Searle, J.R. (1975) Indirect speech acts. In: P. Cole & J.L. Morgan (Eds.) *Syntax and Semantics 3: Speech Acts* New York: Academic Press.

Schuster, Ethel (1988) Pronominal Reference to Events and Actions: Evidence from Naturally-Occurring Data. University of Pennsylvania, Dept. of Computer and Information Science, Tech Rep. MS-CIC-88-13.

Tennant, H. (1979) Experience with the Evaluation of Natural Language Question Answerers, *Proc. IJCAI-79*.

Tennant, H. (1981) *Evaluation of Natural Language Processors* Ph.D. Thesis, University of Illinois at Urbana-Champaign.

Thomas, J.C. (1976) A method for studying natural language dialogue. Technical Report RC 5882, Behavioral Science Group, Computer Science Dept., IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y.

Waltz, D.L. (1978) An English language question answering system for a large relational database. *Comm. ACM.,* **7**, 526-539.

APPENDIX

TABLE 1: Scoring data from the dialogues. HiFi and Travel are advisory and order. PUB is a data base. C line and Wines are advisory systems.

TotalTot %Init %HiFiTot %Init %TravelTot %Init %
INITIATIVES39437,6313043,199032,73
Context Dep16715,9542,397123,5954,623914,1843,33
Context Indep22721,6857,616019,9346,155118,5556,67
RESPONSE50648,3314648,5012746,18
CLARIFICATION434,11103,322,000,73
RESP/INIT1049,9392,995520,00
Mistyping403,7910,18103,327,69114,0012,22
Init%Index%Init%Index%Init%Index%
INDEXICALITY19349,118968,463741,11
Pronouns287,1215,8264,627,6988,8920,51
Ellipsis11429,0164,415441,5469,231820,0046,15
Def descr5112,9828,812922,3137,181112,2228,21
DIALOGUES2153
UTTERANCES1047301275

TABLE 1 contd.

PUBTot %Init %C lineTot %Init %WinesTot %Init %
INITIATIVES3131,318343,235931,38
Context Dep99,0929,033819,7945,78105,3216,95
Context Indep2222,2270,974523,4454,224926,0683,05
RESPONSE4949,499348,449148,40
CLARIFICATION44,04147,29136,91
RESP/INIT1414,1410,522513,30
Mistyping55,0516,1352,606,0294,7915,25
Init%Index%Init%Index%Init%Index%
INDEXICALITY1111,114655,421016,95
Pronouns39,6833,331113,2528,2100,000,00
Ellipsis39,6833,333036,1476,92915,2575,00
Def descr516,1355,5656,0212,8211,698,33
DIALOGUES454
UTTERANCES99192188

es, **3**, 207-231.

Chapanis, A (1981) Interactive Human Communication: Some lessons learned from laboratory experiments. In: B. Shackel (ed.) *Man Computer Interaction: Human Factors Aspects of Computers and People.* Rockville, MD:Sijthoff and Nordhoff.

Clark, Herbert, H. (1979) Responding to indirect speech acts,*Cognitive Psychology*,**11**, 430-477.

Dahlbäck, N. & Jönsson, A. (1986), A System for Studying Human Computer Dialogues in Natural Language, Research Report, Department of Computer and Information Science, Linköping University, LiTH-IDA-R-86-42.

Fraurud, K. (1988) Pronoun Resolution in Unrestricted Text, *Nordic Journal of Linguistics*, 11, pp 47-68.

Gibbs, Raymond W. (1981) Your Wish Is My Command: Convention and Context in Interpreting Indirect Requests. *Journal of Verbal Learning and Verbal Behaviour*, **20**, 431-444.

Gibbs, Raymond, W. (1985) Situational Conventions and Requests. In: Joseph P. Forgas (ed.) *Language and Social Situations*, New York: Springer Verlag.

Good, M. D., Whiteside, J. A., Wixon, D. R. & Jones, S.J. (1984) Building a User-Derived Interface, *Comm of the ACM*, Vol 27, No 10, pp 1032-1043.

Grice, H. Paul,(1975) Logic and Conversation, In: Peter Cole and Jerry L. Morgan (eds.) *Syntax and Semantics (vol 3) Speech Acts.* New York: Academic Press.

