

HOW DO SYSTEM QUESTIONS INFLUENCE LEXICAL CHOICES IN USER ANSWERS?

J. Gustafson¹, A. Larsson¹, R. Carlson¹ and K. Hellman²

¹ Department of Speech, Music and Hearing, KTH
Box 70014, S-10044 Stockholm, Sweden
Tel.:+46 8 790 7879 Fax: +46 8 790 7854 E-mail: {joakim_g | anette | rolf}@speech.kth.se

² Department of Linguistics, Stockholm University
S-106 91 Stockholm, Sweden
Tel.:+46 8 16 2335 Fax: +46 8 15 53 89 E-mail: kicki@ling.su.se

ABSTRACT

This paper describes some studies on the effect of the system vocabulary on the lexical choices of the users. There are many theories about human-human dialogues that could be useful in the design of spoken dialogue systems. This paper will give an overview of some of these theories and report the results from two experiments that examines one of these theories, namely lexical entrainment. The first experiment was a small Wizard of Oz-test that simulated a tourist information system with a speech interface, and the second experiment simulated a system with speech recognition that controlled a questionnaire about peoples plans for their vacation. Both experiments show that the subjects mostly adapt their lexical choices to the system questions. Only in less than 5% of the cases did they use an alternative main verb in the answer. These results encourage us to investigate the possibility to add an adaptive language model in the speech recognizer in our dialogue system, where the probabilities for the words used in the system questions are increased.

1. BACKGROUND

If natural language is to be used in spoken dialogue systems, problems can arise from the fact that there are numerous ways of expressing the same thing, by using different lexical items and/or different word order. This can lead to problems in command language systems, where the users are not forced to use only one term per command. Studies show that there is a low probability that two users of a command language system produces the same term for the same command[1].

The preceding argument builds on the assumption that people that engage in conversation indeed utilize their lexical and syntactic repertoire in a varied and perhaps even unpredictable way. Whether this assumption is valid is of course an empirical question, the answer of which will be important in the design of human-computer dialogue systems. There are many theories about human-human dialogues that could be useful for the design of spoken dialogue systems.

Humans use language to perform many communicative functions. In daily life spoken language mostly has an interactional function - to establish and maintain personal relationships, while written language mostly has a transactional function - to transfer information[2]. There are commonly agreeable facts about human-human dialogues, for example: mostly one speaker talks at the time; the order of speakers is not pre-determined; topic shifts; conversations have variable length; there are pauses and gaps in talks[3]. The main feature of a dialogue that distinguishes it from a monologue is that there are at least two partners who are contributing to the discourse. The dialogue consists of *turns*, where turns can be seen as spaces where speakers are allowed to speak, and that are marked off by a speaker-shift. Turns can have various components, from a single phone to several utterances[3,4]. In dialogues there are regularities in the ordering described as *adjacency pairs*, for example Question-Answer. This simple structure is not always applicable, there is often an insertion-sequence that delays the Answer-part to a Question-part, until some other question has been answered. There are two simultaneous information channels in a dialogue: the information channel from the speaker, and the back-channel feedback from the listener. The back-channel feedback indicates attention, attitudes and understanding, and its purpose is to support the interaction.

The Speech Act theory was based on Austin's studies on performatives, utterances that can be used to perform acts. The theory was further developed by Searle, who defined *illocutionary acts*, like requesting, informing and promise[5, 6]. Sinclair & Coulthard defined a discourse grammar that used *exchanges* instead of turns as basic unit of discourse, which have *acts* and *moves* as their single-speaker units. Moves are interactive units that indicate what an utterance does in the discourse; and they consist of one or more acts that indicate what the speaker means at a specific point in discourse. Their acts are different from the speech acts used by Searle and Austin in that they emphasize the role of the situation[7]. A turn is what the speaker says as long as he holds the floor, while a move is what the speaker does in a turn[8].

Another fairly well agreed upon finding is that most human dialogues are characterized by co-operation. Grice defined the Co-operative Principle: "*Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged*", which is manifested in the maxims of Quantity, Quality, Relation and Manner[9]. To this end, the participants establish a *common ground* from their past conversations, their immediate surroundings and the current dialogue. Speakers co-ordinate their use of language with other participants in a language arena in two phases: first an utterance is presented, which is then accepted when the receiver signals that he has got the information. The acceptance is acknowledged by feedback words like "OK", paraphrases of the presented utterance, or by implicit acknowledgement, which could be made by reusing the terms the participant used. The participants in a dialogue try to minimize their collaborative effort, the work both do from the initialisation of a contribution to its mutual acceptance[10, 11, 12, 13].

