Talking to a computer is not like talking to your best friend

Arne Jönsson
Nils Dahlbäck

Natural Language Processing Laboratory
Department of Computer and Information Science
Linköping University, S-581 83 LINKÖPING, SWEDEN
Internet: NDA@LIUIDA.SE, ARJ@LIIUDA.SE

Abstract
When developing natural language interfaces, it is important to know the characteristics of the language used. We have developed a set of tools for conducting human-computer natural language dialogue simulations (Wizard of Oz experiments) and for analyzing the data obtained. We report methods used and results obtained from a series of such experiments. The focus of the study was on occasional users and on the structure of the dialogue. Three different background systems were used; one database and two advisory systems of different complexity. The results point to the need for mechanisms for handling connected discourse in interfaces for this user group. They also indicate that there are different classes of dialogue situations requiring discourse representations of different computational complexity.

1. Introduction

The major part of the AI research on dialogues has taken human dialogues as its starting point, based on the assumption that (spoken) dialogues are the most natural form of communication. A corollary assumption is that the most natural form of communication between a human user and a computer should resemble this as much as possible. The assumptions underlying this research can, somewhat simplistically, be described as follows: Natural Language interfaces should make it possible to communicate with computers in natural language. Further, "We can therefore assume that the input to a natural language program is a character string typed into a CRT terminal or found on a computer tape" (Lehnert & Ringle, 1982, p xiii). On this level of abstraction, this is presumably an uncontroversial position. But the seemingly "obvious" corollary to this, namely that the natural language dialogue between human and computer should as much as possible resemble a dialogue between two people, and that therefore the research strategy should be to study human dialogues, and then try to give the NLI as many as possible of the human abilities needed to participate in such a dialogue, is the point we want to question.

There are two main problems with this corollary. The prevalent view of language and communication in linguistics is, despite an increased interest in spoken language and in communication, to a large extent based on a written language bias (Linell, 1982), i.e.

the tendency to regard the written language as the norm, the "correct" language, as it were, and consequently to regard other forms of language as "ill-formed" or "ungrammatical". This "taken for granted" assumption has in our opinion led to an overestimation of the "ill-formed" quality of the input to natural-language interfaces.

There are also some obvious and important differences between human dialogues and man machine dialogues in areas known to influence the form and/or content of the dialogue. The social situation is different — will a user have a need to be polite to a computer by using indirect speech acts? The communication channel has unique limitations and possibilities — it is difficult to make a computer use background knowledge in interpreting an utterance; the output rate on a CRT is faster than through the human mouth. The number of possible actions are limited. How will all this affect the optimal interaction patterns? Present day knowledge in linguistics does not make it possible to say how this will affect the language used, but there is no doubt that it will do so. This has also been noted by researchers conducting empirical studies of the language used in man-machine communication by means of natural language (Grosz, 1977, Guindon, Shuldberg & Conner, 1987, Reilly, 1987a). Similar observations have been made by linguists working on sublanguages (e.g. Kittredge and Lehrberger, 1982).

So we have a problem here: If the language used to communicate with a NLI will differ from the language used in dialogues between humans, we want to build an NLI to cope with this language, and not with the language used in other situations. But before we have a NLI for people to use, how can we know what the characteristics of such language are. This point has also been stressed by von Hahn who points out that "we have no well-developed linguistics of natural-language man-machine communication." (von Hahn, 1986 p. 523)

We would claim that one step towards the solution of this problem would be to supplement previous research by studying human computer interaction through an NLI from a new set of assumptions; not trying to mimic communication between humans, but as a communicative process between two agents with different capabilities — different strong and weak points — and based on the realization that *natural* communication between them must take these points into consideration. Furthermore, we would want to claim that one way of doing this would be to simulate the dialogue, by letting users communicate with different background systems (databases, expert systems etc) through an interface which they were told were a natural language interface, but which in reality was a person simulating this device.

As part of the work on the LINLIN project (Ahrenberg, Dahlbäck, Jönsson, Merkel, Wiren, 1986; Ahrenberg, 1987), whose purpose is to develop a general Natural Language interface for Swedish, we have conducted a number of such simulations. In the present paper we will first give an overview of previous research in this area, as a background to our own work. Thereafter we will give short descriptions of the dialogue simulation environment (ARNE) and the dialogue analysis system (DagTAg) developed as part of this research. The second part of the paper is devoted to describing our research in more detail; theoretical and methodological considerations, experiments conducted so far, results obtained and conclusions drawn.

