Crosslinguistic Disfluency Modelling: A Comparative Analysis of Swedish and American English Human–Human and Human–Machine Dialogues

Robert Eklund† & Elizabeth Shriberg‡

†Telia Research AB, Farsta, Sweden
‡SRI International, Menlo Park, CA, USA

ABSTRACT

We report results from a cross-language study of disfluencies (DFs) in Swedish and American English human-machine and human-human dialogs. The focus is on comparisons not directly affected by differences in overall rates since these could be associated with task details. Rather, we focus on differences suggestive of how speakers utilize DFs in the different languages, including: relative rates of the use of hesitation forms, the location of hesitations, and surface characteristics of DFs. Results suggest that although the languages differ in some respects (such as the ability to insert filled pauses within 'words'), in many analyses the languages show similar behavior. Such results provide suggestions for cross-linguistic DF modeling in both theoretical and applied fields.

1. INTRODUCTION

In recent years, automatic speech recognition (ASR) systems have attained accuracy levels on constrained tasks that are sufficient for many commercial purposes. However for more open-ended speech input, robustness remains an important issue, both at the level of word recognition, and at higher levels such as natural language understanding. One problem to be addressed is the processing of disfluencies (e.g. filled pauses, repetitions, truncated words, and repairs), or DFs for short, which occur frequently in spontaneous speech. This paper addresses the question of cross-linguistic disfluency modeling by analyzing dialogues in two different languages, Swedish and American English. The analyses examine data in a similar domain, air travel planning, and across two different dyad settings: human–machine (H–M) and human–human (H–H). However, since overall rates may be related to details of the data collection paradigm rather than language proper, our focus is on three types of analyses not inherently tied to the overall rate of DFs. These include relative rates of hesitation forms, the distribution of DFs by position, and surface characteristics of DFs, such as word fragmentation, editing phrases, and retraced words.

2. METHOD

Speech data. The Swedish data were collected as a part of the Spoken Language Translator/Database project [1] at Telia Research AB. The H–M data were collected in a Wizard-of-Oz simulation [5], and the H–H data were collected using real travel agents at the travel agency Nyman & Schultz in simulated booking contexts. The American English H–M data were drawn from the ATIS corpus of air travel planning dialogues with either a Wizard-of-Oz or actual spoken language system [6]. The American English H–H ATIS data came from a corpus of American Express travel dialogues between callers making real travel plans and AMEX agents [4], using only the speech of the callers.

It is important to note that although the corpora are similar in that they all contain ATIS booking data and both are either H–M or H–H, there were some considerable differences in collection paradigms. Both Swedish corpora are telephone conversations, whereas the American H–M corpus used a push-to-talk device. Task and instructions also differed somewhat both across languages and dyad type: Swedish subjects were given tasks in pictorial form, ATIS subjects were given a written scenario to solve, and AMEX subjects made actual travel plans.

DF Annotation. Data in the two languages were analyzed and labeled using common or mappable annotation schemes [7] [3], enabling direct statistical comparisons. From the basic annotations it was possible to automatically classify DFs into types and to pull out simple word-based characteristics. Our experience in annotating these corpora suggests that the annotation method developed in Shriberg [7] indeed seems portable to Swedish. Interestingly, in addition to allowing annotation of most common DFs, the system also seems to apply to Swedish as well as American English for the annotation of "complex DFs" or DFs having multiple interruption points (such as "flights – fares – fares"). In preliminary analyses it appears that in Swedish, like American English, such cases can be analysed as compositional, using chained or nested DF structures.

3. ANALYSES

3.1. Overall Distribution of DFs

Although overall rates are not necessarily indicative of language differences, it is helpful to present some summary statistics to put the remaining analyses in context. Table 1 provides information on the number of speakers, words, sentences, and DFs for each of our four corpora. The set of DFs counted are all cases involving at least one deleted word (including filled pauses), as in [7].

Table 1: Summary Statistics and Overall DF rates.