Furthermore, studies of human-computer dialogue show that variability between conversations is high, but it is relatively low within conversations. Brennan suggests that this is because people mark their shared conceptualizations by using the same terms, *lexical entrainment*. To investigate if these phenomena from human-human conversations also could be found in human-computer dialogues, Brennan performed some Wizard of Oz-tests where people queried a database with written or spoken natural language. If the user referred to an object with a different term than the system, it corrected the user, either by embedding its own term in the answer, or by explicitly asking the user if the different terms were the same. The test showed that people almost always adapted their terms to the systems term if the correction was exposed in an extra dialogue turn. In the case of embedded corrections the adaptation to the system terms was smaller, and lasted shorter. An interesting result was that the adaptation according to the embedded corrections was greater for spoken than for written input[14].

2. ADAPTATION IN WAXHOLM

The inclination for human adaptation, in human-human as well as human-computer dialogue noted by Brennan and discussed above is also found in the Waxholm system, the spoken dialogue system developed at the Department of Speech, Music and Hearing at KTH. In this spoken dialogue system users can ask about the boat traffic and other touristic information in the Stockholm archipelago[15]. Initially in the Waxholm project speech and text data was collected with a Wizard of Oz system where only the speech recognition was simulated. A total of 198 dialogues from 68 subjects were recorded. Studies of these dialogues show that the users were very co-operative when the system asked for information. We

studied the user answers to system questions including the word "åka (go, travel)", for example "Var ifrån vill du åka? (Where do you want to go from?)". There were 503 such dialogue turns, and in 60% of the cases the user answered with an ellipse, such as "från Stockholm (from Stockholm)". In 37% of the cases the subjects answered with complete sentences including a reuse of lexical items from the question, for example "jag vill åka på fredag (I want to go on friday)". Only in 3% of the cases the subjects responded with utterances that did not include reuses or ellipses. In most cases these were answers like "jag vet inte (I don't know)". Only once did the subject answer with an utterance that included the synonymical word "resa" instead of "åka", "jag ville resa på fredag (I wanted to go on Friday)" and only in one case did the subject seem to change the subject by saying "jag vill bo på hotell (I want to stay in a hotel)". This actually was no change of subject, but due to a recognition error in a earlier dialogue turn.

Our results are in accordance with other studies that show that subjects who interact with computers only supply the information that was asked for, using a simple language without politeness items, indirect speech acts, and only use few anaphora or pronouns. This could be because people expect that computer systems only can cope with simple dialogue structures, and that according to their mental model, dialogue systems are only simple retrieval systems where an input by the user retrieves an output from the system. Users will most probably adopt their vocabulary if they can detect the system's vocabulary[16,17,18,19,20,21]. The current trend in speech recognition system is to have very large vocabularies, which might decrease the performance. There have been approaches in recognition to capture the bursty nature of language by letting the recognizer cache words that occurred in the subject's past utterances and increase their probabilities in the system's language model[22,23]. We would like to extend this by also making use of the utterances produced by the system.

3. EXPERIMENT 1

To test if people adopt their lexical choice we designed a Wizard of Oz-test that simulated a dialogue system similar to Waxholm, where the system asked the user about destination, departure place, departure time, number of travellers, way of payment and planned leisure activities. The subjects were given the impression that they used a telephone version of a Waxholm-like system, where all system responses were synthesized speech. The subjects were not aware that the responses were generated by a human researcher who used an interface where all system responses were represented on buttons on the screen. This approach was chosen to make the dialogue faster and to make the systems utterances more consistent. The 35 questions used in the test were designed to be general, varied in lexical choices, varied in syntactic complexity and clear enough to avoid meta-communication.

The most frequently used term for go is "åka", which was noticed in the Waxholm system where 64 of the 198 dialogues begun with a user question that included the word "åka", while only one question contained the word "resa". To check if people would change their preferred term, 12 of the questions in the test contained either the word "resa" or the word "åka". Since we had found that people using the Waxholm system often replied with ellipses without the verb, we included some utterances like "Jag klarar bara av hela meningar (I can only cope with complete sentences)", that would generate responses with a verb. To be able to respond to simple questions from the user some system answers was also included, for example "Båtarna går från Strömkajen (The boats depart from Strömkajen)".

