B L S 28 Berkeley Linguistics Society. 2002. http://dx.doi.org/10.3765/bls.v28i1.3845
Published by the Linguistic Society of America

Discourse-Related Effects on Speech Durations: A Challenge for Models of Speech Production[1]

CAROLINE L. SMITH
University of New Mexico

1. Discourse prosody

A substantial body of work has shown that prosodic characteristics of speech are affected by factors which can be conveniently referred to as "discourse-related". That is, the global organization of the spoken material has consequences for such properties as the pitch, durations and amplitude. Early research in this area focused on prosodic differences between boundaries that were perceived as smaller ("sentence") or larger ("paragraph"), and found differences in F0 and duration (Lehiste 1975, 1979; Kreiman 1982). Recent work has tended to concentrate on how the organization of the discourse is reflected in intonation (e.g., Grosz & Hirschberg 1992; Ayers 1994; Swerts & Geluykens 1994). Other studies have found evidence of an effect of discourse structure in several acoustic dimensions. For example, Herman (2000) compared the realization of the same English sentences produced either discourse-medially or discourse-finally, and found differences in F0, duration of the final pitch-accented syllable and RMS amplitude, even though she compared only those pairs of productions in which the intonational tones were phonologically identical, in order to ensure that any differences did not reflect distinctive variants of the sentence. In a study of direction-giving monologues in English, phrases at the beginning, middle and end of discourse segments were found to differ in speech rate, pause duration and several measures of F0 (Hirschberg & Nakatani 1996).

The studies just described looked at discourse structure by comparing phrases or sentences at different positions in a discourse. Other studies have looked at correlates of prosodic characteristics with different aspects of discourse structure. Studying spontaneous re-tellings of a short Dutch narrative, van Donzel (1999) found that both the magnitude of discourse boundaries and the information structure of the discourse contributed to determining intonational realizations and pause durations. Also in Dutch, Noordman *et al.* (1999) and den Ouden *et al.*

[1] This work was supported by NSF grant BCS-9983106. Many thanks to Lisa Hogan for analyzing the data and assisting with manuscript preparation.

(2000) studied speakers' reading of narrative texts, and found that F0 means, maxima and pause durations were affected by the hierarchical structure of the narrative as determined by theories of discourse structure.

The various studies cited above used different criteria for characterizing the status of an individual phrase or sentence in the larger structure of the discourse. A slightly different approach was used in the present experiment. Rather than attempting to analyze the overall organization of a discourse, the analysis focused on the relation between the topic of one sentence and of the sentence which follows. The scheme used by Nakajima and Allen (1993) for labeling topic boundaries in spontaneous instructional monologues was adapted to the needs of this study, which was based on a written instructional text. This topic-labeling scheme is adequate for describing the organization of instructional material, but is not intended to be a general theory of discourse organization. Its relative simplicity was a virtue for this study in part because it may ultimately be easier to generate an analysis of this type automatically. This would enable speech synthesis systems to analyze the topical organization of a text, then take it into account when determining what prosody to produce.

With the topic labeling providing an analysis of the text's organization, the experiment reported here investigated the relation between the structure of the text and the durational patterns produced when it was read aloud. Durations were chosen for study in part because much previous work has concentrated on F0. In addition, acoustic durations in English are known to be affected by segment identity and local context in a way that F0 is not, so it is interesting to see whether, like F0, durations are also affected by the overall organization of the discourse. The present paper focuses on the interactions between discourse organization and durations, and how these may reveal drawbacks in current models of speech production.

2. The Experiment

This experiment measured the effects of different types of topic transition on several durational properties (more details are reported in Smith and Hogan (2001)). In this experiment a male speaker of American English read a passage drawn from the manual for the computer drawing program Canvas (Deneba Systems 1997). The text was chosen because it offered a relatively well-defined topic structure for the analysis; moreover, such a text typifies the sort of material that speech synthesizers currently read aloud in, for example, Help systems.

