Lund University, Dept. of Linguistics
Working Papers 49 (2001), 2–5

Designing with speech acts to elude disfluency in human–computer dialogue systems

Annika Asp and Anna Decker
Telia Research AB, Farsta, Sweden
Department of Linguistics, Stockholm University

Abstract

We have conducted a pilot-study of one particular aspect of computer system design, namely how the speech acts in a human–computer dialogue system influence the disfluency production of the user. We have analysed two travel-booking corpora with the aim to find out how many disfluencies the various speech acts in the dialogues of the corpora produce. We found that the speech acts do influence the disfluency rates, and that this trend was especially discernible when addressing the 'openness' of the speech acts. Our findings indicate that one way of reducing disfluencies in spoken human–computer systems would be to reduce the number of speech acts which present difficulty in processing for the user, and we propose that this be done in the design of the dialogue manager of the system.[1]

1 Introduction

Human–computer dialogue systems are becoming more prevalent in our everyday life, and we can already use fully automatic dialogue systems in various services. One of the aspects of human speech, which presents an obstacle to current state-of-the-art systems' processing of human spontaneous speech, is disfluency. Disfluencies are phenomena such as repetitions, corrections, and hesitations—the parts of our speech production that are "non-fluent" (cf. Shriberg, 1994). The ideal way of dealing with this problem would be to build a system which would manage all aspects of human speech in general, and disfluencies in particular. This is, unfortunately, not a feasible solution considering the knowledge we possess today. Research has, thus far, primarily focused on how to detect disfluencies, through statistic means or by prosodic and acoustic detection, aiming at "teaching" the machine how to first detect and then process the disfluencies. This field of research is slowly, but steadily, discovering more about the workings of disfluencies in language. Awaiting the advent of the "perfect" system, we need to find a way to make the current dialogue systems more stable. One solution is to try to reduce the number of disfluencies in the user's input to the system, something, which should be attempted already when designing the system.

When designing a dialogue system, one has to employ a variety of devices. One particular device for the system to manage the actual dialogue is the speech act. As the speech acts form the utterances in the discourse, we imagine them to be essential in evoking the response of the user, and are thus integral to the design of the system. Many have investigated the role of the speech act in this context (e.g. Carletta et al., 1997; Stolcke et al., 2000), but few have looked upon the speech acts built into the design of the system as a conceivable source of disfluency.

[1] This article presents a condensed version of our Bachelor's degree (C-level) thesis (Asp & Decker, 2001), which can be found at http://www.ling.su.se/DaLi/uppsats/ .

2 Background

In modern speech act theory, researchers seem to agree that communication is interactive and social at its nature. The cooperative principle, formulated by Grice (1975), is a reoccurring theme in the modern theories, and seems to be a fundamental prerequisite for any human communicative behaviour. Many have argued that communication is an action, which demands coordination and a will to cooperate of its participants.

It seems quite natural to claim that any endeavour towards a system that can effortlessly manage disfluencies must be made in the light of the underlying cognitive factors that cause them, and one important factor in understanding what causes disfluency, is understanding something about the speech production process. According to Levelt (1989), the step in the speech production process, which demands the most cognitive load, is macroplanning, i.e. when the structure of the message is built. Doing this the speaker need to first select which information he wants to go into his utterance, and, second, he needs to plan in which order he wants to present it. When planning the order, he needs to keep in mind what he has already said, i.e. to do "book-keeping". The book-keeping also manifests itself as the need to keep in mind what has happened earlier in the discourse. Disfluencies are believed to occur because of the sometimes too heavy cognitive load allocated to the tasks mentioned. Another step, further down the line in the process of speech production, is the formulating step, where the message is put into words. This is also a spot where disfluency can originate.

Disfluencies have been shown to be triggered in a somewhat predictable pattern and they have a tendency to react to the design of the system. Shriberg (1994) discovered that disfluency rates increase with sentence length, and this finding has been verified by e.g. Oviatt (1995), who reported disfluency rates increasing roughly linearly with sentence length. Bell, Eklund, & Gustafson (2000) compared two corpora regarding various properties of disfluencies and found that the corpus, which had the longer sentence lengths also, had the higher disfluency rates. Both Oviatt (1995) and Bell, Eklund, & Gustafson (2000) show that the structure of the dialogue heavily influences the disfluency production. In Oviatt's study, 60 to 70% of all disfluencies could be eliminated by using a more structured format. It was also discovered that the utterance *type* seems to influence disfluency rates (Bell, Eklund, & Gustafson, 2000; Oviatt, 1995), and it also seems that different types of actions produce different kinds of disfluencies (Shriberg, 1994).

3 Method

The two corpora, 'bionic' and 'woz-2', we used were collected within the frame of the Spoken Language Translator-project at Telia Research (Rayner et al., 2000). The corpora consist of travel booking dialogues between a human user and a "machine" agent. The users all believed that they were talking to a functional automated system, but in bionic, recognition was simulated by a wizard, and in woz-2, an actor who played the wizard was used, i.e. a Wizard-of-Oz set-up was used. A w-o-z set-up typically has the user in one room talking to what he thinks is a computer agent, and the wizard sitting in another room acting as the alleged agent. We counted the number of words and the number of disfluencies in each of the user's utterances. The disfluencies were annotated according to Eklund (1999). The agent's utterances in the two corpora were analyzed into a set of speech acts. The speech acts were first divided into two superordinate groups, *discourse related acts* and *task related acts*. The group of task related speech acts were further divided into three larger groups—questions, confirmations, and information—and these were, in turn, further differentiated to the point where each speech act communicated *one* intension only. These speech acts were analysed with the goal to discover which of the speech acts produced unnecessary high rates of disfluencies.

