Natural Language Engineering **6** (3–4): 273–286. Printed in the United Kingdom

© 2000 Cambridge University Press

Speech technology on trial: Experiences from the August system

JOAKIM GUSTAFSON and LINDA BELL

Centre for Speech Technology, Royal Institute of Technology (KTH) Drottning Kristinas väg 31, 100 44 Stockholm, Sweden
e-mail: {jocke,bell}@speech.kth.se

(*Received 21 September 1999, revised 31 January 2000*)

Abstract

In this paper, the August spoken dialogue system is described. This experimental Swedish dialogue system, which featured an animated talking agent, was exposed to the general public during a trial period of six months. The construction of the system was partly motivated by the need to collect genuine speech data from people with little or no previous experience of spoken dialogue systems. A corpus of more than 10,000 utterances of spontaneous computer-directed speech was collected and empirical linguistic analyses were carried out. Acoustical, lexical and syntactical aspects of this data were examined. In particular, user behavior and user adaptation during error resolution were emphasized. Repetitive sequences in the database were analyzed in detail. Results suggest that computer-directed speech during error resolution is increased in duration, hyperarticulated and contains inserted pauses. Design decisions which may have influenced how the users behaved when they interacted with August are discussed and implications for the development of future systems are outlined.

1 Introduction

Future information systems will benefit from using speech technology components, both in terms of accessibility and user-friendliness. Systems should be designed so that they can be operated by non-specialists using unconstrained language. The experimental Swedish August system was developed to study how non-trained users would communicate with an animated talking agent. One of the objectives of the August project was to collect spontaneous speech from members of the general public, rather than students, paid subjects or users of commercial systems. As a complement to data collected in systems where well-informed and motivated users seek information in a single domain, the August corpus contains computer-directed speech in several domains from a wide range of users from the general public.

This paper is organized as follows: first, the background section offers a brief overview of methods of collecting speech data and discusses user expectations in spoken dialogue systems. The design of the August system is then described. The speech corpus collected in the course of the August project is presented and we report on some of the findings in this corpus. Subsequently, user behavior and user

https://doi.org/10.1017/S1351324900002485 Published online by Cambridge University Press

adaptation during error resolution are discussed. Finally, we summarize some of the things learned from the August project and discuss possible implications for the construction of future spoken dialogue systems.

2 Background

Traditionally, spoken dialogue systems have been designed with a specific and rather narrow task in mind. Many systems have been set up using a structured dialogue model in which all aspects of the human-computer interaction are specified. In a recent study, Heeman, Johnston, Denney and Kaiser (1998) suggested that this is motivated by the fact that these systems presently are the only systems that are relatively simple to build. The interaction in such systems is almost exclusively system-directed, something which places restrictions on the users' linguistic input.

In the development of spoken dialogue systems, Wizard-of-Oz experiments have often been used in an initial phase to collect speech data. Although it can be useful to let a human simulate some part of a system, results from such experiments tend to be difficult to assess. Allen, Miller, Ringger and Sikorski (1996) have argued that unless we have working systems, it becomes almost impossible to make fair evaluations of models and theories on dialogue management. This makes the collection of speech data using alternative methods necessary. The general problem which Thomas (1995: 306) refers to as that of "testing for technologies that do not exist" is definitively a non-trivial one, and it is conceivable that combining several different methods of data collection will be the most productive approach in the long run. Fraser (1997) has described a system-in-the-loop method for the development of spoken dialogue systems. This method is used to collect speech data from real users interacting with an almost finished system. Cheyer, Julia and Martin (1998) report on a promising experiment in which a Wizard-of-Oz simulation was run in parallel with a functional system. Studies by Aust, Oerder, Seide and Steinbiss (1995) and Lamel, Bennacef, Gauvain, Dartigues and Temem (1998) have shown that putting spoken dialogue systems on trial by making them publically available may be useful. Large amounts of data from real users have already been collected in telephone-based systems such as Jupiter (Zue, Seneff, Glass, Hetherington, Hurley, Meng, Pao, Polifroni, Schloming abd Schmid 1997), CMU Communicator (Eskenazi 1999) and RailTel (Lamel, Bennacef, Rosset, Devillers, Foukia, Gangolf and Gauvain 1997). While these systems have successfully been used to collect dialogues in limited domains, they should be complemented by systems which gather data from a wider range of users. In a recent article, Chung, Seneff and Hetherington (1999) argue that future conversational systems should give users access to a wide range of domains within one and the same dialogue. Multi-domain systems in real applications are one of the challenges for next generation spoken dialogue systems.

