the system and informed on the experiment, (ii) four scenarios and (iii) a questionnaire to be filled in and returned immediately after the subject's interaction with the system. Immediately before an experiment one of the system designers called the next subject at work and asked the subject to call the system. Subjects were not told in advance that the system was simulated. In a debriefing telephone interview after the session subjects were in WOZ7 asked whether they thought that they had interacted with a

real system. The majority of external subjects believed that the system was real whereas the colleagues knew in advance that it was simulated.

The two last WOZ iterations each produced a corpus of 47 dialogues. From the seven iterations a total of 125 dialogues were transcribed amounting to about seven hours of spoken language dialogue. 25 early dialogues were never transcribed. 24 different subjects had been used in the seven iterations.

For each iteration the recorded and transcribed dialogues were analysed and evaluated with focus on the extent to which the constraints and goals mentioned in Section 2.1 had been satisfied. Evaluation results were used as a basis for improving the dialogue model before the next WOZ iteration.

Between the fifth and sixth iteration we recorded a corpus of 25 Danish domestic flight reservation dialogues in a travel agency, corresponding to about one hour of spoken human-human dialogue. The original intention was to make these recordings early in the design process but due to practical problems this had not been possible. The structure of the WOZ6 dialogue model was adjusted in the light of typical task order structures identified in the human-human flight reservation dialogues.

2.3 The WOZ evaluation metrics

The evaluation metrics used during the WOZ experiments included measurement of the number of tokens (words) and types (different words), average utterance length, average number of utterances per dialogue exceeding 10 words, the longest turn, average number of turns per dialogue, number of user questions in per cent of the total number of turns (to converge towards zero), vocabulary size, cumulative word type/token ratio for subjects (to converge towards zero, only in WOZ7), average number of types per token in relation to number of tokens used by each subject (only in WOZ7), and the amount and nature of deviations from the normative model of how a scenario should be completed (only used systematically in WOZ6 and WOZ7, cf. below). The occurrence of user questions indicates that the user takes over the initiative. User questions therefore had to be eliminated as far as possible in order to satisfy the constraints on active vocabulary size and user utterance length. Convergence towards zero of the cumulative word type/token ratio is desirable because it indicates that the vocabulary size is sufficiently large for the application and that new users cannot be expected to use words out of the defined vocabulary.

As regards qualitative user evaluation of the system, subjects were asked to fill in a questionnaire, as mentioned above (from WOZ5 onwards). As indicated in parentheses above, some types of measurement were only made for the later WOZ iterations. In the early WOZ iterations some measurement results were much too far from the desired level and the material quite small, which made it irrelevant to study whether, e.g., the user type/token ratio converged.

In the last two WOZ iterations we compared the latest version of the system's dialogue model with the most recent, transcribed WOZ corpus in order to be able to systematically support improvements in system co-operativity. Each transcribed dialogue was plotted onto the graph structure which had system output in the nodes and expected contents of user utterances along the edges (cf. Figure 2). Deviations from the graph structure in terms of unexpected user or system behaviour were marked and the reason(s) for the behaviour analysed. When a deviation did not seem to have been caused by a wizard error, it was regarded as signifying a potential problem to be repaired.

Also at this stage, before each subsequent WOZ iteration we matched the scenarios to be used against the current dialogue structure in order to discover and remove potential user problems. This was done to some extent from WOZ4 onwards. The plotting and matching processes allowed identification of both actually occurring and potential user problems during dialogue. Actual user problems are such that actually occurred during user-system dialogue in the WOZ experiments. Potential user problems are problems discovered by the designers when putting themselves in the place of the users.

2.4 WOZ results

Each WOZ iteration produced quantitative as well as qualitative data. The quantitative data were used for measuring the extent to which the technological constraints were satisfied. Both quantitative and qualitative data were used for measuring usability constraint satisfaction. An important indicator of the degree of satisfaction of usability constraints is the number of user problems identified.

The technological constraints on maximum and average user utterance length were satisfied in WOZ7 (cf. Figure 3). Similarly, the task structure that had been developed appeared to make it possible to meet the constraint of a maximum active

vocabulary of 100 words. This, however could only be achieved at the expense of user initiative. The dialogue model of WOZ7 was entirely system-directed, cf. Figure 4 [Dybkjær et al. 1993, Dybkjær et al. 1995a].

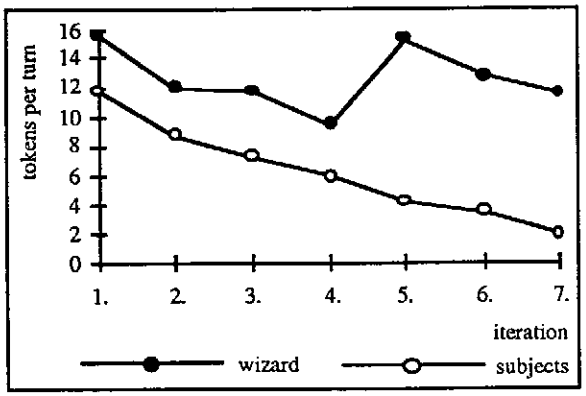


Figure 3. Average length of wizard and subject utterances in terms of tokens per turn.

The dialogue model was made system-directed by having the system conclude all its turns by a non-open question in order to preserve dialogue initiative. Non-open questions are questions which address a well-defined topic and ask for a specific piece of information. The non-open questions used by the system may be categorised as being of four types.

One type invites a yes/no answer, e.g.: "Do you want a return ticket?"

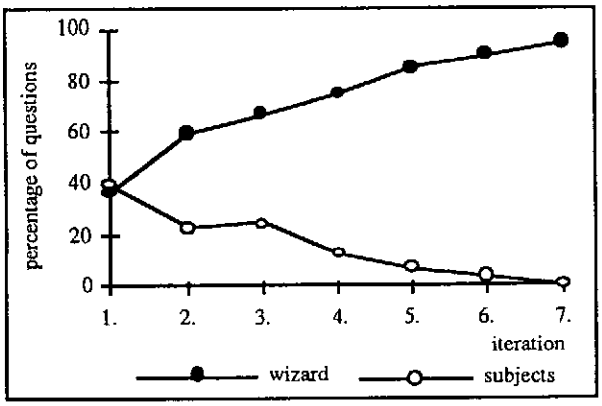


Figure 4. Number of questions in per cent of total number of turns.

The second type is a multiple choice question according to which the user is expected to choose an element from an explicit list of alternatives, for instance: "Is the ticket to be mailed or will the traveller pick it up at the airport?"

The third type of question invites the user to state a proper name or something similar, such as the name of an airport or an id-number. The application uses id-numbers instead of person names which cannot be dealt with because of vocabulary limitations. Users' names are looked up in the database by using the id-number as key. For instance: "Please state the id-number of the traveller."

The fourth type is the most open type or the one which allows the broadest variety of expressions in reply but which still concerns a specific topic, such as date of departure. For instance: "On which date will the journey start?"

None of these types of question invites the user to take over the initiative from the system.

During WOZ, a dialogue model was developed for ticket reservation as well as flight information and change of reservation. However, whereas "pure" reservation is a well-structured task, the information and change of reservation tasks are not. In a well-structured task there is a prescribed amount of information to be exchanged between the dialogue partners and the order in which this information is to be exchanged is often also prescribed to a certain extent. Complex ill-structured tasks such as the information task, on the other hand, are characterised by having a large number of optional sub-tasks. Each of these sub-tasks may be well-structured in itself but the overall task becomes ill-structured because of the optional character of the many sub-tasks it includes. This means that the system cannot make use of a valid stereotypical model that tells which sub-tasks the user wants to accomplish and possibly in which order [Bernsen et al. 1994a, Bernsen et al. 1994b, Dybkjær et al. 1995b].

Complex ill-structured tasks require mixed-initiative dialogue to be acceptable to users. Our heavy technological and feasibility constraints did not allow us to address the challenging task of designing mixed-initiative dialogue for a complex task such as the information task. It was therefore decided to implement only the reservation task which, because of its stereotypical structure allowed system-directedness and usability to co-exist. Thus, our recordings of human-human reservation dialogue in a travel agency showed that in reservation tasks the travel agent typically takes over after the initial customer turn and asks for the missing information piece by piece [Dybkjær and Dybkjær 1993].

As regards vocabulary size it was our hypothesis that 500 words would not be sufficient for the domain. The data from the WOZ experiments confirmed the hypothesis since the WOZ vocabularies did not clearly converge, not even the one in WOZ7. A 500 word vocabulary for the reservation task was defined mainly on the basis of

the WOZ data. The user test of the implemented system was expected to provide more data on the sufficiency of the vocabulary.

With respect to evaluation of usability constraints, a large amount of work went into the identification and repair of actual and potential user problems. As mentioned in Section 2.3, we plotted transcribed dialogues onto the graph structure representation of the dialogue model and we matched scenarios against the dialogue model to be used next.

The work on identifying and repairing user problems was systematised at the end of the WOZ design phase. The user problems found during the entire WOZ experiment were analysed, classified and represented as violations, made by the dialogue system, of principles of co-operative dialogue. The result was a set of co-operative principles for human-machine dialogue derived from a WOZ corpus of realistic task-oriented (simulated) human-machine dialogue. Adherence to each principle should guarantee that a certain class of usability problems can be avoided in SLDS design more generally. [Bernsen 1993, Bernsen et al. 1994a, Bernsen et al. 1995b]

In order to have users evaluate the dialogue model, the WOZ subjects received a questionnaire, cf. Section 2.2. Figure 10 in Section 4.4 shows subjects' opinions of the dialogue system they had interacted with in WOZ7 and in the user test, respectively.

On the whole, subjects evaluated the system fairly positively in the WOZ questionnaires. The positive answers on robustness (few errors) and reliability in WOZ7 (see Figure 10) are probably due to the fact that the wizard did not simulate misrecognitions. In three cases there is no doubt that the WOZ7 system was evaluated negatively. Subjects found the system boring, perhaps because of the monotonous and slow voice used by the wizard in order to make subjects believe that they were interacting with a real system. Subjects also found the system inflexible and certainly the dialogue structure had become rigid and system-directed. Finally, it was quite clear that the subjects would prefer to talk to a human travel agent instead of the system. Probably the main reasons were the rigid dialogue structure and the correct impression that such a system has limited capabilities and cannot cope with non-routine matters.

Questionnaire results from the user test will be discussed in Section 4.4.

3 DIALOGUE IMPLEMENTATION AND DEBUGGING

3.1 Implementation

The reservation task was implemented in DDL (Dialogue Description Language) which is an event-driven recursive flow chart language [Dybkjær and Dybkjær 1994, Dybkjær et al. 1995a]. Compared to the initial formalisation of the dialogue task provided by the graph representation, the implementation task had to face two types of shortcoming. Firstly, some dialogue elements had not been simulated in the WOZ experiments at all and others had not been simulated in sufficient detail. Secondly, the graph representation was still far from possessing the formal rigour required of the implemented system and realised in the DDL flow chart representation. In more detail, the shortcomings were the following:

- *Task structure.* In the WOZ experiments only the structure of the hour task had been defined in some detail. The exact structure of other tasks had to be figured out during implementation. Moreover, in task-oriented dialogues most sub-tasks have a common basic structure and differ only on points such as the exact phrasing and the specific piece(s) of information they concern. This commonality had not been exploited in the WOZ graphs.
- *Meta-communication.* Focus in the WOZ graphs was on task communication, i.e. on turn-taking in the direct course of task execution. However, many turns in ordinary conversation are about the dialogue itself, i.e., they are turns of meta-communication. The possibilities of meta-communication were only rudimentarily expressed and never really used during the WOZ experiments.
- *Domain.* The different pieces of information, rules and constraints needed in the system's domain representation had no prior representation in the graphs, and the interface between domain representation and dialogue was only implicit. For example, it had not been clearly defined which actions should be taken with respect to the system's domain representation when the user provided information on, e.g., the day of departure.
- *Dialogue state.* The overall as well as the local state of the dialogue had not been represented in the graph, including values of information slots, their status etc.

The above points had to be formalised during implementation through expanding and detailing the WOZ specification. The task structure required a new representation as described below, thus abandoning DDL's dialogues-as-graphs paradigm.

An outline of the main components of the implemented dialogue description or dialogue handler is presented in Figure 5 [Dybkjær et al. 1994].

The dialogue handling is task-oriented. There are two classes or levels of tasks:

- *Atomic tasks* concern one item of information, where an item is a value from the application or user domains. Atomic tasks are tagged with current system, user, and domain status, dialogue focus, and alternative values. Moreover, all user exchanges are done within atomic tasks, as explained below.
- *Compound tasks* manage the temporal structure of sets of atomic tasks. Examples are the reservation task and the overall dialogue frame. Compound tasks are modelled via a task dialogue structure represented as a graph where nodes are atomic tasks and edges are static links to other atomic tasks in a default template structure. The choice between links is made dynamically on the basis of Task Record and Dialogue History.

The atomic tasks follow a fixed scheme:

- check preconditions, i.e. if all required items are established;
- user-system exchange loop until item status is OK for both user and system:
  - ask the user:
    - for a value (and wait for answer), or
    - to select a value from a list (and wait for answer), or

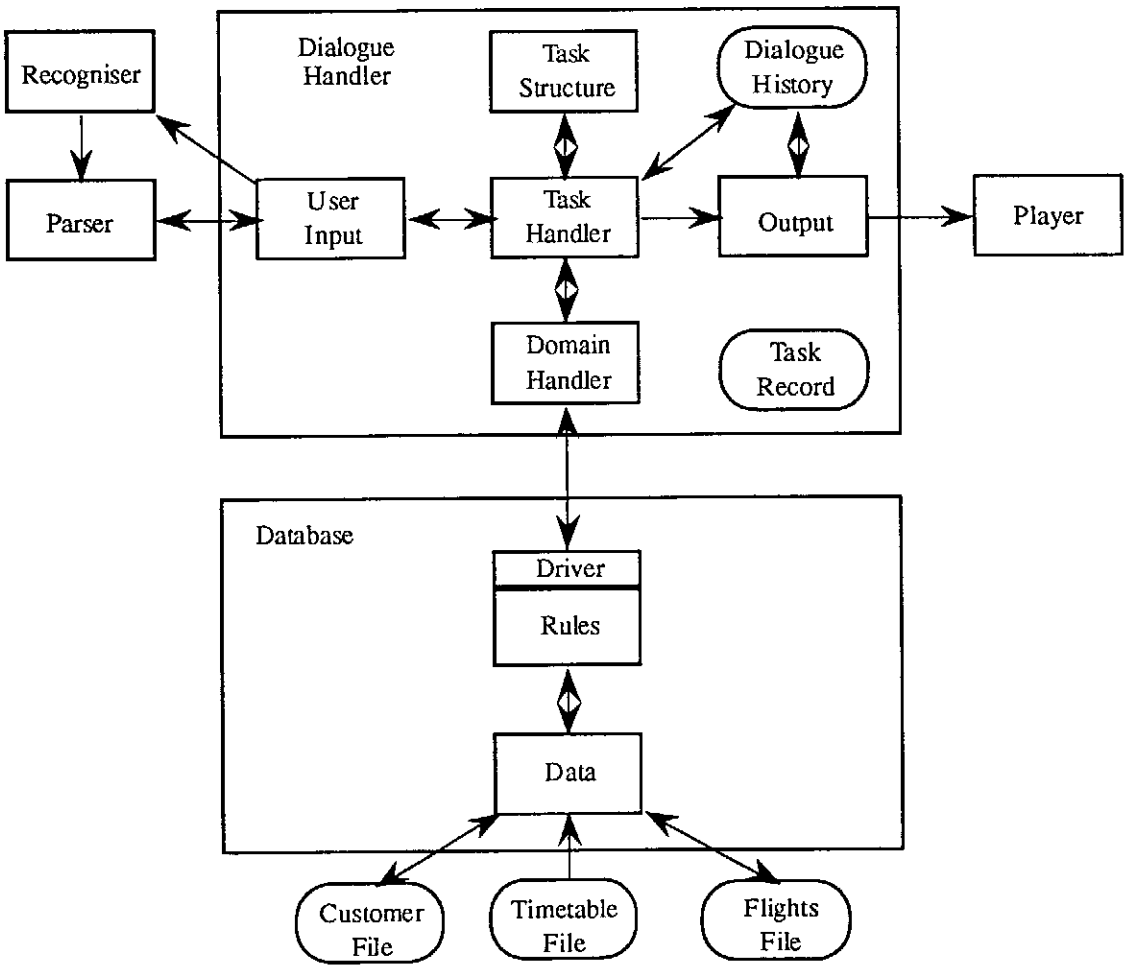


Figure 5. The communication structure of the recogniser, parser, player, dialogue handler and database with a detailed view of the dialogue handler and the database which represents domain knowledge. In the Dialogue Handler the Task Record is used by all processes. Rounded boxes indicate data and rectangular boxes indicate processes.

    - if a given value is desired (and wait for answer);
  - check the domain integrity of the value;
  - give feedback to users consisting in:
    - the accepted value, or
    - an error message;
- check post-conditions, i.e. if any other items are affected.

All checks and user-system exchanges are parametrised with respect to the items. In both pre- and post-checks and after user responses the Task Handler may jump directly to other tasks, thus circumventing the Task Structure.

As an example of the dialogue handling consider the following piece of dialogue in which the hour of departure is determined from S1b to S3a:

S0: On which date will the journey start?
U0: On Friday.
S1a: Friday May 19th.
S1b At which time of day?
U1: In the morning.
S2a: In the morning there are flights at 6:30 and 7:30.
S2b: Would you like one of these flights?
U2: Yes, 7:30.
S3a: 7:30.
S3b: On which date will the return journey start?

After S1a the Task Structure decides that the next item to be determined is the hour of departure. Control is transferred to the Task Handler which first checks if all other items required (route and date of departure) have been determined already. Then the exchange loop is entered and the system asks for time of day (S1b). The user answer (U1) is checked with the database which answers that there are two possible departures in the morning. In S2a this is given as feedback to the user. In S2b a new question is asked. The user answers the direct as well as the indirect question (U2). Since 7:30 has already been checked with the database, feedback is given without consulting the database again (S3a), the post-conditions are checked, control is transferred to the Task Structure, and a new cycle begins.

3.2 Debugging

A blackbox test was performed on the implemented dialogue model embedded in the entire system except the recogniser. The recogniser was disabled in order to make it possible to reconstruct errors. Internal communication between system modules was registered in logfiles. We created a number of test files all containing user input for one or more reservations of one-way tickets and return tickets with or without discount.

A test sequence always had to include an entire reservation involving several interdependent system and user turns. In a query-answering system a task will often only involve one user turn and one system turn. Hence one may ask a question and simply from the system answer determine if the system functions correctly for the test case. In a task such as ticket reservation which involves several turns, the system's reactions to the entire sequence of turns must be correct. An apparently correct system reaction, as judged from the system's immediate reaction, may turn out to have been partly wrong when we inspect the sequence of interdependent system reactions. Hence to test our dialogue model it was not sufficient to test, e.g., isolated transactions concerning customer numbers, possible destinations, or a selection of dates. Also the combinations of the test data had to be considered. Furthermore, each test reservation can only test a limited amount of cases so we had to create a long series of test reservations.

The blackbox test was not entirely exhaustive. In particular, it was not exhaustive as regards various interesting combinations of test data. However, the test did reveal a number of problems. Some of these were due to disagreements between the dialogue model specification and the implementation. But the majority of problems were such that had not been taken into account during specification.

Resources were not available for implementing solutions to all discovered problems. It was therefore considered, for each problem, how time consuming the implementation of a solution would be and how important it was. The hard problems were in many cases due to the fact that system-directed dialogue is not entirely sufficient to handle the cases in question. Solutions to such problems were not implemented because they would probably be sub-optimal anyway as long as the system-directed dialogue paradigm is maintained. Examples are round-trip tickets and reservations concerning, e.g., one passenger travelling out alone but going back together with another person. Both examples deviate from the standard reservation task and in the present system they have to be carried out as two separate reservation tasks. A round-trip ticket must be booked as two one-way tickets and the second example would have to be resolved by booking one return ticket and one one-way ticket.

