DISTRIBUTION OF DISFLUENCIES AND ERRORS IN ENGLISH DISCOURSE

Nanette Veilleux†, Alejna Brugos‡, Stefanie Shattuck-Hufnagel, Alicia Patterson**

†Simmons College, ‡Boston University, *Massachusetts Institute of Technology
veilleux@simmons.edu, abrugos@bu.edu, stef@speech.mit.edu

ABSTRACT

Discourse boundaries have been associated with an increased rate of disfluent events. It is hypothesized that the reason for this increase is the heavy processing requirement incurred either in planning the next chunk of discourse or in the introduction of many new or high perplexity entities. In a sample of academic lecture speech, we find that non-error disfluencies (such as filled pauses) occur preferentially shortly after (but not right at) the beginning of a new discourse segment. This suggests that the processing load may not increase just at the boundary but instead somewhat later, i.e. that the speaker can make use of the results of earlier planning during the first portion of the new segment. In contrast, errors of selection or serial ordering of grammatical elements do not show a boundary-related peak in their distribution across a discourse segment, supporting the hypothesis that this second kind of nonfluent event arises at a different point in the speech production planning process.

Keywords: Disfluencies, discourse structure, speech errors.

1. BACKGROUND

Discourse boundaries have been associated with a variety of acoustic events, including an increased rate of disfluencies [4][10]. For example, in an experimental study of spontaneous Dutch monologues, Swerts et al. [10] found that the distribution of filled pauses varied by strength of discourse boundary: they were more prevalent in the Intonational Phrase just after a strong discourse boundary (one labelled by more than 75% of labeller subjects) than after a weak boundary (one labelled by fewer than 75% of labellers). Watanabe found similar results for informal Japanese spontaneous speech [14], although not in academic lectures and prepared conference talks [13].

More generally, Arnold and colleagues [1] found a correlation between disfluencies and new information in English, and showed that disfluencies (including repairs, repeats, metalinguistic comments as well as filled pauses) occur more frequently at the beginnings of utterances. A subsequent perceptual study [2] revealed that disfluencies cue listeners to the presence of new, and presumably more difficult to process, information.

These studies focused on dialogs comprised of short utterances. However, an association between disfluencies and the onsets of larger discourse segments would also be expected, since theories of discourse, such as Grosz et al.'s Centering theory [6][7], would place more new (or re-introduced but still higher perplexity) information at the beginnings of new discourse segments. In a study of a longer dialogue in English, Veilleux [12] described instabilities, i.e. general areas of disfluency (as well as the presence of shorter discourse segments) in regions between long, stable discourse segments. In contrast to the work cited above, she examined both sides of the discourse boundary, and found instability on both sides. She postulated that these regions were "bridges" between stable discourse segments, i.e. discourse regions where participants in a dialogue negotiate what new topic will follow.

These results suggest two questions:

1. Does nonfluent speech occur more frequently in discourse segment initial or final positions than elsewhere?
2. Do different kinds of disfluencies behave differently in this regard?

This work explores these two questions by examining two types of nonfluent speech (lexical and non-lexical) to determine whether the likelihood of a nonfluency changes across a discourse segment, and whether the distribution pattern varies for different types of nonfluency.

2. METHODOLOGY: DATA AND ANNOTATION

The work reported here makes use of professionally produced videotaped monologues that are nevertheless spontaneous and unrehearsed. The corpus consists of 10 lectures, each about 30 minutes long, from a longer series of lectures by a single professional lecturer. Printed transcripts of these monologues were available, and paragraph boundaries as labeled by the professional transcriber were used as an estimation of discourse segmentation. While using the judgment of a single transcriber does not provide a mechanism for distinguishing relative sizes of discourse segment boundaries (as in e.g. [10]), it does provide convenient segmentation of a large corpus of spontaneous speech. However, it is unlikely that another transcriber, even another professional, would agree exactly on any given segmentation.