During human-computer interaction in current spoken dialogue systems, it is often the case that user expectations fail to coincide with system capabilities. This discrepancy can partly be attributed to the fact that people's linguistic behavior in human-human as well as human-computer interaction has been insufficiently analyzed. As a consequence, many errors and misunderstandings appearing on

https://doi.org/10.1017/S1351324900002485 Published online by Cambridge University Press

different levels of spoken dialogue systems cannot yet be adequately handled. As reported by Kennedy, Wilkes, Elder and Murray (1988), Zoltan-Ford (1991), Brennan (1996) and Gustafson, Larsson, Carlson and Hellman (1997), people often adapt their language during man-machine interaction to meet the demands of the system. Oviatt, Cohen and Wang (1994) and Pargellis, Hong-Kwang and Lee (1999) have discussed the need to match users' behavior to current system capabilities. In an overview of different system prompting techniques for speech-only interfaces, Yankelovich (1996) has demonstrated that directing users only works to a certain extent. It is also important to allow users to become familiar with the technology, something which has been shown by Kamm, Litman and Walker (1998). For certain applications, typically one-domain information retrieval tasks, a combination of implicit and explicit prompts which make system limitations clear for the user are likely to be successful. However, from the user's point of view, a serious drawback of this approach is that the dialogues resulting from such interactions tend to feel rather unnatural. Speech interfaces should be as natural as possible for people to prefer to use them. Making the system interface human-like may be one way of making things easier for users, since we are all accustomed to conversing with other humans. According to Nass and Steuer (1993), conversational systems that exhibit certain social characteristics will also encourage users to respond socially. This should be taken into account as one attempts to anticipate user reactions to any system using spoken or written language as its output.

3 The August system

In the following section, the August system and its components are described. Some of the practical questions addressed in the development of this system are also discussed. August is a Swedish multi-modal spoken dialogue system featuring an animated agent (named after the author August Strindberg) with whom the user interacts (Gustafson, Lundeberg and Liljencrants 1999). The system was designed with a number of simple domains rather than a single complex one. August provided users with general information about the Royal Institute of Technology (KTH), speech technology and Stockholm. By extracting information from web-based yellow pages, the location of restaurants and other facilities in the city could be presented. In addition, some basic facts about the life and works of the agent's namesake Strindberg were available. Finally, because the users did not really know what to say to August and since his appearance was rather human-like, we expected that they would want to exchange greetings with him. The ability to handle and respond to some of these social utterances was also built into the system.

In the construction of the August system, the Wizard-of-Oz approach was not used. Instead, this experimental system was exposed to real users from the general public from the beginning. The users of the system were not given any explicit instructions. To suggest possible topics of conversation, facts related to the domains of the system were randomly selected and read out loud by the animated agent. Although the system's capacities were limited, this was not directly made known to the users. The system's responses differed both in length and complexity, from simple

https://doi.org/10.1017/S1351324900002485 Published online by Cambridge University Press



Fig. 1. The entire August system (left), closeups of the animated agent and the thought balloon (middle) and the street map (right).

single-word utterances to long phrases. As long as a user asked something within one of its domains, the system often gave a reasonable response. This resulted in a system that sometimes appeared to handle almost anything and which managed to generate quite human-like responses, while it sometimes did not seem to 'understand' much at all.

The system was put together between January and August of 1998. It was then available six days a week during six months to all visitors at the Cultural Center in downtown Stockholm. August was set up in an exhibition area with high levels of background noise from other equipment and visitors. It was therefore necessary to install a click-to-talk mechanism for recording the speech input. Since the system was unsupervised and the equipment had to be protected, a directional microphone secured in a metal grid box was installed. The metal box introduced some sound deterioration but experiments indicated that this would not effect the speech recognition significantly (Gustafson, Lundeberg and Liljencrants 1999). The system had two computer screens, as can be seen in Figure 1. The first screen featured the animated agent who communicated with synthetic speech and a combination of facial expressions and head movements. In addition, there was a 'thought balloon' in which textual information and suggestions on topics of conversation were displayed. The second screen was used for presenting tables and a street map, where locations of items that matched the users' requests were shown.