	Swedish H–M (WOZ-2)	Swedish H–H (Nymans)	AmE H–M (ATIS)	AmE H–H (AMEX)
No. subjects	22	8	523	66
No. sentences	1,728	970	18,483	1,755
No. sentences excl. 1-wd-sents	1,531	590	18,331	1,335
No. words	11,375	5,630	185,749	11,592
No. disfl. sents	454	227	1,227	423
No. DFs	672	360	1,457	594
% disfl. sents of tot. no. sents	26.3%	23.4%	6.64%	24.1%
% disfl. sents of tot. no. sents. excl. 1-w. sents.	29.6%	38.5%	6.69%	31.7%
Tot. no. DFs / tot. no. words given as %	5.9%	6.4%	0.78%	5.1%

Three observations can be noted from Table 1. First, rates in the Swedish data are higher overall than those in the American English H–H data. This is possibly a consequence of task details, but pending more detailed analysis, we prefer to leave the issue open at this time. Second, within each language, DFs were more frequent in the H–H than in the H–M dialogue, suggesting that people modify their speech when they converse with a machine. To allow truly natural interaction, future systems should be able to handle DF rates on par with that of H–H conversation. Third, DF rates are markedly lower in the American English H–M data. As pointed out earlier, this is likely to be an artefact of the data collection method, which differed from the other corpora particularly in that subjects were able to plan their utterances and when ready to push the "talk" button; in all other corpora they exchanged turns in real time. As a result, DFs should be limited since there is no need to "hold the floor", and also because there is increased time to plan for the successful completion of an utterance [7]. In addition, for a large portion of the ATIS interactions, subjects used a fully automatic spoken language system rather than a wizard system. In such contexts subjects saw the often-imperfect output of the ASR system, and thus tended to speak more carefully and use shorter sentences than in the other three corpora.

3.2. Relative Rate of Hesitation Forms

As just discussed, DF production is sensitive to task details. Thus unless one has perfect control of such details (which was not true in our case), it is preferable to examine metrics not inherently tied to overall rates in order to explore differences that might be related to the language.

Filled Pauses and Repetitions. One such meaningful metric is the relative rate of the use of DFs that have a similar function. In both Swedish and American English, speakers conveyed hesitation using at least two different forms involving extra words: filled pauses (FPs) and repetitions (REPs). Table 2 shows the relative rate at which speakers used the two forms.

Table 2: Relative Rate of Filled Pauses and Repetitions.

	Swedish H–M (WOZ-2)	Swedish H–H (Nymans)	AmE H–M (ATIS)	AmE H–H (AMEX)
Total DFs	672	360	1,457	594
Tot. no. FPs	381	146	593	353
% of tot. no. DFs	57%	40%	41%	59%
Tot. no. REPs	43	75	206	116
% of tot. no. DFs	6%	21%	14%	20%

As shown, FPs make up a large percentage—between 40% and 60%—of all DFs. One possible explanation for the higher prevalence of FPs is that regardless of language, FPs do not require the speaker to commit yet to the word content. Thus FPs can be used for example at the start of an utterance or turn, where the speaker wants to simply hold the floor. FPs can also be used in cases in which the speaker has committed to the semantic content, but not yet to the grammatical encoding (or word sequence) of that content. REPs are less frequent, making up only about 20% of DFs with the exception of the Swedish H–M data (which awaits explanation).

Prolongations. Although our overall analyses included only those DFs involving extra words, we were also interested in the usage of other hesitation forms by which an ASR system might be affected, in particular prolongations (or phones extended in duration due to hesitation). Rates for prolongations are provided in Table 3.

The Swedish data were coded in their entirety for such hesitations. For the American English H–M data, a subset of 4,739 utterances were used; no such annotations were available for the American English H–H data. Separate figures are given for word-initial, word-medial and word-final phones. For each of these categories, the percentages of vowels or sonorants (liquids, nasals, laterals) are given. Note that the percentages of initial/medial/final phones do not sum exactly to 1.0, since cases of prolongations on single-phone words (such as "I") were counted as both initial and final.

As shown, the overall rate of prolongations is quite similar in the H–M data across languages. Furthermore, although the overall rate of prolongations is somewhat higher in the H–H data, all three corpora show similar relative rates of prolongations by position, with roughly a 30-20-50 distribution for initial, medial, and final position, respectively.

Interestingly, while we had expected that prolongations would tend to be word-final and to involve a vowel or sonorant, such cases were by no means the majority of cases.

Table 3: Rate of Prolongations by Position and Phone.

	Swedish H–M (WOZ-2)	Swedish H–H (Nymans)	AmE H–M (ATIS)
No. Prolongations	81	106	299
% of Words	0.7%	1.8%	0.5%
% of Sentences	4.9%	11.2%	6.3%
% Initial phone	31%	24%	32%
% vowels/sonorants	(16%)	(35%)	(43%)
% Medial phone	18%	17%	22%
% vowels/sonorants	(7%)	(33%)	(63%)
% Final phone	49%	58%	50%
% vowels/sonorants	(90%)	(87%)	(77%)

Close to half of the prolongations in both Swedish and American English occurred in non-final position, and of these a large percentage involved phones other than vowels or sonorants. Swedish was particularly notable in this regard, with a far lower rate of vowel/sonorant prolongations in these positions, even though vowels and sonorants were the most common type of prolongation for final position.