The speaker was recorded reading the text aloud along with a set of "control" sentences ten times at intervals of approximately one week. The control sentences were constructed so that the final word of each sentence in the Canvas text—the "target" word—occurred in a sentence-medial position in the control sentence. For example, one sentence ending with *box* in the original text was as follows [italics not in original text]:

(1) You can search and replace character strings that you specify using the Find dialog *box*. (Deneba 1997:224)

The corresponding control sentence for the target word *box* is given in (2). In all control sentences, the target word was placed in the sentence so that three syllables preceded it and eight syllables followed it:

(2) The dialog *box* lists all settings in the program.

The Canvas text passage encompassed 60 sentences which ended in 38 different target words. Three durational measures were made on each sèntence in the text:

- Sentence-final lengthening, measured as the increase in the duration of the target word when it occurred in sentence-final position in the text as compared to its duration sentence-medially in the control sentence
- Speech rate, measured as the number of syllables per second in the interpausal speech runs at the end of each sentence and the beginning of the following sentence
- Pause durations, measured as the length of time that elapsed between sentences.

In the analysis of the relation between these durational properties and the organization of topics in the text, the transition from one sentence of the text to the next was classified into one of four categories:

- ***Topic Shift***, if the following sentence introduced new material
- ***Topic Continuation***, if the following sentence continued the topic, advancing the narrative
- ***Elaboration***, if the following sentence provided additional detail about the preceding sentence
- ***Text Marker***, if the following sentence were an overt indicator of textual organization (an example of a Text Marker is *Note:*).

The classification was done by five linguists; in cases of disagreement, the labeling preferred by the majority was used. Below is a brief extract from the text (Deneba 1997:226) with the topic transitions labeled and the target words in italics:

(3) ...Otherwise, choosing Interactive will turn this feature *on*. (Topic Shift)

Spell checking *text* (Topic Continuation)
You can check the spelling of highlighted blocks of text, a selected text objects, or an entire *document*. ...

The analysis of the durational measures for each type of topic transition yielded these results:

- Topic Shifts occurred with significantly longer pauses than other types of transitions. Speech rate tended to be slower at a Topic Shift than at other transitions, but did not change at the transition from one sentence and the next.
- Sentence-final lengthening was similar in Topic Shifts, Continuations and Text Markers. However, at a Topic Continuation, speech rate increased significantly between the end of the first sentence and the beginning of the following sentence.
- Topic Elaborations had significantly less final lengthening than other transitions. Pauses were of similar duration for Elaborations, Continuations and Text Markers. Speech rate was faster for Elaborations and Text Markers, but did not change at the transition.

All three measures of duration were affected by the type of topic transition from one sentence to the next. Additionally, different types of transition were associated with different configurations of the three durational measures. Although these results apply to one speaker only (analysis of additional speakers is underway), the effects were substantial and statistically robust. For example, the amount of sentence-final lengthening at Topic Shifts was over twice what occurred at Elaborations. Therefore, these patterns ought to be incorporated into any model representing the factors which potentially contribute to the process of speech production.

But the explanation of these effects must be more complex than the usual explanations of segmental context and phrasal organization—the relationship between the topic of one sentence to the next causes differences in the durational properties of the end of the first sentence, the beginning of the second sentence, and the pause between them. The temporal extent of these differences means that the transition between sentences is not completely localized at the boundary between the sentences. This finding presents problems for models of speech production.

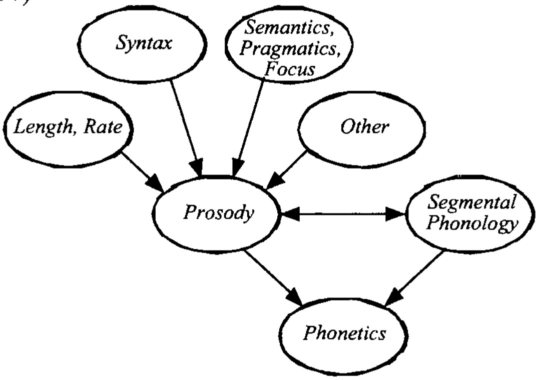
3. Models of speech production

One of the difficulties with the findings reported in section 2 is that the effects of topic organization relate to a subsequent sentence as well as the current one, so information outside the current prosodic domain must be accessed. In addition, these effects relate to the utterances' content and meaning, not their structure, and content is not usually considered to be the kind of information that is accessible to the prosody. The remainder of this paper considers the problems that these considerations raise for incorporating the results of this study into two representative models of speech production. The two models are based on extremely different representations of linguistic knowledge. One, referred to here as the "modular" model, assumes that grammar is divided into separate modules, each of which has access to only a subset of linguistic structure. Processing by different modules brings together the information needed to produce speech. The other model, referred to here as the "exemplar" model, does not differentiate among different aspects of linguistic structure. All the information needed to produce speech is combined in the representation, but complex processing is required to access it.