1.1 Previous studies

We are by no means the first to suggest or use some kind of simulations of an interface as a part of the development process, or in evaluating the capacity of existing systems. Early examples focussing on natural language are Malhotra (1975, 1977), Thomas (1976), and Tennant (1979, 1981). Good, Whiteside, Wixon & Jones (1984) used a similar technique in the iterative development of a command language for an e-mail

system. Before presenting our own work, we will give a short overview of other studies. Our aim is not to give a complete overview, but instead to give the reader a glimpse of the diversity of methods used and results obtained, as a background to a discussion of our own approach. Since our main interest concerns the 'higher' linguistic levels, we will concentrate on this aspect in what follows.

The work of Grosz (1977, 1981) is perhaps the most well known in this area. It was conducted as a part of the development of the SRI Speech understanding system (Walker, 1978). It was also published in Deutsch (1974) and Grosz (1977). Grosz' main interest was not in the unique features of man-machine dialogue. Instead, the focus of the study was on the dialogue structure of a task oriented dialogue, since this type of dialogue resembles possible applications of Natural Language communication with computers. The familiar notion of *focus space* was first presented in this work.

Grosz was well aware that this type of dialogue differed in many respects from the dialogue with a data base, and therefore included some simulations of such dialogues as well. A total of 10 task oriented dialogues and 5 data base dialogues were collected under different conditions. The main analysis was done on four task oriented dialogues, two in which the users knew that they were communicating with a human, and two "pure" simulations. Grosz is aware of the fact that it is a small number of subjects, and that therefore speaker idiosyncrasies may limit the generalizability of the data, but she points out that it still is a vast amount of data to analyze. (So it is understandable but regrettable that much research in this area has used Grosz' corpus instead of collecting new material)

The data are analyzed on a number of linguistic levels. We will here only mention those parts most relevant for our current interests. One basic finding was that the structure of the task influenced the structure of the dialogues; subdialogues in the task oriented dialogues reflected subparts of the problem to be solved, and in a sense basically the structure of the object assembled. The data base dialogues did not have any complicated global structure. Most of these dialogues consisted of sequences related locally through the use of ellipsis etc. Openings and closings of these subdialogues were hard to detect linguistically. Grosz concludes that "there seems to be a continuum (...) from the highly unstructured table-filling dialogues to the highly structured task dialogues." (*ibid* p. 33)

As mentioned above, Grosz is not primarily interested in differences between human dialogues and human computer dialogues in natural language, so data from the two types of dialogues are mostly analyzed together. She notes, however, that the language used in the real simulations differed from the other dialogues. She describes it as more 'formal' (1977, p.69), while at the same time noting that it is hard to point out exactly what it is that gives them this quality. Grosz concludes that this is an important area for future research.

Another question raised but not answered is how much to quality of the output of the system influences the language used by the user. There are some indications that users adapt to the system's language, especially as regards vocabulary.

Guindon, Shuldberg, and Conner (1987) have an interesting motive for conducting simulation studies. They want to find a "domain-independent subset of grammatical and ungrammatical structures, to help design more habitable natural language interfaces to advisory systems". That is, instead of limiting the range of necessary linguistic capacities in a NLI by studying sublanguages for different domains, they want to find the core common to all advisory dialogues, thus making the development of a system in a new domain easier.

This is a real simulation, sometimes called a *Wizard of Oz-experiment*, i.e. the subjects thought that they were communicating with a computer, but were in fact communicating with a human. The background system was a real statistical package and a (simulated) advisory system. The subjects were 32 graduate students with basic statistical knowledge and unfamiliar with the statistical package. No information is given on the ss's familiarity with computers. Nor is any information given on the linguistic quality of the output from the system.

The data are compared with those from two other studies; Thompson (1980) and Chafe (1982). Thompson compared three different dialogue-types: Spoken face-to-face, Typed human-human, and Human-computer with the REL NLI. Chafe compared informal spoken language (i.e. dinner conversations) with formal written language (i.e. academic papers). The data are compared on four dimensions — Completeness and formality of user's utterances; Ungrammaticalities; General syntactic features; Features due specifically to the user advising situation.

The results show that for completeness and formality, the language used in this situation resemble human-computer dialogues and formal written language more than spoken face-to-face or typed human-human dialogues. 24% of the utterances were fragmentary. The authors conclude that the "users in typed user-advisor dialogues seem to expect the interface to be unable to handle fragmentary input such as found in informal spoken language" (Guindon *et al*, p 42). This is even more interesting when one considers that the "system" hardly ever rejected or misunderstood any input. The conclusion drawn by the authors from this is important when evaluating data from dialogue studies that are not pure Wizard of Oz studies: "It appears that a priori beliefs about the nature and abilities of the adviser (i.e. that he is not a human) can determine the characteristics of the language produced by the user, even when the task and linguistic performances by the adviser were not negatively affected by fragmentary language from the user". (*ibid* p 42) An interesting question that cannot be answered on the basis of the published information is how "formal" the output from the system was, and therefore to what extent the quality of the users' input was a reflection of the language used by their conversation partner.