The August system was developed using existing KTH speech technology components (Bertenstam, Beskow, Blomberg *et al.* 1995). Figure 2 is an overview of these components. The HMM based continuous speech recognizer (Ström 1997) had an

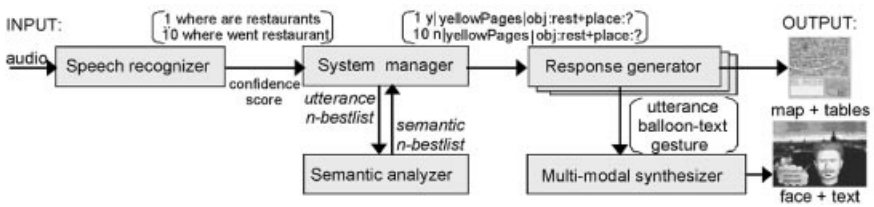


Fig. 2. The components of the August system.

https://doi.org/10.1017/S1351324900002485 Published online by Cambridge University Press

initial lexicon of about 500 words and idiomatic phrases. The recognizer generated an n-best list of utterance hypotheses, as well as a confidence score. The confidence score was computed by comparing the recognition scores from two engines run in parallel: one with a lexicon of the words used in the system, and one with all permitted syllables in Swedish. This score was used in conjunction with a semantic analysis to identify poorly recognized utterances.

To make it easier to update and extend the system, the semantic analysis was based on a data-driven method (Lindberg 2000). The analyzer server was built around Timble (Daelemans, Zavrel, van der Sloot and van den Bosch 1998), a freely available memory-based machine learning system. An utterance hypothesis produced by the speech recognizer was given a semantic analysis according to similar examples in an annotated example database. This analysis consisted of a simple feature-value structure. Three main fields made up the semantic analysis, each of which was filled out by an independent classifier. The first field stated whether an utterance was acceptable or not (*y* or *n*). The second field predicted the domain of the utterance (e.g. *strindberg*, *kth*) and the third field was instantiated with a flat feature-value representation of the utterance (e.g. {*object:restaurant, place:mariatorget*}). Since the complexity of the system was found in covering a number of simple domains instead of a single complex one, the dialogue handling was kept very simple. To make the system appear less deterministic, a number of possible answers was generated in response to each semantic analysis.

A new lip-synchronized 3D talking head was developed for the August project (Lundeberg and Beskow 1999). To give the agent a 'personality', he was made to resemble the author August Strindberg. Strindberg is a well-known character, and the choice of this persona also indicated that the system had some knowledge about Stockholm, history and literature. When designing the agent, it was important that he should not only be able to generate convincing lip-synchronized speech, but that he would also have access to a repertoire of non-verbal behavior. Therefore, facial movements which displayed emotions and directed the user's attention to physical objects in the environment were developed. Furthermore, the synthetic speech output was accentuated using gestures to stress certain focussed words and phrases. Speech synthesis parameter trajectories were generated by the KTH audio-visual text-to-speech system. The speech output was generated using an Mbrola synthesizer (Dutoit, Pagel, Pierret, Bataille and van den Vreken 1996), with a Swedish voice created by Gösta Bruce and Marcus Filipsson at Lund University.

One important issue in the August project was that of extending the coverage of user utterances, so that the system could be improved by adding a new domain, or altering an existing one. After the August system had been on display for three months, all user utterances recorded so far were transcribed and analyzed. The system was then updated in accordance with the most frequently observed user behavior. In this process, we modified the recognition lexicon, the semantic analyzer and the system output. Furthermore, our analyses implied that the users sometimes had difficulties understanding that the system was processing their input. To enhance the perceived reactivity of the animated agent, listening and thinking gestures were created.

https://doi.org/10.1017/S1351324900002485 Published online by Cambridge University Press

4 The August speech corpus

In this section, the August corpus and some of the linguistic analyses of this database are presented. The August corpus has been previously described in (Bell and Gustafson 1999a, 1999b). All 10,058 utterances of the database were transcribed orthographically, part-of-speech tagged, parsed and labeled with some basic speaker characteristics. The distribution of users, utterances and words in the corpus can be seen in Figure 3. The 39,594 words make up a lexicon of 2915 word forms, out of which half occurred only once. The 200 most frequently used word forms cover 81% of all words in the database. A majority of utterances only contained a single clause, while coordinated or subordinated clauses rarely occurred. The average number of words per utterance was about four, which partly explains the relatively small number of syntactically complex structures in the corpus. The syntactic parsing resulted in phrase level sentence structures, for example (NP) (VP) (NP) (PP). More than half of all the utterances in the database could be covered by ten such structures. The number of utterances originating from a single speaker in the corpus ranged from 1–49, but the average was 4.1 utterances for men, 3.3 for women and 3.5 for children.

All utterances in the database were also labeled according to the presumed intentions of the users. The purpose of this categorization was to get an overview of what the users wanted to convey when interacting with the system. Were they trying to retrieve information or were they merely interested in socializing with the animated agent? The utterances in the database were therefore categorized in accordance with a simplified pragmatic model containing six categories.