Grosz, B.J. (1977) The Representation and Use of Focus in Dialogue Understanding. Unpublished Ph.D. Thesis. University of California, Berkely.

Guindon, R., Shuldberg, K. & Connor, J., (1987) Grammatical and Ungrammatical structures in User-Adviser Dialogues: Evidence for Sufficiency of Restricted Languages in Natural Language Interfaces to Advisory Systems, *Proc, 25th ACL*, Stanford, CA.

von Hahn, W., (1986) Pragmatic considerations in man-machine discourse. *Proc. Coling 86*, Bonn.

Hauptman, Alexander G. & Rudnicky, Alexander I. (1987). Talking to Computers: An Empirical Investigation. Technical report. CMU-CS-87-186.

Hobbs, Jerry (1978). Resolving Pronoun References., *Lingua* , **44** .

Jarke, M., Stohr, E., Vassiliou, Y., White, N. H. & Michielsen, K. (1985) A Field Evaluation of Natural Language for Data Retrieval, *IEEE Transactions on Software Engineering*, Vol, SE-11, No 1, January.

Jönsson, A. & Dahlbäck, N. (1988) Talking to a Computer Is not Like Talking to Your Best Friend, *Proc. of the First Scandinavian Conference on Artificial Intelligence*, Tromsø, Norway.

Kelly, J. F. (1983) An empirical methodology for writing User-Friendly Natural Language computer applications, *Proc. CHI '83*

Kittredge, R. & Lehrberger, J. (1982) *Sublanguage. Studies of Language in Restricted Domains.* Berlin: De Gruyter.

Linell, Per (1982) *T*he written language bias in linguistics, Studies in communication 2 (SIC 2). Department of Communication Studies, Linköping University.

Linell, P., Gustavsson, L. & Juvonen, P. (1988) Interactional Dominance in Dyadic Communication. A Presentation of the Initiative-Response Analysis. *Linguistics, 26*(3).

Malhotra, A. (1975) Design Requirements for a Knowledge-Based English Language System: An Experimental Analysis. Unpublished Ph.D. Thesis, Sloan School of Management, MIT.

Malhotra, A. (1977) Knowledge-Based English Language Systems for Management: An Analysis of Requirements. *Proc. IJCAI-77*.

Perrault, C. Raymond and Allen, James F. (1980) A Plan-Based Analysis of Indirect Speech Acts. *American Journal of Computational Linguistics.*, **6**, 167-182.

Reilly, R. (1987) Ill-formedness and miscommunication in person-machine dialogue. *Information and software technology, 29*(2),69-74,

Richards, M. A. & Underwood, K., (1984) "Talking to Machines. How are People Naturally Inclined to Speak?", In, *Contemporary Ergonomics*, (Ed) Megaw, E.D., Taylor & Francis.

Task and dialogue structure

When developing NL-technology, it is important to try to assess the applicability domain of a system. As mentioned above, the major dividing line between different classes of systems in our corpus seems not to be between database and expert (advisory) systems. But there are important differences between these and the third class used in this study, the advisory-and-order systems. In these cases more than one task can be performed, asking for information and giving an order. This means not only that the discourse representation needs to be more complicated, which in turn causes problems when trying to find the referent of referring expressions, but that it becomes necessary to understand the illocutionary force of the utterance. As was shown in the Planes system (Waltz 1978) when all the user can do with the system is to request information, all input can be treated as questions, thus simplifying the analysis of the input considerably. But this is of course not possible in these cases. The problem this causes becomes especially clear in dialogues where the user follows Grice´s quantitative maxim as much as possible, something which occurs in some of our HiFi dialogues, where one or two word utterances are very common. From a communicative point of view this is a very natural strategy—if one is engaged in an information seeking dialogue sequence requesting information about the price of different tuners, there is no need to say anything more than the name of one of them, i.e. specify the referent, but taking the illocutionary force and the predicate to be given. And when one is satisfied with the information, and wants to order the last one, why say something more than **order**, i.e. only specify the illocutionary force? What makes this problematic is of course that in some cases what is ordered is not only the last mentioned item, but a number of them, namely the set defined by the last mentioned tuner, amplifier, turn-table and loudspeakers. But realizing this requires knowledge of what constitutes as HiFi set.

