DISFLUENCIES IN SWITCHBOARD

Elizabeth Shriberg

Speech Technology and Research Laboratory
SRI International, Menlo Park, California, USA
ees@speech.sri.com

ABSTRACT

Disfluencies ("um," repeats, self-repairs) are prevalent in spontaneous speech, and are relevant to both human speech communication and speech processing by machine. Although disfluencies have commonly been viewed as 'noisy' events, results from a large descriptive study indicate that disfluencies show regularities in a number of dimensions [9]. This paper reports selected results on Switchboard and two comparison corpora of spontaneous speech. Results illustrate the systematic distribution of disfluencies, and highlight differences as well as universals across corpora and speakers.

1. INTRODUCTION

Disfluencies (e.g., "um" and "uh", repeats, self-repairs) are prevalent in spontaneous speech, and are relevant to modeling both human speech communication and speech processing by machine. Although historically disfluencies have been viewed as noisy events, and have received relatively little attention, a more recent focus on spontaneous speech has directed increased interest to disfluencies in both theoretical and applied fields [1,4,6,8,9,10,11].

The goal of the present work is to illustrate that disfluencies are not 'noise' but rather show systematic distributions in various dimensions. This paper summarizes results from a large descriptive study aimed at revealing and modeling trends in the distribution and form of disfluencies in three large corpora of spontaneous speech, including Switchboard, and two comparison corpora differing in speech style.

Selected results include: (1) overall rate, (2) number of edited words, and (3) rate by disfluency type. The paper focuses on differences and similarities across corpora and speakers. Such data should guide and constrain cognitive models of speech processing; they can also aid the development of better models for automatic recognition and understanding of spontaneous speech.

2. METHOD

Data were drawn from the Switchboard (SWBD) corpus of informal telephone conversations on prescribed topics [3]. In addition, to examine how rates and types of disfluencies vary in different styles of speech, results for two additional corpora were analyzed. One comparison corpus, ATIS (Air Travel Information System, [7]), consists of human-computer dialog in the air travel planning domain. While spontaneous, it represents a quite different style of speech from SWBD, both in dyad (human-human versus human-computer) and task dimension. The second comparison corpus, AMEX, consists of recorded telephone speech from SRI employees speaking to travel agents while making actual travel plans [5]. Subsets of data were selected from these corpora, and hand-labeled for disfluencies and for sentence boundaries as described in [9]. The sizes of the hand-labeled datasets are indicated in Table 1.

	SWBD human-human free conv.	ATIS human-computer air travel	AMEX human-human air travel
words	40,515	188,327	12,762
sentences	4,583	18,675	1,821
disfluencies	2,586	745	1,694
speakers	30	66	523

Table 1: Size of Hand-Labeled Datasets

The disfluencies labeled were of the following general types, illustrated in Table 2. In the second through fifth classes, the class name reflects the relationship between corresponding words before and after the interruption point (marked by "." in Table 2). The interruption point is preceded by "edited" words (shown in strike-through font), that is, words that have to be removed to obtain the intended fluent utterance. The classification system is described in further detail in [9].

Disfluency class	Example
filled pause	she . ~~uh~~ liked it
repetition	~~she~~ . she liked it
substitution	~~she~~ . he liked it
insertion	she ~~liked~~ . really liked it
deletion	~~it was very~~ . she liked it
speech error	~~shle~~ . she liked it

Table 2: Disfluency Classification

This research was supported by DARPA and NSF under NSF Grants IRI-9314967 and IRI-8905249. The views herein are those of the author and should not be interpreted as representing the policies of DARPA or NSF.

3. RESULTS AND DISCUSSION

3.1. Rate Overall

Figure 1 shows the probability of a sentence being fluent (not having any disfluencies) as a function of the length of the sentence. As shown, each corpus exhibits a roughly exponential decay. Similar decay patterns were observed when data were examined by individual speakers, although different speakers had different absolute rates of disfluencies (consistent with past work, as reviewed in [9]).

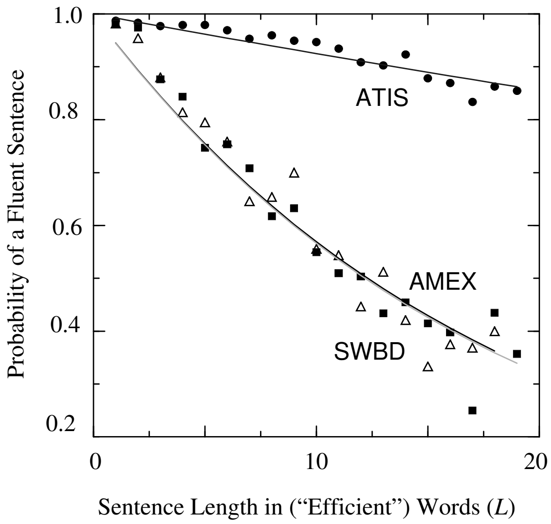

Figure 1. Probability of a fluent sentence (no disfluencies) by sentence length.