Nonfluent events in this speech sample were labeled as either *disfluencies* or *errors*, as part of another study on speech production errors [7]. *Errors* included events in which a linguistic element (e.g. sound segment, morpheme, word) was produced (or omitted) in a location which could be interpreted as unintended by the speaker. These events were presumed to occur during the process of selecting and serially ordering the abstract lexico-grammatical elements for a planned message, i.e. during an intermediate span of the utterance planning process. Other *disfluencies* could be characterized as either more intentional (i.e. planned and intended events, such as filled pauses) or less intentional (i.e. unplanned and unintended events, such as slurring, stuttering or unidentifiable mispronunciation). These kinds of disfluent events were presumed to occur either early in the message formulation process (e.g. filled pauses at the discourse planning level, where computational resources might be challenged by the demands of a new or complex discourse segment), or at a later motor-control process (e.g. slurring at the stage of implementing a well-formed motor command). In contrast, lexico-grammatical errors were viewed as planned events subsequent to an unintended processing error, in which e.g. an element was mis- selected or mis-ordered. Both kinds of nonfluent speech, i.e. errors and disfluencies, were detected and labeled by a single trained labeler who assigned a label and time stamp in per second intervals.

These labels correspond generally to Shriberg's disfluency classes [9]. In particular, in both that work and in [10], nonfluencies that do not alter the orthography of a transcription of the words, such as a prolongation of *the* in *theeeeeeee king*, are not included. (In contrast, prosody annotation systems such as ToBI would use a disfluency diacritic such as a 2p, etc. for such prolongations.) [3]

Because we presume that *errors* occur largely as the result of mis-selection among similar grammatical elements and locations during the serial ordering process [7], we hypothesize that they will occur more randomly across a discourse than *disfluencies* which reflect a speaker's response to challenges to the discourse planning or motor implementation system. That is, disfluencies might cluster in regions of a discourse where computational loads are particularly high. In the following discussion, *nonfluent speech* will mean both of these types collectively. When the term *error* is used, it will refer to the lexical-unit errors described above. *Disfluencies* will refer to those disfluencies that do not involve lexico-grammatical units, although they may involve word-like filled pauses (*um, uh*); discourse-marking interjections (e.g. *well, now* produced as isolated IPs) were not included here as nonfluencies.

The 10 lectures included in this study had a total of 411 paragraphs. Paragraphs varied in duration from 2 to 101 seconds and were 41.8 seconds on average with a standard deviation of 18 seconds. Nonfluent speech was comparatively rare: 1.5 nonfluent labels/ paragraph on average. Over a quarter (28.5%) of the paragraphs had no labeled nonfluent speech events. Most paragraphs (257 or 62.5%) were labeled with one to three, inclusive, nonfluent labels and the remaining 9% of the paragraphs had 4 – 9 nonfluent labels. Shorter paragraphs have fewer nonfluent labels on average. Errors as defined here are half as frequent as disfluencies: in the 10 lectures, 201 errors were labeled compared to 418 disfluencies. (5 tokens were discarded from analysis due to labeled time inconsistencies.)

3. ANALYSIS AND RESULTS

The working hypothesis behind this study is that the distribution of disfluencies is not uniform across speech, but associated with discourse structure. In particular, nonfluent regions may occur preferentially in discourse-segment-initial or final regions. One can define what is meant by the

initial region of a discourse segment in several possible ways. i.e. in terms of a fixed time, in terms of initial prosodic or syntactic constituents, or as a proportion of the discourse segment. Swerts [11] examined events in the first Intonational Phrase (henceforth IP) of a discourse segment. While Watanabe [14] reported on disfluencies in all IPs, the proximity of disfluencies to all larger discourse boundaries was only considered in terms of the first IP immediately following those boundaries. However, Veilleux [12] found that the disfluent interval surrounding the discourse segment boundary was quite variable in dialogues and could spread over several intonational phrases.

