A Comparison of Disfluency Distribution in a Unimodal and a Multimodal Speech Interface

Linda Bell†, Robert Eklund‡ and Joakim Gustafson†

† Centre for Speech Technology, Royal Institute of Technology, Stockholm, Sweden
‡ Telia Research AB, Farsta, Sweden and
NLPLab, Dept. of Computer and Information Science, Linköping University, Sweden

ABSTRACT

In this paper, we compare the distribution of disfluencies in two human–computer dialogue corpora. One corpus consists of unimodal travel booking dialogues, which were recorded over the telephone. In this unimodal system, all components except the speech recognition were authentic. The other corpus was collected using a semi-simulated multi-modal dialogue system with an animated talking agent and a clickable map. The aim of this paper is to analyze and discuss the effects of modality, task and interface design on the distribution and frequency of disfluencies in these two corpora.

1. INTRODUCTION

In human–human as well as human–computer dialogue, spontaneous spoken language contains disfluencies (pauses, truncations, prolongations, repetitions, false starts etc.), or DFs for short. For spoken dialogue system applications, DFs can be problematic, since current automatic speech recognition is limited in its ability to process them. Depending on the type of discourse or task involved, the type and frequency characteristics of DFs will vary. In general, we need to increase our knowledge of how the setting, task, timing and overall fluency of the human–computer dialogue affects DF distribution. Previous studies have shown that DF rates and the frequency and distribution of particular types of DFs vary according to the scenario and task details. Furthermore, longer, more spontaneous utterances tend to be more disfluent than briefer, more structured utterances [9, 13]. Moreover, individual predispositions are important. It has been shown that some speakers are consistently more disfluent than others [2, 13]. Other factors, such as planning difficulties, speech rate, confidence, social relationships and gender have also been discussed in conjunction with DFs [3, 13]. Furthermore, user expectations and previous experience with spoken dialogue systems might play a role.

In a study where multimodal interaction was compared with a system that supported speech alone, Oviatt reported that multimodal interaction tended to contain briefer and simpler language [8]. Multimodal interaction has also been shown to be advantageous from the point of view of error handling, since users tend to switch from one modality to another when their interaction with the computer becomes problematic [10].

Are there any differences in disfluency rates and distribution when unimodal and multimodal interaction is compared? While several studies have been devoted to single-channel applications, such as telephone-based services, or speech-governed screen-based applications, no studies—to the best of our knowledge—have yet compared the occurrence of DFs in a unimodal telephone-based system with a system with multimodal input possibilities. This paper compares DFs in two Swedish corpora of human–machine interaction, a single-channel corpus collected at Telia Research [4] and a multichannel corpus, collected at KTH [1].

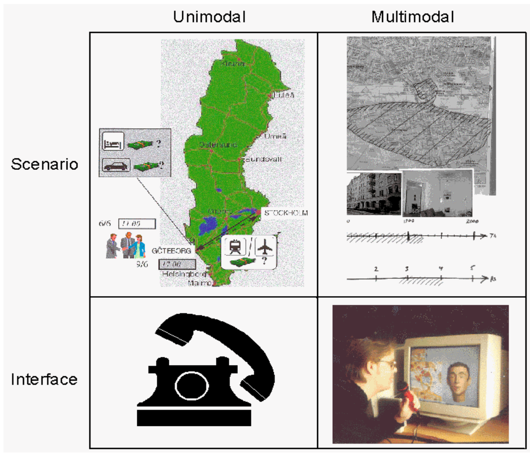

Figure 1: The scenarios and modi for the unimodal and the multimodal corpus.

2. METHOD

2.1. Data

The scenarios and collection setting for the two corpora are shown in Figure 1.

Unimodal/Human–Machine This corpus (UC) contains human–machine business travel booking dialogues, collected over a telephone line. A wizard was used to simulate speech recognition, while all other components were authentic. The corpus consists of 16 speakers (9 male, 7 female). The subjects were all Telia employees, and were used to the task of booking business trips. In order to avoid linguistic bias, the subjects were given the tasks in pictorial form, and they were also given some time to prepare the task. All subjects believed they were talking to a functional system.