For all corpora, a good fit is provided by a simple exponential decay function, $C * b^L$, where L is the length of the sentence in "efficient" (excluding edited) words. Regressions show that for all three corpora, the value of C was close to 1.0. This is the value of the y-intercept if there is no vertical scaling of the function. The function could therefore be simplified to a single-parameter function, b^L.

The parameter b represents the fluency rate; 1-b is thus an overall rate of disfluencies per word (over all positions and assuming independently distributed disfluencies; for further discussion of violations of these assumptions see [9]). Results for b were .0553 for SWBD, .0547 for AMEX and .0078 for ATIS. Statistical tests showed no difference between SWBD and AMEX, while both were more disfluent than ATIS. Thus, these results suggest that conversational partner (human versus computer) is an essential factor in disfluency rate, while whether or not the dialog is goal-oriented (AMEX) or free conversation (SWBD) appears irrelevant to overall rates on this measure.

However, the rate was not uniform over disfluency position. Further analysis divided disfluencies into two simple categories based on location with respect to linguistic sentences. Results showed that in all corpora, disfluencies were significantly more probable at sentence beginning than elsewhere, as indicated in Figure 2. This result is consistent with past work showing disfluencies to be more likely to occur early in a phrase (e.g. [2]). Results also have implications for speech applications; for example, the language model could assign different prior probabilities to disfluencies, depending on position relative to a sentence boundary.

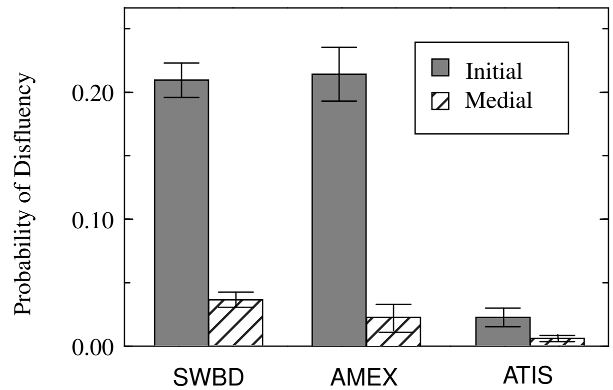

Figure 2. Probability of disfluency by location in sentence.

Interestingly, both initial and medial disfluency rates depended on the length of the sentence; both rates increased for longer sentences. This explains why the trends in Figure 1 pass through 1.0: since both initial and medial rates depend on L, both are rolled into the base in the function, leaving no sentence-level effect.

3.2. Number of Edited Words

A second study examined the distribution of the number of edited words in disfluencies in each of the three corpora. This statistic applies across all disfluency types.

The distribution of the frequency of edited words is shown in Figure 3. This measure yields a strikingly regular trend in two respects. First, each corpus shows a roughly uniform rate of decay; this can be modeled by the simple function $y = q^k$, where y is the percentage of disfluencies in a corpus having k edited words, and q is a free parameter corresponding to the rate at which frequency decays with increasing k. This is a single-parameter model. C is a deterministic function of q since the sum of percentages for all deletion lengths within a corpus must equal 1; C can be rewritten as $(1-q)/q$.

This model is consistent with results from a previous study [1], shown as the fourth dataset in Figure 3. In the previous work it was observed that most disfluencies edit only one or two words. The present, more detailed analysis shows that this is the case because frequency decays rapidly with k.

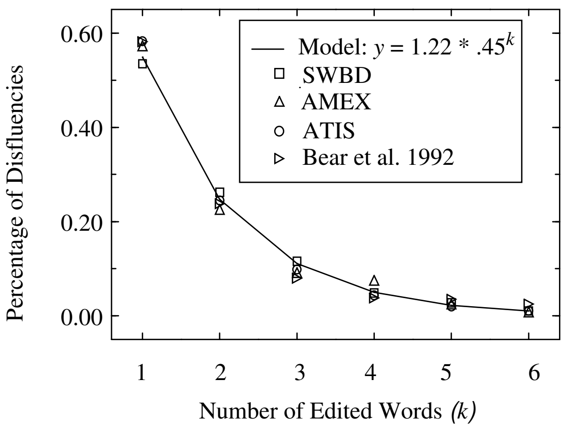

Figure 3. Distribution of edit lengths by corpus. Model reflects average value of parameter across corpora, since parameter does not differ significantly across corpora.