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 domains indicated by the system. 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**. These six categories were then brought together into two main groups. The first one, *socializing*, included the categories social, insults and test while the second one, *information-seeking*, included the categories domain, meta and facts.

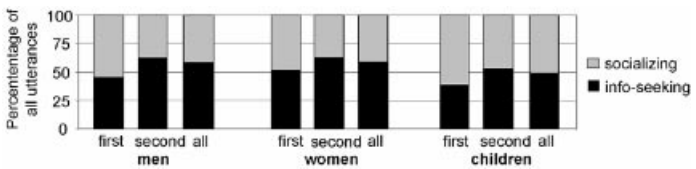


Fig. 3. The distribution of users, utterances and words in the August corpus.

https://doi.org/10.1017/S1351324900002485 Published online by Cambridge University Press

Table 1. *The utterance types in the August database*

Socializing	Translated examples	*Info-seeking*	Translated examples
Social	Hello August!	**Domain**	Where can I find restaurants?
Insult	You are stupid!	**Meta**	What can I ask you?
Test	What is my name?	**Facts**	What's the capital of Finland?

Some lexical and syntactical differences between the two main utterance categories were observed in the analyses. As many as 67% of the words only found in the socializing category occurred only once, compared to 20% in the information-seeking category. The lexical overlap between the two categories mostly consisted of common function words and main and auxiliary verbs. The syntactical patterns of the sentence structures of the utterances found in the socializing category appeared more uniform than the ones in the information-seeking category. The utterances categorized as meta contained the largest number of different sentence structures, probably because these utterances were used during error resolution. To cover 75% of the meta utterances, 185 structures were needed, compared to the social category, where 18 structures were sufficient. The analysis of the August corpus indicated that the addition of a small number of words and sentence structures would be enough to handle most of the users' socializing utterances.

5 User behavior as a result of system design

This section covers some aspects of user behavior in the August system as it is manifested in the speech data. We will discuss how certain features of the system design may have influenced the users' linguistic behavior. The distribution of utterance types in the corpus is also discussed, and the system's responses are briefly analyzed.

Several aspects of the design of the August system probably contributed to the outcome of the dialogues and the user behavior observed in the corpus. The animated agent, with its human-like appearance and 'personality', may have encouraged users to emphasize the social aspects of the interaction. The system was installed in a public place, and was intended to attract the attention of visitors and bypassers by means of its animated agent. The agent announced randomly selected bits of information linked to its domains, so as to give the users some idea of its capabilities. This explains that the number of domain-related utterances was rather large, given the fact that the users were not explicitly informed of what they could talk about. August was designed to be easily accessible in a public location, and the users of the system were therefore often surrounded by other people. This made the use of speech detection impossible. Because the users had to click to talk and since they could be overheard by others, their spoken input was probably less spontaneous than it would otherwise have been. Moreover, the system's limited lexicon covered certain domains quite well. Utterances that were out of domain, on the other hand,

https://doi.org/10.1017/S1351324900002485 Published online by Cambridge University Press

Table 2. *Nine turns of a dialogue taken from the August corpus*

	User (U1–U10) and System (S1–S10) Interaction
U1	*What is your name?* (Social)
S1	I call myself Strindberg, but I don't really have a surname
U2	*When were you born?* (Social)
S2	Strindberg was born in 1849
U3	*What do you do for a living?* (Social)
S3	I can answer questions about Strindberg, The Royal Institute of Technology and Stockholm
U4	*How many people live in Stockholm?* (Domain)
S4	Over a million people live in the Stockholm area
U5	*Do you like it here?* (Social)
S5	People who live in glass houses should not throw stones
U6	*Yes, that was a smart thing to say* (Meta)
S6	Hello, I come from the department of Speech, Music and Hearing
U7	*The Royal Institute of Technology* (Domain)
S7	The information is shown on the map
U8	*Thank you* (Social)
S8	You are welcome
U9	*Goodbye* (Social)
S9	Perhaps we will meet soon again
U10	*Yes, perhaps we will* (Meta)
S10	Strindberg was married three times

were often not adequately responded to. Some users may therefore have gotten the impression that the system's ability to interact was rather well-developed, while others thought the opposite. The example dialogue shown in Table 2 is an excerpt from a longer dialogue, marked with utterance types. Many of the words in this user's utterances were in the recognition lexicon.

