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 assumption 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 definite 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 prosecing systems. *Computational Linguistics*, 16(3):175–181.

Pattabhiraman, T. (1994). Review of "user modeling in text generation" by cécile 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.