Without pursuing the examples further, we wish to make two comments on this. The first is that delimiting the classes or subsets for which NL-technology with different capabilities are suitable seems to depend more on the task situation than on the computer technology of the background system. The second is that since the communicative behaviour described in the previous section can be seen to be in accordance with established theories of dialogue communication, and since it, in spite of the terseness of the utterances, seems to present no problems to the human dialogue participants, it seems somewhat strange to classify such utterances as ill-formed or in other ways deviant, something which is not uncommon. Chapanis (1981, p 106) claims that "natural human communication is extremely unruly and often seems to follow few grammatical, syntactic and semantic rules". And Hauptman and Rudnicky (1987, p 21) takes this to be supported by Grosz (1977) "whose protocols show incomplete sentences, ungrammatical style, ellipsis, fragments and clarifying subdialogues". Perhaps these examples demonstrate an extreme form of the written language bias, but in our opinion any analysis showing that a large part of a communicative event breaks the rules of communication should lead to a questioning of the validity of the formulated rules. Perhaps present day analysis of the structure of language in dialogues (including our own) is too much influenced of the traditional linguistic analysis of isolated utterances, and a shift of perspective is required for a breakthrough in this area.

A Final Remark

As can be seen in the tables in the appendix, there are differences between the different background systems, for instance the use of pronouns in the PUB dialogues is as frequent as the use of ellipsis, while Wines have no pronouns. There are also differences between different users, ranging from very condensed one word phrases to small essays on two to three lines. This indicates that when designing a NLI for a specific application it is important to run simulations, preferably with the real end users (cf. Kelly 1983 and Good *et al* 1984). We intend to proceed in that direction and develop a method for design and customization of NLI´s based on Wizard of Oz experiments.

Acknowledgements

We thank all our friends at NLPLAB for creating an intellectually and socially rewarding environment. Special thanks to Lars Ahrenberg for comments on an earlier version of this paper. Bernt Nilsson has implemented the ARNE-2 experimental environment. Ulf Dahlén and Åke Pettersson have implemented the tagging system DagTag used in the analysis. We also thank our students for their work with the data collection.

REFERENCES

Beun, R.J. and Bunt, H.C. (1987) Investigating linguistic behaviour in information dialogues with a computer. In: *IPO Annual Progress Report*

Bosch, Peter (1988) Representing and Accessing Focused Referents, *Language and Cognitive Process-*

Task structure

The results concerning task structure are interesting. It is perhaps not too surprising that the task structure in a data base application is simple. Here one task is introduced, treated, finished, and dropped; and then another is introduced. A basically similar pattern is found in the advisory systems.

The advisory-and-order systems, however, shows a completely different picture. These systems are in an important sense more complicated, since two different types of actions can be performed; obtaining information or advice, and ordering. The collected dialogues show that these two tasks are executed in parallel, or rather that they are intertwined. The consequence is that we have *two* active tasks at the same time. For instance, in the HIFI simulations the interlocutors shift rapidly between discussing the ordered equipment, its total price, etc, and discussing technical information about available equipment. 7% of the initiatives are task shifts in this sense. The problem is, that while it presents no difficulty for the human reader to follow these task shifts, it is difficult to find any surface cues indicating them. The computational mechanisms for handling this type of dialogue will therefore presumably be more complex than for the other applications that we have studied. In our opinion this confirms Grosz' (1977) observation that there are different types of dialogues with different task structure. It also indicates that categories such as data base and expert systems are not always the most relevant when discussing application areas for NL-techniques.

System initiatives

The system's linguistic behaviour seems to influence the language used by the user in an important sense. The utterance type Resp/Init reflects how often the system not only responds to an initiative, but also initiates a new information request. This is used more frequently in three simulations. This ought to result in the number of Context Dependent initiatives being lower than in the other dialogues, because the user has here already provided all the information needed. This hypothesis is corroborated in two of the three simulations (PUB and Wines). They have 17% respective 29% context dependent initiatives compared to the average of 42%. (We do not tag whether a response is context dependent or not.) The result is interesting, because it indicates that this is a way of 'forcing' the user to use a language which is computationally simpler to handle, without decreasing the habitability of the system, as measured in the post-experimental interviews.