The second striking observation from Figure 3 is that the data for all three corpora, as well as the data from the additional study, show nearly identical values for q. A statistical comparison of slopes across corpora yielded no significant differences; thus a single value for q can be used to predict all data points in all corpora.

The finding that q does not differ over corpora should be of interest to both theoretical and applied work on disfluencies. One possible interpretation of q for theories of human speech communication is that it reflects a universal constraint on how much material can be processed in producing or comprehending disfluencies. Such an interpretation is consistent with q being invariant across corpora, since one would not expect corpus factors to interact with a limitation on processing. These results are also useful for speech applications, in particular to model sentence restarts that are difficult to detect by other means such as matched words [1]. The results suggest that such cases could be modeled as exponentially less likely, as the hypothesized interruption point occurs additional words away from the beginning of the sentence

3.3. Type Distributions

Figure 4 shows the rate of disfluencies in the corpora by disfluency type, where type is described in Table 2. The interesting observation from Figure 4 is that the large difference in overall rate of disfluencies between ATIS and the human-human corpora observed earlier (Figures 1 and 2) is attributable to only three of the disfluency types: filled pauses, repetitions, and deletions. The rate of the remaining types is roughly similar for all three corpora.

The rate differences by type suggest that different types of disfluencies arise from different underlying factors. Errors may reflect basic problems in formulation or encoding of a message; the three types that are more prevalent in human-human dialog may function to coordinate exchanges with a conversational partner. Further evidence for this possibility comes from looking at position; the first three types are highly correlated with positions relevant to turn exchanges such as turn or sentence initially. Such results also have implications for modeling disfluencies in automatic speech processing. For human-human dialog, disfluency processing should probably be focused on these three highly frequent types [10].

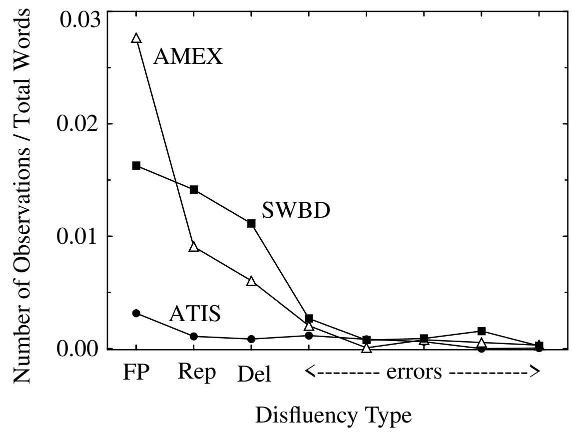

Figure 4. Rate of disfluencies by type. “FP”= filled pause, “Rep” = repetition, “Del” = deletion; see Table 2. All other types (substitutions, insertions, articulation errors and combinations of these) are indicated as “errors”.

A further noteworthy result was discovered when the type distribution for SWBD was examined by speaker. Figure 5 illustrates the results for the data shown in Figure 4 for SWBD, excluding filled pauses, and plotted by speaker.

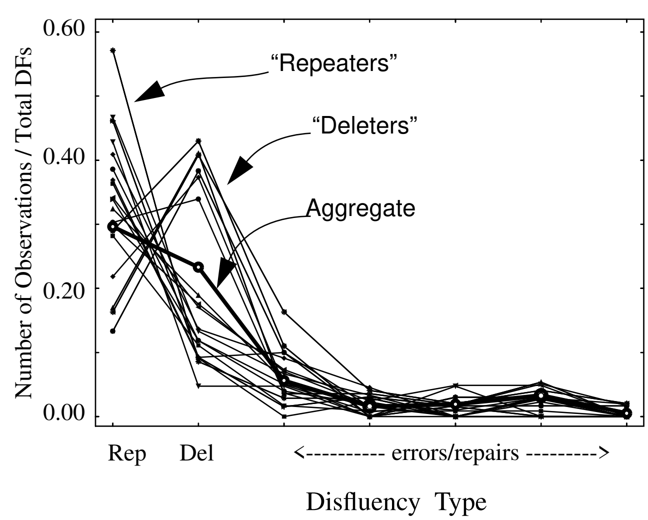

Figure 5. Proportion of types of disfluencies by speaker in SWBD. “Rep”=repetition, “Del”=deletion. Lines connect points for individual speakers.

As shown in Figure 5, the type distribution over speakers is not a good reflection of type distributions for any particular speaker. All speakers show high frequencies for the Rep and Del classes; however, speakers clearly fall into one of two patterns. One set of speakers, called "repeaters", shows a much higher rate of repetitions than of deletions. Another set of speakers, called "deleters", shows the opposite behavior. Rates of Rep and Del also showed a significant inverse correlation within speakers, and internal consistency when rates were examined over multiple conversations from the same speaker.