Multimodal/Human–Machine The multimodal AdApt corpus (MC) contains speech and graphical data from users who interacted with a semi-simulated multimodal dialogue system [1]. AdApt is an experimental dialogue system which is used to retrieve information about apartments in downtown Stockholm. The system's graphical interface consists of an animated talking head, an interactive map and a table. The MC corpus consists of 16 speakers (8 male, 8 female), each of whom performed two dialogues with a Wizard-of-Oz version of the system. Subjects were informed that they could use either speech or graphical input at any time during the dialogues. They were given pictorial tasks in which they were asked to look for apartments in different Stockholm neighborhoods. While about half of the subjects reported that they had previous experience of web-based tools in the real-estate domain, none had interacted with a multimodal dialogue system before. Post-experimental interviews showed that all users had been unaware of the fact that this was not a real system.

2.2. Disfluency Annotation

All corpora were labeled according to an annotation scheme described in Eklund [4]. This system draws on the annotation scheme developed by Shriberg [13], with some extensions and minor changes. Both UC and MC were labeled by the second author. The following DFs were covered:

Filled pauses (FPs) Also called "filler words" in the literature, most often realized as "eh" or "öh" in Swedish.

Unfilled pauses (UPs) Silent parts in fluent speech. An example would be "I want a ……. flight to Kiruna".

Prolongations (PRs) Segments which are markedly longer than in normal, fluent speech, e.g. "Ennnnn trea eller fyra" (A three-room or four-room [flat]).

Explicit Editing Terms (EETs) Words or phrases like "Sorry", "No, wrong", "I mean…" and so on.

Truncations (TRs) Interrupted words, either in repairs or caused by an intervening system/agent, e.g., "Book the fli…"

Mispronunciations (MPs) Words with the wrong pronunciation, e.g. "Är den nyredo … nyrenoverad?" (Is it newly renovated?).

Repairs (REPs) A sundry variety of self-corrections, including substitutions (I want to find a train plane to Malmö), repetitions (Please find me … find me a ticket to Stockholm), insertions (I want a ticket a cheap ticket to Östersund) and others. In this paper, each interruption point counted as one REP, regardless of whether the repair was simplex or complex (employing nested structures).

2.3. UP: DF or not?

It is sometimes cumbersome to decide whether or not a specific item is a sign of disfluent speech, and UPs are often excluded from DF statistics. A likely reason for this is that in English UPs do not present a major problem to recognizers. One argument for not including UPs in DF analyses is that their number heavily depends on the definition of an utterance. We would like to argue, however, that UPs occur on a scale from authentic hesitation phenomena, to planned breaks in-between different "utterances". One obvious case where UPs must be considered is when they occur inside words, which has been observed, in both Swedish [5], and German [6]. In our data, UPs occur inside roots, e.g. in the word "fö… UP ...re" (be… UP ...fore). UPs also appear between lexical morphemes in compounds. An example from MC is the word: "fyrarums… UP ...lägenhet" (four room... UP ...apartment). In Swedish compounds are normally written as one word, and UPs inside compounds consequently constitute a problem to recognizer lexica. The word "konferens.. eh UP eh ..lokalen" (the conference… eh UP eh …hall) includes both FPs and a UP. A weaker case is when UPs occur between words, but in positions where they indicate hesitation due to planning, e.g. "en tur-och-retur till... UP ...Borås" (a round trip to... UP ...Borås). A difficult case is when UPs occur between constituents, e.g., "Finns det nån som är byggd före 1850… UP ...på hela Södermalm" (Is there one which was built before 1850… UP …on the whole of Södermalm), where the part preceding the UP forms a complete sentence. Such cases could result from the user reacting, by giving additional information, when the system does not respond fast enough.

3. RESULTS

3.1. Corpus Statistics

Overall corpus statistics are given in Table 1. The differences are statistically significant both when one-word utterances are included ($p = 0.004$, chi-square), and when one-word utterances are excluded ($p < 0.001$, chi-square).