The two main utterance categories, socializing and information-seeking, were examined so that the users' strategies could be studied more in detail. In the course of the dialogues, these strategies may have become influenced by the system's limited capabilities. However, the first two utterances from each user should indicate what the users expected the system to be able to handle. Figure 4 shows the distribution of the categories in all the users' first and second turns, as well as the distribution for the entire database. Some tendencies can be observed in the patterns of utterance types. It appears as if women were more inclined to begin their interaction with the system by seeking for information. Children frequently began by socializing with the system,

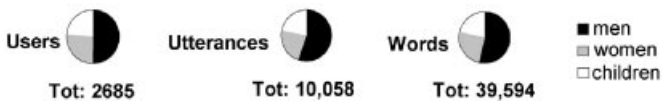


Fig. 4. Distribution of utterance categories in the users' first and second turns, compared with the distribution for all the utterances in the database.

https://doi.org/10.1017/S1351324900002485 Published online by Cambridge University Press

and often remained in that category, probably because they were not interested in the information offered by the system. A tendency in the longer dialogues was that adults often went from some initial socializing utterances to information-seeking. The statistical analyses of the corpus also showed that the utterances mostly retained the same number of words as the users' interaction with the system went on. Very few users turned to a more 'telegraphic' command language.

An overview of the utterance types and the system's ability to handle and respond to the spoken input can be seen in Figure 5. Figure 5(a) shows the distribution of the utterance types in the entire August database. All words in the input utterances were also compared to the words in the recognition lexicon. As shown in Figure 5(b), the coverage of the recognition lexicon sufficed to handle two thirds of the social utterances and about half of the domain utterances to the system. The utterances categorized as test and facts were most difficult to deal with, since they often contained word forms that occurred only once in the corpus. All system responses were also assessed, so that we could determine whether the user had received a reasonably correct answer or not. Figure 5(c) shows to what extent the input utterances generated an adequate system response. It is clear that the children more seldom received correct system responses than the adults, independent of utterance type. The speech recognizer used in the August system was trained almost exclusively on adult speech, which partly explains these results.

As the utterance categories were analyzed in the longer dialogues, we could see that those users that had received an accurate response to a socializing utterance tended to remain for a greater number of turns. On average, they remained for two extra turns when compared to users who were searching for information or those who did not receive an accurate response to a socializing utterance. An important question was whether it was possible to encourage the users to talk about the selected domains of the system instead of merely socializing. Accordingly, the 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. A test was carried out to see how these prompts influenced the users when they were mistakenly generated due to recognition failure. Results from this test showed that 63% of the users actually conformed to the system by immediately asking about one of the topics mentioned in the preceding prompt.

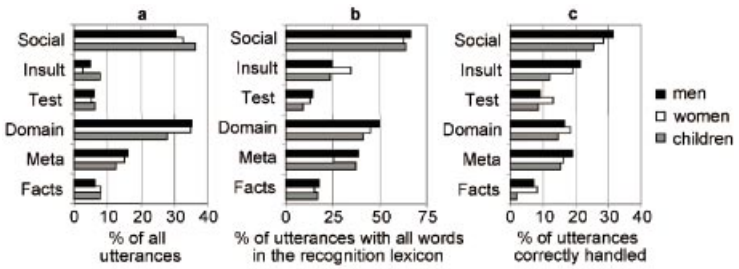


Fig. 5. Distribution of the utterance types, their coverage in the recognition lexicon and how they were handled by the system.

https://doi.org/10.1017/S1351324900002485 Published online by Cambridge University Press

The number of domain-related user utterances before these prompts appeared was 26%. The number of utterances in the socializing category decreased significantly after such a prompt.

6 User adaptation during error resolution

The question of how the users reacted when their interaction with the system became problematic will be addressed in the following section. More specifically, repetitive sequences in the database will be analyzed from the point of view of user adaptation during error resolution.

As described above, the users of the August system had not been explicitly informed about how to speak to the system or what kind of sentences it could handle. Since the system's recognition lexicon was quite limited, it often failed to handle the input. When the system failed, the users tried to resolve the errors that had occurred and sometimes adapted their speech to be better understood. For instance, users often repeated or rephrased what they had just said to the system. Repetitions and near-repetitions should therefore be interesting from the point of view of user adaptation during error recovery. These repetitive sequences constituted about 10% of the August corpus.