The solutions which were implemented influenced not only the implementation but also the specification including the order of the dialogue structure. This again implied that the test files had to be revised to bring them in agreement with the specification. This is caused by the fact that the reservation task involves not only one user-system exchange but a whole sequence of exchanges which have to be made in a certain order.

The revised dialogue model was blackbox tested with the revised test files. Bugs were corrected but no major new unknown problems were revealed.

4 USER TESTS

When the system had been debugged we performed two series of user tests. In the first test the system was used with a simulated recogniser, in the second, the real recogniser was used. At the time of writing, the second test has not yet been completed and the analysis of results from the first test is in progress. Therefore, only first results from the simulated-recogniser user test are presented below. The setup and material used in the second test are the same as were used in the first test, cf. Section 4.1.

4.1 User test with a simulated recogniser

The system including a simulated recogniser was tested with naive users, i.e. users who had no previous knowledge of the system. A wizard keyed in the users' answers to 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 the real speech recogniser. The 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 vocabulary and which conformed to the recogniser's grammar rules.

Ten external and two in-house subjects were used. Ten of them were secretaries. The percentage of secretaries approximately corresponds to the percentage of secretaries among the customers who called the travel agency in which we recorded our human-human dialogue corpus.

Each subject received an envelope containing (i) a letter informing on the experiment, (ii) a colour brochure introducing the system, (iii) four scenarios, and (iv) a questionnaire. The dialogues were conducted over the telephone as in the WOZ experiments. Immediately after interaction with the system, subjects received a telephone interview. In this interview all subjects stated that they believed that the system was real.

4.2 Scenario design

The two different sets of scenarios used in the WOZ experiments (Section 2.2) 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. Whereas the domain coverage of these scenarios was reasonable, meta-communication was not simulated. The scenarios did not give subjects incorrect information and subjects were not otherwise asked to simulate situations in which errors occurred. This proved to be a drawback during implementation since we had no information on users' meta-communicative reactions to work from. The conclusion is that the WOZ scenarios should have covered the same ground as should the input cases in a black-box test.

The scenario set used in the user test corresponds to the notion of *evaluation and test* scenarios. Based on the WOZ scenario experiences, we carefully considered what to test and why. We decided not to do user testing on a number of possible but unlikely cases of communication failure. These have been tested instead in the black-box test during system debugging. 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. This process generated a set of 20 scenarios.

The later WOZ experiments had shown that subjects tended to copy the temporal vocabulary used in the scenario descriptions, i.e. the expressions of date and hour of departure. Yet the sub-language vocabulary of the dialogue system was derived from the scenario-based WOZ dialogues. This constitutes 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

To explore how to avoid this effect and elicit a more realistic sublanguage, two groups of test subjects were formed each of which received a different version of the scenario material. One group received standard travel descriptions of the kind likely to be copied by subjects, whereas the second group received a new version of the scenarios in which the copying effect had been effectively blocked [Dybkjær et al. 1995c]. Each group consisted of six subjects.

We had carefully considered which information to mask in the scenarios, and how. For this purpose we used the categorisation of system questions into the four types mentioned in Section 2.4: yes/no questions, multiple choice questions, questions asking for a proper name or something similar, and questions asking for date or time.

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 in 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 important to prevent them from modelling the scenario representations. System questions in this case 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 am.' 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, for instance: 'they want to arrive early in the evening'. This means that the 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 and generate a linguistic expression representing the result of that inference, thus excluding the possibility of priming.

All 20 scenarios were represented in two different versions. The masked version combines language and analogue graphics (cf. Figure 6) whereas the control group version uses standard linguistic text (cf. Figure 7) and roughly corresponds to the style of 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 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

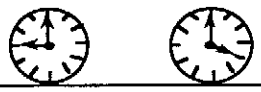


Figure 6. An analogue graphic scenario representation.

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 7. A text scenario representation corresponding to the graphic scenario of Figure 6.

4.3 User test evaluation metrics

The evaluation metrics used includes all relevant measurements, quantitative as well as qualitative, which were applied to the WOZ data. These measurements will allow comparison with the WOZ results and show if the user test data conform to the technological constraints which were the crucial parameters in the WOZ experiments. In addition, measurements have been made on meta-communication, system performance, other data

pertinent to usability evaluation, and sub-language differences between the two groups of subjects which received different scenario versions.

4.4 Overall user test results

The user test produced a corpus of 57 dialogues. 32 dialogues were based on text scenarios and 25 dialogues were based on graphic scenarios. Subjects sometimes repeated a scenario if they did not succeed the first time. Figure 8 provides a translation from the Danish of a transcribed dialogue from the user test. The dialogue is based on the scenario shown in Figure 7.

The test results show that the average user utterance length is still within the prescribed limits, cf. Figure 9. The prescribed maximum 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 'repeat' but were expressed in free style by users. Two long utterances were produced by subjects who took over the initiative

S1: Hello, this is the DanLuft reservation service for domestic flights. Do you know how to use this system?

U1: No.

S2: The system can reserve tickets for Danish domestic flights. You use it by answering the system's questions. In addition you may use the two special commands "repeat" and "correct" to have the most recent information repeated or corrected. The system will only understand you when you answer its questions briefly and one at a time. Please state your customer number.

U2: 4

S3: Customer number 4. How many people will travel?

U3: 4 people.

S4: 4 people. Please state the id-number of the first person.

U4: Id-number 1.

S5: Please state the id-number of the second person.

U5: Id-number 4.

S6: Please state the id-number of the third person.

U6: Id-number 6.

S7: Please state the id-number of the fourth person.

U7: Id-number 7.

S8: Id-numbers 1, 4, 6, and 7, Jens Hansen, Marie Hansen, Steen Sørensen, and Jane Sørensen. Where does the journey start?

U8: From Copenhagen.

S9: From Copenhagen. Where does the journey end?

U9: Aarhus.

S10: Aarhus. Do you want a return ticket?

U10: Yes please.

S11: The red discount is possible for certain departures on weekdays and the green discount is possible for weekend departures. If you choose red discount the return journey cannot start until two nights after the arrival date. Are you particularly interested in discount?

U11: No thanks

S12: On which date will the journey start?

U12: The 26th of January.

S13: Thursday the 26th of January. At which time of day?

U13: 7:20.

S14: 7:20. On which date will the return journey start?

U14: The 26th of January.

S15: Thursday the 26th of January. At which time of day?

U15: 17:30.

S16: 17:30. You have now booked 4 tickets for Jens Hansen, Marie Hansen, Steen Sørensen, and Jane Sørensen from Copenhagen to Aarhus on Thursday the 26th of January at 7:20 returning on Thursday the 26th of January at 17:30. Your reference number is 49. The price is 4760 kroner. Are the tickets to be mailed or will the travellers pick them up at the airport?

U16: We would like to pick them up at the airport, please.

S17: The tickets will be sent to the airport and can be picked up at check-in desk number 19. The tickets must be picked up 20 minutes before departure at the latest. Do you want anything else?

U17: No thanks.

S18: Goodbye.

Figure 8. A translated example of a dialogue from the user test.

when asked 'Do you want anything else?'. Finally, in some cases subjects provided more information in their answers than had been asked for. All long utterances, therefore, were produced when subjects took over the initiative against the principles on which system-directed dialogue is based.

The dialogue is entirely system-directed as appears from the example in Figure 8, and this actually did prevent users from asking questions as was also the case in the later WOZ experiments. In the user test, only four out of 998 user utterances were questions. One question was asked because the subject had misread the scenario text. The three other 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.

As predicted, the system's vocabulary is not sufficient, in particular as regards quantitative time expressions, cf. Section 4.5.

The system's task domain coverage is substantial but limitations exist exactly at points of maximum domain complexity where system-directed dialogue comes close to its limits, cf. Section 3.

Figure 10 compares answers from the WOZ7 questionnaires with answers to the user test questionnaires. In many cases there is no real difference between the two sets of answers. The negative development with respect to subjects' opinion on how easy it is to make corrections is probably due to the fact that misunderstandings were not simulated in WOZ7. This meant that hardly any 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.

This concludes our presentation of the general data obtained in the user test. Additional data and a comprehensive analysis will be presented in [Bernsen et al. 1995a].

	WOZ7		User test	
Total number of subjects	12		12	
Total number of dialogues	47		57	
	User	System	User	System
Total number of turns	881	905	998	998
Total number of tokens	1633	10495	2468	12185
Total number of types	165	350	188	189
Longest turn	12	92	23	87
Total number of turns > 10 tokens	3	272	17	253
Average number of tokens per turn	1.85	11.59	2.47	12.20
Average number of types per turn	0.19	0.39	0.19	0.19
Average number of turns per dialogue	18.74	19.26	17.51	17.51
Average number of turns > 10 tokens per dialogue	0.06	5.79	0,30	4,44
Average number of tokens per dialogue	34.74	223.30	43.30	213.77
Average number of types per dialogue	3.51	7.45	3.30	3.32
Total number of questions	4	-	4	-
Number of questions in per cent of total number of turns	0.45	-	0.40	-
Average number of types per token	0.10	0.03	0.08	0.02

Figure 9. Comparison of results from WOZ7 and the user test. The number of system questions were not calculated. All system turns except for the closing phrase contained a question.

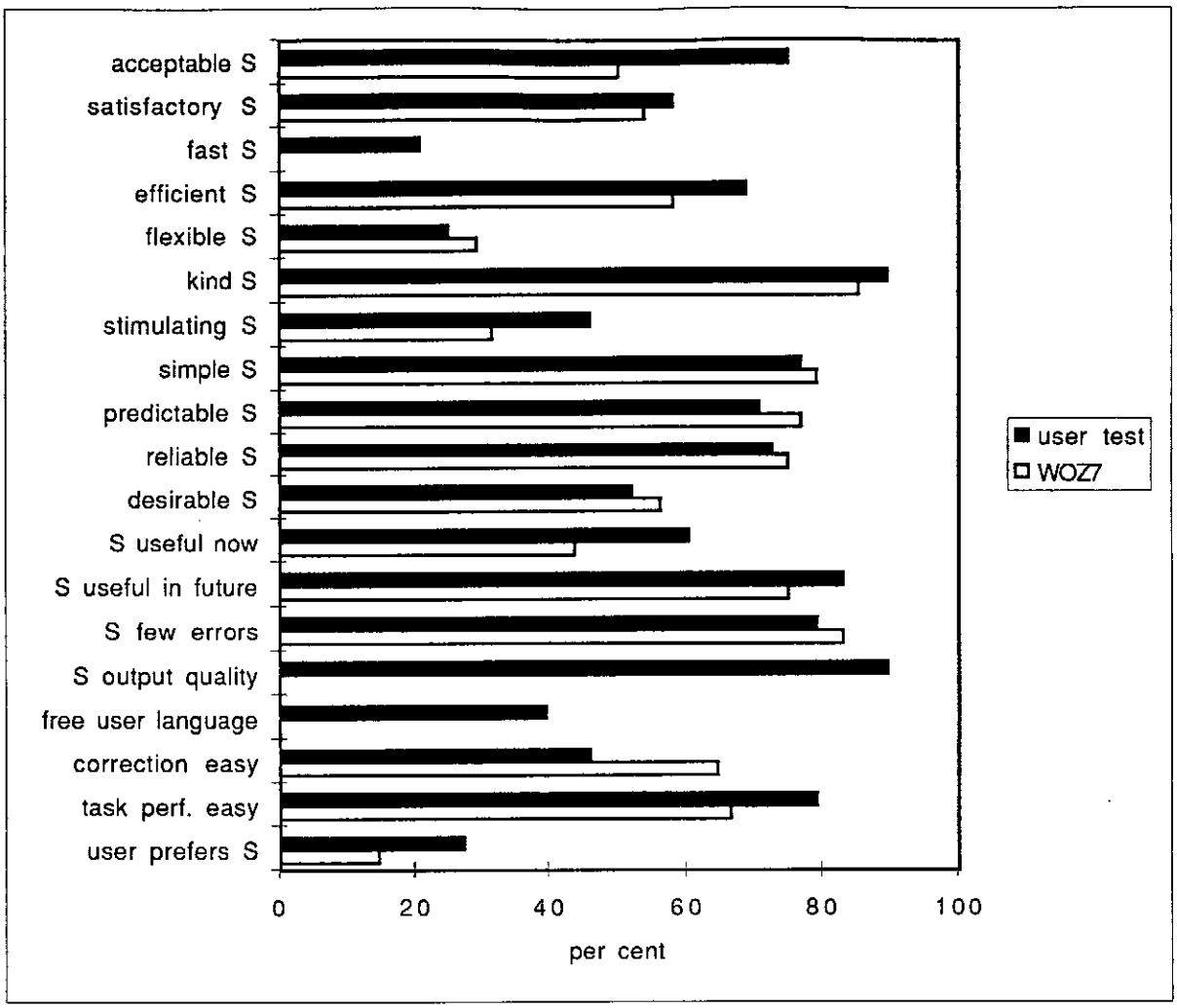


Figure 10. Subjects' answers to the questionnaires from WOZ7 and the user test in per cent of the maximum possible score. A score of less than 50 per cent indicates a negative opinion of the system. 'S' in the left-hand column refers to the system.

4.5 User test results related to scenario versions

The test results presented in Figures 9 and 10 above are based on analysis of the entire user test corpus. Figure 11 presents comparative data on the dialogues based on the graphic and text scenarios. Our hypotheses, as regards 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 sub-language vocabulary than those based on text scenarios in terms of (i) total number of different words and (ii) out-of-vocabulary words. The first hypothesis was confirmed whereas the second was not. In addition, we had an unexpected result which could provide a strong argument in favour of using graphic scenarios for SLDS development.

4.5.1 Priming Effects

As expected, we found a massive priming effect from the text scenarios and virtually none from the graphic scenarios. The first row of Figure 12 expresses the "cleaned" number of user turns for which priming from the scenarios was possible. 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 the dates and times of out and home journey departures. In these cases there is no immediate priming from the expressions used by the system itself and figures are not influenced by repeated or changed user answers.

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 considered as non-primed. For example, if the scenario said 'Friday the second of January' then 'the second of January' and 'Friday the second' would count as primed but not 'the second of first' which is a common Danish calendar expression.

	text scenarios	graphic scenarios
no. of subjects	6	6
no. of different scenarios	20	20
no. of dialogues	32	25
no. of user turns	547	451
no. of user turns*	181	178
no. of user tokens	1606	862
no. of user tokens*	705	451
no. of user word types	151	94
no. of user word types*	85	63
average user utterance length	2.94	1.91
average user utterance length*	3.90	2,53
longest turn	23	11
number of turns > 10 tokens	16	1

Figure 11. Data on the dialogues based on two different scenario types. * indicates that the figures only concern the dialogue parts on date and time.

	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%	-

Figure 12. Priming effects in WOZ7, and for text and graphic scenario-based dialogues, respectively.

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 second of January'. Furthermore, there are several frequent date expression formats. The system's feedback was of the form 'the second of first'. The decrease from 45% to 23% partly seems to be due to the fact that users changed from modelling the scenario text to modelling the system's feedback when answering the question about home date, and partly to the use of relative dates such as 'the same day'.

Throughout the WOZ scenarios the date format 'Friday the second 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 in WOZ7.

4.5.2 Vocabulary Effects

The use of graphic scenarios did not result in a significantly richer vocabulary than use of the text scenarios, nor in the elicitation of more new words. On the contrary, dialogues based on graphic scenarios contained fewer different words, cf. Figure 11. The scenario sets generated no out-of-vocabulary dates and only nine new words for times.

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

4.5.3 An Unexpected Result

We found a significant difference in tokens (words) per turn between dialogues based on text and graphic scenarios, respectively, cf. Figure 11. 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, therefore, 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 unsuitable 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.

5 CONCLUSION AND FUTURE WORK

Some preliminary conclusions on our dialogue model development process and the resulting dialogue system are:

The WOZ prototyping method is a powerful tool for dialogue model development although it does not eventually produce a model which is sufficiently formalised for implementation purposes. The quality of the produced model strongly depends on how well the simulations have been planned, trained, executed and iteratively evaluated. The main weaknesses of our own WOZ process were the lack of some form of tentative meta-communication simulation and the absence of formalisation details which therefore had to be developed during implementation. Overall, however, the WOZ development process has been successful in so far as there is reasonable correspondence between the final WOZ results and the results obtained during the user test.

The resulting dialogue system is entirely system-directed. This is primarily because of the strong constraints on active system vocabulary and user utterance length. A second important reason, however, is that we still lack a solid science base for developing mixed-initiative SLDSs for complex tasks [Bernsen et al. 1994b, Dybkjær et al. 1995b, Peckham 1993]. System-directedness makes task completion somewhat less efficient than might have been the case had mixed-initiative dialogue been feasible. As argued above, our corpora makes it clear that, for some sub-tasks of the reservation task, system-directed dialogue comes very close to its limits.

The system's qualitative time vocabulary is insufficient, as was expected. Its meta-communication apparatus, although functionally adequate, presents difficulties for novice users. However, users appear to quickly adapt to the system.

In addition to completing user testing and data analysis, we have begun to pursue two new directions of research. Both directions aim at consolidating a technologically and scientifically sound basis for building SLDSs for complex tasks. The first direction of research explores how the task of *informed reservation* might be formalised and implemented through the use of mixed-initiative dialogue [Dybkjær et al. 1995d]. An alternative to the use of mixed-initiative dialogue is to use multimodal technology. So, the second direction explores how the combined use of spoken input/output and graphic output may help overcome the limitations of system-directed dialogue in the performance of complex tasks.

ACKNOWLEDGEMENTS

The work described in this paper was carried out under a grant from the Danish Government's Informatics Research Programme whose support is gratefully acknowledged.

REFERENCES:

[Baekgaard et al. 1995] Baekgaard, A., Bernsen, N.O., Brøndsted, T., Dalsgaard, P., Dybkjær, H., Dybkjær, L., Kristiansen, J., Larsen, L.B., Lindberg, B., Maegaard, B., Music, B., Offersgaard, L., Povlsen, C.: The Danish Spoken Dialogue Project - A General Overview. *Proceedings of the ESCA workshop on Spoken Dialogue Systems*, Vigsø, Denmark, May 30 - June 2, 1995.

[Bernsen 1993] 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.

[Bernsen et al. 1995a] 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.

[Bernsen et al. 1995b] Bernsen, N.O., Dybkjær, H. and Dybkjær, L.: Cooperativity in Human-Machine and Human-Human Spoken Dialogue. Submitted to *Discourse Processes*, 1995.

[Bernsen et al. 1994a] Bernsen, N.O., Dybkjær, L. and Dybkjær, H.: Task-Oriented Spoken Human-Computer Dialogue. *Report 6a from the Danish Project in Spoken Language Dialogue Systems*. Roskilde University, February 1994.

[Bernsen et al. 1994b] 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.

[Dybkjær et al. 1993] 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.

[Dybkjær et al. 1995a] Dybkjær, H., Bernsen, N.O. and Dybkjær, L.: Dialogue Development and Implementation in the Danish Dialogue Project. To be published in *European Speech Projects*, Springer Verlag 1995 (in press).

[Dybkjær and Dybkjær 1994] Dybkjær, H. and Dybkjær, L.: Representation and Implementation of Spoken Dialogues. *Report 6b from the Danish Project in Spoken Language Dialogue Systems*. Roskilde University, May 1994.

[Dybkjær et al. 1994] Dybkjær, H., Dybkjær, L. and Bernsen, N.O.: Database Access via Spoken Language Interfaces. *Proceedings of FQAS '94*, Workshop on Flexible Query Answering Systems, Roskilde, November 1994, 69-79.

[Dybkjær et al. 1995b] 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).

[Dybkjær et al. 1995c] 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.

[Dybkjær et al. 1995d] Dybkjær, L., Bernsen, N.O., Dybkjær, H. and Papazachariou, D.: On the Use of Context in Building Spoken Language Dialogue Systems for Large Tasks. *IJCAI Workshop on Context in Natural Language Processing*, Montreal, August 1995.