As mentioned above, this pattern is not found in the third system, the travel advisory system. This system belongs to the advisory-and-order class. We cannot at present explain this difference, but would still claim that the result obtained is interesting enough to deserve a thorough follow-up, since databases and advisory systems presently are the largest potential application areas for NLIs.

Indirect speech acts

Indirect speech acts (Searle, 1975) have been one of the active areas of research in computational linguistics. It can perhaps be of interest to note that there are only five indirect speech acts in our corpus, all of which use standardized expressions (*Can you tell me ...?* etc). Beun and Bunt (1987) found a higher frequency of indirect requests in their corpus of terminal dialogues (15%). However, this frequency was considerably lower than in their control condition of telephone dialogues (42%). Taken together, these results seems to support our belief that some of the reasons for using indirect means of expression does not exist in man-computer dialogues in natural language (c.f. Dahlbäck and Jönsson, 1986).

The lack of variation in the expression of indirect speech acts is perhaps not all that surprising when viewed in the light of psychological research on their use. Clark (1979) expanded Searle's (1975) analysis by distinguishing between *convention of means* and *convention of forms* for indirect speech acts; the former covers Searle's analysis in terms of felicity conditions and reasons for performing an action, the latter the fact that *can you open the window?* is a conventional form for making an indirect request, whereas *Is it possible for you to open the window?* is not. Gibbs (1981, 1985) demonstrated then that what counts as a conventional form is dependent on the situational context in which it occurs. There is therefore in our opinion good reasons to believe that indirect speech acts can be handled by computational methods simpler than those developed by Perrault and co-workers, something which in fact seems compatible with the discussion in Perrault and Allen (1980). In conclusion, we believe that indirect speech acts are not as frequent in man-computer dialogues as in human dialogues, and that most of them use a small number of conventional forms which suggests that computationally tractable and cost-effective means of handling them can be found.

comparative constructions without expression of the comparative object e.g. *Wines:4:9 Is there any cheaper white wine [Finns det något billigare vitt vin].*

However, in spite of the fact that we have not used an explicit grammar, we have also regarded syntactic incompleteness as a ground for tagging an utterance elliptical. Certain questions like HiFi*:3:12 price sondek [pris sondek]* are tagged elliptical for syntactic reasons. On the other hand imperative utterances like *HiFi:3:28 Order Sondek [Beställ Sondek]* are not tagged context dependent and thus not indexical at all. This might seem inconsequential, but is in fact a reflection of the characteristics of our assumed grammar.

Results and discussion

There are 1047 utterances in our corpus. Of these, 38% are Initiatives, 48% Responses, 10% Resp/Init, and 4% Clarifications. Table 1 and 2 in the appendix summarize some of our results. 58% of the Initiatives are Context Independent, i.e. utterances that can be interpreted in isolation. However, of these about 10% are dialogue openings. This means that only 48% of the Initiatives *within* the dialogues can be interpreted in isolation.

Context Dependencies

The complete set of data concerning the number of context dependent utterances and the distribution of different types of context dependency are presented in the appendix. While we believe that the data presented here give a correct overall picture of the qualities of the language used in the dialogues, the previously mentioned caveat concerning the theory dependency of the data, especially as regards ellipsis and definite descriptions, should be kept in mind. We will for the same reasons in this paper concentrate our discussion on the usage of pronouns in the dialogues.The number of Context Dependent utterances are 167 or 42%. Thus, when the users are given the opportunity to use connected discourse, they will — even when the response times (as in our case) occasionally seem slow.