Table 1: Summary corpus statistics and overall DF rates. The percentage of disfluent utterances is provided both including and excluding one-word utterances.

	UC	MC
No. subjects	16 (9M/7F)	16 (8M/8F)
No. utts.	602	847
No. utts. excl. 1-word utts.	413	799
No. words	4,013	5,829
No. disfl. utts.	252	291
Disfl. utts / total no. utts.	41.8%	34.3%
Disfl. utts./ total utts. excl. 1-word utts.	61.0%	36.4%

3.2. Disfluency Statistics

In Table 2, the various types of DFs are broken down by type.

Table 2: Summary of DF rates. For both corpora, the numbers and percentages are given, broken down by DF type. The number is divided by the total number of utterances and words in the corpora, respectively. The number of DFs is also divided by the number of utterances excluding one-word utterances.

	UC	MC
Total no. FPs (per word)	197 (4.9%)	151 (2.6 %)
Total no. UPs (per word)	385 (9.6%)	441 (7.6%)
Total no. PRs (per word)	93 (2.3%)	52 (0.9 %)
Total no. TRs (per word)	93 (2.3%)	43 (0.7%)
Total no. MPs (per word)	8 (0.2%)	6 (0.1%)
Total no. EETs (per word)	11 (0.3%)	34 (0.6%)
Total no. REPs (per word)	172 (4.3%)	65 (1.2%)
• no. DFs incl. UPs (per word)	959 (23.9%)	792 (13.6%)
• no. DFs excl. UPs (per word)	574 (14.3%)	351 (6.0%)

Overall Figures As can be seen in Table 2, speakers in MC produced, on average, 6.0% DFs per word with UPs excluded. If PRs, TRs and MPs are excluded, which is the case of most previous studies, this figure drops to 4.2%. The UC corpus exhibits higher figures: 14.3% when UPs are excluded. When PRs, TRs and MPs are excluded, the figure drops to 9.4%. For both corpora, the figures are more similar to the figures reported for human–human communication, and slightly higher than the figures normally given for human–machine communication. Part of the explanation for this could be that both UC and MC open microphones rather than push-to-talk, which forced the users to plan their contribution while speaking. Eklund [4] reports on four different corpora, including UC, as well as two WOZ corpora and a human–human corpus using the same tasks, and finds that while the two WOZ corpora and the human–human corpus are similar with regard to DF rates, UC contains a higher rate of DFs. This could imply that UC may not be fully representative for structured human–computer dialogue.

Filled Pauses FPs are more common in UC than in MC. However, the number of utterance-initial FPs is about the same in UC and MC: 43.1% and 42.4%, respectively. Shriberg [12] and Bortfeld et al. [3] report that men produce significantly more FPs than do women. These results are not corroborated in our study. Women produced 1.34% FPs as divided by the total number or words, while men produced 1.25%. This difference is not significant.

Unfilled Pauses The number of UPs are significantly higher in UC than in MC ($p = 0.001$, chi-square).

Prolongations PRs are significantly more common in UC ($p < 0.001$, chi-square).

Truncations TRs are more also common in UC. One possible cause could be that roughly 25% of the TRs in UC are system-interruptions, but the difference is still significant ($p < 0.001$, chi-square).

Mispronunciations MPs are rare in both corpora, 0.2% in UC and 0.1% in MC and the difference is not significant ($p < 0.2$, chi-square). This confirms previously reported analyses, and re-establishes the fact that MPs are indeed a rare phenomenon.

Explicit Editing Terms The rate of EETs is slightly higher in MC than in UC. Although this difference is statistically significant when including one-word utterances, it is not significant when one-word utterances are excluded. Since one-word explicit editing phrases are hard to conceive, one can conclude that EETs do not differ between the two corpora.

Repairs The number of REPs is significantly higher in UC than in MC ($p < 0.001$, chi-square).

3.3. Underlying Factors

Sentence Length

As can be seen in Figure 2, there is a difference between the two corpora with regard to the number of DFs as a function of utterance length.