[Dybkjær and Dybkjær 1993] 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.

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

[Peckham 1993] Peckham, J.: A New Generation of Spoken Dialogue Systems: Results and Lessons from the SUNDIAL Project. In *Proceedings of Eurospeech '93*, Berlin 21-23 September, 1993, 33-40.

Mutuality Strategies for Reference in Task-Oriented Dialogue

David G. Novick & Brian Hansen
Center for Spoken Language Understanding
Oregon Graduate Institute
P.O. Box 91000, Portland, OR 97291-1000 USA
novick@cse.ogi.edu, brianh@cse.ogi.edu

ABSTRACT

How do people coordinate the production and understanding of complex referring expressions in conversation? Analysis of a series of actual dialogues in a simple domain suggests that people employ strategies that involve levels both of domain structures and of acceptance. Our model of mutuality strategies extends Clark and Schaefer's (1989) model of discourse by explaining which domain entities will be referenced and which levels of acceptance will be employed.

1. INTRODUCTION

This paper is about the way people refer to things. In a task-oriented conversation, like ordering a meal in a restaurant or explaining how to assemble a mechanical device, participants coordinate the production and understanding of referring expressions so that the conversation stays "on track." To do this, the participants must maintain, as the conversation progresses, a degree of mutuality of knowledge (Clark & Marshall, 1981) sufficient for each participant. Using a corpus-based approach, we explore the processes that conversants employ to achieve this coordination. Our principal claim is that people form and use *mutuality strategies*, which provide a kind of comprehensible structure to patterns of referring. The mutuality strategy model (MSM) combines elements of discourse structure and domain structure to make predictions about referents and their expression in succeeding utterances in a dialogue.

In a task-oriented conversation, a usually joint domain goal guides the predominant focus of the interaction. Such a task orientation is typical of interactions using computer-based spoken-language systems. Examples might include scheduling an appointment, ordering a pizza, or filling out a questionnaire. Other kinds of conversations, which are often oriented toward more subtle social goals such as building friendships, are far less likely to be suitable for computer-human interaction, oral or otherwise. Characteristics of task-oriented conversations include discourse structure that reflects the structure of the domain task (Grosz & Sidner, 1986), negotiation of referring expressions (Clark & Wilkes-Gibbs, 1986), and an inherent requirement for mutuality of knowledge.

We have collected and otherwise have access to a variety of conversational corpora, including task-oriented dialogues in the domains of air-traffic control (Novick & Ward, 1993), directions for motorists (Novick & Sutton, 1994), visual puzzle solving (Marshall & Novick, 1995), and work-process reengineering (Wynn & Novick, in press). In these conversations, as in most others, conversants expend a high degree of effort in making their beliefs about the conversation mutual. For example, the 21 subjects in the vehicular navigation corpus produced 1107 acknowledgments in 2499 turns. This constitutes a substantial part of the interaction.

Further analysis of such corpora suggests that conversants vary the ways they present evidence that they have understood a prior utterance. Clark and Schaefer (1989), who developed strength-of-evidence levels to characterize acceptance of presentations, used the acceptances to form contribution trees that reflect the organization of the entire conversation. They did not, however, reach a full explanation of why particular kinds of acceptance evidence were used in any given conversational context; Clark and Brennan (1991) began to address this problem by describing the costs associated with different kinds of interactive behaviors. Analysis of task-oriented corpora also suggests that conversants vary the scale of the domain entities to which they refer. That is, they refer to details, aggregates or wholes as the conversational situation demands. Our 1988 model (Novick, 1988) began to address the question of granularity by hypothesizing the use of sub-sequence as a discourse structure. While our sub-sequence hypothesis successfully explained much of the observed dialogue, other patterns occurred that it could not explain. At this point, we lack a strong theory as to when and how conversants choose an

appropriate scale or grain of reference, given some context.

Accordingly, this paper will present evidence from a corpus of free conversation in a simple domain that choices of evidence and scale can be explained in terms of strategies adopted by the conversants to achieve conversational goals. The paper's *mutuality strategy model* (MSM), is particularly useful for task-oriented dialogue precisely because the conversational goals of are generally explicable. We will describe our experimental domain and corpus, briefly review Clark and Schaefer's levels of evidence of acceptance, and present a number of examples from the corpus. We will motivate a computational representation for dialogue knowledge about the use of referring expressions. Finally we will present the MSM and its implications.

2. DOMAIN AND CORPUS

To examine the relationship between acceptance and granularity, we collected a new corpus of dyadic conversations designed to make clear the elements of this issue. Clark and Schaefer's (1989) analysis was based on the London-Lund corpus, for which context is unavailable and the intentions of the conversants are unknown. Accordingly, we designed a laboratory task to elicit task-oriented mixed-initiative conversations with known context and intentions. Because we are interested in examining meta-discourse effects, the domain was deliberately kept as simple as possible. Furthermore, we wanted the task to be short enough that complete conversations could be analyzed at a detailed level. We chose for our study a mental task, that of jointly reconstructing a random sequence of letters.

Two subjects were each given a card on which was printed a sequence of 17 letters and blanks (see Figure 1). The cards were prepared by producing a random sequence of letters and then introducing blanks so that (1) no more than seven letters occurred without an intervening blank, and (2) a given position was blank for at most one of the cards. Thus, the subjects could reconstruct the entire sequence only by collaboration. Each subject was instructed to memorize the sequence on his or her card without looking at the other subject's card. The cards were then put out of sight and the subjects were instructed to work together to reconstruct the complete sequence. Subjects were given as much time as they wanted to complete the task.

Preliminary testing showed that this task is difficult enough to demand a high degree of task focus and even to raise the possibility of error, but not so difficult that subjects are likely to become frustrated. At the same time, the domain of discourse is simple enough to permit modeling of conversants' goals and belief states at a detailed level. Although the task imposes a high cognitive load on the participants, the utterances themselves tend to be very simple linguistically, often consisting of only the name of the next letter. While the domain of this conversational task is quite simple, it nonetheless captures some of the flavor of many real-world conversations, with each conversant cooperatively attempting to reach mutual understanding in the face of missing information.

Our collection procedure matched pairs of participants by gender and approximate age. All were unacquainted adults and all were native speakers of American English. Subjects were paid a small amount for their time; payment was not contingent upon completing the task successfully. A total of four pairs of subjects (two male, two female) each completed two letter sequences for a total of eight dialogues. Subjects were seated approximately three feet apart, at a 90° angle to each other with a low coffee table between them. Three cameras were used to record the sessions. One camera was positioned in front of the subjects to captured the interaction as a whole. The other two cameras were positioned to the sides, one focused on the face of each subject to capture gaze direction and facial expressions.

Two transcripts were created for each dialogue. One transcriber prepared a detailed record of the hand gestures, gaze direction and utterances of each speaker. A second transcriber, working from the video tapes and the detailed transcripts, prepared an explication in a narrative style similar to that suggested by Cook (1990). These narrative transcripts attempted to capture the experimenters' sense of what

Subject A:	O	_	O	S	E	_	A	G	F	H	C	W	_	E	_	_	Z
Subject B:	_	S	O	S	_	X	_	G	F	H	_	_	L	E	B	Y	_

Figure 1 . Sample letter sequences

had transpired during the conversation. Figure 2 contains an excerpt of a detailed transcript and its corresponding narrative for one of the conversations in the corpus. Each set of transcripts was checked by the other transcriber. Discrepancies were resolved informally. The detailed transcript depicts the move-

Time	Events#	Left Subject				Right Subject			
		Face/ body	Hands	Eyes	Verbal	Verbal	Eyes	Hands	Face/ body
7:59:10	38				O				
	39			to					
	40						away		
	41					S			
	42				O	O			
	43			away			to		
	44				C				
	45				No				
	46				S E				
	47			to					
	48		R down				away		
	49					X			
	50		R fiddles w/shoe	away			to		
	51				A				
	52				G				
	53				F				nods
	54			to	H				
	55				C				
	56				W				
	57	nod				L	away		
	58	nod				E			
	59	nod				B			
	60	nod				Y	to		
8:23:15	61				Z				

38-9 LEFT starts out with O and then looks at RIGHT.

40 RIGHT takes the turn and looks away briefly to concentrate.

41-42 When he is ready, RIGHT says the next two letters, S and O, the last of which LEFT repeats. LEFT seems to be confirming what RIGHT said when he repeats the O.

43 When RIGHT looks at LEFT, LEFT looks away, taking the turn. He is not ready to supply the next letters immediately, and pauses briefly.

44-46 LEFT utters the letter C, but then corrects himself saying "No, SE," with an emphasis on the S, to indicate that the C is being replaced with the S.

47-48 Left looks at RIGHT to signal that he is through, and RIGHT looks away momentarily. LEFT rather aimlessly drops his hand down toward his left foot which is resting on his right knee.

49-50 RIGHT supplies the next letter, X, with an even tone and looks steadily at LEFT. RIGHT's steady glare causes LEFT to look away and fidget nervously with the heel of his shoe. He appears to need just a bit of time to remember the next letters. He seems to feel pressured because RIGHT is looking at him.

51-53 LEFT says "A G F," and RIGHT nods in agreement when left says "F." RIGH T was listening particularly for the part where he noted the error in 30, and seems to be satisfied with the adjustment made by LEFT.

54-56 LEFT continues, looking at RIGHT and supplying H, C and W.

57-60 RIGHT, realizing that it is his turn, looks away and says, "L E B Y." With each letter LEFT nods in acknowledgment of RIGHT's contribution. LEFT may remember this part of the sequence from the first time through.

61 LEFT confidently supplies the final letter, Z, nodding.

Figure 2. Detailed transcript and narrative

ments of the face, body and eyes as well as the verbal behavior of two laboratory subjects carrying out the letter sequence task. The sequences are the same as those shown in Figure 1. In transcribing we were concerned with capturing synchrony and sequence of events occurring in multiple channels. The passage of time is shown as a progression of events going from the top of the transcript to the bottom. Simultaneous and overlapping events are depicted as occurring at the same or overlapping event numbers.

3. CONTRIBUTION AND ACCEPTANCE

Speech act theory (Austin, 1962; Searle & Vanderveken, 1985) suggests that we can motivate dialogue and explain conversational coherence by modeling conversation in terms of the conversants' goals and their plans for reaching those goals. This approach accords well with findings that the structure of discourse about a particular task closely follows the structure of the task itself (e.g., Oviatt & Cohen, 1988; Cohen & Perrault, 1979; Grosz & Sidner, 1986). In this view, language is just another tool to be used in accomplishing some goal, and utterance planning becomes incorporated into the larger task planning (e.g., Power, 1979; ; Litman & Allen, 1987).

For both domain and meta-conversational goals, mutuality is a key element. Clark and his colleagues have proposed a theory of conversation as a process in which conversants collaborate in building a mutual model of the conversation (e.g., Clark and Wilkes-Gibbs, 1986). The conversants' beliefs about the mutuality of their knowledge serves to motivate and explain the information that conversants exchange. The collaborative view of conversation offers an explanation for conversational coherence by viewing conversation as an ensemble work in which the conversants cooperatively build a model of shared belief (Clark & Schaefer, 1989).

In earlier work we have modeled dialogue and conversational control acts at an abstract level (Novick, 1988; Novick & Ward, 1993; Novick, Walton & Ward, 1993). Our computational model of dyadic and multiparty dialogue encompasses relations among acts, utterances, and beliefs. Figure 3 summarizes the conceptual model. Conversants are modeled as autonomous agents. An agent's beliefs may include beliefs about another agent's beliefs, depicted as smaller areas within an agent's belief space. Agents communicate by forming intentions to perform an act, then instantiating that act in terms of some physical action, e.g., an utterance. Agents interpret actions as acts based on their own beliefs. Note that if B's beliefs about A are in error or incomplete, B may infer a different act than A intended and so misunderstand A's action.

Our model, then, is based upon the collaborative view of conversation: conversation is seen as an attempt to establish and build upon mutual knowledge using speech acts. This synthesis was first proposed by Novick (1988) to explain conversational control or "meta-locutionary" acts. More recently, Traum and Hinkelman () have developed a similar model of mutuality maintenance as part of their theory of "conversation acts." The underlying conceptual model is similar in style to that proposed by Allen (1991) and Traum (1991), in that a distinction is made between privately-held and shared knowledge in understanding a task-oriented conversation. Allen's TRAINS model, however, is built around the negotiation and recognition of plans.

Our current work is largely inspired by that of Clark and Schaefer (1989), who proposed a model of discourse that characterizes conversation through

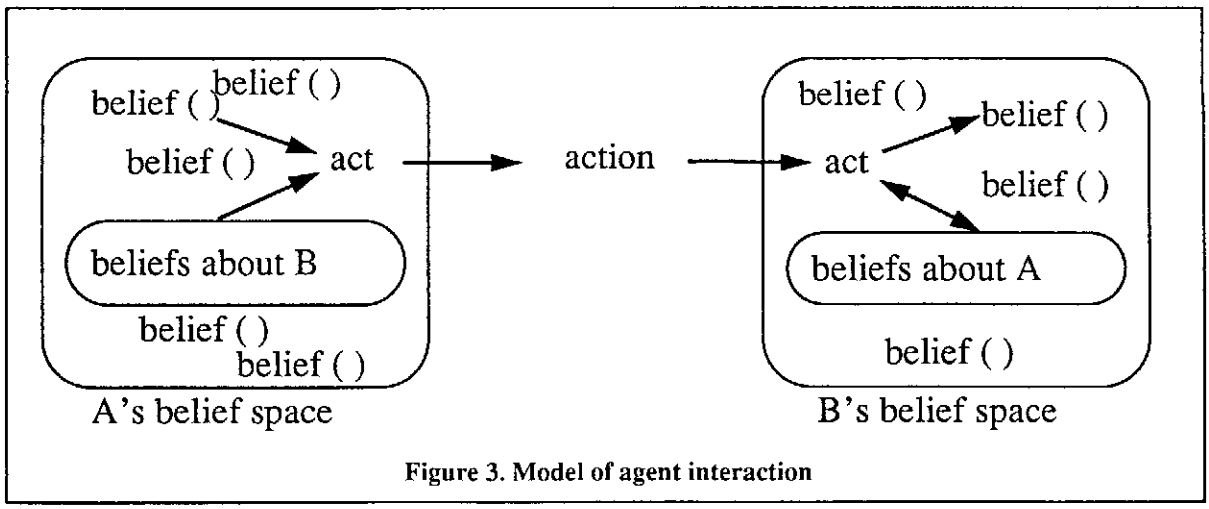


Figure 3. Model of agent interaction

contributions, where each contribution comprise a presentation and an acceptance. A key element of Clark and Schaefer's model is that participants provide different strengths of evidence that they are accepting each other's presentations. These levels of evidence are listed in Table 1. Continued attention involves non-verbal, "backchannel" responses, such as gaze. Acceptance via next relevant contribution is achieve by presenting a new element in the conversation that explicitly or implicitly indicates understanding of a preceding utterance; relevant contributions, in effect, move the conversation forward. Acknowledgments explicitly indicate understanding. Acceptance via demonstration involves performing an action, verbal or otherwise, that indicates understanding; an example would be if someone handed you the salt when you asked at a meal for someone to pass the salt. Display involves verbatim reproduction of an utterance in a way that suggests understanding; this may depend on prosody for interpretation.

4. ANALYSIS OF CORPUS

The appeal of the contribution model, and particularly of the acceptance evidence levels, is that it accounts for relationships among utterances in terms of collaborative control. Using the concept of acceptance levels, one can explain utterances in a dialogue as serving a systematic functions of grounding the discourse. These functions can be observed in the letter-sequence corpus; moreover, there are shifts between groups of similar acceptances. The utterances in the corpus also show shifts between groups of referring expressions with similar granularity. We now explore these levels and shifts as they occurred in the corpus.

From the perspective of a single conversant, the domain of the letter-sequence task is naturally broken down into three levels of granularity: the single letter, the sub-sequence of letters bounded by blanks, and the entire sequence. Earlier simulations (e.g., Novick, 1988) centered on the interaction pattern called the "sub-sequence hypothesis" in which a conversant retained the conversational turn as long as he or she had positive information to convey (the non-blank sub-sequence) and ceded the turn upon reaching a blank.The other conversant then took up the turn and followed the same principle. Although successful in portraying the pattern of interaction that occurred in several of the experiments, in itself, the sub-sequence hypothesis did not predict how a listener would react to the speaker uttering a single letter within a sub-sequence, nor the means by which turn exchange might be accomplished. Furthermore, in many instances other patterns occurred. Conversants sometimes got stuck and entered into lengthy repair sub-dialogues, they nodded and verbally acknowledged individual letters, they echoed the speaker's utterances, and sometimes even chanted along.

The letter-sequence corpus contains patterns of reference that tie together multiple utterances with common combinations of acceptance and grain level. For example, in the following exchange early in one of the conversations (where the joint sequence began "O S O S E ...," participant R echoes the L's production letter-by-letter. The evidence of acceptance in this case is display.

R: I had O and then you had—

L: S.

R: S.

L: O.

R: O. 'N I had ...

Later in the dialogue, the conversants return to this part of the sequence. This time through they again use display for evidence acceptance. The grain-size of the referent, though, has shifted to the sub-sequence, as R echoes back a two-letter sub-sequence and then contributes another. In this case, it appears that the conversants are now more confident about the letter sequence and are willing to risk greater uncertainty in mutuality for the advantage of conversational efficiency.

R: Okay, I had O, then you had what?

L: S O.

R: S O. (And) I had S E...

The conversation partially reported in Figure 3 also involves the sequence beginning "O S O S E ..." At event 42 in the transcript, the conversants pro-

Level
1. Continued attention
2. Next relevant contribution
3. Acknowledgment
4. Demonstration
5. Display

Table 1. Evidence of acceptance

duce the same referring expression simultaneously. R and L effectively "chant" the letter O.

R: Let's try that again.

L: O.

R: S // O. //

L: // O //

Beyond individual letters and subsequences, conversants often refer to fuller sequences. In the following excerpt, conversant R presents an extended sequence, and L acknowledges this.

L: Now is that right

R: Okay so I had a blank, S, O, two blanks, X, blank, G, F, blank.

L: All right ...

These fragments are typical of the conversations in the letter-sequence corpus. The conversations exhibit a variety of grain sizes, acceptance evidence levels, and combinations of these. Shifts of acceptance and grain levels are typically associated with boundaries of larger discourse structures, such as "passes" through the sequence or repair episodes. Within a structure or segment, the levels are usually consistent, as typified by the excerpts discussed above.

5. COMPUTATIONAL REPRESENTATION

As a final prelude to presentation of the mutuality strategy model, we illustrate our methodology for validation of discourse control theories. This method involves the simulation or replication of human-human dialogues—such as those in the letter-sequence corpus—by computational agents with interpretable representations of the model. As our research project plans to simulate the MSM with such agents, we need to assure that the model's expression can be represented computationally. To introduce this, then, we review the representations we use in validating the sub-sequence hypothesis in (Novick, 1990).

An agent's understanding of the world is represented as a set of beliefs of the form

believe (*agent*, *proposition*, *truth-value*, *mutuality*).

The *proposition* term represents the item about which a belief is held. The *agent* term represents the agent to whom the belief is ascribed, which may be either agent. The *truth-value* term represents the truth value that the *agent* assigns to the *proposition* (as understood by the agent in whose memory the belief appears). The mutuality term represents the various (possibly trivial) groups that mutually believe the proposition with truth value. For example, the following belief in agent A's memory:

believe (Agent_A, turn(Agent_B, true, [[Agent_A]])).

would indicate that Agent A believes that it is now Agent B's turn and that only A believes this.

In addition to knowledge of beliefs, agents can also hold beliefs about actions. These are represented as predicates of the form

act (*agent*, *action*, *truth-value*, *mutuality*).

For example, the action

act (Agent_A, give_turn(Agent_B), true, [[Agent_A, Agent_B]])

indicates that A gives the turn to B and that they mutually know this to be the case.

In the simple domain used in our current study, agents exchange information using the acts *request*, *assert*, and *respond*.

