# INTERACTION WITH AN ANIMATED AGENT IN A SPOKEN DIALOGUE SYSTEM

*Linda Bell and Joakim Gustafson*

Centre for Speech Technology,
Department of Speech, Music and Hearing, KTH
SE-100 44 Stockholm
{bell, joakim_g}@speech.kth.se

## ABSTRACT

The study reported in this paper is based on results from a Swedish database of spontaneous computer-directed speech. This database was investigated to determine how people adapt their language when they interact with computers. A spoken dialogue system with an animated agent, August, was installed in a public location in downtown Stockholm. Members of the general public were invited to interact with the system and dialogues were recorded. The database was collected during a period of six months and consists of transcriptions of more than ten thousand spontaneous utterances. The domain restrictions in this spoken dialogue system were minimal, and the users were not explicitly told what they could expect the system to understand. In this paper the users' communicative strategies, as they are manifested in the input utterances, are studied. The influence of the interface design on user expectations is also discussed. Results indicate that user adaptation, as reflected in the corpus, comprises lexical as well as syntactical aspects

Keywords: spontaneous computer-directed speech, user strategies, linguistic adaptation

## 1. INTRODUCTION

When people interact with spoken dialogue systems, they try different approaches to make their communication successful. For instance, people who interact with computers often adapt their linguistic behavior to meet the demands of the system [1]. This adaptation takes place regardless of whether the user's input to the system is spoken or typed [2]. Users are even more likely to modify their language when the interaction with the system does not run smoothly, i.e. during human-computer error resolution. In a study of multimodal human-computer interaction, Oviatt and VanGent [3] have shown that users distinguish repetitive utterances from original failed input by means of linguistic contrasts. Most experienced users of spoken dialogue systems are aware of system limitations in terms of lexicon, and are thus likely to limit their vocabulary. Inexperienced users, on the other hand, may expect spoken dialogue systems to handle large or even unlimited vocabularies. Studies have also shown that people who interact with computers tend to reuse words and structures used by the system itself [2, 4, 5]. It is assumed that if a specific lexeme, phrase or idiomatic expression occurs as system output, it should be possible to recognize as user input as well. Other user strategies include keeping utterances short and avoiding anaphoric pronoun use [1]. When the interaction with the system goes wrong, users are likely to attempt other, more advanced strategies for resolving errors.

The system interface design also contributes to people's expectations and may indirectly influence their linguistic behavior as they interact with the system. A central feature of the present system was that the August's synthetic face was intended to appear human-like and that he was given a "personality". The agent's synthetic speech output was lip-synchronized and extra attention had been given to make the prosody of the voice sound natural. Moreover, August's face was given a varied set of extra-linguistic gestures and expressions, such as eyebrow-movements, movements of the head and movements of the eyes and eyelids indicating that he was "thinking" [6]. The August system was constructed to handle a number of simple domains rather than one complex domain. The animated agent's human-looking appearance made us anticipate that the users would want to exchange greetings and socialize with him. The ability to handle and respond to these social utterances was built into the system. It could be argued that the animated agent's face, with the above-mentioned features, influenced the way in which the users interacted with the August system.

In the present paper, some of the linguistic strategies employed by the users of the August system are investigated. The aim is to examine whether the type of user adaptation that is characterized by short utterances, avoidance of anaphora and lexical limitations is prevalent in the current database. Lexical aspects of the August corpus will be discussed and word and utterance statistics will be presented. Adaptations in syntactical structure will also be investigated, especially in the context of repetitive utterances spoken during error resolution.

## 2. MATERIAL

### 2.1 Data

A spoken dialogue system with an animated agent was set up in a public location and spontaneous human-computer interactions were recorded during a period of six months [7]. The people who interacted with the August system were mostly inexperienced users of spoken dialogue systems and they were given little or no information as to what they should expect the system to understand. The material analyzed in this paper consists of 10,058 utterances of computer-directed speech. All of these utterances were transcribed orthographically and some basic speaker characteristics were manually labeled, so that men, women and children among the users of the system could be roughly divided into groups. No attempt at estimating the individual ages of the subjects was made. The total number of users was 2685, out of which 50% were men, 26% women and 24% children.

The average number of utterances per user in the database was 4.1 for men, 3.3 for women and 3.5 for children. The utterances were then divided into the categories *information-seeking* and *socializing*, respectively. Typical utterances in the information-seeking category include "Hur mycket är klockan?" (What time is it?) and "Var finns det restauranger i Stockholm?" ("Where are there restaurants in Stockholm?) while the socializing category can be exemplified by "Hej August!" (Hello August!) and "Hur gammal är du?" (How old are you?). This categorization of the corpus was performed by the present authors and involves a subjective element. The utterance categories are comprehensively described in [8].