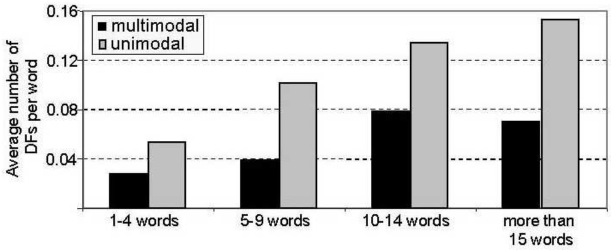

Figure 2: Number of DFs per word as a function of utterance length.

In UC, the figures are in line with previously reported studies in that the number of DFs increases as a more or less linear function of utterance length. MC deviates from the norm by displaying fewer DFs in the utterances that were more than 15 words than those that were 10-14 words. As will be shown below, this can partly be attributed to the function of the utterances in which the DFs occur, see Figure 4.

Individual Variation According to Shriberg's report on individual 'styles' of disfluency [13], certain speakers are more likely to use repetitions while other speakers exhibit a relatively high number of deletions. Furthermore, Branigan et al. [2] show that frequent occurrences of one type of disfluency for an individual speaker often correlate with high frequencies of another type of disfluency. Thus, some speakers seem to be more disfluent than others, regardless of the type of DF. In the present study, a few of the speakers in both the unimodal and multimodal corpus exhibited a strikingly high number of disfluencies, relatively speaking. As can be seen in Figure 3, these individual differences are apparent even in turns of average length. There are even two speakers in the unimodal corpus and two speakers in the multimodal corpus who were not disfluent at all. Individual variation thus exceeds most other kinds of factors in explaining DF rates. In our data, factors such as gender, age and computer skill had no effect on DF rates.

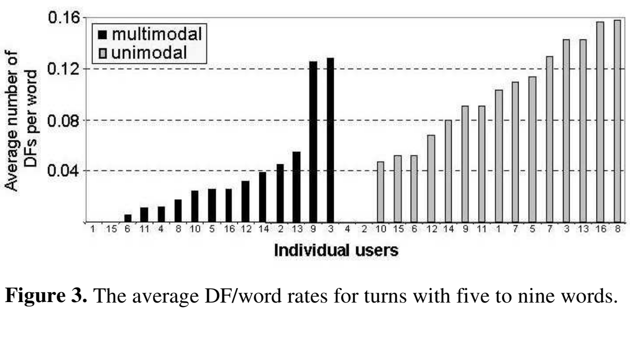

Figure 3. The average DF/word rates for turns with five to nine words.

Dialogue State According to Oviatt [9], the structure of the dialogue affects the manner in which users interact with a spoken or multimodal dialogue system. A system that employs an unconstrained format will encourage its users to produce utterances with higher information-per-utterance ratio than users who are prompted for more specific information. In UC, the system greeted the subject with an open question like "Welcome to the travelling service. How may I help you?", while in MC, the opening utterance from the system was the more constraining: "Hello my name is Urban. I can help you find apartments in Stockholm. Where would you like to livc?" This could explain that the average length of the first user utterance in UC is 17 words, while the first user utterance in MC is 10 words on average.

Another factor which is likely to have affected the collected data is that the wizard of the UC system was not explicitly instructed to limit the number of words in an utterance that he should 'understand', nor was he instructed to misunderstand fragmented or otherwise problematic utterances. Similarly, the wizard in MC 'understood' long and fragmented utterances within the domain of the system. However, the MC wizard did not 'understand' utterances with complicated syntax or out-of-domain words.

As is indicated in Figure 4, disfluency rates in UC are highly dependent on the utterance type in the dialogue in which they occur. In some cases, the utterance type appears to be even more influential than utterance length as a way of explaining DF distribution.