With this representation of belief and action, we developed a set of dialogue rules characterizing the letter sequence task and its conversational realization. We defined meta-locutionary, speech-act, and domain rules to explain the course of the dialogues. These rules are summarized in Table 2.

Domain rules capture the discourse planning and interactions that are peculiar to a particular domain. In this case, the rules at this level implement this task's goals of confirming and exchanging information about individual letters and substrings of letters.

Meta-locutionary rules and speech-act rules attempt to capture the domain-independent notions of performing speech acts (assert, request, and acknowledge) and meta-acts (turntaking).

We built a working implementation of the rule set described above and tested it in *saso* (Novick, 1990), a rule-based shell developed in Prolog as a tool for modeling multi-agent problems involving simultaneous actions and subjective belief. The conversants are represented by computational agents (rules implemented as *saso* operators) that communicate using the acts defined in the model. *Saso* uses forward-chaining control to simulate the parallel execution of multiple rule-based agents with separate belief spaces. A user-defined set of extended STRIPS-style operators is used to specify the behavior of agents in terms of preconditions and effects.

Agents communicate through acts; when an agent performs an act, clauses representing the action are posted to the memories of all agents in the system.

To simulate each dialogue, an initial state corresponding to the initial beliefs and intentions of the conversants was defined. The simulation was allowed to run to completion (*saso* halts when there are no more operations to execute) and the results were compared with the transcripts of the original conversation.

While these rules could account for much of the letter-sequence task and produced realistic exchanges of speech acts, the model's approach to acceptance and grain levels was too simple to produce the patterns of referring expressions observed in the corpus. For example, the agents only produced or confirmed sequence elements on a letter-by-letter basis. Consequently, we recognized a need to extend the model and the representation to account for the regularities in acceptance and grain levels in a way that accounted for the collaborative functions of the level-combination groups and of the shifts between these.

6. MUTUALITY STRATEGY MODEL

To explain the patterns of referring expressions observed in the data, we developed the idea of a mutuality strategy: The conversants effectively devise or settle on an understood n-tuple of strategy elements, where each element is combination of granularity and acceptance level that is the most effective scheme for assuring mutuality in that context. By "effective" we mean that (a) mutuality is maintained to a degree that the conversants' uncertainty about references is sufficiently minimized while (b) the pace of the conversation is not unduly slowed by overcareful and redundant discourse. The number of tuples depends on the number of domain grain levels. Some elements in a strategy for a given task may be missing, ambiguous, or superfluous.

In the course of our analysis, we developed a variation on Clark and Schaefer's (1989) acceptance evidence model. Conversants carrying out the letter-sequence task did not seem to make any use of *demonstration* acceptances. Further, while the listener's echoing of a speaker's utterance counted as a display acceptance, the case where both speakers spoke at the same time (chanted) seemed not to fit into Clark and Schaefer's model. Following Clark and Schaefer, then, our adapted forms of acceptance (ordered from weakest to strongest) are continued attention, next relevant contribution, acknowledgment, echo, and chant. To avoid confusion with Clark and Schaefer's model, we denote our adaptations as *confirmation* levels.

	Rule	Purpose
Meta-Locutionary Rules	• Give turn • Recognize my turn • Recognize other's turn	Perform and recognize turn-taking
Speech Act Rules	• Do assert • Assertion received	Perform and recognize assertion acts
	• Do request • Request received	Perform and recognize request acts
	• Do acknowledgment • Acknowledgment received	Perform and recognize acknowledgment acts
Domain Rules	• Goal: confirm next subsequence • Next subsequence confirmed • Last subsequence confirmed	Confirm a subsequence of several letters
	• Goal confirm next letter • Next letter confirmed • Last letter confirmed	Confirm a letter
	• Goal: obtain next subsequence • Next subsequence obtained • Last subsequence obtained	Obtain the next subsequence from the other agent
	• Goal: obtain next letter • Next letter obtained • Last letter obtained	Obtain the next letter from the other agent
	• Goal: assert next letter	Offer the next letter
	• Goal: request next letter	Request the next letter
	• Goal: respond next letter	Respond to a request for the next letter

Table 2. : Dialogue Rules for the Letter Sequence Task

The space of possible combinations of granularity and confirmation levels is depicted in Table 3. The cells of the table contain "strategy elements" labelled with names meant to suggest how a conversant using it would behave. The letters "ABC" correspond to any sub-sequence, and the letter "A" is meant to stand for any single (non-blank) letter within a sub-sequence. Following this convention, an interaction in which conversants use the "ABC-okay" strategy element would, for example, proceed as follows: the first conversant utters the letters of their sub-sequence and other conversant acknowledges. If they were using the "run-through" strategy element, then the first conversant would utter his or her entire sequence (blanks and all), and the other conversant would respond with their entire sequence. This is similar to the business letter example mentioned earlier; the lack of copresence encourages the conversants to use an interaction style that minimizes turntaking.

By combining the strategy elements from each level of the domain, one can describe the overall mutuality strategy of a conversant. A mutuality strategy for the letter sequence task would be a 3-tuple consisting of a confirmation level to be used for each level of the domain: the letter, sub-sequence and full sequence levels.

7. DISCUSSION

Validation of the MSM depends on (1) the robustness of the model in interpreting the observed conversations and (2) the power of the model to predict confirmation and grain levels. The model's predictive power is the subject of current research, using the dialogue simulation technique described above. We now turn to the problem of determining the robustness of the model's explanatory value.

We determined that evidence-of-acceptance levels and domain granularity could be reliably coded from the transcriptions. We gauged interrater reliability in a study of five coders who had a basic familiarity with the notions of levels of acceptance. The coders were graduate school faculty, staff and students in computer science. This analysis principally addressed the coding of the verbal aspects of the interaction. After a short training session, the coders worked independently at coding from a transcript. The process of coding involved coding the level acceptance; the level of granularity of the domain task did not require judgment. Because the sub-sequences of each conversant are known in advance, the sub-sequence acceptances can be derived from the transcripts. A separate coding was produced for each speaker, so each conversation produced two analyses.

Although there are many possible ways of performing the letter-sequence task, subjects invariably worked by making one or more passes through their sequences until they felt that they had achieved their joint goal. After each pass, the conversants typically engaged in what we have called a diagnostic phase, in which they discussed where things had gone wrong and how they should proceed. These diagnoses were not coded because the MSM does not make specific predictions as to their contents.

In the study, acceptance levels were coded in terms of Clark and Schaefer's (1989) levels of evidence, as all coders had prior experience with this

	domain level		
confirmation level	letter	sub-sequence	full-sequence
1. attend	A-attend	ABC-attend	sequence + attend
2. contribute	alternators	sub-sequence hypothesis	run-through
3. acknowledge	A-okay	ABC-okay	run-through+okay
4. echo	A-echo	ABC-echo	echo whole sequence
5. chant	A-chant	ABC-chant	chant whole sequence

Table 1: Mutuality strategy elements for the letter-sequence task

model. We extended the model to include a category for explicit dis-acceptance (for example, where a conversant says "No," "Wait," "That's not right," or "What?"). While most of the acceptance levels are straightforward, it is important to note that a question can serve as a next relevant contribution. For example, the utterances "My next letter is J" and "What did you have next?" are both coded as next relevant contributions.

The results of the codings indicated that of the 62 utterances within the scope of the study, 58 were coded identically by all five coders, producing an interrater agreement of 93 percent. The Kappa coefficient for the interrater agreement is 0.75.

Although we obtained a fairly high level of agreement, instances where coders' judgments differed suggest possible problems with the level-of-acceptance model. The most significant divergence in coding occurred in the display and next-relevant-contribution levels. In the case where one conversant echoes the other's utterance, one coder pointed out that the conversant who echoed may not have intended to do so; they may not have realized that the other would continue to speak and so they intended to make the next relevant contribution. One implication of this line of reasoning would be to require that there be tow levels of acceptance associated with each conversant: the intended and the achieved. If the echoed conversant had taken the echo to be the next relevant contribution, he or she would eventually have to conclude that there had been a divergence of beliefs about the sequence and that efforts to repair might be needed. Such a divergence would constitute a mismatch of mutuality strategies. Or, the echoed conversant may have deduced that the echo was inadvertent. In the actual conversations, the fact that in these cases there was no attempt at repair is suggestive evidence that echoes of this kind should be treated as instances of the display evidence of acceptance.

One way to distinguish between real and inadvertent echoing is by comparing the times at which the utterances occurred. If the utterances were produced simultaneously, then they are both considered to be next relevant contributions. But after some period of time has passed, the echoer must have known that the other conversant had continued to speak and their echo would be considered an instance of display.

A second divergence in coding occurred at the beginning of a pass over the sequence. Conversant "Right", whose first letter was a blank, said "You start." Conversant "Left" then said the first letter of the sequence, "O." Some coders judged Left's utterance as a demonstration; others coded it as a next relevant contribution. Although clearly a demonstration (by definition), this utterance also functions as a next relevant contribution because it introduces the next conversational element. This suggests either that an utterance may have more than one form of acceptance or that the categories of evidence need to be refined.

Because the coders were asked to make judgments only on the conversants' verbal behaviors, nonverbal behaviors such as gaze and gesture were not included in this study.

8. OPEN ISSUES

Beyond the question of whether the MSM's implementation as an agent-based simulation will replicate the strategy elements, combinations and shifts, the model inspires a set of deeper issues. These issues include application of the model to other domains, why and how conversants change strategies, use of the model in other modalities, extension of a model about acceptance techniques to include dis-acceptance strategies, and the general problem of validation of dialogue models. We discuss each of these issues in turn.

Granularity in other domains

One of the characteristics of letter-sequences is that the domain structures have a fairly clear granularity. Are features of the MSM limited to some unique quality of the domain? Consideration of other tasks typically encountered in spoken-language systems suggests that many other domains also exhibit an organization of grain structure. So, for example, in a retail sales task such as ordering a take-out pizza levels of granularity of domain entities would include the order as a whole, individual pizzas within the order, and the various toppings on a pizza. In an air-travel reservations task, levels of domain entities would include a complete itinerary, the flights that comprise the itinerary, and the arrival and departure times for each flight. Many other domains can be similarly decomposed.

The fact that domains can be characterized by hierarchically related entities does not, by itself, make it certain that conversants will use grain levels as part of a mutuality strategy. The "natural" structure of non-laboratory tasks, though, is likely to lead to coherent groupings of utterances as the structure of the discourse follows the structure of the task.

Transition rules

The MSM, as currently expressed, stands for the proposition that strategies exist and can be identified in discourse. The model does not yet include rules for transitions among strategies or, within a strategy, among strategy elements. We expect that uncertainty about mutuality would lead to use of strategy elements with smaller grain size and higher strength of evidence. An open question is whether conversants pick strategies from a repertoire, where the strategy elements are well-known, or if they instead negotiate the choice of a strategy and its constituent elements in a manner analogous to that described by Clark and Wilkes-Gibbs (1986) for individual referring expressions. Strategy negotiation would differ from that for individual referring expressions in that mutuality strategy is a dynamic conversational process.

Modality effects

In developing the MSM, we used highly detailed transcripts of face-to-face conversation. Although we were able to model successfully a significant amount of nonverbal expression such as gaze, for future work we are collecting new protocols using telephone technology so that he can eliminate nonverbal interaction for future analyses that hold the promise of more direct application to human-computer interaction. The treatment of continued attention as evidence of acceptance in a telephone conversation remains an open issue; from the point of view of one conversant, is the other's silence acceptance via continued attention or non-acceptance via not listening? Do conversants develop different mutuality strategies for conversations in different modalities?

Dis-acceptance

In the case where breakdowns occur and efforts at conversational repair are undertaken, there is a clear need to produce a better model of dis-acceptances as a parallel to Clark and Schaefer's model of acceptances. Clark himself later proposed a framework for understanding the use of repair as being of three types (preventatives, warnings, and repairs) working over three levels (vocalization and attention; presentation and identification; and meaning and understanding), but this scheme does not seem to capture the kinds of distinctions that Airenti et al.'s (Airenti et al., 1993) five phases of comprehension and production of communicative acts (understanding literal meaning, understanding a speaker's meaning, communicative effect, reaction, and response generation), nor those outlined in Novick and Sutton's (Novick & Sutton, 1994) description of the functional uses of acknowledgment.

Validation methods

The general problem of validation of dialogue models remains difficult. While the interrater agreement study for the MSM shows that the acceptance behaviors in the corpus can be consistently classified, it is much harder to prove that the model can accurately predict uses of strategies and shifts among strategy elements. The simulation technique that we use for replicating observed conversations depends on virtually complete knowledge of the context, but there is much of the context to which we do not have access. Even for a domain as simple as letter sequences, there is a high degree of variation in the behaviors of individual conversants. No two of the conversations in the corpus were identical; presumably this was due to variations in cognitive capacity and memorial processes. Given that the initial conditions for every conversation were nominally identical, how can the model and its rule-based implementation account for the observed differences in the dialogues? Which particular dialogue should be replicated?

We are attempting to address these questions by developing metrics for assessing a "distance" between the conversational acts produced by the agents in the simulation and those produced by the human conversants in their original conversation. Because each utterance in a conversation becomes part of the conversation's emerging context, the notion of distance becomes increasing meaningless as the simulation diverges from the original. Consequently, we propose that the metric be applied at each acceptance and the simulation be "reset" if necessary to conform to the original conversation. The distance values at each step of the dialogue would then be combined to produce an overall score for the model's ability to predict the grain and acceptance levels used by the conversants. This technique should be broadly applicable to conversational control models.

9. REFERENCES

Airenti, G., Bara, B. & Colombetti, M. (1993). Failures, exploitations and deceits in communication, *Journal of Pragmatics*, 20(4), 303-326.

Allen, J. F. (1991). Discourse structure in the TRAINS project, *Proceedings of the Fourth DARPA Workshop on Speech and Natural Language.*

Austin, J. L. (1962). *How to do things with words.* Oxford: Oxford University Press.

Clark, H. & Brennan, S., (1991). Grounding in Communication, *Shared Cognition: Thinking as Social Practice*, APA Books.

Clark, H. & Marshall, C. (1981). Definite Reference and Mutual Knowledge, *Elements of Discourse Understanding*. Cambridge: Cambridge University Press, 10-63.

Clark, H. & Schaefer, E. (1989). Contributing to Discourse, *Cognitive Science*, 13, 259-294.

Clark, H. & Wilkes-Gibbs, D. (1986). Referring as a collaborative process, *Cognition*, 22, 1-39.

Cohen, P. R. (1984). The pragmatics of referring and the modality of communication. *Computational Linguistics*, 10(2), 97-146.

Cohen, P.R., & Perrault, C. R. (1979). Elements of a plan-based theory of speech acts, *Cognitive Science*, 3(3), 177-212.

Cook, G. (1990). Transcribing infinity: Problems of context representation, *Journal of Pragmatics*, 14, 1-24.

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse, *Computational Linguistics*, 12(3),175-204.

Litman, D. J. & Allen, J. F. (1987). A plan recognition model for subdialogues in conversations. *Cognitive Science*, 11, 163-200.

Marshall, C. R., & Novick, D. G. (1995). Conversational effectiveness in multimedia communications. *Information Technology & People*, 8(1), 54-79.

Novick, D. G. (1988). *Control of mixed-initiative discourse through meta-locutionary acts: a computational model*. Technical Report CIS-TR-88-18, Department of Computer and Information Science, University of Oregon.

Novick, D. G. (1990). Modeling belief and action in a multi-agent system, *Conference on AI, Simulation and Planning in High-Autonomy Systems*, Tucson, AZ, March, 1990, 34-41.

Novick, D., Hansen, B., & Lander, T. (1994). *Letter-sequence dialogues*. Technical Report CSE 94-007, Department of Computer Science and Engineering, Oregon Graduate Institute of Science & Technology.

Novick, D., & Sutton, S. (1994). An empirical model of acknowledgment for spoken-language systems, *Proceedings of ACL-94*.

Novick, D., & Ward, K. (1993). Mutual beliefs of multiple conversants: A computational model of collaboration in air traffic control. *Proceedings of AAAI'93*, Washington, DC, July, 1993, 196-201.

Wynn, E., & Novick, D. (in press). Conversational conventions and participation in cross-functional design teams. *Conference on Organizational Computing Systems (COOCS '95)*, Milpitas, CA, August, 1995.

Novick, D., Walton, L., & Ward, K. (1993). *Contribution graphs in multiparty discourse*. Technical Report CSE 93-015, Department of Computer Science and Engineering, Oregon Graduate Institute of Science & Technology.

Oviatt, S. L. & Cohen, P. R. (1988). *Discourse structure and performance efficiency in interactive and noninteractive spoken modalities*. Technical Note 454, SRI International.

Power, R. (1979). The organization of purposeful dialogues, *Linguistics*, 17, 107-152.

Searle, J. R., & Vanderveken, D. (1985). *Foundations of illocutionary logic*. Cambridge: Cambridge University Press.

Traum, D. R. (1991). *Towards a computational theory of grounding in natural language conversation* (Technical Report No. 401). Computer Science Department, The University of Rochester.

Traum, D., & Hinkelman, E. (1992). Conversation acts in task-oriented spoken dialogue. Technical Report 425, Department of Computer Science, University of Rochester. (Also in *Computational Intelligence*, 8(3), August, 1992).

MESSY DATA: WHAT CAN WE LEARN FROM IT?

Norman M. Fraser

Vocalis Limited, Mill Court, Great Shelford, Cambridge CB2 5LD, United Kingdom
Tel: +44 1223 846177, Fax: +44 1223 846178, Email: norman@vocalis.com

ABSTRACT

This paper presents a variety of different empirical data relating to spoken language dialogue models, and particularly to computer models designed to be used to achieve real results in real domains. All of the data presented in this paper relate to the domain of flight information enquiries. An existing flight information service provided by human operators is introduced and some corpus examples relating to it are presented. A brief description of a series of Wizard-of-Oz simulations based on that existing service is provided, together with some example dialogue material collected during the simulation exercise. Data from laboratory trials of an implemented version of the service is examined next. Finally, data relating to the experience of running a very simple version of the system 'live' with real users is reviewed. In each case, the data examined are interestingly 'messy'. The paper seeks through an examination of data of each type, to identify some first lessons to be learned from 'messy' data.

1. INTRODUCTION

The study of language, and the field of theoretical linguistics in particular, has been heavily influenced over the past three and a half decades by the ideas of Noam Chomsky. One of Chomsky's most significant innovations was to earth language study, not in the data of linguistic *performance*, that is behaviour, but rather in speaker/hearer's *competence*, their internal knowledge of their own language. Any normally developed individual is capable of making grammaticality judgements which distinguish between well-formed sentences (such as *'He saw them'*) and ill-formed ones (such as *'Them saw he'*) (Chomsky 1965).

Chomsky's interest when he originally drew this distinction lay firmly in the arena of sentence grammar. How does his work transfer into the domain of dialogue? The answer to this question is not as straightforward as it may at first appear. One might, for example, respond by arguing that the study of dialogue is fundamentally the study of linguistic performance. If this is true then — it could be argued — a Chomskyan approach is doomed from the outset. However, a reasonable defence of the Chomskyan position would be to argue that people possess knowledge (competence) about what constitutes a well-formed dialogue, and this knowledge is separable from knowledge about how to apply it in actual situations of language use. So, for example, Grice's Co-operative Principle (Grice 1975) might form part of our dialogue competence.

The way we address this question will be conditioned, in large part, by the kind of model of dialogue we wish to construct. People with predominantly theoretical interests will tend to conceive of competence models of dialogue, whereas people whose focus is the development of a usable spoken dialogue computer system will tend to have a more integrative model, which may blur any putative distinctions between dialogue competence and dialogue performance. Whichever approach is taken, the same fundamental question will need to be answered: on which data should the models be founded?

An important historical feature of the Chomskyan programme in syntax is that it led to a diminished interest in the collection of empirical data. Since every speaker had access to their own linguistic competence, and since this had the additional advantage of being free from the 'noise' introduced by performance, there was no motivation to collect language corpora for this purpose. Could it be that empirical data is likewise redundant when constructing a model of dialogue?