There was no significant difference in overall disfluency rates between the two groups, nor were there significant differences in sentence-related measures such as mean sentence length. A correlation was found, however, for speaking rate. Deleters produced on average significantly more words per second than repeaters, even after adjusting for differences in the number of words in the disfluencies themselves.

One possible interpretation of such results is that faster speakers "get ahead of themselves", and thus often have to retract provisional starts and begin anew, whereas slower speakers take more time to plan, increasing hesitations such as repetitions but reducing the need to retract false starts. Results also suggest that it would be useful to represent speakers or speaker-types in automatic disfluency processing, not only to allow for overall rate differences but also to model differences in type distributions.

Finally, mention should be made of results for filled pauses (not included in Figure 5). Filled-pause rates by speaker did not correlate with any of several sentence-related factors such as mean sentence length or rate of disfluent sentences. Filled-pause rate also did not correlate with speaking rate. Filled-pause rate did, however, correlate with gender. Men produced significantly higher rates of filled pauses than women, as measured by rate of filled pauses per word.

These results reveal an interesting difference between filled pauses and the other types of disfluencies examined; whereas the latter were either similar over speakers or correlated with characteristics of the speech produced, filled pauses were correlated with a sociolinguistic variable. The higher rate of filled pauses for men is interesting in light of the view that filled pauses may serve to "hold the floor" (see [9] for review); an inference is then that men may tend to control the floor more than women (although this does not imply that men spend more time speaking in conversation). However, it would be premature to infer a direct connection between gender and floor-holding based on these data alone, since gender is confounded with other sociolinguistic variables in SWBD (e.g., education level and occupation). In addition, further analysis is needed to assess the location of filled pauses and to take into account the gender of the listener as well as the speaker.

4. Conclusion

The selected results illustrate that disfluencies in SWBD and other spontaneous-speech corpora are not 'noisy' events; rather, they show regular trends that can be modeled by simple functions. Comparisons of parameter values across corpora show significant differences across corpora on some measures (such as overall rate), while values on other measures (e.g., number of edited words) may be universal across corpora. The distribution of disfluency types shows that the increased rate of disfluencies in human-human over human-computer dialog is attributable to only three disfluency types. Individual speakers group into two classes (repeaters and deleters) based on systematic profiles in the relative frequency of using these two frequent disfluency types; furthermore, deleters consistently speak more quickly than repeaters. Rates of the third frequent disfluency type, filled pauses, do not correlate with characteristics of the speech produced, but do correlate with a sociolinguistic variable (speaker gender). Such results must be accounted for in models of human language processing. They can also guide and constrain the development of new methods for the automatic processing of spontaneous speech.

5. REFERENCES

1. Bear, J., Dowding, J. and Shriberg, E. "Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog," *Proc. ACL*, 56-63, 1992.
2. Beattie, G.W. "Planning units in spontaneous speech: Some evidence from hesitation in speech and speaker gaze direction in conversation," *Linguistics*, *17*, 61-78, 1979.
3. Godfrey, J.J., Holliman, E.C. and McDaniel, J. "SWITCHBOARD: Telephone speech corpus for research and development," *Proc. ICASSP*, 517-520, 1992.
4. Heeman, P. and Allen, J. "Detecting and correcting speech repairs," *Proc. ACL*, 295-302, 1994.
5. Kowtko, J.C. and Price, P.J. "Data collection and analysis in the air travel planning domain," *Proc. DARPA Speech and Natural Language Workshop*, 119-125, 1989.
6. Levelt, W.J.M. *Speaking: From Intention to Articulation.* Cambridge, Mass: MIT Press, 1989.
7. MADCOW. "Multi-site data collection for a spoken language corpus," *Proc. Fifth DARPA Speech and Natural Language Workshop*, 7-14, Morgan Kaufmann, 1992.
8. Nakatani, C.H. and Hirschberg, J. "A corpus-based study of repair cues in spontaneous speech," *Journal of the Acoustical Society of America*, 95(3), 1603-1616, 1994.
9. Shriberg, E.E. Preliminaries to a theory of speech disfluencies. Ph.D. thesis, University of California at Berkeley, 1994.
10. Shriberg. E.E. and Stolcke, A. "Word predictability after hesitations: A corpus-based study," *Proc. ICSLP*, 1996.
11. Stolcke, A. and Shriberg, E.E. "Statistical language modeling for speech disfluencies," *Proc. ICASSP*, 405-408, 1996.