3.3. Location Effects

We also looked at the distribution of DFs with respect to the location of sentence and word boundaries. Because location is dependent on DF type, here we restrict our summary and report location for only the most frequent DF type across corpora, FPs.

Filled Pauses and Sentence Boundaries. Across corpora, FPs were not evenly distributed in sentences, but rather more likely to appear in certain positions than others. For simplicity we examined the rate of FPs in two positions with respect to sentence boundaries: sentence-initial position and sentence-medial position. Table 4 shows the distribution of FPs by position. In all corpora, speakers use FPs more often when in sentence-initial position than once they have begun the sentence. FPs in this position could reflect global planning as opposed to more local phenomena such as lexical search. For speech applications, initial position should be given higher probability of a FP than elsewhere. If one wants to add FPs to a grammar in a limited fashion, sentence-initial position would be the optimal position in which to allow them.

Table 4: Position of FPs and REPs. *P(init)* = probability of an FP in initial position (i.e. no. of initial FPs divided by no. of sentences.) *P(med)* = probability of an FP in non-initial position (i.e. no. of non-initial FPs divided by no. words minus no. sentences).

	Swedish H–M (WOZ-2)	Swedish H–H (Nymans)	AmE H–M (ATIS)	AmE H–H (AMEX)
P(init)	209/1,728 0.121	35/970 0.036	249/18,675 0.013	224/1,821 0.123
P(med)	172/9,647 0.018	111/4,660 0.024	344/169,652 0.002	129/10,941 0.012

Filled Pauses and Word Boundaries. Another observation we made is that there seems to be a difference between the languages in that a small number of FPs occur *within* words in the Swedish data. While unfilled pauses (not treated in this paper) and prolongations frequently occur within words, word-internal FPs are rare in Swedish. In the material labelled so far, only three examples have been found. Moreover, although they undeniably occur within words, they do not occur within morphemes (the way unfilled pauses and prolongations do). An example is the word[1] "tågförbindelse", ("train connection") which appeared with a distinct filled pause, occurring between the two morphemes, thus:

tåg EH *förbindelse*

In American English, no cases of within-word FPs were found despite large amounts of data. This suggests that languages with significant productive word-compounding, such as Swedish or German, may allow FPs between morphemes. This difference has implications for psycholinguistic models of the status of 'words' in different languages. For ASR systems, the implication is that in languages like Swedish, word-models may have to allow for within-word FPs. It also has implications on how recognition lexica should be organized.

3.4. DF Characteristics

Disfluencies come in various shapes and forms, and speakers have various ways of effecting them, including cutting off the word before the interruption point, using editing terms such as FPs, and recycling words in the repair region, as illustrated below:

Which flights leave bef- after noon?
Which flights leave uh after noon?
Which flights leave (pause) *leave after noon?*

The presence of such characteristics can serve as cues to DF detection in speech applications, thus it is useful to explore the extent to which cues are present across languages and speech task.

Intra-word Cut-offs ("Fragments"). Fragments make up a low percentage of overall words (between 0.2 and 1.1% in the corpora) and also sentences (ranging from 2.6% in ATIS to 7.2% in WOZ-2), since DFs already have a fairly low percentage at the word level. However, of the set of DFs, fragments are frequent. They are more frequent for REPs than for other types of repairs, indicating that speakers are stopping for some reason *other* than an overt error. Fragments appear more frequently in H–M than H–H dialogue, although not by much in the Swedish data (34.4% vs. 11.1% in American English data; 18% vs. 15% in Swedish data). This could indicate a generally higher covert error rate in H–M dialog and/or an increase in the speaker's monitoring for error, resulting in words cut off earlier than in H–H dialog.

[1] That we are dealing with one word here is clear for prosodic reasons, much as American listeners would be able to distinguish between "a black uh bird" = compound noun, and "a black uh bird" = adjective + noun.