The emphatic answer offered by this paper is 'No!' Empirical data is indispensable. There are several ways in which this position could be

defended. For example, a theoretical argument could be constructed to show that certain aspects of dialogue knowledge are not accessible to introspection. A practical argument could be offered to show that use of empirical data can significantly speed up the process of producing a dialogue model. However, rather than trying to argue the case in the abstract, in this paper I shall present dialogue data of various kinds to illustrate how rich and complex they can be. I shall focus particularly on what might be called 'messy' data, that is, data which appear at first glance to include random performance 'noise'. Careful study of messy data often reveals an unexpected underlying orderliness which would never have been discovered through introspection and which would have been overlooked if the data had been subject to an early and simplistic division into 'clean' data for further analysis and 'messy' data to be discarded.

All of the data presented in this paper relate to the domain of flight information enquiries. Section 2, section 3 and, to some extent, section 4 draw on work carried out as part of the SUNDIAL project (Peckham 1993, Peckham and Fraser forthcoming). In section 2, an existing flight information service supplied by human operators is introduced. Section 3 briefly describes a series of Wizard-of-Oz (WOZ) simulations based on that existing service. In section 4, some initial human-computer data from laboratory trials of an automatic version of the service is examined. Section 5 reports on the experience of running a very simple version of the system 'live' with real users. In each case, the data examined are interestingly 'messy'. We shall seek through an examination of data of each type, to identify some first lessons to be learned from 'messy' data.

2. HUMAN-HUMAN DIALOGUE DATA

British Airways operates a flight information telephone service distinct from its ticket booking service. A corpus of telephone calls between members of the public and British Airways agents was recorded and transcribed. From these, a sample of 100 dialogues was selected, again at random, for detailed analysis. Some of these dialogues turned out to be reasonably brief and orderly. For example, one of the simplest flight enquiry dialogues in the corpus is shown in Extract 1.

Extract 1: a simple human-human flight enquiry dialogue (A = agent; C = caller)

```
A: Flight Information. Can I help you?
C: Yes, could you tell me the arrival
   ti- the expected arrival time of
   British Airways two five eight from
   Caracas, pleas?
A: Yes, certainly. Can you hold the
   li-ne, please?
C:   -Thank you.
   (4.5)
A: Yes, it's expected now at
   thirteen thirty.
C: Thirteen thirty.
A: Terminal four #h Heathrow Airport.
C: Thank you very much indeed.
A: Thank yo-u. Bye bye.
C:          -Thank you. Bye.
```

Note that, in spite of its simplicity, this dialogue still contains a number of 'messy' features. For example, the agent's initial utterance *'Can I help you?'* could be understood as a direct speech act (= 'Do you think I am able to help you?') or as an indirect speech act (= 'Tell me what you want'). The caller appears to respond to both possible speech acts in one utterance, in the process producing a similarly ambiguous act (*'Yes, could you tell me...'*). In return, the agent also answers both possible acts (*'Yes, certainly?'*) The agent then produces yet another ambiguous act (*'Can you hold the line please?'*). This time the caller has no difficulty in interpreting the second part of the agent's utterance as an indirect act in which the agent is asking the caller to hold the line rather than questioning the caller's ability to hold the line. One might imagine that an appropriate response to such a request might be for the caller to answer *'Yes'* or *'Okay'*; instead, the caller answers *'Thank you'*. Being made to hold the line is scarcely the kind of activity likely to engender thanks in the person being made to do so. In certain contexts expressing thanks for an unpleasant experience would have to be interpreted ironically. Here that reading does not spring readily from the context. It is much more likely that the caller reads the agent's 'hold the line' request as evidence that the agent is actively engaged in trying to resolve the caller's query; for this ongoing effort, or in anticipation of a positive result, the caller thanks the agent.

Thus, in this simplest of dialogues, the agent and the caller appear systematically to use ambiguous speech acts. (In certain other dialogues,

the same or similar words are used, but the alternate readings appear to be intended.) Furthermore, fairly sophisticated reasoning appears to be going on, with linguistic turns being used to refer to events as given which have not been linguistically introduced.

Answers to the puzzles raised by these observations may well lie in the constraints inherent in this circumscribed domain of telephone flight information enquiries, but, if so, they will not have trivial answers. Not all dialogues proceed like this one, and we are still a long way away from producing a theoretical account which adequately covers all the speech act phenomena in the corpus. One of the striking features of these relatively straighforward corpus materials is that a few minutes' inspection reveals a host of basic problems which fall outside current theoretical understandings in pragmatics.

Notice that there are numerous other problematic phenomena in this short extract. There is a self-correction (*'...the arrival ti- the expected arrival time...*), an audible inbreath (*#h Heathrow Airport*), instances of talk-in-overlap, and a potentially confusing temporal reference (*'It's expected now at thirteen thirty'*).

It is understandable that some researchers arriving for the first time in the field of dialogue modelling might set out with the goal of designing dialogue models which are, in some information theoretic sense, optimal. But it cannot be emphasised strongly enough: *human dialogue is not optimised in such a way as to minimise ambiguity and maximise the rate of information transfer.*

Consider what this simple dialogue might look like if it were re-written according to some standard of 'optimality'. A first attempt is offered in Extract 2.

Extract 2: an (invented) optimal flight enquiry dialogue

```
A: This is British Airways' Flight
   Information Service. Please state
   your request.
C: Tell me the arrival time of BA two
   five eight.
A: Hold on.
   (4.5)
   That flight is expected at
   thirteen thirty.
C: Okay.
```

This dialogue is efficient, but unpleasant. Real human dialogue is messy and at least some of the 'messiness' is non-optional if the dialogue is to seem pleasant and natural for ordinary people (i.e. people who have not been specially trained to accept very brusque modes of interaction).

So far we have examined only a very short, very uncomplicated example. Some dialogues in the corpus extend over dozens of turns, and involve misunderstandings, negotiations, expressions of emotions such as frustration and humour, and other complex data. All of these will need to be captured to some extent in any reasonable model of human behaviour in this very limited domain.

There is another sense in which real task-oriented dialogues can be messy, namely when one participant tries to discuss a topic which falls outside the prescribed scope for dialogues of that type. In some cases, though the caller's immediate request cannot be resolved, it is possible to provide helpful information. An example of this type is shown in Extract 3.

Extract 3: agent handling inappropriate enquiry

```
A: Flight Information. May I help you?
C: Er::m, yes, I'm actually trying to
   make an enquiry about (.)
   Gibraltar Airways (.) which I
   think you (.) handle as well
   (.5)
A: Er, not on this number, sir. It's
   handled on seven five nine one
   eight one eight.
C: One eight one eight. Thank you
   very much.
A: That's alright, sir. Bye -bye.
C:                          -Bye.
```

Here the caller explicitly states the belief that the query falls within the scope of the service. However, the very act of doing so may indicate that this belief is not held very firmly (or at all!). On other occasions, callers are much more open about their confusion. In Extract 4, the caller launches straight into a detailed account of what she requires, pausing only to note that *'I'm not quite sure where to start'*.

Extract 4: complex inappropriate enquiry

```
A: Flight Information. Good afternoon.
C: Oh, hello (.) Erm, I've been asked
   to erm (.3) warn someone and I'm
   not quite sure where to start, but
   three passengers on a British
   Airways flight to er (.) New York
   will be cutting it a little bit
   fine this evening, uhm, but they
   should be there within sort of (1)
   couple of minutes of uhr:m (.7)
   check-in time (.) So who do I
   speak to about that?
   ...
```

This dialogue extends to thirty three turns before ending without the caller's problem having been solved. On the way it ranges over topics as diverse as the density of traffic in Central London, the ticket class of the delayed passengers and the normal procedures for check-in. It is much easier to envisage devising heuristics for identifying and partially satisfying queries of the type illustrated in Extract 3 than those of the type illustrated in extract 4.

We shall return to a different range of issues associated with out-of-domain tasks in section 5, below.

3. WIZARD-OF-OZ SIMULATION DATA

The requirement which originally motivated the collection of the human-human dialogue corpus introduced in the previous section was to furnish empirical data on the domain of flight information enquiries leading, ultimately, to the construction of a computer flight information system. The next step was to use that data to design a series of WOZ simulations in which subjects believed they were using a system, when they were, in reality, conversing with an experimenter whose voice had been disguised in order to sound synthetic. The basic WOZ technique is described in Fraser and Gilbert (1991a); the simulations themselves are described in Fraser and Gilbert (1991b) and MacDermid (1993).

In the first instance, a parallel corpus of WOZ dialogues was collected which could be compared closely with the human-human dialogue corpus. A detailed comparative study can be found in Wooffitt et al. (forthcoming). The WOZ corpus shows many simplifications in comparison to the human-human corpus across a broad range of issues including, lexis, syntax, turn-taking and dialogue organisation. For example, consider the relative orderliness of Extract 5.

Extract 5: simple WOZ dialogue
(S = simulated system; U = user; 'please wait' is the standard holding message used throughout this iteration of simulations)

```
S: Flight Information. Can I help you?
   (1)
U: Yes, I was wondering if you could
   tell me when British Airways flight
   number four eighty one (.3) is
   expected to arrive from Barcelona
   today
   (4.5)
S: Please wait.
   (15)
   Flight BA four eight one from
   Barcelona to London Heathrow
   Terminal One arrives at eleven
   twenty
   (1.3)
U: Thank you very much (.) B'bye.
   (3.5)
S: Thank you. Good bye.
```

It is now widely attested across different domains and languages that human-(simulated) dialogues are much more constrained than human-human dialogues collected in similar circumstances. However, they are not without interest, especially when the 'messy' phenomena are examined.

For example, consider the simulated system's opening question in Extract 5, *'Can I help you?'*. As in the human-human example of Extract 1, the user provides an explicit response to this before proceeding to formulate their enquiry. In the earlier discussion it was suggested that this kind of behaviour could result from a hedging of bets between reading the opening utterance as a direct or indirect speech act. However, this analysis does not stand up well when examined against a larger database. It turns out that there is a

strong tendency to preface the initial enquiry to a flight information service with some additional material. The nature of the material seems to be less important than its presence. It is almost as though *anything* will do, so long as it's not *nothing.*

For example, the following represents just a small sample of the range of first words produced by users in the simulations:

Yes, I'm just enquiring...
#h Erm, yes, I wonder if...
Ehm, can you tell me...
Yes, hello, erm, BA nine oh two...
Oh, yes please, uhm, I'm just...
Hello? Ehm, Can you tell me...
#h Er, I'm flying...

We shall call prefatory elements such as 'yes' and 'yes please' 'response tokens'. Elements such as 'erm' and '#h er' we shall call 'hesitation items'. It is not our task here to speculate on the subject of why they occur. Rather, we simply wish to note first, that they do occur non-randomly and, second, we wish to understand their patterning more clearly.

In our corpora, response tokens occurred in 87% of human-human dialogues, but in only 46% of WOZ dialogues. In the WOZ corpus, there was a significant tendency for individual users to produce response tokens in most dialogues (>80%) or in hardly any dialogues (<20%). Gender is significant here: typically males produce response tokens and females do not.

A somewhat different pattern emerges in the case of hesitation items. These occur in 35% of human-human dialogues and, surprisingly, in 35% of WOZ dialogues. There is a slightly greater tendency amongst males to produce hesitation items.

These may seem like very minor matters by comparison with larger issues such as those relating to the overall structure of the dialogue, and no doubt they are. However, it is important to note that what appeared at first sight to be no more that noise, no more than messy data, turns out to be structured and partially predictable. Initial user utterances play an extremely significant role in the flight information domain, so any data which can help us to build improved models for initial utterance understanding deserve to be weighed carefully.

In this section I have used a limited and relatively minor phenomenon to illustrate a general point. There are many more complex areas in which careful analysis reveals orderliness at the heart of what, at first sight, appeared to be chaotic data. Wooffitt et al (forthcoming) provides an extended investigation along these lines of issues relating to dialogue openings, request formulations, turn taking, overlap, closings and, perhaps most importantly, dialogue repairs.

4. HUMAN-COMPUTER DIALOGUE DATA

Having collected a human-human corpus and carried out a series of WOZ simulations, the next step was to produce a laboratory prototype dialogue system. (In fact, we practised and would advocate a methodology in which progressively more and more of the system is incorporated into a 'bionic' simulation.) Once an operational system has been developed, it too can be used to collect corpora for detailed analysis. Practical experience suggests that careful attention to qualitative analysis of data at this stage can yield invaluable insights which may be applied in improving the dialogue model and hence yielding significant performace enhancements.

Extract 6 shows part of a relatively successful dialogue in which a naive (untrained) user interacted with the system over a private exchange telephone line.

Extract 6: a spoken dialogue system performing well with an untrained user

```
S: Welcome to British Airways Flight
   Information Service. How can I
   help you?
U: Can you tell me which terminal at
   Heathrow BA flight five eight oh
   flies from?
S: Did you say British Airways flight
   five eight zero?
U: Yes.
S: British Airways flight five
   eight zero leaves from London
   Heathrow Airport Terminal 1 at
   ten fifteen. It arrives at Turin
   City Airport Terminal 1 at
   thirteen ten. Do you have another
   inquiry?
U: Yes.
S: What would you like to know?
U: Can you tell me what time the
   flight that leaves Madrid at
   thirteen hundred hours today
```

```
   arrives at Heathrow?
S: Do you know the flight number?
U: No.
S: Where does the flight leave from?
U: Madrid.
S: Did you say from Madrid?
U: Yes.
S: Where is the flight going to?
U: Heathrow.
S: Did you say to Heathrow Airport?
U: Yes.
S: What is the time of departure?
U: Thirteen hundred hours.
S: Did you say thirteen hundred hours?
U: Yes.
S: British Airways flight four six
   seven leaves from Madrid Barajas
   Airport at thirteen hundred hours.
   It arrives at London Heathrow
   Airport Terminal 1 at fourteen
   fifteen. Do you have another
   inquiry?
```

This dialogue was constructed in such a way that (for this trial) all major task parameters were confirmed explicitly, a precaution which proved to be tiresomely unnecessary in cases such as this. Part of the reason for the success of this example is that the user behaved in a highly constrained fashion, always answering exactly and literally the question asked in as few words as possible. Though the system was capable of coping with less constrained behaviour, it tended to respond better to more constrained than less constrained user behaviour.

As a general principle of dialogue system testing, I would argue that dialogue data decreases in interest the more orderly and successful it is. A wonderfully successful dialogue may draw gasps of admiration, but the real interest is to be found in analysing and understanding why dialogues have failed. It is an understandable fact of human nature that people do not enjoy publicising their failures. However, in the hope that it may encourage others to do likewise, I offer for scrutiny the painfully problematic exchange shown in Extract 7. In this extract, an untrained subject calls the system over long distance telephone lines. Whether there were problems on the line or problems in the recogniser, the fact is that the resulting recognition performace was atrocious, and this brought to the surface problems with the dialogue design which might otherwise have gone unnoticed. The user of the system also plays his part in the creation of this debacle as he persistently fails to learn how to maximise the system's performance. To aid understanding of what is going on, explanations will appear in italics throughout the extract.

Extract 7: a spoken dialogue system failing miserably with an untrained user

This was preceded by earlier dialogue material. This marks the beginning of a new task.

```
S: What would you like to know?
C: The arrival time of a plane leaving
   Madrid to Heathrow.
```

Complete misrecognition leads to understanding of an incorrect query.

```
S: Was that British Airways flight
   eight eight two?
C: I don't know.
```

Recognised 'no'

```
S: Do you know the flight number?
C: No.
S: Where does the flight leave from?
C: Madrid.
S: Did you say from Madrid?
C: Yes.
```

Heard nothing.

```
S: Was that from Madrid?
C: Yes.
```

Heard nothing.

```
S: Please answer yes or no. Was that
   from Madrid?
C: Yes.
S: Please tell me the destination.
C: Heathrow.
S: Was that to Heathrow Airport?
C: That's correct.
```

System is set up to recognise 'yes', 'no' or a destination name at this point — it is not able to recognise 'That's correct', which did not occur with significant frequency in any of the previous corpora. Mistakenly recognises 'Athens'.

```
S: Did you say to Athens?
C: No.
S: Did you say to Heathrow Airport?
```

If the user had confirmed 'Athens' it would have ruled out 'Heathrow' as a possibility. However, 'Heathrow' remains in the system with a status which is neither confirmed nor disconfirmed.

C: Yes, that's correct.

Misrecognised as 'Gatwick', a potential correction of 'Heathrow'.

S: Was that to Gatwick Airport?
C: No.

Misrecognised as 'Berlin' which, if confirmed, would disconfirm both 'Gatwick' and 'Heathrow'.

S: Did you say to Berlin?
C: No.

Correctly recognised disconfirmation of 'Berlin' thus bringing focus back to confirmation of 'Gatwick'.

S: Was that to Gatwick Airport?
C: No.

System has now lost confidence in 'Heathrow', which was first recognised but never confirmed too many turns ago. It opts to start this phase of the dialogue afresh.

S: Please tell me the destination.
C: Heathrow Airport.
S: Was that to Heathrow Airport?
C: Yes, that's correct.

A simple 'yes' would probably have been recognised, but the user's preference for verbose affirmative answers initiates a sequence of misrecognitions and failed confirmations which takes in 'Gatwick', 'Athens' and 'Berlin' all over again.

S: Was that to Gatwick airport?
C: No.
S: Was that to Athens?
C: No.
S: Was that to Gatwick Airport?
C: No.
S: Did you say to Berlin?
C: No.
S: Was that to Gatwick Airport?
C: No.

The system resets itself for a third time.

S: Please tell me the destination.
C: Heathrow.
S: Was that to Heathrow Airport?
C: Yes! Yes! That's correct.

At last the destination has been correctly recognised and the system can move on.

S: What is the time of departure?
C: Thirteen hundred hours.
S: Did you say thirteen hundred hours?
C: That's correct.

Success has been short lived — a verbose positive answer once again derails the dialogue, causing it to recognise an alternative time.

S: Did you say oh five fifty?
C: Could you repeat that?

Misrecognised as yet another time expression.

S: Was that oh eight oh three?
C: No.

The system resets itself.

S: Please tell me the departure time.
C: Thirteen hundred hours.
S: Was that thirteen hundred hours?
C: Yes, that's correct.

Here we go again — the system misrecognises the verbose positive as a time expression.

S: Did you say oh eight thirty?
C: No.

Heard nothing.

S: Did you say oh eight thirty?
C: No.

Heard nothing.

S: Sorry, I missed what you said. Was
   that oh eight thirty?
C: No.

System resets itself.

S: What is the time of departure?
C: Thirteen hundred hours.
S: Was that thirteen hundred hours?
C: Yes.
S: That flight arrives at fourteen
   fifteen. Do you have another
   enquiry?
C: yes

This is not a dialogue to be proud of, but it does illustrate some important points. First, it powerfully emphasises the difference between succeeding in performing some task and having a successful dialogue. Though the system manages to deliver exactly the information desired by the

user (even succeeding in getting the focus of the answer correctly fixed on the arrival time rather than some other task parameter), the dialogue is nonetheless a complete failure from the perspective of usability. Somewhere between the comfortable efficiency of Extract 6 and the painful convolutions of Extract 7 lies a point beyond which further progress is pointless, because the user has already been so alienated that even a successful resolution to the user's query will not result in a successful dialogue. This is especially important for public-facing commercial systems in which alienation of customers is to be avoided wherever possible.

This point had already been taken into account in the design of the system illustrated in Extract 7. It had a simple heuristic built into it which ensured that if the system got into the same state more than three times, it would terminate the call and pass the caller directly to a human agent. However, that eventuality never occurred in this dialogue. The thing which renders this dialogue frustrating is the repetition not of isolated states, but of larger patterns of behaviour. These may not be exactly identical on each iteration, but they are sufficiently similar to engender a feeling of inevitability of failure in the user. It can be extremely difficult to design a system so as to recognise such recurring patterns.

A simpler solution inspired by the tedious length of Extract 7 might be to assign a maximum number of turns to a dialogue of each type in a given domain. Assuming that the type of dialogue can be correctly identified, the dialogue can be abandoned if the maximum number of turns is exceeded.