The most common forms of indexicality are ellipsis (64%) and definite descriptions (29%). The use of pronouns is relatively rare, only 16%. The limited use of pronouns is not something found exclusively in our corpus. Similar results were found by Guindon *et al* (1987), where only 3% of the utterances contained any pronouns. While being to small an empirical base for any conclusive results, this does suggest that the use of pronouns are rare in typed man-computer dialogues in natural language. Some suggestions why this should be the case can be found in a study by Bosch (1988) on the use of pronouns in spoken dialogues. He argues for a a division of the focus structure into two parts, explicit and implicit, and claims that "explicit focus is typically, though not exclusively, accessed by means of unmarked referential expressions (typically de-accented anaphoric pronouns), while implicit pronouns focus is accessed only by marked devices, including accented pronouns"(Bosch, 1988, p 207). What is interesting with this analysis in the present context, is that para-linguistic cues (accent) is used to signal how the pronoun should be interpreted. Since this communicative device is absent in written dialogues, this could explain why the subjects refrain from using pronouns.

We believe this to be an expression of a general principle for the use of pronouns. Since a pronoun underspecifies the referent compared to a definite description, there is every reason to believe that language users following Grice's (1975) cooperative principle should only use them when the listener/reader effortlessly can identify the intended referent. This is supported by data from Fraurud (1988), who analyzed the use of pronouns in three different types of unrestricted written Swedish text. She showed that for 91% of the 457 singular pronouns a very simple algorithm using only syntactical information could correctly identify the antecedent, which in 97.4% of the cases were found in the same or preceding sentence. Similar results have also been obtained by Hobbs (1978).

We obtained results similar to those of Fraurud (1988) as regards the distance between the pronoun and its antecedent. All our antecedents where found in the immediate linguistic context, except for one problematic category, the pronoun *man (one/you)*, excluded in her study which often refers to some global context, e.g. *C line:5:10 Does* ***one*** *read mechanics [Läser* ***man*** *mekanik]*.

We will by no means conclude from this that it is a simple task to develop a computational discourse representation for handling pronouns. As pointed out by Shuster (1988), it is often unclear whether a pronoun refers to the whole or parts of a previously mentioned event or action. While this underspecification in most cases seems to present no problems for human dialogue participants, it certainly makes the computational management of such utterances a non-trivial task.

forms of indexicality will be handled here, e.g. ellipsis and pronouns that can be resolved by available surface structure linguistic information. The third module uses a case-frame like representation of the current discourse domain (task)[1]. Here utterances whose interpretation requires background knowledge can be interpreted. One consequence of the use of this latter module is that it is necessary to specify the task structure of the discourse domain in advance of the analysis. This approach differs from linguistically oriented approaches to discourse analysis, where the task structure of the dialogue is found through the linguistic analysis.

ANALYSIS categories

We divide our utterances into four different categories (c.f. Linell, Gustavsson and Juvonen, 1988): **1) Initiative** means that one of the participants initiates a query. **2) Response** is when a participant responds to an initiative, such as an answer to a question. **3) Resp/Init** is used when a *new* initiative is expressed in the same utterance as a response. Typical situations are when the system has found an answer and asks if the subject wants to see it. The utterance type **4) Clarification** is used in reply to a Response of type Clarification request and indicates what type of clarification is used. Jönsson and Dahlbäck (1988) describe and discuss the analysis categories in more detail.

Task and Context

Initiatives are analyzed ("tagged") for *Context Dependence* which concerns the *interpretation* of an utterance. We tag an utterance Context Dependent if it cannot be interpreted without information in the immediate context. Every utterance that is complete enough to be interpreted without context is tagged Context Independent, regardless of the possible existence of a usable context in the previous utterance. Initiatives are tagged *Task Dependent* if background knowledge is required for their interpretation.

Indexicality

We tag our Context Dependent utterances for indexicality using three main categories: pronoun, ellipsis and definite description. It is important to note that there is a difference between these types, since they vary in their dependence of a specific theory or discourse representation model. What counts as a pronoun can be determined lexically, and presents no major problem. But what counts as an ellipsis is dependent on the grammar used in the analysis, and to count a definite description as context dependent simply because there exists something in the previous text that could be seen as its antecedent seems somewhat dubious. In our opinion such an utterance should be called context dependent only if knowledge of the preceding linguistic context is necessary for finding its referent in the discourse representation, i.e. that the antecedent is necessary for determining the referent. And this is obviously dependent on the qualities of the discourse representation and the process working on it.

