Utterance types in the August dialogues

Linda Bell and Joakim Gustafson

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

ABSTRACT

August, a Swedish multi-modal spoken dialogue system featuring an animated agent, was used to collect a database of spontaneous computer-directed speech. The system was designed with several simple domains rather than a single complex one. It was installed in a public location in the center of Stockholm and was available for the general public during six months. The people who interacted with the system were given little or no information on what they could expect the system to understand. In this paper, we draw on the experiences gathered in course of the development of the August system and discuss findings in a database of more than 10,000 utterances. The aim of the paper is to address the issue of how the different types of utterances found in the database reflect the demands and strategies of the users. The categorization of the speech input into utterance types and the subsequent analysis of these types will be discussed, and possible implications for modeling future dialogue systems will be covered briefly.

Keywords: spontaneous computer-directed speech, utterance types, multi-modal spoken dialogue systems

1. INTRODUCTION

Future information systems will benefit from using speech technology components, both in terms of accessibility and user-friendliness. These systems should be designed so that they can be operated by non-specialists. The experimental Swedish August system was developed to study how non-trained users would communicate with an animated agent using spoken language.

Traditionally, spoken dialogue systems have been designed with a specific and rather narrow domain in mind. Many spoken dialogue systems have been set up with a structured dialogue model that specifies every aspect of the human-computer interaction. In a recent study, Heeman et al. suggest that this is motivated by the fact that structured systems presently are the only systems that are relatively simple to build [1]. The human-computer interactions in such systems are almost exclusively system-directed and place limits on the users' linguistic input to the system.

In the development of spoken dialogue systems, Wizard-of-Oz experiments are often used in an initial phase to collect speech data. However, it could be argued that these experiments are insufficient substitutes for genuine man-machine interaction. Allen et al. have argued that unless we have working systems, it becomes nearly impossible to make fair evaluations of models and theories on dialogue management [2]. In the construction of the present system, the Wizard-of-Oz-approach was not used. Instead, the experimental August system was exposed to the general public from the very beginning. The system had a recognition lexicon of 500 words and idiomatic phrases. However, the system responses had been manually pre-processed, integrating prosodic information as well as head- and facial movements [3]. This resulted in a system that sometimes appeared to handle almost anything and which generated quite human-like dialogues, while it sometimes failed almost completely.

One of the aims of the August system was to collect spontaneous speech input from people who had little previous experience of spoken dialogue systems. The corpus has been studied from two perspectives: Firstly, user reactions during error resolution have been analyzed, revealing how users changed their way of speaking when the dialogue fails [4, 5]. Secondly, dialogues where the system responses were adequate in most dialogue turns, making the system appear reasonably intelligent [4]. Many of the people who interacted with the August system seemed to be more interested in making the system respond to their spoken input rather than searching for information. This resulted in a large number of what in our analysis is referred to as **social** utterances. In addition, the users asked factual questions that were clearly out-of-domain, commented on the system itself and previous dialogue turns and sometimes even tried to deceive the system. Dialogue management that is based on the notion of a structured dialogue with a clearly defined task and a single complex domain will not be able to deal with all of these types of utterances. The August system was designed with a number of simple domains instead of a single complex one, and one of these domains handled greetings and other social utterances. Nonetheless, it is clear that the performance of the system did not always match the users' expectations. This is also reflected in the current database, which contains a number of utterances that are referred to below as **insults**.

In this paper, we describe the categorization of the spoken input in the August database into utterance types. The aim of this paper is to address the issue of how these utterance types reflect the expectations and strategies of the users of the spoken dialogue system. Lexical and syntactic aspects of the August corpus are also briefly examined to see whether the utterance types are distinguishable in terms of linguistic complexity. Differences between how adults and children interact with the spoken dialogue system are considered, and implications for future dialogue systems are suggested.

2. METHOD

This study is based on speech data collected during the six-month period when the August system was functional in public, and consists of recordings and transcriptions of spontaneous computer-directed speech. The system had several simple domains, and the users were not explicitly told what they could expect the system to understand. These simple domains comprised facts about the agent's namesake August Strindberg (a 19th century Swedish author), Stockholm, KTH and the locations of restaurants and other facilities in the city. Moreover, a number of frequently used socializing idiomatic expressions were added to the lexicon and appropriate system responses were supplied. No users were supervised or observed in the process of actually interacting with the system. The study is therefore based solely on the recorded soundfiles of these man-machine interchanges.