There was a high frequency of ungrammatical utterances. 31% of the utterances contained one or more ungrammaticalities. The most common of these were fragments (13%) missing constituents (14%) and lack of agreement (5%). The authors leave the issue open as to whether this quality of the language used is due to the fact that there are some ungrammaticalities that are difficult to avoid, in spite of attempts to use a 'correct' language, or if it is a reflection of a tendency by the subjects to use a 'telegraphic' language assumed to be better understood by the 'system'. However, as is discussed further below ungrammaticality can only be defined in relation to an explicit grammar, and since there is a tendency in linguistics to use the written monologic rather than the spoken dialogic language as the norm, it is unclear if for instance all the fragmentary sentences should be regarded as ungrammatical. This point gains some support from the authors' analysis of general syntactic features, where it is concluded that the language used in this respect resembles informal spoken language. It is unfortunate that the grammar used when classifying utterances as ungrammatical or fragmentary is not included, but most likely this is due to the limited allotted space in conference proceedings.

One interesting feature of the corpus was the infrequent use of pronouns (3% of the utterances) and the frequent use of complex nominals with prepositional phrases (e.g. "a record of the listing of the names of the features"). Such constructions were found in 50% of the utterances.

Reilly (1987a, 1987b) has reported from a simulation experiment conducted as a part of the development of a natural language understanding dialogue system that can cope

with a wide variety of ill-formed input and can deal adequately with miscommunication. The theoretical starting points for this project are the work by Ringle & Bruce (1980) on communication failure in dialogues, the discourse theory of Grosz & Sidner (1986), and Burton's (1981) discourse analysis model.

In the main dialogue simulations, the users had to fill out a form requesting information contained in a database. The users typed the queries to an expert who accessed the database. The users, 24 females and 2 males, knew that they were communicating with a person. They had no familiarity with computers, nor with the task they were asked to perform. The sessions lasted on the average 25 minutes. (This research group conducted a number of preliminary dialogue studies in a pilot phase, varying background system, interaction mode, subjects etc. These pilot studies (Eagan, Forrest *et al* 1986) were in some respects more thorough than some of the studies published in the literature.)

The focus of the analysis is, in line with the research goals of the project, on the ill-formedness of the input. 58% of the users' utterances were ill-formed. The most common types were misspelling, fragmentary input and ellipsis. It was also noted that, compared with ordinary dialogues, there was an almost complete lack of management moves such as affirmations, go-ahead signals and other back-channeling devices. There was also a noted lack of requests for help or explicit signals of misunderstanding from the users. (Reilly (1987a) therefore argues for the development of some substitutes that fits into this communicative situation, since the lack of such devices greatly increases the risk of miscommunication.)

On the basis of the analysis of the dialogue corpus, Reilly (1987a, p 73) concludes that three areas should have the highest priority in the development of a robust natural language dialogue system, namely treatment of misspelling, treatment of fragmentary input, and treatment of ellipsis.

The syntactic variation in the corpus was rather limited, and Reilly therefore considers the investment of effort in developing a parser capable of handling unusual syntactic constructions to be a wasted one if the application is database interfaces. Instead focus should be on lower levels, such as morphology, and higher levels, such as user modeling.

These studies are by no means the only ones that have used some sort of simulated human-computer dialogue in natural language as a part of the development of a natural language interface. Bates and Sidner (1983) used a similar method for studying the concurrent use of natural language and graphics in an interface; Tennant (1981) used simulations as part of the evaluation of the PLANES system (Waltz, 1978, Waltz & Goodman, 1977); Fineman (1983) simulated the interaction with a voice-input driven interface with limited vocabulary (50 words) and syntactic capability to mention just a few. However, most of these have not been pure Wizard of Oz simulations. This is unfortunate, since those that have conducted such simulations have noted that the language used differs from the language used between humans, even though the differences are difficult to describe. Grosz (1977) describes it as 'formal', Reilly (1987a) calls it 'computerese', and notes its telegraphic quality, as do Guindon *et al* (1987).

Summarizing the results from the studies mentioned above (and other similar ones), the following observations can be made: The syntactic variation is rather limited, and presumably well within the capacity of current parsing technology. Only a limited vocabulary is used, 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. 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. However, this latter result is the most difficult to evaluate, mainly because such studies often have been made without an explicit description of the dialogue representation used. An utterance can only by ill-formed relative to a formal specification of well-formedness. With some hesitation the exclusion of such a specification can be accepted as far as syntax is concerned. Both linguistic theory and our linguistic intuitions are adequately developed to guarantee some consensus on what counts as ungrammatical. But when it comes to dialogical 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 interpretation of the current one, is very common.