Sequences of near-repetitions (e.g. Hi, what's your name? → Hello, what's your name?) were analyzed to see how the users adapted their syntactical and lexical patterns during error resolution. This sub-group made up 4% of the utterances in the entire database. Table 3 is an overview of the changes in the near-repetitions when compared to the original input to the system. The following features were judged: the exchange of one lexical item for another, changed word order, insertion/deletion of a word or phrase and increased/decreased syntactic complexity. In addition, a small number of repetitions were categorized as 'other'. As can be seen in Table 3, the most commonly observed modification from original to repetition is the exchange of one lexical item for another. Generally, the users seem to be testing different strategies in their attempts to interact successfully. When the subjects repeated the same utterance more than once, they often alternated one feature such as increased/decreased complexity or insertion/deletion of a word or phrase (e.g. How are you? → How are you, August? → How are you?). This pattern of linguistically contrastive pairs occurred in 41% of these sequences.

To study how the users adapted their pronunciation during error resolution,

Table 3. *Changes in % of all lexically different repetitions*

	Changed lex. item	Changed word order	Inserted word/phrase	Deleted word/phrase	More complex	Less complex
Men	29	11	16	12	10	14
Women	41	13	10	16	4	9
Children	35	13	13	14	13	7

https://doi.org/10.1017/S1351324900002485 Published online by Cambridge University Press

lexically identical repetitions were subjected to a detailed acoustic and phonetic analysis. 1162 utterances (527 originals and 635 repetitions) were manually extracted from the database. Original utterances and repetitions were compared and the following features were assessed: duration, degree of articulation, loudness, focal shifting and inserted pauses. When these identical repetitions were compared to the original input to the system, a clear tendency was that both adults and children users moved towards clearer articulation, see Figure 6. This adaptation of the pronunciation also affected the speech rate. The duration was on average 26% longer for repetition, which is somewhat more than previously reported figures for repetitive sequences (Oviatt, Levow, MacEachern and Kuhn 1996). In more than half of all cases, the second repetitions were even longer than the first repetitions of the same utterance. The second repetitions were also distinguished by the fact that they more frequently contained inserted pauses between words. While an increased loudness rarely occurred in the adults' repetitions, this was fairly common in the children's repetitions. Focus shifting occurred more often in the adults' repetitions than in the children's, as can be seen in Figure 6. Hyperarticulated speech and speech with inserted pauses has been shown to be more difficult to handle for continuous speech recognizers than normal speech (Alleva, Huang, Hwang and Jiang 1997; Bell and Gustafson 1999b).

The utterances studied in this section were often taken from longer dialogues. While some of these exchanges worked quite well, others contained only failures. To see how this affected the pronunciation, both repetitions and original utterances were examined in a wider context. Hyperarticulation hardly occurred in utterances taken from dialogues where most of the system responses were accurate. Utterances that were taken from dialogues with no correct responses, on the other hand, were often clearly articulated. Hyperarticulation was especially frequent with children in unsuccessful dialogues. A decrease in speech rate was also quite common in the difficult dialogues. The largest decreases appeared after the system had been unable to record the speech input and had asked the user to speak into the microphone.

7 Discussion

The users of the August system were members of the general public who were not instructed on how to interact with a spoken dialogue system. We anticipated the animated agent's human-like appearance to raise user expectations of the system's capabilities. One of the most interesting overall findings in the corpus was therefore

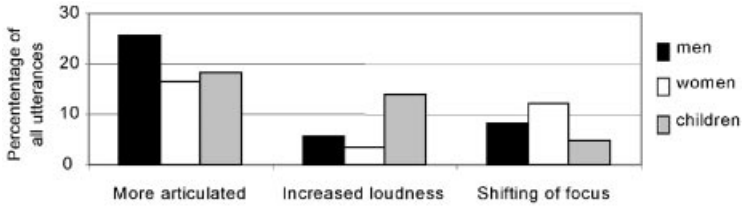


Fig. 6. Changes in lexically identical repetitions.

https://doi.org/10.1017/S1351324900002485 Published online by Cambridge University Press

the relative simplicity of the users' linguistic input. The utterances in the August database were generally quite short and simple, and a small number of phrasal combinations covered almost all of the syntactical structures in the database. Although a number of odd words and expressions appeared in the material, there were not that many unique words in the corpus. A clear tendency in the data analyzed was that when complex structures or unusual words occurred, it was when the users became frustrated with the system, attempted to resolve errors or were deliberately testing the limits of the system.

Given the choice of a topic of conversation in human-computer interaction, users of a spoken dialogue system could be expected to immediately want to access some sort of information. However, approaching a human-looking animated agent without exchanging some initial social greetings could for some users feel rude. In our analyses of the August corpus, we could observe that almost half of the utterances in the August corpus were categorized as socializing. This can partly be explained by the fact that the users were not given a specified task and did not really know what to say to the agent. Many users socialized with the system for a few turns before moving on to the information-seeking categories. Children tended to stay in the socializing phase, perhaps because they were not interested in the information offered by the system.