The material analyzed in the present paper consists of 10,058 utterances. These utterances were transcribed orthographically and labeled with some basic speaker characteristics. The total number of speakers was 2685, out of which 50% were judged to be men, 26% women and 24% children. The average number of utterances per user was 4.1 for men, 3.3 for women and 3.5 for children. The number of utterances originating from a single speaker ranges from one to forty-nine. The number of words per utterance in the August corpus also varied greatly, although most utterances were relatively short. The average utterance contained four words, but this figure varied from a single word to twentytwo words.

3. ANALYSIS

3.1. Lexicon and syntax

The database consists of almost 40,000 words, out of which close to 3000 are unique. Approximately half of these unique words occurred only once in the database. The 200 most frequently occurring words in the database covered about 80% of all the words in the corpus. When the August data was compared with an equally large amount of data taken from a corpus of Swedish newspaper texts, it appears as if the number of different words follows the same type of logarithmic growth pattern. In Figure 1, the actual datapoints of these two corpora as well as the calculated growth functions for the number of different words are plotted. As can be seen in Figure 1, the different words in the written corpora are roughly four times as many as in the August corpus, and this relationship appears to be constant as these corpora are increased in size.

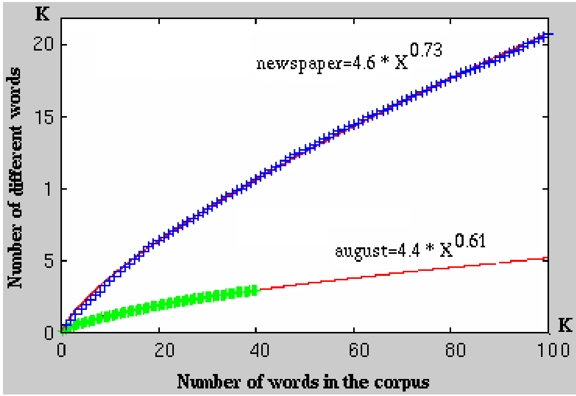


Figure 1. The correlation between the number of different words and the size of the corpora

The input utterances were mostly short and their syntactic patterns were seldom complex. As reported in [4], 188 sentence types covered 80% of all the utterances in the database. The number of sentence types differed greatly between the different categories of utterances in the database. 35 sentence types covered 80% of the utterances used for seeking information from the system while 26 sentence types covered 80% of the utterances used for socializing with the animated agent. The largest number of different sentence types was used during error resolution and by users who were testing the limits of the system.

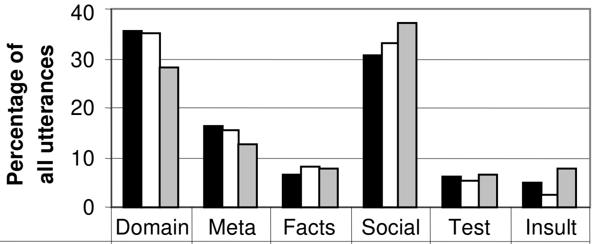
3.2. Utterance types

In the analysis of the August database, the utterances were labeled according to the presumed intentions of the users. The purpose of this categorization was to get a better picture of the kinds of things the users wanted to convey when interacting with the system. Were the users trying to retrieve information or were they merely interested in socializing with the animated agent? The concept of communicative intention is a difficult one, both in human-human and human-computer interaction [6]. Categorizing of utterances always involves an arbitrary element, as one and the same utterance may express different communicative intentions depending on the context. Moreover, which and how many categories to use can be a problematic issue. It should be noted that natural language utterances in human-human dialogues often have been divided into a much larger number of groups than the present model contains [7]. Nevertheless, the utterances in the database were categorized in accordance with a simplified pragmatic model containing six major categories.