4 Results

Our results show that the design of the speech acts does influence the disfluency rates. The task related speech acts—questions, confirmations, and information—evoked higher disfluency rates than did the discourse related speech acts, something which was expected since the task related speech acts are the ones directly addressing the task at hand.

Speech acts with a more open and unconstrained nature seem to evoke higher disfluency rates, which can be due to the cognitive load that responding to open speech acts demands of the user. The more constrained speech acts of the system direct the user more precisely what to think of, which decreases cognitive load and thereby evokes fewer disfluencies. Examples from the corpora are the question *time?*, as in 'vilken tid vill du resa?' (What time do you want to leave?), which is more constrained and less open at its nature than the question *when?*, as in 'när vill du resa?' (When do you want to leave?). The question *time?* is to a higher degree directing the user what to think of, i.e. *time* of day and not date or even month, than the question *when?*, a relation which can be found in the differences in disfluency rates for the bionic corpus—41% disfluent words in response to a question of *when?* compared to 14.7% disfluent words succeeding a question of *time?*. Another speech act with disfluency evoking openness was the discourse related speech act "is there anything else you want to book?" which occurred several times in the course of one dialogue, often when closing a subdialogue. It presented high disfluency rates, probably because here the user has the opportunity to freely ponder what more to book and, additionally, retrace what has already been booked—the book-keeping process. These results speak in favour of a more constrained format where the speech acts of the system demand as little consideration as possible.

It was apparent in the analysis of the speech acts that they had a tendency to cluster; i.e. they did not always occur in isolation. Instead they formed clusters of typically two or three speech acts acting together. It was shown that the number of disfluencies evoked by a cluster might be boosted above the sum of the disfluencies evoked by the participating speech acts, when these occur in isolation. This was particularly evident when the speech acts in a cluster were open speech acts.

As for the speech acts labelled information, it became clear that there seems to exist a difference in disfluency evocation between two major groups; positive/neutral information versus negative information not in line with user's expectations. The speech acts displaying negative information reached a figure of 27% succeeding disfluent words compared to 13.6% disfluent words for the group of positive/neutral information. In both corpora, there existed what we call 'dead-ends', where the system uttered negative information, such as 'I cannot book a taxi', without any further attempt to satisfy the request of the user. In human–human interaction this kind of negative information is often, according to the cooperation principle, accompanied by helpful information which is aimed at fulfilling the user's request to some extent, e.g. by adding a phrase like 'but there are buses leaving every hour'.

Occasionally, the wizards in bionic and woz-2 corpora made inferences, which often is made in human–human communication and makes the dialogue run smoothly. Nevertheless, there was some tendency for such speech acts to cause disfluency. In examining the data more closely, we found a scenario where the user seemed positively surprised by the system's capacity and started to take the initiative. The system did not allow the user to take the initiative, and this made the user confused and disfluent.

5 Conclusion

Our results show the importance of carefully choosing a computer dialogue system's set of speech acts. Speech acts with a nature of openness within provoke high disfluency rates, due to the cognitive load they demand of the user. Additionally, one needs to take into account the boosting-effect on disfluency of combining two open speech acts. Another point is to aim at providing helpful and cooperative information when a request of the user cannot be fully satisfied, i.e. avoiding 'dead-ends'.

It goes without saying that in a computer based dialogue system there need to exist a balance between the hard prompting of the user and the spontaneity of human speech. The goal in a computer based system is foremost that the user should succeed in the task the system is used for, but also that he should want to use it again. In the light of these facts, it seems naive to propose a system, which only allows constricted speech, acts. A better solution would be a system, which combines the approach of reducing disfluency with the approach of detecting and understanding disfluency.

Acknowledgements

This study was carried out at Telia Research AB in Farsta, Stockholm, Sweden. We would like to thank our supervisor Robert Eklund for his support.

References

Asp, A., & Decker, A. 2001. *Reducing Disfluency through Speech Act Design.* Bachelor's degree thesis. Dept. of Linguistics, Stockholm University.

Bell, L., Eklund, R., & Gustafson, J. 2000. A comparison of disfluency distribution in a unimodal and a multimodal speech interface. *Proceedings ICSLP 2000,* Beijing, China, 16–20 Oct. 2000. Vol **3**: 626–629.

Carletta, J., Isard, S., Doherty-Sneddon, G., Isard, A., Kowtko, J.C., & Anderson, A.H. 1997. The reliability of a dialogue structure scheme. *Computational Linguistics*, **23:** 13–31.

Eklund, R. 1999. A comparative analysis of four Swedish travel dialogue corpora. *Proceedings of Disfluency in Spontaneous Speech Workshop.* UC Berkeley, CA. pp. 3–6.

Grice, P. 1975. Utterer's Meaning and Intentions. *Philosophical Review,* **78:** 147–177.

Levelt, W. 1989. *Speaking: From Intention to Articulation.* Cambridge, Mass., MIT Press.

Oviatt, S. 1995. Predicting spoken disfluencies during human–computer interaction. *Computer Speech and Languages*, **9(1)**: 19–35.

Rayner, M., Carter, D., Bouillon, P., Digalakis, V., & Wirén, M. 2000. *The Spoken Language Translator.* Cambridge, Cambridge University Press.

Shriberg, E. 1994. *Preliminaries to a Theory of Speech Disfluencies.* PhD thesis, UC Berkeley, USA.

Stolcke, A., Coccaro, N., Bates, R., Taylor, P., van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., & Meteer, M. 2000. Dialogue act modelling for automatic tagging and recognition of conversational speech. *Computational Linguistics,* **26(3).**