3.1 Conventional "modular" model

Current theories about the organization of speech production most often assume a model somewhat like that shown in (4), which is adapted from Shattuck-Hufnagel and Turk (1996). In this model, a variety of grammatical and extra-grammatical factors contribute to the determination of the prosody, but the model permits only the prosody and the segmental phonology to affect the phonetics (which presumably includes determination of durations).

(4) "Modular" model, adapted from Figure 5 in Shattuck-Hufnagel and Turk (1996:237)



The problem with this model is that topic organization would be able to affect durations only via the prosody. In other words, different types of topic transitions would have to be associated with differences in the prosodic structure which in turn would create differences in durations. The question becomes what prosodic unit(s) delimit the boundaries at which the topic transitions occur. The version of the prosodic hierarchy discussed by Shattuck-Hufnagel and Turk (1996) does not include any units larger than the intonational phrase, but many of the sentences in this experiment included more than one intonational phrase (IP). This is a problem if the IP is the largest unit available to associate with the boundaries at the transitions. However, if the Strict Layer Hypothesis must be modified so that

IPs can be nested within larger IPs, as Ladd (1996:244) argues, then a prosodic structure could be constructed so that each written sentence is coextensive with a (possibly nested) IP. This is one approach to the problem of relating sentences in a written text to prosodic units in speech. The assumption that topic transitions occur at sentence boundaries is appropriate in written text because a sentence contains a complete proposition. However, nesting IPs in order to make sentence boundaries coincide with IP boundaries would not, in itself, predict the observed differences at different types of transitions.

Two ways in which IPs might form the basis for an alignment of topic organization with prosodic structure are as follows: (i) IPs could be categorized in some way according to the type of topic relation that holds between adjacent phrases; or (ii) they could be organized into larger, superordinate units, and their position within the larger structure would correlate with topic organization. Both of these possible solutions attempt to introduce a reflex of structure in the semantic/discourse domain into prosodic structure. Proposal (i) would violate basic assumptions of the hierarchical model of prosody by distinguishing among sister units solely on the basis of explicit labels, rather than distinguishing them by their structure. Identifying stress feet as strong or weak might seem to be an accepted usage of labels on prosodic units, but this distinction is actually made on the basis of the position and/or syllabic content of each foot. The labels on stress feet are for convenience, whereas the proposed labels on IPs designating their transition types would encode crucial differences.

Although Shattuck-Hufnagel and Turk (1996) do not mention it, the phonological utterance is the top level of the prosodic hierarchy in many versions of prosodic theory. Under the right circumstances, an utterance can include more than one sentence, so proposal (ii) might seem to pose little difficulty. Sentences which are closely related could be connected into a single phonological utterance, which would predict minimal marking of the transition between them. A less closely-related pair of sentences could be in separate utterances, in essence placing a larger boundary between them, which predicts greater marking of the transition. But even this proposal, in the spirit of the prosodic hierarchy, does not capture the complexity observed in the durational marking of transitions between sentences.

First of all, even in the very simple transition-labeling scheme used in this study, there were four types of transition, each with its own durational characteristics. In order to distinguish among four types, it would be necessary to posit more different structures than just the distinction between a sentence which ended at an utterance boundary and one not at an utterance boundary. A four-way distinction would require expanding the prosodic hierarchy. More significantly, statistical tests showed that the various durational measures were very rarely correlated. Each appears to vary independently of the others so they do not combine to make a consistently “bigger” boundary at a Topic Shift, for example, than at a Topic Continuation. Rather, these are different types of boundary. For this reason, the topic effects could not be handled by treating Topic Shifts as the delimiter of a prosodic unit “topic”, with the other transition types delimiting smaller units included within a “topic”. Such a structure would fail to represent the differences among Topic Continuations, Elaborations and Text Markers. A hierarchical prosodic structure offers no tidy way to represent this.