In future multi-domain systems with large vocabularies, being able to adapt one or several parts of the system to meet the demands of the user will be important. An initial socializing phase could then be used to adjust the speech recognizer and dialogue manager to achieve higher accuracy. Such a phase could perhaps also be used to determine what kind of user the system is interacting with: One that requires prompting and prefers system-initiative or one that is more independent. As our analysis indicated, a successful socializing interaction could make users' less prone to quit if (or when) errors occur later in the dialogue.

8 Conclusion

This paper has described the experimental August spoken dialogue system. The construction of the system was motivated by the need to collect data from real users in a public environment. The August database contains spontaneous computer-directed speech from more than 2500 members of the general public, out of which one fourth were children. This database could serve as a complement to speech corpora containing more task-oriented dialogues in restricted domains. Specific aspects of the system design that contributed to the outcome of the dialogues were considered, so that the users' behavior could be adequately analyzed. These aspects included the absence of a single domain and explicit task as well as the divergence between the animated agent's human-like appearance and his relatively small vocabulary. The speech input was categorized into utterance types and it was observed that many utterances were used for socializing with the system. These socializing utterances were quite easy to handle, and could be useful to add to a future multidomain system. Moreover, user adaptation during system failure was examined in conjunction with repetitive sequences in the database. Adaptive strategies included

https://doi.org/10.1017/S1351324900002485 Published online by Cambridge University Press

lexical and syntactical modifications as well as changes in the pronunciation in the lexically identical repetitions. Our analysis indicated that both adults and children moved from conversational to clear speech in the lexically identical repetitions. We could observe that children tended to speak louder during repetition, and that they frequently hyperarticulated in dialogues where most of the turns were unsuccessful.

This research has implications for the development of future systems intended for public use, since the August database could be used to predict the behavior of uninformed users interacting with an animated agent. Especially, we believe that the experiences from the August system could be put to use in future systems that enable users to browse a number of domains in a publically available information kiosk or interact with an animated agent in an educational setting.

Acknowledgements

The authors would like to thank the people who contributed to the development of the August system. We are especially grateful to Nikolaj Lindberg, who generously helped with some technical details in the process of writing this paper, and whose comments and suggestions have been valuable.

References

Allen J. F., Miller, B. W., Ringger, E. K. and Sikorski, T. (1996) Robust understanding in a dialogue system. *Proc. of 34thmeeting of the Association for Computational Linguistics.*

Alleva, F. Huang, X., Hwang, M.-Y. and Jiang, L. (1997) Can continuous speech recognizers handle isolated speech? *Proc. of Eurospeech '97*, pp. 911–914.

Aust, H., Oerder, M., Seide, F. and Steinbiss, V. (1995) The Philips automatic train timetable information system. *Speech Communication* **17**(3–4): 249–262.

Bell, L. and Gustafson, J. (1999a) Interaction with an animated agent in a spoken dialogue system. *Proc. of Eurospeech '99*, pp. 1143–1146.

Bell, L. and Gustafson, J. (1999b) Repetition and its phonetic realizations: Investigating a Swedish database of spontaneous computer-directed speech. *Proceedings of ICPhS '99*, pp. 1221–1224.

Bertenstam, J., Beskow, J., Blomberg, M., Carlson, R., Elenius, K., Granström, B., Gustafson, J., Hunnicutt, S., Högberg, J., Lindell, R., Neovius, L., Nord, L., Serpa-Leitao, A. and Ström, N. (1995) The Waxholm System – A Progress Report. *Proc. of ESCA Tutorial and Workshop on Spoken Dialogue Systems*, pp. 81–84.

Brennan, S. (1996) Lexical entrainment in spontaneous dialog. *Proc. of ISSD*, pp. 41–44.

Cheyer, A., Julia, L. and Martin, J. C. (1998) A unified framework for constructing multimodal experiments and applications. *Proc. of CMC '98.*

Chung, G., Seneff, S. and Hetherington, L. (1999) Towards multi-domain speech understanding using a two-stage recognizer. *Proc. of Eurospeech '99*, pp. 2655–2658.

Daelemans, W., Zavrel, J., van der Sloot, K. and van den Bosch, A. (1998) TiMBL: Tilburg Memory Based Learner, version 1.0, Reference Guide, *LK Technical Report 98–03.*

Dutoit, T., Pagel, V., Pierret, N., Bataille, F. and van der Vreken, O. (1996) The MBROLA Project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. *Proc. of ICSLP '96, Philadelphia*, pp. 1393–1396.