As mentioned above, these remarks should be regarded as preliminary. The main reasons for this is that many of the studies have not been pure Wizard of Oz experiments, that only a very limited number of subjects have been used, and only a limited number of background systems have been used. We need to know more about individual differences and what their causes are (linguistic habits, typing skill, familiarity with computers, etc), and more about how the background system and the task to be done affects the language used, before we can know how the results can be generalized to other users and systems than those used in these studies.

Since lack of knowledge is notable in many areas of natural language communication with computers, but, as was noted above, mostly so regarding 'higher' linguistic levels, we have focused our current research effort on dialogue phenomena. If the language used in communication with a computer does not exhibit all the variability of natural language as far as vocabulary and syntax is concerned, and therefore hopefully can be handled by (relatively) simple, or at least known computational systems, perhaps this is the case for dialogue, too. We have therefore started a dialogue simulation project aimed at answering that question.

2. Method

As noted above, there are many reasons to believe and some experimental data to support the notion that the language used with a computer will differ from the language used between humans. It is therefore our firm belief that dialogue studies should focus on pure Wizard of Oz experiments, where the subjects believe that they are communicating with a computer, and furthermore that there are a number of precautionary measures that should be taken to strengthen this illusion (see below). If the research interest is mainly to ascertain the similarities and differences between human dialogues and dialogues with computers, the experimental design should of course also include non-simulated dialogues, but if the aim of the study is to specify which linguistic capabilities to include in a specific system, the point above is of utmost importance.

It seems very likely that the output from the system influences the language of the user. Therefore it is important to give the output a 'computerized' quality, especially in two respects: First, It should have the mechanical quality of present-day computers. How much it should vary is open to dispute (partly based on one's assessment of the progress in text generation research), but it should not be as stylistically varied as a human would make it. Second, the response times should be fast (or at least as fast as possible). That accidental misspellings and other traces of human fallibility and imperfection should be avoided is perhaps to obvious to be mentioned.

In order to achieve this goals, and in order to simplify the collection of dialogues, we

have developed an environment for conducting Wizard of Oz experiments, called ARNE (Almost Realistic Natural language Equipment) (Dahlbäck & Jönsson, 1986). ARNE is run on a Xerox Lisp machine. The 'wizard' simulating the natural language interface has at his disposal one terminal window connected via a 'talk' link to the user's terminal. Another window is connected to the background system, which can be a database on another computer, a database on the Lisp machine or simply some prestored texts to be used to simplify the work when the wizard is also simulating the background system, or parts of it.

There are two more main windows. One contains an editor and the other is used when parsers or other modules are available for testing or just to make the simulation faster. All windows contain menues with canned text to enhance the possibility of giving fast and consistent answers, something not always done in previous studies.

The background system should be something that can run on a computer today - or at least tomorrow. While more futuristic uses of computers are also of considerable interest, it is an important enough task to improve the interface of the systems of today. This has the further advantage of setting realistic constraints on the communicative situation. This is important, since the task and the capabilities of the background system will influence the language used, and hence the language the developed interface will need to handle.

We have so far used three background systems of different complexity; a library database used at our department and two simulated advisory systems. One of these contains information about the computer science courses at Linköping University, to be used by student advisers. The other contains information about high quality HIFI equipment. In the latter system the user can also order equipment that he wants to buy.

We have developed for each of these background systems a scenario for the subjects. The users of the library system are being asked to find books for courses given at the department. The student advisory system is used for answering a letter from a student requiring information on the Computer Science course. The users of the HIFI advisory system are asked to compile and buy equipment for the home, with the restriction that the total price may not exceed 100 000 SEK ($16 000).

So far we have collected dialogues from 17 subjects. Approximately half of them were students, all of them with limited previous experience with computers.

2.1 The tagging system

For the purpose of analyzing our dialogues, we have developed a tagging system, called DagTag (Ahrenberg & Jönsson 1987). Here we will only give a very brief description of some of its main features.

In DagTag, tagging can be done on five different levels: (1) Dialogue, (2) Sequence, (3) Utterance, (4) Clause/Move, (5) Phrasal constituent, i.e. the system assumes that a Dialogue can be analyzed into a sequence of Sequences, which in turn can be analyzed into a sequence of Utterances, and so on. The category of Phrasal constituent is recursive. Thus, a tree structure is imposed on the dialogue during the analysis.