Tagging a **pronoun** is usually straightforward, but there are some utterances which are ambiguous. For instance, the Swedish pronoun *det (it)* may act as an anaphoric pronoun or as a formal subject in various types of constructions, e.g.*Travel:1:26 What does **it** cost?*[2] *[Vad kostar **det**?].* This is a question to a previous response suggesting a hotel to live in. The *it* in Travel:1:26 can be interpreted either as pronoun referring to the hotel, or it can be a formal subject and then the utterance is elliptical. There are five utterances tagged ambiguous (all from the travel dialogues) and they are not included in the results.

Definite descriptions are definite NP's or other definite referents like demonstratives, e.g. *HiFi:1:5 What is the price for a complete hifi system with **these models.**[Vad blir priset för en komplett hifi-anläggning med **dessa modeller.**].* Proper names are not tagged as definite descriptions.

Ellipsis is a problematic category, cf. above. Our basic criterion is semantic incompleteness, thus one word phrases, except for some imperatives and expressives (*Yes*, *Help, Thanks etc*), are tagged ellipsis e.g. *C line:4:5 prerequisites? [förkunskaper]* as response to a list of courses . We also use ellipsis for

1. We use the term Task in this paper. The notion is similar to what we previously called Topic (Dahlbäck and Jönsson 1988, Jönsson and Dahlbäck 1988).

2. All examples are from our corpus. The first field indicate the dialogue, the second subject and finally utterance number. The bold face does not occur in the dialogues. The corpus is in Swedish and translated into English striving for verbatim rather than idiomatic correctness.

theory and our linguistic intuitions are adequately developed to guarantee some consensus on what counts as ungrammatical (though the *written language bias in linguistics* (Linell, 1982), i.e. the tendency to regard the written language as the norm, and to view other forms as deviations from this, has in our opinion lead to an overestimation of the ill-formedness of the input to natural language interfaces also in this area). But when it comes to dialogue aspects of language use, we lack both theory and intuitions. What can be said without hesitation, however, is that the use of a connected dialogue, where the previous utterances set the context for the interpretation of the current one, is very common.

It is therefore necessary to supplement previous and on-going linguistic and computational research on discourse representations with empirical studies of different man-computer dialogue situations where natural language seems to be a useful interaction technique. Not doing so would be as sensible as developing syntactic parsers without knowing anything about the language they should parse.

Other researchers have proposed the use of field evaluations as they are more realistic. However, doing so requires a natural language interface advanced enough to handle the users language otherwise the evaluation will only test the NLI's already known limitations, as shown by Jarke, Turner, Stohr, Vassilou & Michielsen (1985).

Method

We have conducted a series of Wizard of Oz experiments. There are two important aspects to consider when developing the experimental situation. The first concerns the background system. It should in our opinion be something that could run on a computer using the technology of today — or at least tomorrow — to ensure that the influence of the situation does not invalidate the use of data and results when developing a natural language interface. Great care should also be given to the design of the *scenario,* i.e. the task given to the subjects. Obviously, any simple task which only requires a few interactions between user and system will not give us much data to analyze. Our experience shows that one should either give the subjects a task for which there does not exist a single correct answer, but where the subjects own preferences determines what counts as a satisfying goal, or by having a task where there exists more than one way to achieve the goal.

When conducting a Wizard of Oz experiment it is important to ensure that the subjects believe they are using a computer. To achieve this we have developed an experimental environment with a number of tools. The use of windows gives easy access to all relevant systems.The 'wizard' has at his disposal windows monitoring the user, the background system, an editor and windows with parsers or other modules developed for the current application. Menus with prestored (partial) answers guarantee a consistent, fast output with a 'computerized' quality (Dahlbäck and Jönsson, 1986).

Generalizability of results requires experiments with a variety of background systems, scenarios and many subjects. We have used five different scenarios for five background systems of varying complexity[1]; one library database used at our department and four simulated advisory systems: one student advisory system; one wine selection advisory system and two advisory-and-order systems — one for HIFI equipment and one for travel. We have collected dialogues from 21 subjects. Approximately half of them were students. The subjects' previous experience with computers were limited or nonexistent.

The discourse model

The collected corpus should be analyzed with an explicit formalism in mind. Our goal is not to develop a general discourse model, but instead to find the simplest possible usable model for natural language interface applications (or some subclass of such applications).