### 2.2 Processing of data

After all the utterances had been transcribed and categorized as described above, they were automatically sorted according to their frequency in the database. The entire corpus was part-of-speech tagged and parsed. All unique words were extracted from the database and the number of words per utterance was counted.

The words that had been given the part-of-speech tag indefinite pronoun were excerpted from the corpus. A number of these indefinite pronouns were believed to refer to back to previous turns in the dialogue and could therefore constitute instances of anaphora. Because these expressions can be ambiguous in Swedish, they had to be manually labeled to make sure only genuinely anaphoric references were included. Therefore, it was necessary to listen to the individual input utterances in which the indefinite pronouns occurred as well as the utterances spoken just before and after these.

When users of a spoken dialogue system fail to make themselves understood, they often repeat or rephrase what they have just said. Repetitions and near-repetitions should therefore be interesting from the point of view of user adaptation. Exact and approximate repetitions together constituted 12% of all utterances in the current database. Results from a phonetic investigation of exact, lexically identical repetitions in the August database have been reported in [9]. However, only those approximate repetitions displaying lexical variation were examined in the present study. This sub-group made up 4% of all utterances in the corpus, or 402 utterances. The purpose of studying repetitive utterances in this paper was to get a clearer picture of any changes of syntactical patterns that occurred during error resolution.

## 3. ANALYSIS

### 3.1 Word statistics and lexicon

The total number of words in the database was 39,230, out of which 23,604 words belonged in the information-seeking category and 15,626 in the socializing category. 2918 word forms in the corpus were unique and half of them were hapaxes, i.e. words that occurred only once in the corpus. In the Waxholm spoken dialogue system, which provided information on boat traffic, the number of unique words was 600 [10]. The 200 most frequently used words in the Waxholm database covered 92% of all words. By contrast, the 200 most frequently occurring words in the August database covered 81%. This can be seen in Figure 1 below.

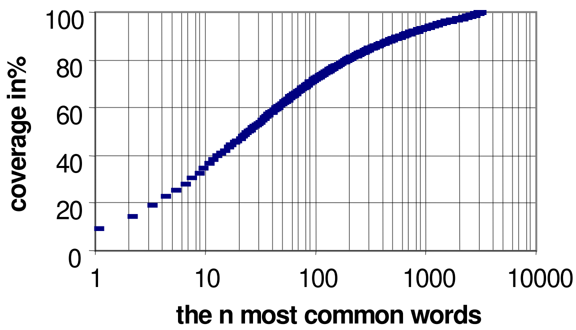


*Figure 1. Word coverage as a function of frequency.*

Some differences between the lexicons of the two utterance categories in the corpus can be seen. 1431 of the words occurred only in the information-seeking category, and 20% of these were hapaxes. The corresponding figure for the socializing category was 852 words, out of which as many as 67% were hapaxes. The overlap between the two categories was 632 words, many of them common function words and main and auxiliary verbs. This distribution can be seen in Figure 2 below. The most frequently occurring words in the August database were then compared with the most frequent words in the KTH corpora, a database consisting mainly of large quantities of Swedish newspaper text (approximately 150 million words). When the 200 most frequent words in the KTH corpora had been listed, it could be noted that only 14 of those words were not present in the August corpus. The corresponding figure for the Waxholm system was 75 words. However, 99 of the words that occurred in the August corpus could not be found in the large KTH corpora. Most of these words were names, expletive expressions (often including swear words) and nonsense words.

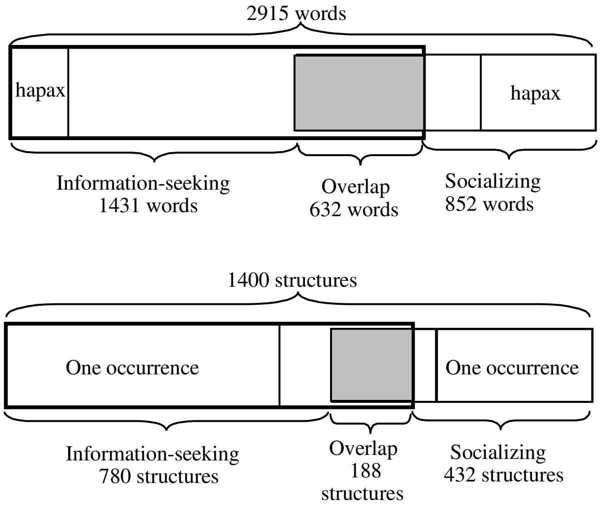


*Figure 2. The lexical and syntactic distribution of utterances in the corpus according to their category.*