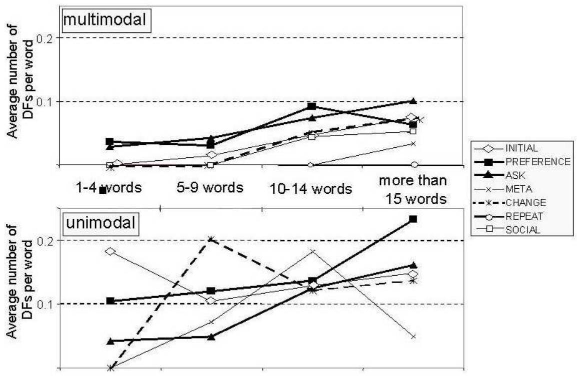

Figure 4: Number of DFs per word as a function of utterance type. INITIAL: the initial turn of each task; PREFERENCE: preferences like destination and number of rooms; ASK: question within the task; META: question about the system capabilities; CHANGE: changing of features such as departure time of a suggested trip; REPEAT: the user asks for repetition (only in MC); SOCIAL: Greeting (only in MC).

In MC, utterance type does not seem to affect DF distribution in a significant way. However, in ten-word sentences or longer, there is an increase in DF production irrespective of utterance type. In UC, it was clear that certain stages in the dialogues required a lot of planning on the part of the user. In particular, this was the case when the users were asked to specify the departure times. The scenarios specified scheduled times for meetings, conferences etc. and the users had to figure out for themselves when they had to arrive at the destination in order to make it on time for their appointment. Naturally, this required more planning and effort than the simple 'slot-fillling' questions that were frequent at other places in the dialogues. Consequently, the peaks for 'preference' for UC in Figure 4 above can be explained by the elevated DF figures for these specific turns. When the users of UC suggested a departure time, the system sometimes proposed, in detail, a trip with too late an arrival time for the user to be able to make his or her appointment. This lead subjects to enter into a clarification subdialogue with the system, and these attempts to negotiate with the system often yielded long and highly disfluent user utterances. These utterances are labeled as 'change' in Figure 4. This tendency can also be seen in MC, albeit to a lesser extent.

At certain points in UC as well as in MC, open questions from the system could be assumed to have encouraged the subjects to express themselves with some verbosity. The average utterance length in both corpora was about 7 words. However, after an open question from the system the average utterance length in UC increased to 13 words, while the corresponding figure for MC was 9.5 words. Thus, the tendency to become more verbose after unconstrained questions appears to be more accentuated in UC. The reason for this is probably that the open questions in MC were within the task at hand, while the open questions in UC initiated a new (sub-)task. In MC a typical open question was "What else do you want to know about the apartment", while a typical open question in UC was "I have booked a flight from *A* departing at *T1* to *B* arriving at *T2*. What else do you want to book?". It is likely that the subjects were affected by the system's verbose summary of the booking. Since the system 'understood' long and informationally dense utterances from the start, the users may have been implicitly encouraged to supply the system with as much information as possible in a single turn.

Topicalization On the grammatical level, one notable difference between the corpora is the occurrence of topicalized utterances in MC, that are not found in UC. A total number of 28 utterances in MC have the form "Den gröna fastigheten, har den balkong?" ("The green building, does it have a balcony"), rather than the standard "Har den gröna fastigheten balkong?" ("Does the green building have a balcony"). These topicalized sentences are characteristic in that the fronted item is followed by either a FP or an UP, e.g., "Eh den röda fastigheten på Swedenborgsgatan, eh har den balkong?" ("Eh the red house on Swedenborgsgatan, eh does it have a balcony?"). The fact that the users of MC have the discourse objects visually available, at least during certain stages of the interaction, seemingly has an effect on both the grammar and the DF distribution.

4. DISCUSSION AND FUTURE WORK

A number of factors contributed to the differences in DF rates between the two corpora. The scenarios were not identical, and the time-planning feature of the UC dialogues can be assumed to have influenced the results significantly. There was a greater number of very long sentences in UC, which raised the DF rates in this corpus. These, and probably other factors, contribute to the differences in results reported for the corpora. However, some of the observed dissimilarities can be ascribed to the modality used in the collection. Oviatt [9] reports that telephone speech is more disfluent that face-to-face conversations. This could explain the overall higher DF rate in UC as compared to MC. Adding a face seems to increase the naturalness of the interaction. Despite the fact that the animated face in MC was not a real human face, the MC corpus contains a higher degree of social and conversational behavior than UC. Although Nass & Gong [7] point out that channel consistency is crucial in human–computer interaction, we believe that the higher DF rate found in UC could also be explained in terms of the "Computers Are Social Actors" hypothesis [11], i.e. that people basically treat everything human-like in the way they treat a real human being.