3.1. Results of experiment 1

The results of analysis of the 128 answers from 9 subjects show that people reuse terms from the system questions. An analysis of the answers to all types of questions shows that 36% contained reuse of words from the question, about 38% of the answers contained ellipses or were simple answers to yes/no questions. Only in 17% of the cases the users' answers did not include any word from the question, and in 9% the subject did not answer the question. The responses to the questions including the word "åka" had the same number of ellipses and reuses as in the analysis of the Waxholm database. The most promising result of the experiment was the responses to questions including the word "resa". The subject reused the word "resa" in 35% of their responses, while they only used their preferred word "åka" in 19% of the responses. In 8% of the cases the subjects did not answer the question, but said something like "jag hörde inte frågan (I didn't hear the question)". Encouraged by our preliminary results we designed a second experiment to further investigate these matters.

4. EXPERIMENT 2

The first Wizard of Oz experiment that simulated an information retrieval system, led to problems because subjects tended to take the initiative and question the system themselves instead of answering the system's questions. In a second experiment we decided to make the task more guided towards answering questions by letting 26 subjects use a system that simulated a questionnaire about how they would like to spend their vacation. The system was said to be fully automatic using speech recognition and synthesis. In this experiment we wanted to investigate the reuse of the main verb in the question. In order to do this 39 questions were designed that included a verb that could be varied, for example "Hur ofta brukar du *vandra/ströva* i skogen? (How often do you *hike/stroll* in the woods?) ". Some of the questions included unusual choices of the desired verb like "Skulle du vilja *luncha* på en skärgårdsbåt?". The system prompted the questions with either pre-recorded synthesized speech or human speech. In the experiment

the subjects first got an oral introduction by the chosen voice, where they were asked to use complete sentences in their answers since the recognizer preferred these. In order to get used to the system voice the subjects were then asked 5 general questions about age, sex, origin, occupation and place of residence. The subjects used a graphical interface with a push-to-talk button. If the recording was too loud or soft, a third system voice told them to repeat. In the actual experiment they were asked a selection of 30 of the 39 questions with a delay of about 1 second between each question.

4.1. Results of experiment 2

Analysing the 771 answers from the 26 subjects revealed an even stronger trend than the one found in the first experiment. This was probably due to the more restricted task of only answering prompted questions. In this way most of the utterances produced by the subjects were actually answers to the system questions. Another difference was that the subjects in the second experiment got an oral introduction where they were asked to use complete sentences. The general result was that people often were very co-operative in their answers, and they actually answered the question 98% of the cases. The most co-operative subjects simply remodelled the question into an answer by changing the word order of the question, only adding a few words or phrases, for example:

[yes / no] [I would [not] like to] [REORDERED QUESTION]

Three subjects mostly answered with ellipses instead of the complete sentences, that they were instructed to use in order to make the task easier for the computer. This is not surprising since it is the simplest way to answer and the most commonly used by the subjects in the first experiment. Some subjects used other types of answers with a more varied language. Most of these, 60% of cases, where a simple yes or no answer, where the subject added some phrases for example: "nej, *det tror jag inte* (no, *I don't think so*)". They said that the reason for this was that they had been instructed to answer with complete sentences, and this was the only way they could construct a complete sentence as an answer to these questions. The total distribution of answer types is shown in Table 1.

Table 1. The total number of answers of different types.

| Type of answer | Percentage of all answers |
|----------------|---------------------------|
| Reuse | 51 |
| Ellipse | 18 |
| Other | 24 |
| No reuse | 4 |
| No answer | 2 |

There were 36 different main verbs in the system's questions that was supposed to be adopted by the subjects. The subjects only used another main verb as reply to questions with 11 of these. In about one third of these cases the subject used the phrase "tycker om" instead of the word "gillar".

The variability on the answers from the different subjects was large, but the variability on the answers from each subject was relatively low, which corresponds to the findings of Brennan. Figure 1 shows the distribution of the different types of answers from the subjects. As can be seen in this figure most subjects tried to adjust their answers by reusing large parts of the question, while only a few use their preferred way of answering using ellipses and sentences including yes or no.

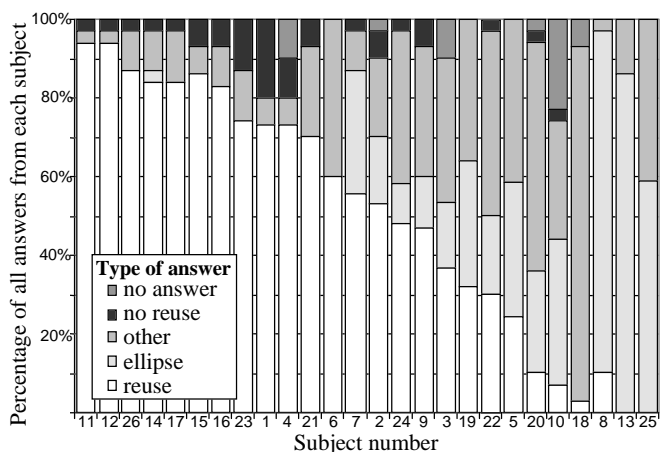


Figure 1. Distribution of answer types from each subject.