Eskenazi, M., Rudnicky, A., Gregory, K., Constantinides, P., Brennan, R., Bennett, C. and Allen, J. (1999) Data collection and processing in the Carnegie Mellon communicator. *Proceedings of Eurospeech '99*, pp. 2695–2698.

https://doi.org/10.1017/S1351324900002485 Published online by Cambridge University Press

Fraser, N. (1997) Assessment of Interactive Systems. In: Gibbon, D., Moore, R. and Winski, R. (eds.), *EAGLES Handbook of Standards and Resources for Spoken Language Systems.*

Gauvain, J. L., Bennacef, S., Devillers, L., Lamel, L. F. and Rosset, S. (1995) The spoken language component of the mask kiosk. *Proc. of Human Comfort & Security Workshop.*

Gustafson, J., Lundeberg, M. and Liljencrants, J. (1999) Experiences from the development of August – a multimodal spoken dialogue system. *Proc. of IDS '99*, pp. 81–85.

Gustafson, J., Larsson, A., Carlson, R. and Hellman, K. (1997) How do system questions influence lexical choices in user answers? *Proc. of Eurospeech '97,* pp. 2275–2278.

Heeman, P. A., Johnston, M., Denney, J. and Kaiser, E. (1998) Beyond structured dialogues: Factoring out grounding. *Proc. of ICSLP '98,* pp. 863–866.

Kamm, C. A., Litman, D. J. and Walker, M. A. (1998) From novice to expert: The effect of tutorials on user expertise with spoken dialogue systems. *Proc. ICSLP '98,* pp. 1211–1214.

Kennedy, A., Wilkes, A., Elder, L. and Murray, W. (1988) Dialogue with machines. *Cognition* **30**: 73–105.

Lamel, L., Bennacef, S., Gauvain, J. L., Dartigues, H. and Temem, J. N. (1998) User evaluation of the Mask Kiosk. *Proc. of ICSLP '98*, pp. 2875–2878.

Lamel, L. F., Bennacef, S. K., Rosset, S., Devillers, L., Foukia, S., Gangolf, J. J. and Gauvain, J. L. (1997) The LIMSI RailTel System: Field trial of a telephone service for rail travel information. *Speech Communication* **23**(1–2): 67–82.

Lindberg, N. (2000) Data driven methods in natural language processing – Two applications. *Licentiate Thesis*, Royal Institute of Technology, Stockholm.

Lundeberg, M. and Beskow, J. (1999) Developing a 3D-agent for the August dialogue system. *Proc. of AVSP Workshop*, pp. 151–154.

Nass, C. and Steuer, S. (1993) Voices, boxes, and sources of messages: Computers and social fctors. *Human Communication Research* **19**(4): 504–527.

Oviatt, S. L., Cohen, P. R. and Wang, M. Q. (1994) Toward interface design for human language technology: Modality and structure as determinants of linguistic complexity. *Speech Communication* **15**(3–4): 283–300.

Oviatt, S. L., Levow, G. A., MacEachern, M. and Kuhn, K. (1996) Modeling hyperarticulate speech during human-computer error resolution. *Proc. of ICSLP '96*, pp. 801–804.

Pargellis, A., Hong-Kwang, J. K. and Lee, C-H. (1999) Automatic application generator matches user expectations to system capabilities. *Proc. of IDS '99*, pp. 37–40.

Ström, N. (1997) Automatic Continuous Speech Recognition with Rapid Speaker Adaptation for Human/Machine Interaction. *PhD Thesis*, Royal Institute of Technology, Stockholm.

Thomas, J. C. (1995) Human factors in lifecycle development. In: Syrdal, A., Bennett, R. and Greenspan, S. (eds.), *Applied Speech Technology*. Boca Raton: CRC Press.

Yankelovich, N. (1996) How do users know what to say? *ACM Interactions* **3**(6).

Zoltan-Ford, E. (1991) How to get people to say and type what computers can understand. *Int. J. Man-Machine Studies* **34**: 527–547.

Zue, V., Seneff, S., Glass, J., Hetherington, L., Hurley, E., Meng, H., Pao, C., Polifroni, J., Schloming, R. and Schmid, P. (1997) From interface to content: Translingual access and delivery of on-line information. *Proc. of Eurospeech '97*, pp. 2227–2230.

https://doi.org/10.1017/S1351324900002485 Published online by Cambridge University Press