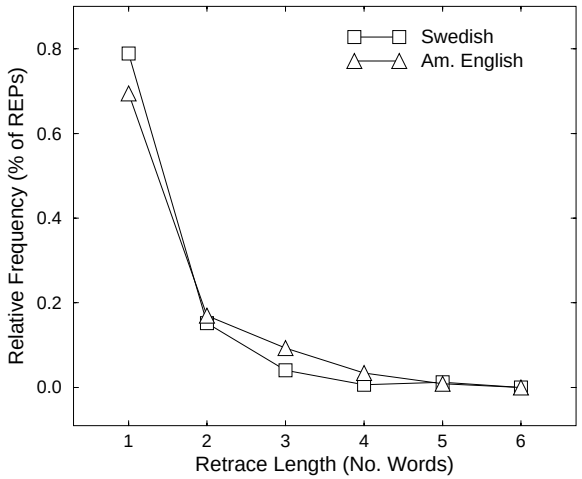

Figure 1: Distribution of Retrace Lengths by Language

Editing Phrases. Although it is commonly assumed that DFs are accompanied by editing phrases such as FPs and terms like "sorry" or "rather", in actuality such overt editing terms are relatively rare. Fewer than 40% of repairs in the Swedish H–M data were marked this way, and this is likely to be an overestimate because subjects may have been directing some of these terms to an experimenter (who monitored the sessions from the next room), rather than to the system. Even lower rates were found for the other three corpora. For the Swedish H–H data, fewer than 25% of repairs contained an editing term; for both American English corpora, the rate was below 20%. Clearly then, the rate of editing phrases is low across corpora, implying that automatic DF processing cannot rely on such phrases for DF detection. Furthermore, not all editing terms are reliable cues to repair. In all corpora, the most frequent editing term was the FP; however in both Swedish and American English, as well as German [2], FPs more often appear alone than as the editing term of a repair.

Retraced Words. After an interruption, both Swedish and American English speakers often employ retracing, or the repeating of one or more previous words before continuing. Although a full analysis of retracing should examine retracing with respect to phrasing (e.g. syntactic and prosodic) in the two languages, we noticed that the languages behave similarly in terms of a very simple measure: the probability of retracing back N words. For simplicity, we look only at the case of REPs in the two languages. A example retrace length of $N = 1$ is "the the", an example of $N = 2$ is "on the on the", and so on. Figure 1 shows the distribution of retrace lengths in the two different languages (collapsed over the H–M and H–H corpora). In both languages, the falloff in frequency by retrace length can be modeled as an exponential decay; overall there is a uniform probability that a speaker will retrace an additional word, as was found earlier for American English [7].

4. SUMMARY AND CONCLUSION

We conclude that many aspects of DFs show similarities across Swedish and American English. Although overall differences in DF rates are more likely to be explained by the specifics of the data collection set-up (for example whether there is a push-to-talk button present), relative rates of hesitation phenomena, certain positional effects, and many surface characteristics of DFs are similar across the two languages. Our results are preliminary to a more detailed analysis of the languages, speaker effects, elicitation paradigms, and so on. Nevertheless they suggest that overall, speakers appear to be using DFs in similar ways both languages, that H–M dialog differs from H–H dialog in ways that show up in DF distributions, and that there is a potential for applying similar DF modeling techniques across languages for speech applications.

5. ACKNOWLEDGEMENTS

Work of the second author was supported by NSF Grants IRI-9314967 and IRI-9619921.

6. REFERENCES

1. Agnäs, M-S., Alshawi, H., Bretan, I., Carter, D., Ceder, K., Collins, M., Crouch, R., Digalakis, V., Ekholm, B., Gambäck, B., Kaja, J., Karlgren, J., Lyberg, B., Price, P., Pulman, S., Rayner, M., Samuelsson, C. & Svensson, T. *Spoken Language Translator: First Year Report.* SRI Technical Report CRC–043. 1994.
2. Batliner, A., Kießling, A., Burger, S. & Nöth, E. Filled pauses in spontaneous speech. *Proceedings of the XIIIth International Congress of Phonetic Sciences* Vol. 3, 472–475, 1995.
3. Eklund, R. Interaction between prosody and discourse structure in a simulated man–machine dialogue. *Journal of the Acoustical Society of America*, Vol. 102, No. 5, Pt. 2, December 1997, 3202 [Abstract].
4. Kowtko, J.C. & Price, P.J. Data collection and analysis in the air travel planning domain. *Proc. DARPA Speech and Natural Language Workshop*, 119–125, 1989.
5. MacDermid, C. & Eklund, C. *Report on the First WOZ Simulation for the SLT-DB Project.* Internal Report, Telia Research AB, 1997.
6. MADCOW. Multi-site data collection for a spoken language corpus. *Proc. of the Fifth DARPA Speech and Natural Language Workshop*, Morgan Kaufman, 1992.
7. Shriberg, E. *Preliminaries to a theory of speech disfluencies.* PhD thesis, University of California, Berkeley, 1994.