Interviews with the subjects revealed that the subjects found it hard to answer the questions with complete sentences. Only half of the subjects noticed that they reused the vocabulary of the system. Most of the subject did this because they said it was the easiest way to produce an answer that the system understood.

5. CONCLUSIONS

Our experiments show that people adapt their answers to the system questions, using both the vocabulary and the structure. In the second experiment the task was to answer questions using complete sentences, if possible. This made the subject think that the linguistic capacity of the system was low and they had to adapt their language accordingly. This computer adjusted language could be seen as a special case of receiver-adjusted talk like *motherese*, the language use by mother to their babies. This computer-adjusted talk could consequently be called *computerese*. This way of talking is highly adapted to the capacity of the used system: if the system seems to prefer complete sentences the subject reuses large parts of the questions to construct the answer; and if the system uses simple sentence structures the answer is constructed mainly by ellipses. It would be interesting in a future study to compare this Computerese with other receiver-adjusted speech like *Motherese* and *Elderspeech*.

6. REFERENCES

[1] Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987) "The vocabulary problem in human-system communications", In ACM, 30, 964-971

[2] Brown, G & Yule, G. (1983) "Discourse analysis", Cambridge University Press

[3] Sacks, H., Schegloff, E. A. & Jefferson G. (1974) "A simplest systematics for the organisation of turn-taking for conversation", Language 50: 696-735.

[4] Schenkein, J. (1978) "Studies in the organisation of conversational interaction", New York: Academic Press.

[5] Austin, J. L. (1962) "How to do things with words", Oxford: Oxford University Press.

[6] Searle, J. R. (1969) "Speech Acts", Cambridge University Press.

[7] Sinclair, J. M. & Coulthard, R. M. (1975) "Towards an analysis of Discourse: The English used by teachers and pupils", Oxford University press.

[8] Stenström, A-B (1984) "Questions and responses in English conversation", Lund studies in English, Liber Förlag Malmö

[9] Grice, H. P. (1975) "Logic and conversation." In P. Cole & J. L. Morgan (Eds.), Syntax and semantics, volume 9: Pragmatics (pp. 113-128). New York: Academic Press.

[10] Clark, H. H., and Wilkes-Gibbs, D. (1986) "Referring as a collaborative process", Cognition, 22, 1-39

[11] Clark, H. H. & Schaefer, E. F. (1989) "Contributing to discourse", Cognitive Science, 13, 259-294.

[12] Clark, H. H. & Brennan, S. E. (1991) "Grounding in communication", In Resnick, Levine & Teasley (Eds.) Perspectives on socially shared cognition (pp 127-149). Washington, DC: APA Books.

[13] Traum, D. R. & Allen, J. F. (1992) "A speech acts approach to grounding in conversation", In Proceedings of ICSLP-92, pp 137-40.

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

[15] Carlson, R., Hunnicutt, S. and Gustafson, J. (1995) "Dialogue management in the Waxholm system" Proc. Spoken Dialogue Systems, Vigsø

[16] Dahlbäck, N. (1991) "Representation of Discourse - Cognitive and Computational Aspects", Doctoral Thesis, Linköping University

[17] Guindon, R. (1988) "A multidisciplinary perspective on dialogue structure in user-advisory dialogues", in Guindon "Cognitive Science and its Application for Human-Computer Interaction", Lawrence Erlbaum Publ.

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

[19] Zoltan-Ford, E. (1991) "How to get people to say and type what computers can understand", Int J. Man-Machine Studies, 34, pp 527-47

[20] Karlgren, J. (1992) "The Interaction of Discourse Modality and User Expectations in Human-Computer Dialog", Licentiate Thesis, Stockholm University.

[21] Cohen, P. R. & Oviatt, S. L. (1994) "The Role of Voice Input for Human-Machine Communication", National Academy Press

[22] Kuhn, R. & de Mori, R. (1990) "A Cache-Based Natural Language Model for Speech Recognition", IEEE Pattern Analysis and Machine Intelligence v12, n 6, pp.570-583

[23] Lau, R., Rosenfeld, R. & Roukos, S. (1993) "Trigger-based language models: A maximum entropy approach", In ICASSP 93, pages 45--48.