Table 1. The utterance types in August database. The example sentences have been translated into English

Socializing	Examples
Social	*Hello August!* *That's a nice mustache!*
Insult	*You are stupid!* *Is your brain too small*
Test	*What is my name?* *I want to rent a refrigerator*
Info-seeking	**Examples**
Domain	*How many books did Strindberg write?* *What can you study at KTH?* *Where are the restaurants on Kungsgatan?*
Meta	*What can I ask you?* *I told you that already!*
Facts	*What's the capital of Finland?* *What is two times two?*

Table 1 is an overview of the utterance types in the August database. The **social** category consisted of greetings and remarks of a personal kind, while expletive expressions and swear words were placed in the category of **insults**. The category called **test** contained utterances that were spoken with what appeared to be the purpose of deceiving the system. The **domain** category included utterances in one of the established domains which the users had been given some information about. Questions about the system itself and comments about the actual dialogue were grouped in the **meta** category. Factual questions outside the domains mostly turned out to be of an encyclopedic nature and sometimes dealt with things people would expect a computer to be good at, such as calculus. These utterances were categorized as **facts**. Figure 2 shows the distribution of these utterance types in the August database. Some differences between how men, women and children communicated with the system can be observed. For example, children in our study made use of social utterances to a greater extent than adults did. One possible explanation might be that the other domains did not particularly appeal to children. Women rarely used insults, while the children in the present study used them rather frequently.



	Domain	Meta	Facts	Social	Test	Insult
men	35,6	16,4	6,4	30,8	6,0	4,9
women	35,3	15,5	8,2	33,1	5,3	2,6
children	28,3	12,6	7,9	37,1	6,4	7,7

Figure 2. Distribution of the utterance types

In order to be able to get a better overview of the utterance categories in the database, two main groups were created from the above mentioned six. The first one, *socializing,* includes the categories **social**, **test** and **insults** while the second one, *information-seeking,* includes the categories **domain**, **facts** and **meta**. The socializing category constituted 44% of all utterances in the corpus. Figure 3 below points to the different strategies used by those people whose interaction with the August system lasted for more than two turns. It suggests that men more often began by socializing with the system and then turned the dialogue to the area of information-seeking, while women more often focussed on more domain- and fact-oriented questions from the beginning. In contrast, many of the children used only social utterances over the first turns. Very few users alternated between information-seeking and socializing during their first six turns. There seemed to be four distinguishable groups of users: firstly, those who only wanted to socialize, secondly, those who only wanted to seek for information, thirdly, those who began by using some greeting remarks and then turned to information-seeking and remained in that area. The final group was a small one, and it consisted of users who tried to communicate with the system, but failed, and therefore alternated between information-seeking and socializing, trying to get the system to understand.

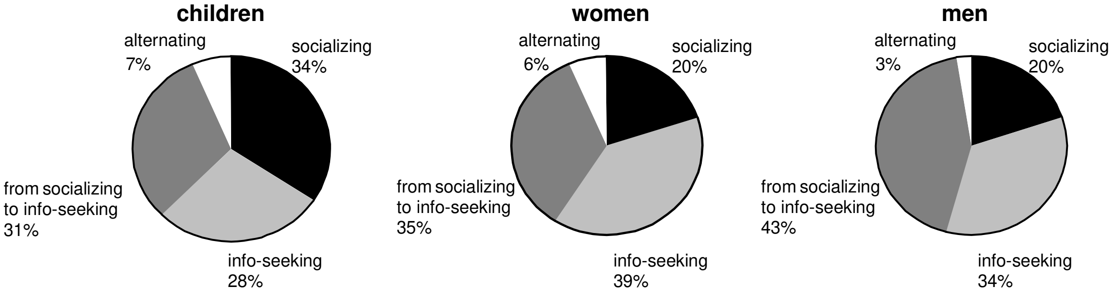


Figure 3. The distribution of speakers with respect to their usage of the utterance categories socializing *and* info-seeking. *The statistics are based on the first utterances (up to six) from all users that said more than two utterances to the system. These constitute 67% of all utterances from children and women, and 58% of all utterances from men*

4. EFFECTS OF SYSTEM PROMPTING

The analysis of the present database indicates that the users of the August system can be divided into different groups depending on their dialogue strategies. The users seem to either want to socialize with the system or search for information. An important question was whether it was possible to make the users talk about the selected domains of the system instead of merely socializing. In order to study this in the August corpus, those utterances that occurred immediately before and after certain system prompts were analyzed. These selected system prompts were supposed to be generated when the users had asked what they could say to the system. The animated agent then responded either: *I know where certain streets are located* or *I know things about Strindberg, KTH and Stockholm*. Figure 4 below shows how these prompts influenced the users when they were mistakenly generated due to recognition failure.