The average number of words per utterance was 3.8 for men as well as children and 4.3 for women. The average utterance in the database was thus rather short. An interesting finding was that the utterances did not become shorter and more "telegraphic" as the interaction with the system went on, but rather retained the same length or even became longer. As Figure 3 illustrates, there were no relevant differences between the user groups in this aspect.

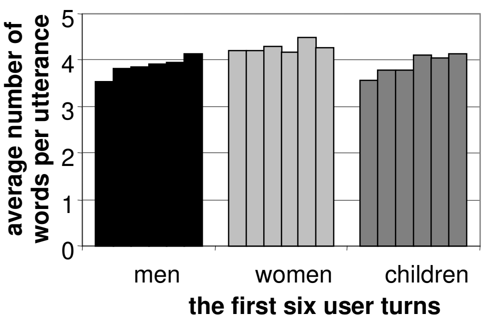


*Figure 3. The average number of words in the first six user turns.*

## 3.2 Syntactic aspects

Judging from the average number of words per utterance, there were not many instances of syntactically complex structures in the present corpus. The parsing results yielded lists of phrase level structures, for example (NP) (VP) (NP) (PP). It turned out that more than half of all the utterances in the database could be covered by ten such syntactical structures. The syntactical patterns of the phrase structures of the utterances found in the socializing category were more uniform than the ones in the information-seeking category. 780 structures occurred only in the information-seeking category while the corresponding figure for the socializing category was 374. Most of these, however, occurred only once, see Figure 2.

The overlap between the two categories, 188 phrase structures, covered 80% of all the utterances in the database. Most utterances contained only one verb phrase. Furthermore, in the cases were two verb phrases were found, the heads of these phrases were usually one main and one auxiliary verb. A majority of utterances were made up of one single clause and coordinated or subordinated clauses rarely occurred. To sum up, most utterances in the database can be described as syntactically simple.

Indefinite pronouns that were used anaphorically occurred in 1.6 % of the utterances in the corpus. Even though these anaphorically used pronouns were not very frequent, it is interesting that they occur. Their existence in the database does point to an interesting observation: users talk about the system and refer back to previous turns in the dialogue. In 10% of these meta-utterances, anaphoric indefinite pronouns could be found. An example of a dialogue with two anaphoric pronouns can be seen in Figure 4 below. This excerpt from the corpus illustrates anaphoric uses of pronouns that refer to entire utterances rather than to previously mentioned noun phrases. The anaphoric pronouns in the present database often had an entire utterance as their antecedent. Moreover, the dialogue excerpt in Figure 4 below indicates the difficulty of handling utterances that refer back to previous turns in the discourse or the dialogue itself.

## 3.3 Repetitive utterances

The results from the study of repetitive utterances in the database were not uniform. Generally speaking, most users seemed to use repetitive utterances in a linguistically contrastive manner. The original input to the system was distinguished from the repetitive utterance(s) by means of changing of word order, exchanging of lexical items, the addition or subtraction of a word or phrase, and topic shifting.

In about 11% of the repetitions, the complexity of the original utterance was reduced. If the original utterance had been a full sentence, the repetitive utterance would in most of these cases contain no more than a single word. In about 9% of the repetitions, nonetheless, the opposite happened. Here, an original utterance consisting of a single word was changed to a full sentence as it was repeated. In 13% of the repetitions, the users deleted a word or phrase, while an additional word or phrase was added in just as many cases. When the subjects repeated the same utterance more than once, they often alternated one feature such as complexity or insertion/deletion of a word or phrase. This pattern of linguistically contrastive pairs occurred in 41% of these cases. If different features were used, something which occurred in 21% of these repetitive sequences, it was mostly the case that the first repetition displayed a simpler sentence structure than the original input, and that a lexical item was changed in the second repetition.