A clear difference in the interface modality dimension is that a telephone interface puts heavier demands on the buffer memory of the user when the system presents information than does a graphical interface. This could explain the higher frequency of DFs in UC in interactional stages where the user has to react to information output from the system, while at the same time keeping, and accessing, the required information in their working memories. As has been shown, the occurrence of topicalized utterances in MC shows that the way information is presented to the user affects the syntax of the users' responses, and consequently also DFs distribution.

Future work includes a further exploration of the advantages of multimodal interaction from the point of view of DFs. More specifically, we intend to examine how multimodal interfaces can be used to lessen the cognitive load of a user, thus decreasing DF rates, by displaying parts of the information (e.g. time tables) graphically rather than verbally. A combination of verbal and graphical channels for conveying information should be the most efficient design for human–machine interaction.

5. ACKNOWLEDGEMENTS

The authors wish to thank Eva Gustafson for labeling the dialogue states in UC.

6. REFERENCES

1. Bell, L., Boye, J., Gustafson, J. & Wirén, M. 2000. Modality Convergence in a Multimodal Dialogue System. *Proc. Götalog 2000*, Fourth Workshop on the Semantics and Pragmatics of Dialogue, pp. 29–34.
2. Branigan, H. Lickley, R. & McKelvie, D. 1999. Non-Linguistic Influences on Rates of Disfluency in Spontaneous Speech. *Proc. ICPhS'99*, pp. 387–390.
3. Bortfeld, H., Leon, S.D., Bloom, J.E., Schober, M.F. & Brennan, S.E. 1999. Which Speakers Are Most Disfluent In Conversation, And When? *Proc. Disfluency in Spontaneous Speech Workshop*, Berkeley, California, 1 July 1999, pp. 7–10.
4. Eklund, R. 1999. A Comparative Study of Disfluencies in Four Swedish Travel Dialogue Corpora. *Proc. Disfluency in Spontaneous Speech Workshop*, Berkeley, California, 1 July 1999, pp. 3–6.
5. Eklund, R. & Shriberg, E. 1998. Crosslinguistic Disfluency Modeling: A Comparative Analysis of Swedish and American English Human–Human and Human–Machine Dialogues. *Proc. of ICSLP'98*, Sydney, Nov. 30–Dec. 5, vol. 6, pp. 2631–2634.
6. Lüngen, H., Pampel, M., Drexel, G., Gibbon, D., Althoff, F., & Schillo, C. 1996. Morphology and Speech Technology. *Proc. ACL–SIGPHON Conference*, Santa Cruz, pp. 25–30.
7. Nass, C. & Gong. L. 1999. Maximized Modality or Constrained Consistency? *Proc. AVSP'99*, Santa Cruz, August 7–10, pp. 1–5.
8. Oviatt, S.L. 1997. Multimodal Interactive Maps: Designing for Human Performance. *Human–Computer Interaction* 12, pp. 93–129.
9. Oviatt, S.L. 1995. Predicting spoken disfluencies during human–computer interaction. *Computer Speech and Language*, *9*, pp. 19–35.
10. Oviatt, S.L. & VanGent, R. 1996. Error resolution during multimodal human–computer interaction *Proc. ICSLP'96*, vol. 1, pp. 204–207.
11. Reeves, B. & Nass, C. 1996. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge University Press/CSLI, New York.
12. Shriberg, E. 1996. Disfluencies in Switchboard. *Proc. ICSLP'96*, Philadelphia, 3–6 October 1996, vol. addendum, pp. 11–14.
13. Shriberg, E. 1994. *Preliminaries to a theory of speech disfluencies*. PhD thesis, University of California, Berkeley.