Though there are many lessons to be drawn from data as messy as these, the last we shall consider in this section relates to the diversity of user behaviour. The reason why the system was so ill-equipped to deal with the user's verbose positive utterances (*'Yes, that's correct'*) was because the empirical evidence up to that time was that users tended to use highly constrained forms for functions such as confirmation, and this constraint tended to increase if misrecognition provoked repetitions. Unfortunately, the initial data set was too small and a larger data set shows that there are significant variations in behaviour from speaker to speaker.

The behaviour of speakers in general can only be judged on the basis of very large data sets. Exactly how large is an open research question, but they will undoubtedly need to consist of many more than is common today, where empirical materials based on tens, or at most hundreds, of speakers or dialogues are all that are available.

Incidentally, a single dialogue or dialogue extract does not furnish enough data to inform any judgements about the overall effectiveness of a given dialogue system. The abject failure of Extract 7 continued immediately with the (only slightly flawed) success of Extract 8.

Extract 8: the same spoken dialogue system performing rather better with the same user

```
S: What would you like to know?
C: The departure time from Heathrow
   of flight BA seven three two to
   Geneva.
S: Was that Air France flight eight
   eight seven?
C: No, BA seven three two.
S: Do you know the flight number?
C: Yes.
S: Please tell me the flight number.
C: BA seven three two.
S: Was that British Airways flight
   seven three two?
C: Yes, that's correct.
S: British Airways flight seven three
   two leaves from London Heathrow
   Airport Terminal 1 at nineteen ten.
   It arrives at Geneva at twenty one
   forty. Do you have another enquiry?
C: No.
S: Thank you for calling. Goodbye.
```

---

5. REAL SYSTEM UTILISATION DATA

The final step is to take a dialogue system out of the laboratory and into the real world, where it is used by ordinary people in large numbers to achieve goals for which they are highly motivated but completely untrained. I predict that over the next few years, real systems in the field will become a major focus for empirical investigation. There are several reasons for this. First, the amounts of data available for analysis will be vast, far beyond anything we are used to at present. For example, Vocalis currently has a simple dialogue system deployed in the field which is automating in the region of 200,000 telephone interactions per day (Peckham 1995). Second, the reliability of the data is likely to be greater than

we are currently used to. Data collected in real environments from real users is bound to be more *realistic*. Third, commercial requirements to evaluate investments in technology and develop systems with ever-increasing functionality will ensure that researchers will have particularly strong motivations to examine real-world systems.

In this section, I report briefly on some results obtained when a simple (i.e. fairly constrained) flight information dialogue system developed by Vocalis was made available to real users by an airline for a trial period. To protect the identity of the airline, I shall refer to them as 'XY Airlines'.

The focus in this section is on issues to do with the way in which the system was utilised and problems associated with evaluating the performance of the system in a real world context. This study was undertaken because XY Airlines reported that less than 20% of callers to the system were successfully completing dialogues. This led to the recording and analysis of a corpus of 122 dialogues.

The figure of less than 20% successful dialogue completion had been calculated automatically on the basis of counters in the dialogue system which were incremented very time the dialogue was started and every time the user explicitly accepted the answer offered. However, detailed analysis of the corpus of dialogues revealed the classification shown in Table 1.

The dialogue result categories should be interpreted as follows:

- Success - dialogue fully completed: the user followed the dialogue all the way through to the end and got the correct answer.
- Success - put the phone down: the user got the correct answer but put the phone down before the application terminated.
- Success - flight number not known: the system asked whether the caller knew their flight number, the caller said 'no' and was correctly transferred to an agent without further ado.
- Success - asked for agent: the caller asked to speak to an agent and was instantly transferred. In all but one case this happened in the first utterance. In the remaining case the dialogue had proceeded smoothly until the caller realised he didn't know the date of travel and asked to speak to an agent.
- Success - system performed well until user put phone down for no apparent reason: there

Dialogue Result Category	%
Success - dialogue fully completed	15.8%
Success - put the phone down	8.3%
Success - flight number not known	19.2%
Success - asked for agent	7.5%
Success - system performed well until user put phone down for no apparent reason	5.0%
Success - system responded to incomprehensible initial user utterance by switching caller to an agent	0.8%
Failure - inappropriate user behaviour	9.2%
Failure - misrecognition	6.7%
Loud background noise - can't tell what user said	1.7%
Inaudibly quiet speech	2.5%
Silence	22.5%
Can't interpret what happened	2.5%

Table 1: summary of dialogue outcomes

Success categories	77.6%
Failure categories	22.4%

Table 2: overall dialogue success results (adjusted)

	Well-formed
Flight numbers	83%
Dates	86%

Table 3: users' ability to follow instructions

is no evidence that anything on the systems' part led to the termination of the dialogue.

- Success - system responded to incomprehensible initial user utterance by switching caller to an agent: it seems reasonable to treat cases of this kind as instances of successes on the grounds that the system correctly identifies that something is going on which it is unlikely to be able to handle, so it passes the call to an agent immediately.
- Failure - inappropriate user behaviour: the user is acting strangely, e.g. not following explicit instructions.
- Failure - misrecognition: these are simple failures on the part of the technology.
- Very loud background noise - can't tell what user said: the recording is just too noisy to tell what's going on.
- Inaudibly quiet speech: a sound can be detected which is discernibly speech but it is impossible to tell what's being said.
- Silence: XY Airline's telephone system does not allow us to tell when the telephone has been put down. We have no way of detecting this except assuming that a hangup has occurred if three silences in a row are detected. However, it is possible that callers are completely baffled by the system or alienated by it and listen to it in silence until it transfers then to an agent.
- Can't interpret what happened: There is insufficient data to reconstruct what happened in the dialogue.

How well is the system performing? In my analysis it is doing considerably better than the initial figure of around 20% suggested. If we remove the last four categories from the score (because for very loud background noise, inaudibly quiet speech and silence a human couldn't have done any better; the final category is just noise) we are left with the following overall result:

Though these figures are much better than what was originally supposed, they are not good enough. It is necessary to account for the very high number of completely silent calls (22.5%). As mentioned before, one explanation could be that users are so non-plussed by what they hear that they just put the phone down. If this is true then all that is needed is to revise the system prompt design so that it takes more account of human factors, thus making it easier for callers to start talking to a machine.

However, this is not a convincing explanation. In the majority of calls in this sample, users follow the instructions effortlessly. They are, on the whole, very good at saying the required flight number in the rather constrained way required by this system, without any extraneous material such as 'XY' or 'flight number'. The same is true of dates. The results are summarised in Table 3.

In most cases where a user gets the format wrong, they get it right next time (having heard the explanatory prompt repeated). Re-scripting the system messages is unlikely to achieve much here, and the tedium of longer messages is unlikely to be rewarded by vastly increased performance.

Initial utterances are similarly compliant with the instructions, though there is a greater tendency for users to say nothing, hear the prompt being repeated, and then get things right on their first attempt.

A very simple solution lies outside the dialogue system altogether. In our early work on human-human dialogue we examined the nature of the queries being received by British Airways' Flight Information Service. In spite of the fact, that the service exists solely to supply information relating to arrivals and departures of British Airways (and associated) flights, callers attempt to carry out a variety of different tasks. Of these, as many as 30% could be categorised as falling outside the scope of the service. It is thus quite possible, if there is any similarity between the service of British Airways and that of XY Airlines, that the 'silent' dialogues in this real world system can be accounted for straightforwardly in terms of callers

hearing the very clearly focused system greeting message and request for initial input, and realising that they have called the wrong number for the service they require, leading to the telephone being hung up.

The object of this section has not been to establish exactly how well a particular real world system is performing. Rather it is intended to raise the focus of our discussion to the wider context in which dialogue systems will be deployed and evaluated. There are many kinds of data other than linguistic data which will need to be examined before a clear understanding emerges of spoken language dialogue systems as usable artifacts.

6. CONCLUSIONS

This paper has presented data extracts from corpora of various different kinds: human-human dialogue, simulated human-computer dialogue, actual human-computer dialogue in the laboratory, and data associated with the deploying of a human-computer dialogue system in the real world. One of the messages of this paper, then, is a strong affirmation of the value of empirical data at all stages of spoken dialogue system development.

The major message of the paper, however, is that data must not be divided too readily into 'clean' data and 'messy' data, since this is to prejudge the theoretical significance of material for which scant theoretical apparatus currently exists.

I am convinced that the spoken language systems which will succeed in the long run will be those which pay most attention to modelling the 'messy' phenomena of which real world dialogue is largely composed.

ACKNOWLEDGEMENTS

This work builds on the results of a range of different projects in which different people have participated. For their efforts and for the benefit of their wisdom I thank Nigel Gilbert, Catriona MacDermid, Andrew Simpson, Trevor Thomas, Simon Thornton, and Robin Wooffitt.

REFERENCES

Chomsky, Noam (1965) *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Fraser, Norman M. and G. Nigel Gilbert (1991a) Simulating speech systems. *Computer Speech and Language* 5:81–99.

Fraser, Norman M. and G. Nigel Gilbert (1991b) Effects of system voice quality on user utterances in speech dialogue systems. *Proceedings of the 2nd European Conference on Speech Communication and Technology,* Genoa, September, 57–60.

Grice, H.P. (1975) Logic and conversation. In Cole, P. and Morgan, J. (eds) *Syntax and semantics 3: Pragmatics.* New York: Academic Press.

MacDermid, Catriona (1993) Features of naive callers' dialogues with a simulated speech understanding and dialogue system. *Proceedings of the 3rd European Conference on Speech Communication and Technology,* Berlin, September, 955–958.

Peckham, Jeremy (1993) A new generation of spoken dialogue systems: results and lessons from the SUNDIAL project. *Proceedings of the 3rd European Conference on Speech Communication and Technology,* Berlin, September, 33–40.

Peckham, Jeremy (1995) Conversational interaction: breaking the usability barrier. *Proceedings of the ESCA Workshop on Spoken Dialogue Systems: Theories and Applications.* Vigso, May, 1–8.

Peckham, Jeremy and Norman M. Fraser (forthcoming) *Speech understanding and dialogue.* Cambridge, MA: MIT Press.

Wooffitt, Robin C., Norman M. Fraser, G. Nigel Gilbert and Scott McGlashan (forthcoming) *Designing interaction: a conversation analytic study of human-(simulated) computer interaction.* London: Routledge.

Predicting and interpreting speech acts in a theatre information and booking system

Toine Andernach
Parlevink Group, Department of Computer Science
University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands

E-mail: andernac@cs.utwente.nl

Abstract

This paper discusses an approach to speech act analysis in the context of a theatre information and booking system which aims at exploiting as many superficial utterance clues as possible. These clues, combined with contextual information about the dialogue like the dialogue history and predictions derived from the current dialogue and from a test corpus, should enable us to model the course of a dialogue in such a way that a user feels comfortable with it (Dahlbäck & Jönsson 1986).

1 Introduction

In this paper, we present our ongoing work on speech act analysis (communicative functions as we call them) for modelling user intentions in man-machine dialogues. We will give support for our hypothesis that communicative functions of utterances can (at least partially) be determined by using lexical and structural information like words, word order and dialogue history; we will regard communicative functions as partial and multi-dimensional descriptions of what people want in dialogues.

Our empirical basis is a corpus of 64 dialogues collected in a Wizard of Oz environment; the context of the research reported here is Schisma (see section 4), a joint research project of KPN (Royal PTT Nederland) and the University of Twente.

In section 2 we claim that the restrictedness of a dialogue system combined with the awareness of the user of that system enable us to model the course of a dialogue in such a way that a user is satisfied with it.

In section 3 we will discuss some discourse theories, their problems and their attractive aspects and how we profit from the latter.

Section 4 gives an overview of the way we determine the communicative functions of user utterances in our dialogues. We discuss the kinds of communicative functions we consider the ones to be captured in our system and we will discuss some superficial utterance clues; form features of utterances like word order and the presence of a question mark. Other form features we discuss are lexical clues for communicative function.

Section 4.5 gives a quick glance of the rules we will use for mapping form features of utterances with their possible function.

In section 4.6 dialogue structural information like information about the previous utterances is discussed in the light of its use for predicting communciative functions. Special attention is paid to adjacency pairs.

In section 4.7 a method is presented for disclosing relevant information from a corpus and in section 5 we will discuss some future research.

2 Restricted language

In the past decennia, several discourse theories have been proposed. They did not all have the same purpose: some intended to account for the way *people* use (certain phenomena in) language, either in texts, monologues or dialogues. Others aimed at providing computational models for human discourses and yet others aimed at designing discourse models for the development of NLP applications.

The difference between the second and the last is that the latter does not necessarily model man-man linguistic behaviour; users are often aware of the restrictedness of the language by machines; e.g. in a restricted domain the number of (interpretations of) content words, syntactic constructions is limited and there are more semantic and

contextual restrictions. This awareness causes users to adapt their language. As Morel (1989) found in experiments, subjects often adapt their language when they talk to machines; machine-like voice and behaviour do have influence on the linguistic behaviour and the prosody of the users of a dialogue system.

The interplay between man's and machine's behaviour is also observed by Smith, Hipp & Biermann (1992). They assume that the input of the user at a certain point in the dialogue can be predicted in a subdialogue; this subdialogue specifies the focus of the interaction and thereby the number of possible interpretations is limited. More specifically, we follow Waterworth (1987) in his claim that a classification of functions of utterances can be possible and useful as long as we restrict ourselves to identifying and implementing dialogue strategies in a restricted domain.

Therefore, we do not a priori assume that man-machine dialogues proceed the way man-man dialogues do; it is *not* our primary goal to model man-man dialogues in a psychologically and theoretically plausible way although these factors might play a role. People must be able to get information about performances in theatres by using natural language *in such a way that they feel comfortable with it* (Dahlbäck & Jönsson 1986) and that is not necessarily the way they talk to other people.

3 Discourse theories

3.1 Discourse Analysis

In discourse analysis (DA), linguistic techniques are applied to discourse entities larger than the sentence in order to model human linguistic behaviour. An example of a discourse analytic approach is the use of discourse grammars. Basic categories of utterances are identified and concatenation rules are formulated.

According to Levinson (1983) two main problems with discourse analysis are the strict theoretical nature of it and the intuition-based claims of its researchers. These problems could be overcome if we could find generalisations in the structure of *realistic* discourse; discourse grammar rules thus found could be used in a system like ours. The problem however, is that usually, realistic discourses do not obey this kind of rules; the nature of the discourse grammar rules presupposes that the structure of the discourse is fixed. We have seen in our corpus however, that the structure of dialogues is far more flat than often assumed.

3.2 Conversation Analysis

In Conversation Analysis (CA) the emphasis lies on the collection of empirical data while the premature construction of theories is avoided. In naturally occurring conversations, systematic properties of the sequential organisation of conversations are searched for. Conversational analysis is rule-governed and the underlying idea is that shared knowledge of these rules most often enables conversants to have smooth flowing and coherent conversations with one another.

In its main goals CA very well satisfies our needs: first, we base our computational model on empirical data because they simply are more reliable than intuitive data. Second, we think that the dependency relation between an utterance and the utterance immediately preceding it, is more easy to exploit in a dialogue system than the hierarchical structure which is traditionally more emphasized in literature. And third, considering utterances as containing cueing devices used by the speaker seems attractive from an engineering point of view.

3.3 Speech Act theory

The main idea of Speech Act theory (SA) is that utterances do not only have a literal meaning, but perform specific actions (*speech acts*) as well. In SA three aspects of speech acts are distinguished: its *locution*, its *illocution* and its *perlocution* (see (Austin 1962)). The illocution is often considered to be the identifying characteristic of a speech act; it expresses the action executed by the utterance. We will regard the illocution as the most prominent aspect of utterances in dialogues on which the proceeding of a dialogue is based. Therefore, we will concentrate on this aspect here.

Levinson (1983) reports several problems with speech act theory. Among them are:

1. there is no one-to-one mapping between utterances and acts (one utterance can be associated with multiple acts and one act can be performed in multiple utterances)
2. there is no simple form-to-force correlation

We will circumvent these problems by regarding communicative functions as partial and multi-

dimensional descriptions of communicative functions. We assume that every utterance gives at least some clues for these functions.

3.4 Plan theory

Since the late seventies, researchers have tried to apply *Plan Theory* to the generation and interpretation of plans in discourse. A basic assumption among plan theorists is the fact that the linguistic behaviour of agents in information dialogues is *goal-directed*; an agent's goal is to reach a particular state. A term which is often used for this kind of goals is *intention*. A recipe for reaching a particular state is often called *a plan*. It can consist of a number of subgoals each of which can be realised by *a subplan*. Thus, in a plan, goals can be represented in a tree structure in which dominated goals must be reached in order to reach dominating goals (see for instance (Litman & Allen 1990) and (Lambert & Carberry 1991)). Speech acts are considered to be the primitive goals to be met.

A major disadvantage of plans represented as tree structures is that every (non-terminal) node assumes its dominated goals to be fullfilled; all possible plans are fixed for each of their subgoals. That means that we need an extra mechanism to cope with situations in which plans change. Furthermore, Penstein Rosé, Eugenio, Levin & Ess-Dykema (1995) showed that a tree structure is not adequate in cases where dialogues have multiple threads.

Ahrenberg, Jönsson & Dahlbäck (1991) criticised this approach in that they don't consider it to be necessary to model the whole range of plans a user can have; in Grosz & Sidner (1986) intentional structure (i.e. the structure of a user's plans) is isomorphic to dialogue structure and it is as least as difficult to determine the former as it is to determine the latter. Therefore, Ahrenberg et al. (1991) propose a *structural approach* to dialogue modelling; they claim that it is sufficient to use simple discourse plans which consist of two parts, an opening move and a closing move. This idea also stems from CA and fits very well with our idea of dialogue cohesion, which in fact stems from Halliday & Hasan (1976).

4 Communicative functions in Schisma

4.1 Schisma: an introduction

In Schisma we aim at providing a natural language dialogue system which interfaces a database containing information about theatre performances in a certain city or region. The interface should make it possible to ask about performances in general, to tune in to a specific performance and, if desired, make a reservation for this performance. Research until now has concentrated on various aspects of realising such a theatre information and booking system. Among these aspects are the building of a Wizard of Oz environment for the acquisition of a corpus of dialogues for this domain, analysis and tagging of the dialogue corpus, recognition of domain-specific concepts (actors, authors, plays, dates, etc.), syntactic analysis and dialogue modelling.

We are especially interested in the user's goal when he produces an utterance and how he realises that goal in language. We assume that to allow a flexible man-machine dialogue, the *communicative function* of an utterance of a user must be determined. We prefer the term *communicative function* instead of *speech act* because it is a more meaningful term, but we use both terms.

In the Schisma system (see figure 4.1) a special component (the Speech Act Analyser) will be developed. It will get its input from the parser and its output will be transferred to other dialogue managing components.

Like in Conversation Analysis, we assume that there is a strong interdependence between what speakers want and the way they choose their utterances, i.e. between form and function of utterances in a dialogue. We will exploit this interdependence for our system; the more we can rely on superficial information in the utterances for this task, the more computationally attractive this will be. Grosz & Sidner (1986), who were among the first to present a rather integrated computational theory of discourse structure, also stressed the significance of using superficial linguistic clues for identifying structures in discourse. [1]

In an integrated approach to dialogue modelling like (Traum & Hinkelman 1992), traditional speech acts are extended to account for certain

[1] See (Hinkelman 1990) for an illustration of the exploitation of superficial linguistic clues.