The practical work using DagTag consist of four phases: (1) preparing dialogues for tagging, (2) creating tag menus, (3) tagging, and (4) analysis. The first of these phases is not discussed further. The second phase consist of developing tagging categories and a tagging praxis; developing a taggers' manual. Tagging categories, i.e. the attribute — value-pairs, we call *descriptors*, This is a difficult phase; our choice of tagging

descriptors is discussed in more detail below. When the dialogues have been prepared and the menus created, tagging can begin. Tagging is done by marking a part (dialogue, sequence, utterance, or word sequence) in the text and then producing the appropriate tag structure from the menus. These tag structures are similar to what is sometimes called a DAG (Directed Acyclic Graph), (Kay 1979).

After the tagging, the tagged dialogues can be analyzed in several ways. One can count the occurrences of a specific word or any string sequence using wild cards or special string markers. But more important is the use of model dags. A model dag is a dag created as in the tagging phase, but this time used for finding the occurrence of this special construction in the text. For instance, one can specify a sequence of dags for counting the number of questions followed by a clarification request followed by a clarification and finally the appropriate answer, or maybe the first three dags as before but this time followed by a new clarification request, i.e. the subject did not express himself clearly the first time (see below for a discussion on tagging categories). Also we can combine model dags with string search.

3. Tagging categories

Our main effort on the issue of analyzing the dialogues has, so far, been concentrated on developing a useful set of descriptors, i.e. tagging categories. This we believe requires both a grammar and a discourse model. The grammatical tagging, i.e. tagging on the Clause and Phrasal Constituent levels has not been considered in this study, instead we have concentrated on various aspects of dialogue. As a basis we use a simplified and modified version of the IR-analysis (Linell, Gustavsson & Juvonen 1988 and Gustavsson 1988), that is we divide our utterances basically along the dimension initiative — response. Thus, we have only developed descriptors for one of the five possible levels described above, namely the utterance level. We use the following four utterance types:

Utterance Type: Initiative | Response | Resp/Init | Clarification

3.1 Initiative

Initiative means that the subject (background system user) initiates a query, but it could also be that the subject awaits a question from the system. Then the system takes the initiative. We use the following descriptors for initiatives:

Initiative: Topic: Bound | Unbound
Context: Dependent | Independent
Indexicality: Pronoun | Ellipsis | Definite Descriptor |
Pro+Ellipsis | Def descr+Ellipsis | Pro+Def descr

Topic is used to describe how an utterance is *executed* on the database. We require coherence between utterances. The idea is that if an utterance is used for updating the topic representation structure, then it is Topic Bound otherwise it is Topic Unbound. An instance of unbound topic is when the subject accesses a different part of the data base. This forces us to state in advance the various topics that could exist in the particular system being simulated. A typical example of this is in the use of the publication search system (PUB) where different books are discussed concerning different topics. For instance when the user discusses books about linguistics these utterances are coded Topic Bound. If then the next utterance asks for a book on artificial intelligence, then that utterance is coded Topic Unbound.

As the notion of topic is a bit problematic we will give some other examples. The problems with topic is most apparent in the consultation dialogue simulations, for instance the following:

```
6.A>how long is the c-line course?[1]
7.S>THE C-LINE IS 160 POINTS
8.A>is basic used on the course?
9.S>BASIC IS NOT USED ON THE C-LINE
10.A>Do you read mechanics?
```

Here both 8.A and 10.A are considered Topic Bound, we thus regard C-line as topic. Consequently, these dialogues contain very few Topic Unbound utterances. However, it is possible to consider a topic shift in 8.A but still code 10.A as topic bound if one regarded course length as one sub-topic and courses taught as another. This indicates that the delineation of what is and is not a topic is difficult in some dialogue domains (cf Grosz 1977).

Context on the other hand concerns the *interpretation* of an utterance. We code an utterance Context Dependent if it cannot be interpreted without context information. Further, we assume that the expanded form is stored in the context memory, i.e. ellipsis can be interpreted over several utterances, a construction which is very commonly used in telegraphic dialogues:

```
3.A>c-line courses?
4.S>. . .
5.A>prerequisites?
6.S>LOOKING FOR INFORMATION
  . . .
7.A>the structure of the courses?
```

7.A above can be interpreted because 5.A is stored as "Prerequisites for the c-line?" and we thus have the context needed.

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.

Context Dependent utterances are always indexical. A Context Independent utterance may be indexical too. When we have had a context shift and the user types an utterance that is indexical in the previous context, then we have Context Unbound but with Indexicality. This is an indication that a system as simple as the one proposed here is not sufficient.