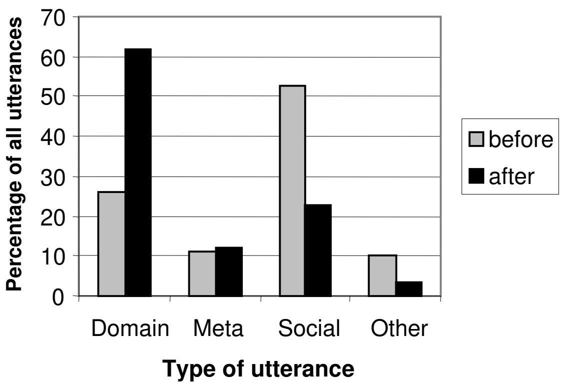


Figure 4. The effects of system prompting

As can be observed, 63% of the users actually conformed to the system by immediately asking about one of the topics mentioned in the preceding prompt. The number of domain-related utterances before these prompts appeared was only 26%. The number of utterances in the socializing category decreased significantly after such a prompt, and only rarely did the users talk about something in one of the other categories.

5. DISCUSSION

Considering the fact that the August system had several different domains and that the dialogue model was not strictly specified, the input utterances in the August database were generally quite simple. People with little experience of spoken dialogue systems have different expectations and make use of a variety of strategies as they interact with such a system. Some users started by looking for information immediately, while others preferred to socialize with the system before going into this mode. Almost half of all the utterances in the August database were categorized as socializing. Children appeared to be especially inclined to socialize with the system, perhaps because their interest in the established domains was low. The animated agent's human-like appearance probably made this social interaction make sense.

In order to build a robust system with multiple domains that can deal with different types of users, a dialogue manager that can handle a variety of input should be included. Since most social interactions are rather simple to handle from a linguistic point of view, such a domain could be added to a system at little extra cost. The dialogue manager should be able to adjust to users with different strategies. A user who is already familiar with the domains of the system and who wants to seek information at once should be able to do so, while a user who wants to socialize with the system should be allowed to do so. A future dialogue system could include a system-directed phase in which some general information about the user, such as gender and age, could be retrieved. The purpose of such a phase would be twofold: firstly, to give the user an idea of the limits of the system and get the dialogue going, and secondly to allow the system to adjust to the user and to determine which domains to suggest. The user could then be guided into asking specific questions in a chosen domain. System prompts could indicate which questions would be possible to ask. Finally, the user could take initiative of the interaction and retrieve information.

6. ACKNOWLEDGEMENTS

The authors wish to thank the transcribers of the August database for their contribution. Nikolaj Lindberg provided us with useful comments and suggestions.

7. REFERENCES

1. Heeman, P. A., Johnston M., Denney J. and Kaiser, E. (1998) Beyond structured dialogues: Factoring out grounding. In *Proceedings of ICSLP '98*
2. Allen, J. F. et al. (1996) A Robust System for Natural Spoken Dialogue In *Proceedings of 34th meeting of the Association for Computational Linguistics*
3. Gustafson, J., Lindberg, N. and Lundeberg, M. (1999) The August Spoken Dialogue System, Submitted to Eurospeech '99
4. Bell, L. and Gustafson, J (1999) Interaction with an animated agent in a spoken dialogue system Submitted to Eurospeech '99.
5. Bell, L. and Gustafson, J. (1999) Repetition and its phonetic realizations: Investigating a Swedish database of spontaneous computer-directed speech. Submitted to ICPhS '99
6. Cohen, P., Morgan, J. and Pollack, M. E. (eds.) (1990) *Intentions in Communication*. Cambridge: MIT Press
7. Stolcke, A. et al. (1998) Dialog act modeling for conversational speech. In *Papers from the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, 98-105.