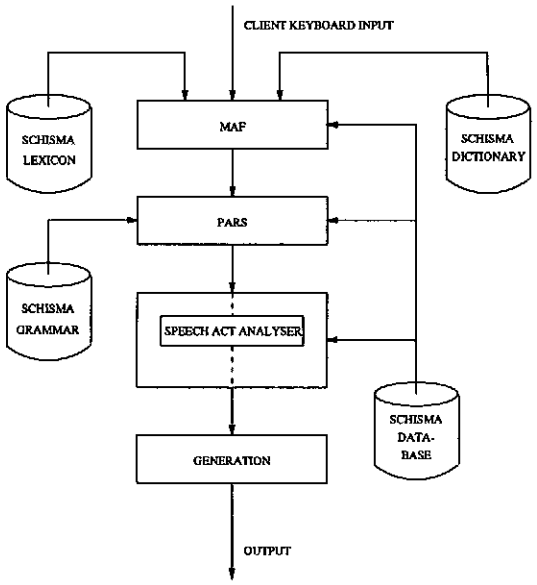


Figure 1: Global architecture of *SCHISMA*

types of coordinated activity that take place between agents in a conversation. Linguistic actions are signalled directly by surface features of the discourse, although usually a combination of surface features and *context* will be necessary to disambiguate acts.

The context in our system at a certain point in the dialogue will consist of a restricted number of possible transitions to communicative functions (determined by a Finite State Automaton), the form and function of previous utterances in the current dialogue, a list of preferred transitions to communicative functions (determined by statistical analysis of a test corpus) and, implicitly, the domain itself, containing certain domain concepts.

4.2 The form-function dichotomy

Since long ago, *the form-function dichotomy* has been recognised as a problem for determining the communicative function of utterances; corresponding utterance forms may have different functions and corresponding functions can be realised by different forms. Examples (1 and (2) taken from the Schisma corpus should make this clearer:

(1) *Kan ik drie kaartjes reserveren voor*
Can I three tickets reserve for
de eerste rij?
the first row?
'Can I reserve three tickets at the first row?'

(2) *Ik wil een kaartje reserveren.*
I want a ticket reserve.
'I would like to reserve a ticket.'

In distinct contexts, (1) can function both as a check for the truth of the proposition and as a request to execute the specified action. Furthermore, (1) and (2) are both requests for reserving (a) ticket(s) while their sentence types differ: (1) is a yes/no question about the speaker's ability to reserve tickets and (2) is a statement.

Crucial in this dichotomy, however, is the definition of the word *form* in this context; usually it is regarded to be the sentence type or mood of an utterance. But would the presumed dichotomy still exist when we take other form features of utterances into account? It is our main hypothesis that the form-function dichotomy can be circumvented by taking into account the form-features and contextual aspects as discussed in the former subsection.

In the next sections we will discuss some form features and our viewpoint of communicative function and how to relate them.

4.3 Communicative functions

After a closer look at the dialogues in our corpus, we found that the main function of all utterances in the corpus are either to *supply* something or to express a *wish* for something (see also (Wachtel 1986)).

We found that the objects of wishing and supplying can be *actions*, *information* and *truth values*. Combining them we get the Cartesian Product of these domain independent dimensions:

1. wish for action
2. supply of action
3. wish for information
4. supply of information
5. wish for truth value
6. supply of truth value

(1) and (2) can have instantiations in a specific domain like *reserve* or dialogue control instantiations like *thank* or *greet*, (3) and (4) concern concepts like *performance* or *actor* and (5) and (6) can for instance be expressed by *yes*, *no* or *ok*.

In an ideal situation, every utterance would give clues for each of these dimensions; in practice however, utterances will give us clues on just a subset of these dimensions. We use the word *dimensions* instead of *levels* because we don't think that a hierarchical classification (i.e. a *taxonomy*) of communicative functions satisfies our needs. Hinkelman (1990)'s taxonomy for instance, expresses that if there is evidence for a certain, say domain-dependent, communicative function, it implies that there is evidence for *all* dominating (more abstract, domain-independent) functions.

In our system, however, we don't want to be forced into a pre-fixed structure of the dialogue. A taxonomy in which domain-independent types of communicative functions dominate more specific domain-dependent functions is not suitable for our purposes because utterances can give clues on each of these dimensions independently:

(3) *En Othello?*
And Othello?
'What about Othello?'

In (3), *Othello* is the name of a performance. The question mark indicates the interrogative force, i.e. in our terminology it is a wish concerning a performance. The word *En* is a clue for the rhetoric relation with the former utterance. However, there is no clue for the kind of question that is meant in (3) (wh or y/n). So, despite of the fact that not every aspect of the domain independent features can be determined, we would still like to be able to account for the domain dependent information in an utterance.

In the following section, we will discuss the superficial clues for the communicative function of an utterance.

4.4 Form features

4.4.1 Sentence type

The first form feature we use for determining the communicative function is *the sentence type* of utterances. Table 4.4.1 is used to determine this sentence type.

The second column labelled with *verb 2nd/1st* indicates whether the finite verb is in second or in first sentence position. The column *subject* indicates the presence of a subject and the column *special* indicates some type-specific features.

The special sentence type *utterance* is introduced for the sentence type of all utterances that cannot be assigned another sentence type. Typical examples are utterances without a finite verb, like noun phrases and other constituents, affirmatives and greetings.

Type	verb 2nd/1st	subject	special
declarative	2nd	+	
imperative	1st	−	imp. verb form
y/n question	1st	+	
wh question	2nd	+	fronted wh-term
utterance			

Table 1: Sentence types

4.4.2 Punctuations

The presence of a question mark is a strong indication for a request. It is a *sufficient* condition in our corpus. Many utterances have a question mark while in the meantime having declarative sentence types:

(4) *2 zei ik toch?*
2 said I, didn't I?
'I said 2, didn't I?'

This observation cannot be generalised; in other contexts for instance, rhetoric questions can occur which in general don't have an interrogative function. See (Beun 1989) for a discussion of so-called *declarative questions* like (4).

On the other side, a question mark does not appear to be a *necessary* condition for an interrogative function: the corpus appears to contain a lot of utterances with interrogative functions, but without a question mark:

(5) *Wanneer is* Silicone Kitty
When is *Silicone Kitty*
'When does *Silicone Kitty* play'

4.4.3 Wh-words

One of the *lexical* clues for the communicative function is the presence of a *wh-word*. In almost all of the cases, the wh-word occupies the first position of the utterance or is part of a preposition phrase in subject position. Of the 62 occurrences of *Wat* for example, 58 are the first word of an utterance. Two of these are used in an exclamatory phrase. Of the other four, two are not interrogative pronouns, one starts with the conjunctive *En* and one is the first word of a subordinate clause.

Examples of utterances containing an interrogative pronoun which is not in first position are (6) and (7):

(6) *Wanneer worden welke voorstellingen*
When are which performances
gegeven van Het nationale toneel
given of *Het nationale toneel*
'When does *Het nationale toneel* perform which plays?'

(7) *Wanneer en hoe laat is*
When and how late is
Under a blue Roof*?*
Under a blue Roof?
'When and at what time will *Under a blue Roof* be played?'

(6) is a special case of a wh-question; two concepts are questioned in one utterance; the same counts for (7) although it is an elliptic utterance, contrary to (6).

4.4.4 Cue phrases

Another kind of lexical clues for the function of utterances in discourses are *cue phrases* (also called *clue words, discourse markers, discourse connectives* or *discourse particles*). Most cue phrases are realised as modal adverbs or adverbial phrases and they are traditionally regarded as explicit indicators of the structure of a discourse. They can e.g. mark a topic introduction, a topic shift (*now*) or a side step (*by the way*).

According to Hirschberg & Litman (1993) structural information conveyed by clue words is crucial to many more tasks:

- anaphora resolution
- inference of speaker intention
- recognition of speaker plans
- generation of explanations and other texts

As, we are mainly interested in the second task and...

> *"...despite the crucial role that cue phrases can play in theories of discourse and their implementation, however, many questions about how cue phrases are identified and defined remain to be examined..."*
> *(Hirschberg & Litman, 1993)*

...we will now have a look at some cue phrases in our corpus, more particularly the words *graag* and *niet*.

Literally, *graag* can be translated as *like to*. In dialogue however, it is often used as a more general politeness marker:

(8) *Ik wil graag naar* Mini en Maxi.
I want very much to *Mini en Maxi.*
'I would like to go to *Mini en Maxi*'

In all 54 cases of *graag* in the corpus it occurs in a declarative utterance. In 36 (67%) of the cases, the word *wil* (want) occurs in the same utterance. 3 other cases the word *zou* has the same function as the word *wil*.

In 10 (18%) of the utterances with *graag*, a verb is lacking and some concept is mentioned as a reply to a question of the Wizard. In 4 cases (7%), *graag* is meant as a confirmation of an immediately preceding yes/no-question of the Wizard. Most of the cases (3) accompanied by the word *ja* (yes). One occurrence of the idiomatic expression *graag gedaan* (it's a pleasure) was found.

A bigram analysis at word level of the utterances of the user yielded the highest frequency for the bigram *Ik wil* while a trigram analyis yielded the highest score for *Ik wil graag*.

To summarise we can say that *graag* supports (strengthens) the wish for information or action; this wish can be implicit (e.g. in the form of a (implicit or explicit) confirmation or choice) or explicit in the form of a wish marker, e.g. the verb *wil*. More specifically:

- (in combination with *ja*) in support of the confirmation of an information or action provision
- in combination with a domain concept in support of the confirmation of a wh or alt question.
- in combination with the explicit wish marker *wil* in support of the request

Another word that can be used as cue phrase is *niet*: in the following examples, *niet* does not function the way it usually to does, as a negation marker:

(9) *Kunt u ze trouwens niet*
Can you them by the way not
opsturen?
send?
'By the way, couldn't you send them to me?'

It seems that if *niet* is omitted, the (logical) meaning remains approximately the same. The

question then is: what does *niet* add in utterances like (9) in the corpus? What they have in common is that they all have interrogative force. This is marked by the y/n question word order.

Let's see what happens if we change (9) in (10):

(10) *U kunt ze trouwens niet*
You can them by the way not
opsturen
send
'You can't send them by the way'

(10) can only have the meaning intended in (9) if it has a rising intonation. With a default declarative intonation, *niet* serves the purpose of negating the proposition expressed in the utterance.

Thus, it seems that the special use of *niet* only occurs in utterances in which the speaker expresses a request. In these directive utterances, the speaker uses *indirectness* techniques to avoid that the speaker will feel forced to obey the speaker. Negating the proposition is one way of doing that. The speaker could also have used the word *misschien* (maybe) which expresses uncertainty by the speaker.

Examples (8) and (9) show that clue words can be very subtle indications for speaker intentions in discourse, very often in combination with other clues in the utterance.

4.5 Formalising the interpretation of communicative functions

Following Hinkelman (1990) we will use rules to determine for a certain input utterance a range of possible partial speech act interpretations. The rule below is an example of the kind of rules given by Hinkelman (1990). It is applicable to (1) above.

```
(S MOOD YES-NO-Q
 VOICE ACT
 SUBJ (NP HEAD ik)
 AUXS {kan}
 MAIN-V +action) => ((REQUEST-ACT ACTION)
                                (SPEECH-ACT))
```

Both structures at the left hand side and the right hand side of the arrow contain features with their values. This rule is applicable if the structure at the left matches (a substructure of) the structure yielded by PARS. The right hand side of the rule is a disjunction of partial descriptions of communicative functions.

4.6 Predicting communicative functions

To optimise the process of assigning communicative functions we could use a Finite State Automaton (FSA) to a priori exclude some communicative functions at a certain point in a dialogue. Such an Automaton is also used in the Verbmobil project (Alexandersson, Maier & Reithinger 1994). In this project, speech acts are both modelled in an FSA which restricts the sequential order in which the speech acts are used and hierarchically modelled in a taxonomy.

The necessary states and transitions in this FSA could be determined by using a test corpus in which the communicative functions are tagged (see section 4.7) or by using common or intuitive knowledge about the sequence of utterances. In the latter case we should be aware of the fact that the word *common* in *common knowledge* does not make the knowledge more reliable.

A rather new way predicting communicative functions is by statistical information (Reithinger & Maier 1995). A finite state model is not sufficient for the prediction task because it is not sufficiently restrictive. Therefore, we will use information about relative frequencies of sequences in a test corpus. This results in information about *adjacency pairs* and *preferred seconds.*

Adjacency pairs consist of two turns each uttered by another speaker. One of the characteristics of the parts of these pairs is their adjacency. Levinson (1983) notices that, instead of occurring strictly adjacent, the parts of an adjacency pair are frequently split up by so-called *insertion sequences* which also consist of adjacency pairs. An example from the corpus is (11):

(11) *S: Hoeveel kaartjes wilt u en met welke reductie?*
C: hoeveel kaartjes zijn er nog?
S: Er zijn nog 400 plaatsen vrij voor deze voorstelling.
C: Doe maar tien

In (11), the second and third turn form an insertion sequence.

Furthermore, Levinson (1983) indicates a problem with the feature of *typedness*, i.e. the fact that a particular first part requires a particular second part. In natural dialogues it is not the case that for instance *an offer* is always followed by an *acceptation*; it can also be followed by a *rejection*. Thus, instead of a strict coupling of

both parts, we could assume a *preference organisation*; an acceptation of an offer is preferred to a rejection of that offer.

This kind of information will be yielded by the statistical analysis of our tagged test corpus.

4.7 Tagging: a systematic way of information disclosure

One way of systematically disclosing the varied amount of information for dialogue management in the corpus is *tagging*; certain characteristics of words, utterances or sequences of utterances are annotated in such a way that common features can be found by statistically processing these annotations.

In order to test our hypothesis that the combination of several superficial clues would enable us to determine (at least some aspects) of its communicative function, we will tag our corpus. To avoid an explosion of feature combinations we will tag three form features: sentence type, presence of a wh-word and presence of a question mark) and four function features (the speaker, the main act (request or provide), the object of the act (information, action or truth value) and the domain-dependent instantiation of the object. The following list gives an overview of the dimensions and their instantiations:

Form:

1. word order
   (a) utterance
   (b) declarative
   (c) y/n
   (d) wh
   (e) imperative
2. presence of a wh word
   (a) yes
   (b) no
3. presence of a question mark
   (a) yes
   (b) no

Function:

1. speaker:
   (a) user
   (b) system
2. domain-independent function classes:
   (a) request
   (b) supply

Domain-independent concept classes with their domain-dependent instantiations:

1. action
   (a) reserve
   (b) annulate
   (c) thank
   (d) greet
2. constant, information
   (a) (ITEM)
   (b) (EMPTY)
3. truth value
   (a) yes
   (b) no
   (c) (NOT KNOWN)

Examples of items are: performance, time, costs, seats, reduction, rank, payment method, address theatre, reservation, information, date, number of people, summary, getting tickets at office

A corpus with utterances characterised this way, can be analysed in several respects: sequences of the tag types, n-grams of clusters of m features of utterances.

Relative frequencies of bigrams of *form* and *function* features of user utterances in the test corpus are used to formulate the rules which map the form with the function features.

Relative frequencies of n-grams of *function* features of both user and system utterances are used in preference rules for predicting communicative functions; after the FSA has restricted the number of potential communicative functions at a certain point in a dialogue, these preference rules will order these functions on a scale ranging from most probable to less probable. This ordening will be used to optimise the process of assigning communicative functions.

5 Future Research

In further research we will test our hypothesis that the communicative function of utterances in man-machine dialogues can be determined using superficial information from the utterances themselves; all utterances will be tagged the way we described in this paper and we will analyse these tagged utterances and improve our current Simulation Environment with a dialogue model based

on the results of this analysis. This environment is used to semi-automatically collect Wizard of OZ dialogues. It will serve as the platform for our eventual prototype; it will be extended with other modules to be developed.

Next step is to collect Wizard of Oz dialogues in a theatre, a more realistic environment, usingthe improved Simulation Environment.

References

Ahrenberg, L., Jönsson, A. & Dahlbäck, N. (1991), Discourse representation and discourse management for a natural language dialogue system, Research report, NLPLAB IDA Linköping University, Linköping, Sweden.

Alexandersson, J., Maier, E. & Reithinger, N. (1994), A robust and efficient three-layered dialogue component for a speech-to-speech translation system, Report 50, DFKI GmbH, Saarbrücken, Germany.

Austin, J. (1962), *How To Do Things With Words?*, Clarendon Press, Oxford.

Beun, R.-J. (1989), The Recognition of Declarative Questions in Information Dialogues, PhD thesis, Instituut voor Perceptie Onderzoek, Eindhoven, The Netherlands.

Dahlbäck, N. & Jönsson, A. (1986), A system for studying human-computer dialogues in natural language, Research report, NLPLAB IDA Linköping University, Linköping, Sweden.

Grosz, B. & Sidner, C. (1986), 'Attention, intentions and the structure of discourse', *Computational Linguistics* **12**(3), 175–204.

Halliday, M. & Hasan, R. (1976), *Cohesion in English*, Longman, London.

Hinkelman, E. (1990), Linguistic and Pragmatic Constraints on Utterance Interpretation, PhD thesis, University of Rochester, Rochester.

Hirschberg, J. & Litman (1993), 'Empirical studies on the disambiguation of cue phrases', *Computational Linguistics* **19**(3), 501–530.

Lambert, L. & Carberry, S. (1991), A tripartite plan-based model of dialogue, *in* '29th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference', ACL, Berkely, California, USA, pp. 47–54.

Levinson, S. (1983), *Pragmatics*, Cambridge University Press, Cambridge.

Litman, D. & Allen, J. (1990), Discourse processing and commonsense plans, *in* P. Cohen, J. Morgan & M. E. Pollack, eds, 'Intentions in Communication', The MIT Press, chapter 17, pp. 365–388.

Morel, M. (1989), Computer-human communication, *in* M. Taylor, E. Néel & D. Bouwhuis, eds, 'The Structure of Multimodal Dialogue', Elsevier Science Publishers B.V., Amsterdam, North-Holland, chapter 24, pp. 323–330.

Penstein Rosé, C., Eugenio, B. D., Levin, L. & Ess-Dykema, C. V. (1995), Discourse processing of dialogues with multiple threads, *in* '33rd Annual Meeting of the Association for Compuational Linguistics', ACL, Cambridge, Massachusetts USA.

Reithinger, N. & Maier, E. (1995), Utilizing statistical dialogue act processing in verbmobil, *in* '33rd Annual Meeting of the Association for Compuational Linguistics', ACL, Cambridge, Massachusetts USA.

Smith, R., Hipp, D. & Biermann, A. (1992), A dialog control algorithm and its performance, *in* 'Proceedings of the Third Conference on Applied Natural Language Processing', ACL, ACL, Trento, Italy, pp. 9–16.

Traum, D. & Hinkelman, E. (1992), Conversation acts in task-oriented spoken dialogue, Technical Report 425, University of Rochester, Rochester.

Wachtel, T. (1986), Pragmatic sensitivity in nl interfaces and the structure of conversation, *in* 'Proceedings of COLING', Bonn, pp. 35–41.

Waterworth, J., ed. (1987), *Speech and Language-Based Interaction with Machines: Towards the Conversational Computer*, Ellis Horwood Limited, Chichester.

PROJECT PARLEVINK

Linguistic Engineering
University of Twente

PARLEVINK Research Topics

The Parlevink Project started in January 1992. It did not start from scratch. In previous years research took place in the area of theory of formal and programming languages (theoretical computer science, compiler construction) and more and more this research became influenced by potential applications in the area of natural language processing. Currently the following three research directions are distinguished:

- research which concentrates on syntactic formalisms and where syntax is the starting point for studying the description and processing of semantic and pragmatic aspects of language;
- research which concentrates on the representation of meaning in dialogue modelling and where syntax is of secondary importance;
- research which concentrates on modelling language behaviour with the help of neural networks and where language learning and integrated use of syntactic, semantic and pragmatic knowledge are the main characteristics.

In 1993 a start has been made with the integration of the different research tracks in the design of a natural language interface that allows a user to ask information about theatre performances in a city. This has taken the form of joint research with PTT-Research.

PARLEVINK Researchers

More than ten researchers, including Ph.D. students, are involved in the project. In 1994 four Ph.D. students are involved in the research. Programming support is provided by the Computing Laboratory of the Department of Computer Science. A large number of computer science students are performing their M. Sci. work in the project. It is not unusual that they spend part of their education in companies in the Netherlands or with research groups in the USA.