3.2 Response

Response is when the system responds to a user initiative, or when the user answers a question from the system. We have the following sub-descriptors:

Response: Response type: Informative | Clarification request | Acknowledgement

Informative is used to tag normal answers. We do not discuss to what extent the system was able to give a correct answer or not. This means that answers such as

[1] The utterances have been translated into English as the simulations are conducted with subjects using Swedish. The numbers indicate the utterance sequence number and A stands for user and S for system.

"Have no information" are regarded as Informative.

Acknowledgement means that the response is not followed up or does not contain any information. Used for instance when terminating or as a confirmation.

Clarification request is used when more information is needed. It is used both when 1) the information given is incomplete: `5.A> Which subjects do you take? 6.S> Please be more specific` 6.S is a Clarification Request. Mistypings that could not be interpreted will also cause a Clarification Request. The other case is 2) when further information is needed, for instance: `11.A>What is the price now if I have a pair of cheaper loudspeakers 12.S> What loudspeakers?` 12.S is coded Clarification Request. Utterances like `Don't understand, Cannot recognize` are also coded Clarification Request.

3.3 Resp/Init

This utterance type is used in situations 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, e.g.

```
1.A> which books are there that treats the subject introductory lisp course
2.S> PUB SEARCHING. There are three books about lisp. Do you want to see them
all?
3.A> yes
4.S>PUB SEARCHING ....
```

Utterance 2.S is tagged as Resp Type: Informative and Init type: Information request. This also means that Resp/Init is used for tagging insertion sequences, as in the coding above: 1.A: Initiative, 2.S: Resp/Init, 3.A: Response, 4.S: Response, i.e. Response 3 answers Initiative 2, while Response 4 answers Initiative 1.

Resp/Init is subclassified in the same way as pure responses and pure initiatives, but without the topic and context descriptors. The result is the following descriptors:

Resp/Init: Resp Type: Acknowledgement | Informative | None
Init Type: Information request

3.4 Clarification

The utterance type Clarification is used as a response to a Response of type Clarification request and indicates what type of clarification is used. If the user changes topic, e.g. (s)he does not follow up his/hers old Initiative, but instead initiates a new query, then this is coded as a new Initiative. Clarification uses the same type of Indexicality as before on Initiative. The following descriptors are used to characterize Clarifications:

Repetition means that the question is repeated more or less exactly as before. Mistypings could be removed or synonyms used, but not synonyms which could be interpreted as changing — expanding — the response.

Expansion is used for coding utterances where the user responds on a Clarification Request with an expanded question, i.e. a question that is interpreted the same as the original question but changed. Example:

```
5.A> A> Which subjects do you take?
6.S> Please be more precise.
7.A> A> Which subjects do you take on the C-line?
```

Here 7.A is coded as an Expansion. It is not exactly the same question, but it has the same communicative meaning; we have a semantic but not a pragmatic expansion. Pragmatic expansion yields Revision.

Revision means that a revised version of the previous question is responded on a Clarification Request. Example:

```
9.A> How does the schedule for year i on the c-line look?
10.S> Do not understand i.
11.A> The schedule for the c-line?
```

As one can believe that the i in 9.A should have been a 1, 11.A is coded as a Revision, because then 11.A is less specific (asking for any c-line schedule) than 9.A (asking for only the first year). Note, that it is also possible to interpret 9.A as just a clumsily formulated question with the same interpretation as 11.A. Then we have a slight mistyping of an i, i.e. the i is just a typing error and should not have been meant as a 1, this then should have been interpreted as an expansion.

We also use Indexicality as described above on the Clarification type utterances.

4. Results and discussion

The 17 dialogues collected contain a total number of 641 utterances. The simulations conducted so far are basically pilot studies. We intend to run further simulations, varying the experimental design based on our experience from these pilot studies. Thus, our results ought to be considered preliminary, but indicate some interesting dialogue features, reported below.

We have carried out postexperimental interviews with our subjects afterwards and concluded that none of them thought that it was a simulation. This is amazing since one subject typed long and very complex sentences such as the following authentical (translated from Swedish) utterance: `15.A>Now I want to change to a better pair of loudspeakers which are at the most 29 400 SEK more than the ones I have now. The new items I'm getting should be as appropriate as possible for a small room.`

4.1 Mistyping

There are 27 mistyped utterances. 21 of these are user initiatives and only one of these contains more than one typing mistake. This indicates that our subjects were anxious to provide a correct input. This is also evident from the unprepared texts which contain a lot of delete sequences on the occasions when the user found a spelling mistake in his/hers utterance.