| User | | System |
|---|---|---|
| What is your name? | → | I call myself Strindberg, but I don't really have a surname |
| When were you born? | → | Strindberg was born in 1849 |
| What do you do for a living? | → | I can answer questions about Strindberg, The Royal Institute of Technology and Stockholm |
| How many people live in Stockholm? | → | Over a million people live in the Stockholm area. |
| Do you like it here? | → | People who live in glass houses should not throw stones. |
| Yes, **that** was a smart thing to say | → | Hello, I come from the department of Speech, Music and Hearing |
| The Royal Institute of Technology | → | The information is shown on the map |
| Thank you | → | You are welcome |
| Goodbye | → | Perhaps we will meet soon again |
| Yes, **it** might be that we will | → | Strindberg was married three times |

*Figure 4. A translated excerpt of the transcription of a dialogue taken from the August corpus. The user is a Swedish male and the dialogue in its entirety went on for 15 turns.*

# 4. DISCUSSION

The following tendencies were observed in the present corpus: Firstly, considering that the users had not been instructed about how to interact with the system and that the system had more than a single domain, the number of unique words in the corpus is not large. This indicates that the users were not very imaginative, but used many common expressions. However, a number of odd words and expressions were found in the database. These words occurred more often in the information-seeking category than in the socializing category. Secondly, a relatively small number of phrasal combinations covered almost all of the syntactic structures in the database. The utterances were mostly quite short and simple. With a few exceptions, users seemed to avoid syntactically complex utterances in their interaction with the system. Most of the complex structures occurred as people referred back to the discourse during error resolution or when they were testing the limits of the system. Thirdly, anaphorically used pronouns did not occur frequently in the database. When these anaphoric constructions were found, nonetheless, their antecedents were entire utterances rather than single noun phrases. An interesting observation was that people not only referred to previous turns in the dialogue but also talked about the system itself. Finally, user strategies in the repetitive utterances included lexical as well as syntactical modifications. These features were often used in a contrastive manner.

To the extent that linguistic adaptation did take place in the August corpus, the influence of the user interface cannot be disregarded. The number of utterances in the database in the socializing category is large, and these social interactions often lasted for several turns. It is reasonable to believe that this, at least in part, can be explained by the existence of the animated agent. Even though the animated agent probably made user expectations rise, it also in a sense made the language input to the system simpler to handle. The reason for this is that the language people use for socializing is largely constructed of idiomatic expressions and prefabricated phrases. Greetings and social phrases contain a limited lexicon and very few linguistically complex structures.

# 5. CONCLUSION

In this paper, results from a database of spontaneous computer-directed speech were presented. It was observed that people who interact with spoken dialogue systems use a range of different strategies to make their communication with the spoken dialogue system successful. However, statistical analyses of the corpus indicate that a system with a limited lexical and syntactical capacity can handle most of the things people say while interacting with an animated agent.


## 6. ACKNOWLEDGEMENTS

The authors wish to thank all the users of the August system and the transcribers of the database for their contribution. Nikolaj Lindberg and Alice Carlberger provided us with tools for the syntactic analysis.


## 7. REFERENCES


[1] Kennedy, A., Wilkes, A., Elder, L. & Murray, W. (1988), Dialogue with machines, *Cognition*, 30 pp 73-105.

[2] Zoltan-Ford, E. (1991), How to get people to say and type what computers can understand, *International Journal of Man-Machine Studies*, 34, pp 527.47

[3] Oviatt, S. and VanGent, R. (1996). Error resolution during human-computer error resolution In *Proceedings of ICSLP'96*

[4] Brennan, S. (1996), Lexical entrainment in spontaneous dialog, *Proceedings of ISSD*, 41-44

[5] Gustafson, J. Larsson, A. Carlson, R. & Hellman, K. (1997), How do system questions influence lexical choices in user answers?, *Proceedings of Eurospeech'97*

[6] Lundeberg, M. and Beskow, J. (1999), Developing a 3D-agent for the August dialogue system, To be pub. in AVSP'99.

[7] Gustafson, J., Lindberg, N. and Lundeberg, M. (1999), The August Spoken Dialogue System, To be pub. in Eurospeech'99

[8] Bell, L. and Gustafson, J. (1999), Utterance types in the August dialogues, To be published in IDS'99, ESCA workshop on Interactive Dialogue in Multi-Modal Systems.

[9] Bell, L. and Gustafson, J (1999), Repetition and its phonetic realizations: Investigating a Swedish database of spontaneous computer directed speech, To be published in ICPhS'99.

[10] Bertenstam, J., Blomberg, M., Carlson, R., Elenius, K, Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., de Serpa-Leitao, A., Nord, L. and Ström, N. (1995), Spoken dialogue data collected in the Waxholm project STL-QPSR 1/1995