PARLEVINK Activities

PARLEVINK participates in the Centre of Telematics and Information Technology (CTIT) of the University of Twente. Research is published in books, journals and proceedings of (international) workshops and conferences (COLING, ICANN, KONVENS, IWPT, etc.). A complete list of publications is available on request. Twice a year a workshop (TWLT: Twente Workshop on Language Technology) is organised. Proceedings of these workshops are available. In 1991 there were workshops on Generalised LR Parsing and on Linguistic Engineering. In 1992: Connectionist Natural Language Processing and Pragmatics in Language Technology. In 1993: Natural Language Interfaces and Parsing Natural Language. In 1994: Computer-Assisted Language Learning and Speech and Language Engineering. Students and project members are informed about research, lectures and other activities during weekly meetings and in the PARLEBODE, a monthly newsletter.

Twente Workshops on Language Technology

The TWLT workshops are organised by the PARLEVINK project of the University of Twente. The first workshop was held in Enschede, the Netherlands on March 22, 1991. The workshop was attended by about 40 participants. The contents of the proceedings are given below.

Proceedings Twente Workshop on Language Technology 1 (TWLT 1)
Tomita's Algorithm: Extensions and Applications
Eds. R. Heemels, A. Nijholt & K. Sikkel, 103 pages.

Preface and Contents
A. Nijholt *(University of Twente, Enschede).* (Generalised) LR Parsing: From Knuth to Tomita.
R. Leermakers *(Philips Research Labs, Eindhoven).* Recursive Ascent Parsing.
H. Harkema & M. Tomita *(University of Twente, Enschede & Carnegie Mellon University, Pittsburgh).* A Parsing Algorithm for Non-Deterministic Context-Sensitive Languages.
G.J. van der Steen *(Vleermuis Software Research, Utrecht).* Unrestricted On-Line Parsing and Transduction with Graph Structured Stacks.
J. Rekers & W. Koorn *(CWI, Amsterdam & University of Amsterdam, Amsterdam).* Substring Parsing for Arbitrary Context-Free Grammars.
T. Vosse *(NICI, Nijmegen).* Detection and Correction of Morpho-Syntactic Errors in Shift-Reduce Parsing.
R. Heemels *(Océ Nederland, Venlo).* Tomita's Algorithm in Practical Applications.
M. Lankhorst *(University of Twente, Enschede).* An Empirical Comparison of Generalised LR Tables.
K. Sikkel *(University of Twente, Enschede).* Bottom-Up Parallellization of Tomita's Algorithm.

The second workshop in the series (TWLT 2) has been held on November 20, 1991. The workshop was attended by more than 70 researchers from industry and university. The contents of the proceedings are given below.

Proceedings Twente Workshop on Language Technology 2 (TWLT 2)
Linguistic Engineering: Tools and Products.
Eds. H.J. op den Akker, A. Nijholt & W. ter Stal, 115 pages.

Preface and Contents
A. Nijholt *(University of Twente, Enschede).* Linguistic Engineering: A Survey.
B. van Bakel *(University of Nijmegen, Nijmegen).* Semantic Analysis of Chemical Texts.
G.J. van der Steen & A.J. Dijenborgh *(Vleermuis Software Research, Utrecht).* Lingware: The Translation Tools of the Future.
T. Vosse *(NICI, Nijmegen).* Detecting and Correcting Morpho-syntactic Errors in Real Texts.
C. Barkey *(TNO/ITI, Delft).* Indexing Large Quantities of Documents Using Computational Linguistics.
A. van Rijn *(CIAD/Delft University of Technology, Delft).* A Natural Language Interface for a Flexible Assembly Cell.
J. Honig *(Delft University of Technology, Delft).* Using Deltra in Natural Language Front-ends.
J. Odijk *(Philips Research Labs, Eindhoven).* The Automatic Translation System ROSETTA3.
D. van den Akker *(IBM Research, Amsterdam).* Language Technology at IBM Nederland.

M.-J. Nederhof, C.H.A. Koster, C. Dekkers & A. van Zwol *(University of Nijmegen, Nijmegen).* The Grammar Workbench: A First Step Toward Lingware Engineering.

The third workshop in the series (TWLT 3) was held on May 12 and 13, 1992. Contrary to the previous workshops it had an international character with eighty participants from the U.S.A., India, Great Britain, Ireland, Italy, Germany, France, Belgium and the Netherlands. The proceedings were available at the workshop. The contents of the proceedings are given below.

Proceedings Twente Workshop on Language Technology 3 (TWLT 3)
Connectionism and Natural Language Processing
Eds. M.F.J. Drossaers & A. Nijholt, 142 pages.

Preface and Contents
L.P.J. Veelenturf *(University of Twente, Enschede).* Representation of Spoken Words in a Self-Organising Neural Net.
P. Wittenburg & U. H. Frauenfelder*(Max-Planck Institute, Nijmegen).* Modelling the Human Mental Lexicon with Self-Organising Feature Maps.
A.J.M.M. Weijters & J. Thole *(University of Limburg, Maastricht).* Speech Synthesis with Artificial Neural Networks.
W. Daelemans & A. van den Bosch *(Tilburg University, Tilburg).* Generalisation Performance of Back Propagation Learning on a Syllabification Task.
E.-J. van der Linden & W. Kraaij *(Tilburg University, Tilburg).* Representation of Idioms in Connectionist Models.
J.C. Scholtes *(University of Amsterdam, Amsterdam).* Neural Data Oriented Parsing.
E.F. Tjong Kim Sang *(University of Groningen, Groningen).* A connectionist Representation for Phrase Structures.
M.F.J. Drossaers *(University of Twente, Enschede).* Hopfield Models as Neural-Network Acceptors.
P. Wyard *(British Telecom, Ipswich).* A Single Layer Higher Order Neural Net and its Application to Grammar Recognition.
N.E. Sharkey & A.J.C. Sharkey *(University of Exeter, Exeter).* A Modular Design for Connectionist Parsing.
R. Reilly *(University College, Dublin).* An Exploration of Clause Boundary Effects in SRN Representations.
S.M. Lucas *(University of Essex, Colchester).* Syntactic Neural Networks for Natural Language Processing.
R. Miikkulainen *(University of Texas, Austin).* DISCERN: A Distributed Neural Network Model of Script Processing and Memory.

The fourth workshop in the series has been held on September 23, 1992. The theme of this workshop was "Pragmatics in Language Technology". Its aim was to bring together the several approaches to this subject: philosophical, linguistic and logic. The workshop was visited by more than 50 researchers in these fields, together with several computer scientists. The contents of the proceedings are given below.

Proceedings Twente Workshop on Language Technology 4 (TWLT 4)
Pragmatics in Language Technology
Eds. D. Nauta, A. Nijholt & J. Schaake, 114 pages.

Preface and Contents
D. Nauta, A. Nijholt & J. Schaake *(University of Twente, Enschede).* Pragmatics in Language technology: Introduction.
Part 1: Pragmatics and Semiotics
J. van der Lubbe & D. Nauta *(Delft University of Technology & University of Twente, Enschede).* Semiotics, Pragmatism, and Expert Systems.
F. Vandamme *(Ghent).* Semiotics, Epistemology, and Human Action.
H. de Jong & W. Werner *(University of Twente, Enschede).* Separation of Powers and Semiotic Processes.
Part 2: Functional Approach in Linguistics
C. de Groot *(University of Amsterdam).* Pragmatics in Functional Grammar.
E. Steiner *(University of Saarland, Saarbrücken).* Systemic Functional Grammar.
R. Bartsch *(University of Amsterdam).* Concept Formation on the Basis of Utterances in Situations.
Part 3: Logic of Belief, Utterance, and Intention
J. Ginzburg *(University of Edinburgh).* Enriching Answerhood and Truth: Questions within Situation Semantics.
J. Schaake *(University of Twente, Enschede).* The Logic of Peirce's Existential Graphs.
H. Bunt *(Tilburg University).* Belief Contexts in Human-Computer Dialogue.

The fifth workshop in the series took place on 3 and 4 June 1993. It was devoted to the topic "Natural Language Interfaces". The aim was to provide an international platform for commerce, technology and science to present the advances and current state of the art in this area of research.

Proceedings Twente Workshop on Language Technology 5 (TWLT 5)
Natural Language Interfaces
Eds. F.M.G. de Jong & A. Nijholt, 124 pages.

Preface and Contents
F.M.G. de Jong & A. Nijholt *(University of Twente).* Natural Language Interfaces: Introduction.
R. Scha *(University of Amsterdam).* Understanding Media: Language vs. Graphics.
L. Boves *(University of Nijmegen).* Spoken Language Interfaces.
J. Nerbonne *(University of Groningen).* NL Interfaces and the Turing Test.
K. Simons *(Digimaster, Amstelveen).* "Natural Language": A Working System.
P. Horsman *(Dutch National Archives, The Hague).* Accessibility of Archival Documents.
W. Sijtsma & O. Zweekhorst *(ITK, Tilburg).* Comparison and Review of Commercial Natural Language Interfaces.
J. Schaake *(University of Twente).* The Reactive Dialogue Model: Integration of Syntax, Semantics, and Pragmatics in a Functional Design.
D. Speelman *(University of Leuven).* A Natural Language Interface that Uses Generalised Quantifiers.
R.-J. Beun *(IPO, Eindhoven).* The DENK Program: Modeling Pragmatics in Natural Language Interfaces.
W. Menzel *(University of Hamburg).* Title.
C. Huls & E. Bos *(NICI, Nijmegen).* EDWARD: A Multimodal Interface.
G. Neumann *(University of Saarbrücken).* Design Principles of the DISCO system.
O. Stock & C. Strapparava *(IRST, Trento).* NL-Based Interaction in a Multimodal Environment.

The sixth workshop in the series took place on 16 and 17 December 1993. It was devoted to the topic "Natural Language Parsing". The aim was to provide an international platform for technology and science to present the advances and current state of the art in this area of research, in particular research that aims at analysing real-world text and real-world speech and keyboard input.

Proceedings Twente Workshop on Language Technology 6 (TWLT 6)
Natural Language Parsing: Methods and Formalisms
Eds. K. Sikkel & A. Nijholt, 190 pages.

Preface and Contents
A. Nijholt (*University of Twente*). Natural Language Parsing: An Introduction.
V. Manca (*University of Pisa*). Typology and Logical Structure of Natural Languages.
R. Bod (*University of Amsterdam*). Data Oriented Parsing as a General Framework for Stochastic Language Processing.
M. Stefanova & W. ter Stal (*University of Sofia / University of Twente*). A Comparison of ALE and PATR: Practical Experiences.
J.P.M. de Vreught (*University of Delft*). A Practical Comparison between Parallel Tabular Recognizers.
M. Verlinden (*University of Twente*). Head-Corner Parsing of Unification Grammars: A Case Study.
M.-J. Nederhof (*University of Nijmegen*). A Multi-Disciplinary Approach to a Parsing Algorithm.
Th. Stürmer (*University of Saarbrücken*). Semantic-Oriented Chart Parsing with Defaults.
G. Satta (*University of Venice*). The Parsing Problem for Tree-Adjoining Grammars.
F. Barthélemy (*University of Lisbon*). A Single Formalism for a Wide Range of Parsers for DCGs.
E. Csuhaj-Varjú and R. Abo-Alez (*Hungarian Academy of Sciences, Budapest*). Multi-Agent Systems in Natural Language Processing.
C. Cremers (*University of Leiden*). Coordination as a Parsing Problem.
M. Wirén (*University of Saarbrücken*). Bounded Incremental Parsing.
V. Kubon and M. Platek (*Charles University, Prague*). Robust Parsing and Grammar Checking of Free Word Order Languages.
V. Srinivasan (*University of Mainz*). Punctuation and Parsing of Real-World Texts.
T.G. Vosse (*University of Leiden*). Robust GLR Parsing for Grammar-Based Spelling Correction.

The seventh workshop in the series took place on 15 and 16 june 1994. It was devoted to the topic "Computer-Assisted Language Learning" (CALL). The aim was to present both the state of the art in CALL and the new perspectives in the research and development of software that is meant to be used in a language curriculum. By the mix of themes addressed in the papers and demonstrations, we hoped to bring about the exchange of ideas between people of various backgrounds.

Proceedings Twente Workshop on Language Technology 7 (TWLT 7)
Computer-Assisted Language Learning
Eds. L. Appelo, F.M.G. de Jong, 133 pages.

Preface and Contents
L.Appelo, F.M.G. de Jong (*IPO / University of Twente*). Computer-Assisted Language Learning: Prolegomena
M. van Bodegom (*Eurolinguist Language House, Nijmegen, The Netherlands*). Eurolinguist test: An adaptive testing system.

B. Cartigny (*Escape, Tilburg, The Netherlands*). Discatex CD-ROM XA.
H.Altay Guvenir, K. Oflazer (*Bilkent University, Ankara*). Using a Corpus for Teaching Turkish Morphology.
H. Hamburger (*GMU, Washington, USA*). Viewpoint Abstraction: a Key to Conversational Learning.
J. Jaspers, G. Kanselaar, W. Kok (*University of Utrecht, The Netherlands*). Learning English with It's English.
G. Kempen, A. Dijkstra (*University of Leiden, The Netherlands*). Towards an integrated system for spelling, grammar and writing instruction.
F. Kronenberg, A. Krueger, P. Ludewig (*University of Osnabruek, Germany*). Contextual vocabulary learning with CAVOL.
S. Lobbe (*Rotterdam Polytechnic Informatica Centrum, The Netherlands*). Teachers, Students and IT: how to get teachers to integrate IT into the (language) curriculum.
J. Rous, L. Appelo (*Institute for Perception Research, Eindhoven, The Netherlands*). APPEAL: Interactive language learning in a multimedia environment.
B. Salverda (*SLO, Enschede, The Netherlands*). Developing a Multimedia Course for Learning Dutch as a Second Language.
C. Schwind (*Universite de Marseille, France*). Error analysis and explanation in knowledge based language tutoring.
J. Thompson (*CTI, Hull, United Kingdom/EUROCALL*). TELL into the mainstream curriculum.
M. Zock (*Limsi, Paris, France*). Language in action, or learning a language by whatching how it works.

Discription of systems demonstrated:
APPEAL (*Institute of Perception Research, Eindhoven*)
Bonacord, Méli-Mélo, ect. (*School of European Languages & Cultures, University of Hull*)
Computer BBS in language instruction (*English Programs for Internationals, University of South Carolina*)
Discatext (*Escape, Tilburg*)
Error analysis and explanation (*CNRS, Laboratoire d'Informatique de Marseille*)
ItalCultura, RumboHispano and IVANA (*Norwegian Computing Centre for the Humanities, Harald*)
It's English (*Department of Educational Sciences, Utrecht University*)
Multimedia course for learning Dutch (*SLO, Enschede*)
Part of CATT (*Department of Computer Engineering and Information Science, Bilkent University, Ankara*)
PROMISE (*Institut für Semantische Informationsverarbeitung, Univerität Osnabrück*)
Speech-Melody trainer (*Institute of Perception Research, Eindhoven*)
The Rosetta Stone (*Eurolinguist Language House Nijmegen*)
Verbarium and Substantarium (*SOS Nijmegen*)
WOORD (*Applied Linguistics Unit, Delft University of Technology*)
FLUENT-II (*George Mason University, Washington*)

The eighth workshop in the series took place on 1 and 2 December 1994. It was devoted to speech, the integration of speech and natural language processing, and the application of this integration in natural language interfaces. The program emphasized research of interest for the themes in the framework of the Dutch NWO programme on Speech and Natural Language that started in 1994.

Proceedings Twente Workshop on Language Technology 8 (TWLT 8)
Speech and Language Engineering
Eds. L. Boves, A. Nijholt, 176 pages.

Preface and Contents
Chr. Dugast (*Philips, Aachen, Germany*). The North American Business News Task: Speaker Independent, Unlimited Vocabulary Article Dictation
P. van Alphen, C. in't Veld & W. Schelvis (*PTT Research, Leidschendam, The Netherlands*). Analysis of the Dutch Polyphone Corpus.
H.J.M. Steenken & D.A. van Leeuwen (*TNO Human factors Research, Soesterberg, The Netherlands*). Assessment of Speech Recognition Systems.
J.M. McQueen (*Max Planck Institute, Nijmegen, The Netherlands*). The Role of Prosody in Human Speech Recognition.
L. ten Bosch (*IPO, Eindhoven, the Netherlands*). The Potential Role of Prosody in Automatic Speech Recognition.
P. Baggia, E. Gerbino, E. Giachin & C. Rullent (*CSELT, Torino, Italy*). Spontaneous Speech Phenomena in Naive-User Interactions.
M.F.J. Drossaers & D. Dokter (*University of Twente, Enschede, the Netherlands*). Simple Speech Recognition with Little Linguistic Creatures.
H. Helbig & A. Mertens (*FernUniversität Hagen, Germany*). Word Agent Based Natural Language Processing.
Geunbae Lee et al. (*Pohang University, Hyoja-Dong, Pohang, Korea*). Phoneme-Level Speech and natural Language Integration for Agglutinative Languages.
K. van Deemter, J. Landsbergen, R. Leermakers & J. Odijk (*IPO, Eindhoven, The Netherlands*). Generation of Spoken Monologues by Means of Templates
D. Carter & M. Rayner (*SRI International, Cambridge, UK*). The Speech-Language Interface in the Spoken Language Translator
H. Weber (*University of Erlangen, Germany*). Time-synchronous Chart Parsing of Speech Integrating Unification Grammars with Statistics.
G. Veldhuijzen van Zanten & R. op den Akker (*University of Twente, Enschede, the Netherlands*). More Efficient Head and Left Corner Parsing of Unification-based Formalisms.
G.F. van der Hoeven et al. (*University of Twente, Enschede, the Netherlands*). SCHISMA: A natural Language Accessible Theatre Information and Booking System.
G. van Noord (*University of Groningen, the Netherlands*). On the Intersection of Finite Staate Automata and Definite Clause Grammars.
R. Bod & R. Scha (*University of Amsterdam, the Netherlands*). Prediction and Disambiguation by Means of Data-Oriented Parsing.

The ninth workshop in the series took place on 9 June 1995. It was devoted to empirical methods in the analysis of dialogues, and the use of corpora of dialogues in building dialogue systems. The aim was to discuss the methods of corpus analysis, as well as results of corpus analysis and the application of such results.

Proceedings Twente Workshop on Language Technology 9 (TWLT 9)

Corpus-based Approaches to Dialogue Modelling

Eds. J.A. Andernach, S.P. van de Burgt & G.F. van der Hoeven, 124 pages.

Preface and Contents

N. Dahlbäck (*NLP Laboratory, Linköping, Sweden*). Kinds of agents and types of dialogues.

J.H. Connolly, A.A. Clarke, S.W. Garner & H.K. Palmén (*Loughborough University of Technology, UK*). Clause-internal structure in spoken dialogue.

J. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon & A. Anderson (*HCRC, Edinburgh, UK*). The coding of dialogue structure in a corpus.

J. Alexandersson & N. Reithinger (*DFKI, Saarbrücken, Germany*). Designing the dialogue component in a speech translation system – a corpus-based approach.

H. Aust & M. Oerder (*Philips, Aachen, Germany*). Dialogue control in automatic inquiry systems.

M. Rats (*ITK, Tilburg, the Netherlands*). Referring to topics – a corpus-based study.

H. Dybkjær, L. Dybkjær & N.O. Bernsen (*Centre for Cognitive Science, Roskilde, Denmark*). Design, formalization and evaluation of spoken language dialogue.

D.G. Novick & B. Hansen (*Oregon Graduate Institute of Science and Technology, Portland, USA*). Mutuality strategies for reference in task-oriented dialogue.

N. Fraser (*Vocalis Ltd, Cambridge, UK*). Messy data, what can we learn from it?

J.A. Andernach (*University of Twente, Enschede, the Netherlands*). Predicting and interpreting speech acts in a theatre information and booking system.

The proceedings of the workshops can be ordered from Vakgroep SETI, Department of Computer Science, University of Twente, P.O. Box 217, NL-7500 AE Enschede, The Netherlands. E-mail orders are possible: bijron@cs.utwente.nl. Each of the proceedings costs Dfl. 30.