4.2 Initiative

There are 264 initiatives. 97 of these are Topic Unbound. In this figure are also included the initial utterances which are always Topic Unbound, and many of the dialogues contain a termination which is Topic Unbound too, but there are still a

number of topic shifts in our dialogues. In one of the dialogues there are also 13 occurrences of Topic Bound Context Independent utterances, but this construction only occurs in this dialogue — the longest dialogue, 156 utterances, with a subject using the system correctly, that is without trying to break in or make it crash.

Another interesting figure is the number of Topic Bound Context Dependent utterances. Here we have 122 utterances, which indicate that the users follow up a previous utterance about 45% of the time. In addition 35 of these 122 utterances are followed up by another Topic Bound Context Dependent utterance.

There is also one utterance that is Context Independent but with Indexicality, an utterance that our proposed discourse model cannot analyze correctly.

Taken as a whole, these results are in accordance with the results from previous studies, showing the need for handling connected dialogue in a Natural Language interface. However, the fact that with the exception of the dialogues with the HIFI system, all but one of the "incomplete" utterances can be interpreted using the immediate linguistic context, seems to indicate that not too complex computational mechanisms presumably can achieve this goal, at least in some application areas.

But even if this should be true for some applications, data from our experiments indicates that this not is true for all cases. An example of this is the 13 Topic Unbound Context Dependent utterances. All of these come from the HIFI dialogues. This is interesting, since this "system" in a sense is the most complex one. It is complex because with it two different tasks are executed; obtaining information and advice on HIFI equipment, and ordering the equipment. These two tasks are executed in parallel, or rather are they intertwined. Therefore, two different topics are active at the same time. For the human reader, it presents no difficulties to understand these utterances in spite of the many topic shifts. But it is sometimes difficult to find any surface cues indicating them. Thus, computational mechanisms for handling this type of dialogues could be more complex than in other cases. The database dialogues, on the other hand, have a simple structure, and the topic shifts are not too difficult to find.

In our opinion this shows the need for following up Grosz' (1977) observation that there are different types of dialogues with different topic structure. We need to know more about how to characterize the different classes, and we need to know more about the computational devices necessary for coping with them.

It should also be obvious from these results that generalizations from one system, be it a Wizard of Oz simulation or a real system, to other applications should be made with caution. We don't know as yet the critical dimensions causing differences in user behavior, but it is becoming clear that the differences can be both large and critical from a computational point of view.

4.3 Indexicality

There are in our study 140 utterances containing indexicality. 136 of them were initiatives and 60 were elliptical. The total number of initiatives were 264, which means that about half of the utterances contained an indexicality and of these half (or totally 25%) were elliptical. But we also found 57 occurrences of Definite descriptions, thus, that construction is equally common. However, the use of pronouns is relative rare, only 13 cases.

The low incidence of pronouns and the relatively high frequency of more or less complex nominal phrases is quite in accordance with the results obtained by Guindon

et al (1987). Perhaps we here have a linguistic feature common to human-computer dialogues in Natural Language across different applications.

4.4 Indirect speech acts

Indirect speech acts (Searle, 1975) have been one of the active areas of AI research on natural language. The computational mechanisms developed for handling them are complex (e.g. Perrault & Allen, 1980). It can perhaps therefore be of interest to note that there are only two indirect speech acts in our corpus, both of the following type: *Can you give me .....",* which perhaps can be regarded as a lexicalized expression.

4.5 Reparation strategies

The analysis of insertion sequences, as defined earlier, produced 16 utterances. However, these utterances are in one group of dialogues and actually reflect the simulator's behavior more than the user's. This then emphasizes the need for controlled experimental settings as discussed above.

5. A Final Comment

As mentioned above, in spite of the results emerging from the simulations conducted so far, these experiments should in some sense be regarded as pilot studies used for developing and refining the simulation method; simulation environment, scenarios, simulation techniques etc, and for developing the analysis method. It is therefore our intent to continue with the dialogue studies with some refinements based on our experiences form previous simulations. The most important improvement is to specify in more detail the linguistic and conceptual capacity of the simulated systems, and hereby improve the consistency of the responses.

There are a number of questions still awaiting an answer. The two most important of these are the sources of variability in user input, and the delineation of the different dialogue situations that require different computational mechanisms for handling topic shifts etc. It is clear that Natural Language interfaces in the foreseeable future will only be able to handle a subset of natural language. The usability of this type of interfaces is therefore dependent on the finding of subsets of natural language that the user can use without experiencing inexplicable "holes" in the systems performance, i.e. subsets for which we can find and handle complete linguistic and conceptual coverage. It is hard to see how such an enterprise can succeed without knowledge of the factors influencing the language used in different situations and by different people.