The interface consists of three modules. One resembles a question-answering system without any dialogue handling capabilities. This will transform the user input into the appropriate query-language command or other background system input, given that enough information is available in the user's utterance. Another (linguistic context) module is used when the input does not contain enough information to form a command to the background system. This module uses the immediate linguistic context, i.e. the user's and the system's last utterance, and tries to complete the fragmentary input. Simple

[1]. This figure does not include pilot studies. We have recently conducted experiments using a combined graphical and NL calendar booking system. Since this communication situation differs from the others, we have excluded these data from the present analysis.

Empirical Studies of Discourse Representations for Natural Language Interfaces

Nils Dahlbäck Arne Jönsson
Natural Language Processing Laboratory
Department of Computer and Information Science
Linköping University, S-581 83 LINKÖPING, SWEDEN
Internet: NDA@LIUIDA.SE, ARJ@LIUIDA.SE
Phone +46 13281644, +46 13281717

Abstract

We present the results from a series of experiments aimed at uncovering the discourse structure of man-machine communication in natural language (*Wizard of Oz* experiments). The results suggest the existence of different classes of dialogue situations, requiring computational discourse representations of various complexity. Important factors seem to be the number of different permissible tasks in the system and to what extent the system takes initiative in the dialogue. We also analyse indexical expressions and especially the use of pronouns, and suggest a psychological explanation of their restricted occurrence in these types of dialogues.

Introduction

Natural Language interfaces will in the foreseeable future only be able to handle a subset of natural language. The usability of this type of interfaces is therefore dependent on finding subsets of natural language that can be used without the user experiencing inexplicable "holes" in the system performance, i.e. finding subsets for which we can computationally handle complete linguistic and conceptual coverage. This points to the need for theories of the 'sublanguage' or 'sublanguages' used when communicating with computers (Kittredge and Lehrberger, 1982). But unfortunately: "we have no well-developed linguistics of natural-language man-machine communication." (von Hahn, 1986 p. 523)

One way of tackling this problem is to simulate the man-machine dialogue by letting users communicate with a background system through an interface which they have been told is a natural language interface, but which in reality is a person simulating such a device (sometimes called a *Wizard of Oz* experiment, see Guindon, Shuldberg, and Conner, 1987). While not being a new technique, early examples are Malhotra (1975, 1977), Thomas (1976), and Tennant (1979, 1981), only a limited number of studies have been conducted so far. A considerably larger number of similar studies have been conducted where the users knew that they were communicating with a person. This is unfortunate, since those researchers who have considered the issue have noted that the language used when communicating with a real or simulated natural language interface has differed from the language used in teletyped dialogues between humans, although it has been difficult to the exact nature of these differences. The language used has been described as 'formal' (Grosz, 1977), 'telegraphic' (Guindon *et al*, 1987), or 'computerese' (Reilly, 1987).

Only a few Wizard of Oz studies have been run, using different background systems and differing in questions asked and methods of analysis used. It is therefore premature to draw any far-reaching conclusions. With some caution, however, perhaps the following can be accepted as a summary of the pattern of results obtained so far: *The syntactic structure is not too complex* (Guindon *et al*, 1987, Reilly, 1987), and presumably within the capacity of current parsing technology. Only a *limited vocabulary* is used (Richards and Underwood, 1984), and even with a generous number of synonyms in the lexicon, the size of the lexicon will not be a major stumbling block in the development of an interface (Good, Whiteside, Wixon, and Jones, 1984). However, it is unclear how much of this vocabulary is common across different domains and different tasks, and the possibility of porting such a module from one system to another is an open question. *Spelling correction* is an important feature of any natural language based system. So-called *ill-formed input* (fragmentary sentences, ellipsis etc) *is very frequent*, but *the use of pronouns seems limited* (Guindon, *et al*, 1987, Jönsson and Dahlbäck, 1988).

However, the results concerning ill-formedness are difficult to evaluate, mainly because they are often presented without an explicit description of the linguistic representation used. An utterance can obviously only be ill-formed relative to a formal specification of well-formedness. With some hesitation the exclusion of such a specification can perhaps be accepted as far as syntax is concerned. Both linguistic