The difficulty in analysis stems from the unknown mechanism that would cause a skewed distribution. For example, if longer discourse segments begin with a longer succession of clauses that each introduce more disfluency-provoking new information than found in shorter segments, or if longer segments represent a larger planning load, analyzing the frequency of disfluencies as a percentage of the time through the discourse segment would show how many disfluencies were early or late, relative to discourse segment size. On the other hand, if disfluencies occur immediately preceding or at the beginning of a discourse segment, regardless of its size or complexity, then the absolute time from the beginning or end of the segment would be the appropriate measure. Of course, if disfluencies occur within the first IP, which, though variable in length, is less variable than the length of the paragraph, then absolute time measures would also capture this.

Since the process behind the distribution is not yet known, three analyses are presented here: frequency of non-fluent speech events in the first or second IP of each paragraph; as a function of the absolute time from the beginning of a paragraph; and as a function of the percentage of time through a paragraph.

3.1. Disfluencies and Errors in the initial and second IP

The first and second full IP boundaries (break level 4) of each paragraph in 5 lectures of the corpus were labeled by an experienced ToBI labeler. In the 193 paragraphs in these lectures that contained nonfluent speech, only 7 nonfluent labels appeared in the first IP and only 6 occurred in the second IP, from a total of 260 nonfluent labels in these lectures. Thus, the occurrence of nonfluent labels in the initial few IPs of a new discourse segment is comparatively rare. These results differ from the Swerts' findings [10] but are similar to those of Watanabe [14] for a corpus similar to the one examined here.

3.2. Distribution of Errors and Disfluencies in absolute time from the beginning of the paragraph

Figure 1 shows that there is a peak in the occurrence of both disfluencies and errors approximately 15 seconds from the beginning of the paragraph. Since the majority of paragraphs (85%) are between 15 and 60 seconds (average: 41.8 s) long, one would naturally expect nonfluent labels to occur more frequently before this time. However, only 10% of paragraphs are 20 seconds or less, so the disfluency peak at 15 seconds does suggest that disfluencies (and to a lesser extent errors) occur earlier rather than later in a typical paragraph. However, the peak is not within the first 10 seconds, which at the typical rate of 3 words per second suggests a substantial delay from the discourse segment onset.

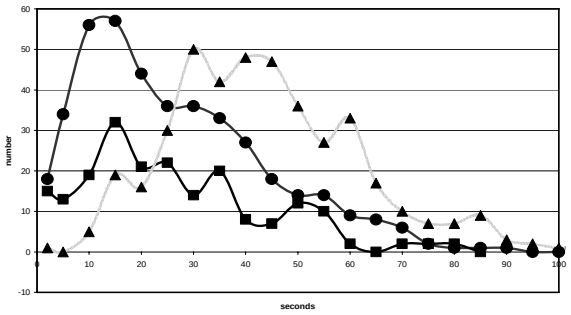

Figure 1: The number of disfluency (●) and error (■) tokens occurring at one-second intervals throughout the paragraph compared to the length of paragraphs in the corpus (▲).

This is consistent with the observation that nonfluencies are rare during the first or second IPs. What could explain a delayed peak in nonfluent events? One hypothesis is that, like the dialogue speakers in Veilleux [12], discourse segments do not appear in discrete sequences but rather speakers in a monologue also use a type of 'bridging' technique. They might make use of already-retrieved information, and so require completion of very little of the message-level planning for the new discourse segment. In support of this hypothesis, a cursory examination of the data found several backward references in discourse initial sentences (e.g. demonstrative

pronouns) that indicate that the speaker is not yet introducing new information. Disfluency and error rates may rise after this bridging period has expired, i.e. after the earlier-planned information is used.

3.3. Distribution of Errors and Disfluencies as a Percentage of Paragraph Length

One interpretation of 'initial' is with respect to the total length of the paragraph. In fact, if longer paragraphs involve more planning, then an initial period of greater nonfluency occurrences might be proportionately longer as well. Figure 2 shows the number of disfluencies and errors normalized for paragraph duration in seconds, accrued over 5% intervals of the total paragraph duration. There is a peak in the number of disfluencies at 15% and 20% (i.e. the interval greater than 10% and less than 20% of the total paragraph length in which the disfluency occurs). This peak, realized over a 10% slice, accounts for 17% of all disfluencies. While the numbers of tokens over the normalized paragraph intervals fluctuates, each of the other 5% interval slices carries roughly 4.6% of the tokens, as would be expected from an otherwise uniform distribution. In contrast, the fluctuations for errors show no significant peak.