It is of course important for the possibility of developing portable flexible dialogue interfaces, that the factors determining the usability of a specific type of discourse representation in a given domain is made known. In our opinion the development of computational theories of discourse should be paralleled by an equally intense effort in determining the characteristics of different dialogue situations through the use of Wizard of Oz studies and in other ways. Not doing this would be as sensible as developing parsers without knowing anything about the language they should parse.

6. Acknowledgements

Lars Ahrenberg, Magnus Merkel, Mats Wirèn, and Ivan Rankin have been striving for years to increase our linguistic and conceptual coverage of conceptual and linguistic issues related to our work. We hope that some traces of this can be found in the

present paper.

This work has been supported by the National Swedish Board for Technical Development (STU).

7. References

Ahrenberg, Dahlbäck, Jönsson, Merkel, Wiren (1986) *Mot ettt dialogsystem för svenska* NLPLAB-memo. Department of Computer Science, Linköping University (*In Swedish)*

Ahrenberg, L. (1987) Parsing into Discourse Object Descriptions. In: *Proceedings of the Third European Chapter ACL Conference, Copenhagen, April 1-3, 1987* 140-147.

Ahrenberg, L. & Jönsson, A. (1987): "An Interactive System for Tagging Dialogues", Research Report LiTH-IDA-R-87-22, Linköping University Department of Computer Science. Also to appear in *Proceedings of the XIV Conference of the Association for Literary and Linguistic Computing, Göteborg, June 1-5, 1987.*

Bates, M. & Sidner, C. (1983) A Case Study of a Method for Determining the Necessary Characteristics of a Natural Language Interface. In: Degano, P: & Sandewall, E. (eds) *Integrated Interactive Computing Systems.* New York: North Holland

Burton, D. (1981) Analysing spoken discourse. In M. Coulthard & M. Montgomery (Eds.) *Studies in Discourse Analysis.* London: Routledge & Kegan Paul.

Chafe, W.L. (1982) Integration and involvement in speaking, writing and oral literature. In D. Tannen (Ed.) *Spoken and written language: Exploring orality and literacy.* Norwood, N.J.: Ablex.

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.

Deutsch (Grosz) B.G. (1974) The structure of task oriented dialogs. *Proceedings of IEEE symposium on Speech Recognition*

Eagan, O., Forrest, M-A., Gardiner, M., Reilly, R., Sheehy, N. (1986) *Deliverable 3, ESPRIT Project P527. Dialogue studies — pilot phase*

Fineman, L. (1983) Questioning the need for parsing ill-formed inputs. *American Journal of Computational Linguistics, 9*(1), 22.

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.

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

Grosz, B.J. (1981) Focusing and Description in Natural Language Dialogues. In A. Joshi, B. Webber & I.Sag (Eds.) *Elements of Discourse Understanding* Cambridge University Press.

Grosz, B.J. & Sidner, C. (1986) Attention, Intention, and the Structure of Discourse, *Journal of Computational Linguistics, 12*(3) 175-204.

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.

Gustavsson, L. (1988) *Language Taught and Language Used. Dialogue Processes in Dyadic Lessons of Swedish as a Second Language Compared with Non-Didactic Conversations.* Ph.D Thesis. University of Linköping. Department of Communication Studies.

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

Kay, M. (1985): "Functional Grammar". *Proceedings of the Fifth Annual Meeting of the Berkeley Linguistic Society.*

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

Lehnert, W. & Ringle, M. (1982) *Strategies for Natural Language Processing.* Hillsdale, N.J. Lawrence Erlbaum Associates.

Linell, P. (1982) *The 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 Mangements: An Analysis of Requirements. In: *Proc. IJCAI-77.*

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

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

Reilly, R. (1987b) Communication failure in dialogue: Implications for natural language understanding. Paper presented at the seminar: *Recent Developments and Applications of Natural Language Understanding*, London, Dec 8-10.

Ringle, M.H. & Bruce, B.C. (1980) Conversation Failure. In: W.G. Lehnert & M.H. Ringle *Strategies for Natural Language Processing* Hillsdale, N.J. :Erlbaum.

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

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 Dep., IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y.

Thompson, B.H. (1980) Linguistic analysis of natural language communication with computers. *Proceedings of the 3rd International Conference on Computational Linguistics*. tokyo, Japan.

Walker, D. (ed) (1978) *Understanding Spoken Language*, New York: New Holland

Waltz, D.L. & Goodman, B.A. (1977) Writing a natural language data base system. *Proc. IJCAI-77*, Cambridge, Mass.: MIT, 1977.

Waltz, D.L. (1978) An English question answering system for a large relational database *CACM* 21(7), july, 1978