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 dialog. 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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No. subjects</td>
<td>22</td>
<td>8</td>
<td>1,138</td>
<td>1,475</td>
</tr>
<tr>
<td>No. sentences</td>
<td>1,728</td>
<td>970</td>
<td>18,483</td>
<td>17,355</td>
</tr>
<tr>
<td>excl. l-wd-sents</td>
<td>1,531</td>
<td>500</td>
<td>18,331</td>
<td>1,235</td>
</tr>
<tr>
<td>No. words</td>
<td>11,075</td>
<td>5,630</td>
<td>185,149</td>
<td>11,629</td>
</tr>
<tr>
<td>No. diff. words</td>
<td>454</td>
<td>220</td>
<td>1,227</td>
<td>433</td>
</tr>
<tr>
<td>No. FPs</td>
<td>92</td>
<td>300</td>
<td>1,457</td>
<td>394</td>
</tr>
<tr>
<td>% diff. words of tot. no. words</td>
<td>26.3%</td>
<td>23.4%</td>
<td>6.64%</td>
<td>24.1%</td>
</tr>
<tr>
<td>% diff. words of tot. no. words excl. l-wd. words</td>
<td>29.6%</td>
<td>38.5%</td>
<td>6.99%</td>
<td>31.7%</td>
</tr>
<tr>
<td>Tot. no. DFs / tot. no. words given as %</td>
<td>5.9%</td>
<td>6.4%</td>
<td>6.78%</td>
<td>5.1%</td>
</tr>
</tbody>
</table>

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 users spoke the two forms.

Table 2: Relative Rate of Filled Pauses and Repetitions.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Tot. DFs</td>
<td>569</td>
<td>960</td>
<td>1,420</td>
<td>784</td>
</tr>
<tr>
<td>Tot. no. FPs</td>
<td>385</td>
<td>146</td>
<td>383</td>
<td>353</td>
</tr>
<tr>
<td>% of tot. no. DFs</td>
<td>57%</td>
<td>40%</td>
<td>41%</td>
<td>59%</td>
</tr>
<tr>
<td>Tot. no. REPs</td>
<td>46</td>
<td>75</td>
<td>206</td>
<td>116</td>
</tr>
<tr>
<td>% of tot. no. DFs</td>
<td>6%</td>
<td>21%</td>
<td>14%</td>
<td>29%</td>
</tr>
</tbody>
</table>

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 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 4,730 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 “F”) 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.
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.** $P(\text{init})$ = probability of an FP in initial position (i.e. no. of initial FPs divided by no. of sentences.) $P(\text{med})$ = probability of an FP in non-initial position (i.e. no. of non-initial FPs divided by no. words minus no. sentences).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>F(init)</td>
<td>2097/1728 (0.121)</td>
<td>36/27/0 (0.026)</td>
<td>249/18.675 (0.007)</td>
<td>224/18.675 (0.007)</td>
</tr>
<tr>
<td>F(med)</td>
<td>172/94/49 (0.018)</td>
<td>117/4/666 (0.024)</td>
<td>34/18.675 (0.002)</td>
<td>129/10.944 (0.012)</td>
</tr>
</tbody>
</table>

**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 "tagförbindelse", ("train connection") which appeared with a distinct filled pause, occurring between the two morphemes, thus: 

$täg förbindelse$

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 (31.4% 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.

\[1\]\n
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.
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-9311997 and IRI-9619921.

6. REFERENCES