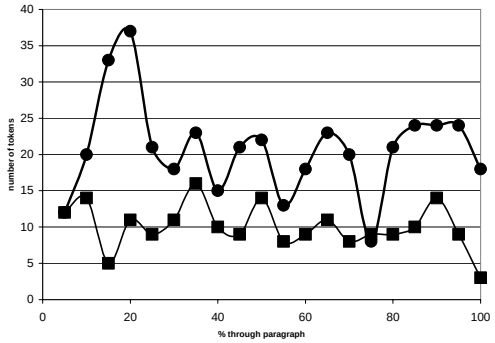

Figure 2: Location of disfluencies (●) and errors (■) as a percentage of time through a paragraph.

4. DISCUSSION

This study shows that nonfluent events in professional planned speech are not limited to the initial phrases of each discourse segment, but occur throughout the segment. However, when errors and disfluencies are analyzed separately, we find that disfluencies are more likely to occur in the region of the segment between 10 and 20% of its duration than one would expect, given a uniform distribution. Lexical-unit errors, on the other hand, appear to be roughly uniformly distributed throughout a discourse segment. These results suggest that speakers may pre-plan the earliest portions of a new segment, at least to some degree, so that computational resources are not strained during the production of the first part of the new structure. They are also consistent with the hypothesis that errors involving mis-selection or mis-ordering of lexico-grammatical elements occur at a different stage of speech production planning and execution than other nonfluencies do.

5. ACKNOWLEGEMENTS

This work was supported by the Simmons College Fund for Research, MIT's Undergraduate Research Opportunities Program, and NIH grants no. R01-DC002978 and R01-DC00075.

6. REFERENCES

[1] Arnold, J., Wasow, T., Ginstrom, R., Losongco, T. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76:1, 28–55.

[2] Arnold, J., Fagnano, M., Tanenhaus, M. 2003, Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research*, 32:11, 25-36.

[3] Beckman, M. E., Hirschberg, J., and Shattuck-Hufnagel, S. 2005. The original ToBI system and the evolution of the ToBI framework. In Jun, S.-A. (ed.) *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press, 9-54.

[4] Clark, H., Fox Tree, J. 2002. Using uh and um in spontaneous speaking. *Cognition* 84, 73 - 111.

[5] Grosz, B., Sidner, C. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12:3, 175-204.

[6] Grosz, B., Joshi, A., Weinstein, S. 1995. Centering: A Framework for Modelling the Local Coherence of Discourse. *Computational Linguistics* 21:2, 203-225.

[7] Grosz, B., Hirschberg, J. 1992. Some intonational characteristics of discourse structure. *International Conference on Speech and Language Processing*. Banff.

[8] Shattuck-Hufnagel, S. 1992. The role of word structure in segmental serial ordering, *Cognition* 42, 213-59

[9] Shriberg, E. 2001. To "Errrr" is Human: Ecology and Acoustics of Speech Disfluencies. *Journal of the International Phonetic Association* 31:1, 153-169.

[10] Swerts, M., Wichmann, A., Beun, R. J. 1996. Filled Pauses as Markers of Discourse Structure. Proc. ICSLP-96, vol. 2. Philadelphia, 1033-1036.

[11] Swerts M. 1997. Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America*, 101:1, 514-521.

[12] Veilleux, N. 2002. Bridges: regions between discourse segments. *Proc. ICSLP-2002*, Denver, 849-852.

[13] Watanabe, M., 2001. The distribution of fillers at discourse segment boundaries in academic monologues in Japanese. *Proc. of the 15th General Meeting of Phonetic Society of Japan*, Kobe, 85-90.

[14] Watanabe, M. 2002. Fillers as indicators of discourse segment boundaries in Japanese monologues. *Proc. Speech Prosody*. Aix-en-Provence. 691-694.