I conclude that in order to represent the types of effects observed in the present study, the “modular” model would have to be significantly altered, and weakened. Either some kind of provision would have to be added for labeling prosodic units according to their semantic/pragmatic/discourse function, or the organization of the model would have to be restructured so that these components have direct access to the phonetics, nullifying the crucial role of the prosody in mediating the many factors that have been shown to play a role in production.

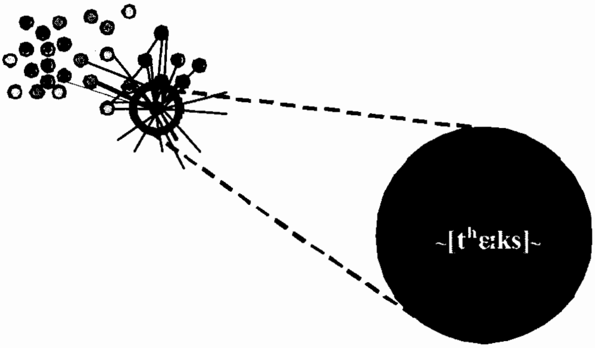
3.2 Exemplar models

The “modular” models described in section 3.1 compartmentalize different aspects of linguistic structure; the model’s predictive power comes, in part, from constraints that the structure imposes on how each component can reference properties under the control of another component. Exemplar models take an alternative view, in which all aspects of linguistic structure have the potential to interact. Exemplar models are somewhat akin to the connectionist models used in psychology to model neural processes (Rumelhart *et al.* 1986). In the linguistic literature, exemplar models have been discussed chiefly as an account of speech perception (Johnson 1997). Bybee (2001) and Pierrehumbert (2001, in press) may be the first to explore how these models can represent speech production.

Different authors have proposed variants of exemplar models for language processing. The description here is a somewhat simplified account based principally on Bybee (2001). The core idea of an exemplar model applied to language is that language users have knowledge of specific instances of linguistic units recorded in memory. These specific instances are the exemplars. Different versions of the theory have different proposals about the size of the unit(s) that are stored; here, for simplicity, I will assume that the stored unit is the word. In this case, each exemplar records a particular pronunciation and usage of a word. In (5), the large circle represents one exemplar of the word *text*, depicting a pronunciation in which the vowel is lengthened but the final [t] is not pronounced. (In this and the next figure, a phonetic transcription is enclosed in square brackets; the ~ ~ notation is being used to symbolize an exemplar, which includes information about meaning and context, not just the pronunciation.)

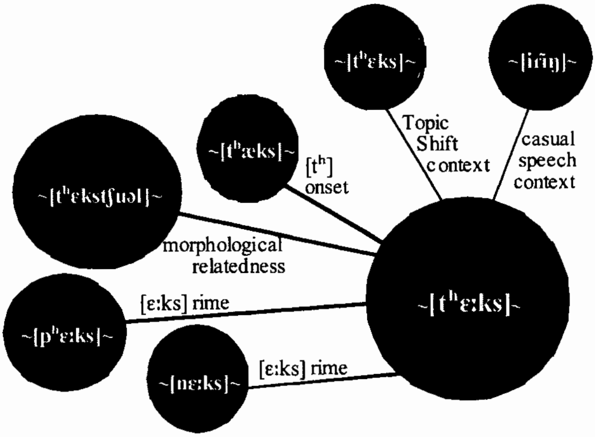
An exemplar is activated each time it is used, with activation decaying over time. Recently-used exemplars are therefore more activated than those not used for some time, and more frequently-used exemplars are more activated than those used less often. Stronger activation is represented in (5) as darker shading.

(5) A group of exemplars, showing some of the connections to one exemplar of the word *text* pronounced with a lengthened vowel



Each exemplar is linked to other exemplars with which it shares some property. The shared property may relate to any aspect of the exemplars—phonological, semantic, contextual, to name a few. A subset of the links to one exemplar is shown in (5) and expanded in (6). This exemplar, with its lengthened vowel, would be linked not only to other exemplars of the same word, but to exemplars of phonologically, morphologically or semantically related words. It would also be linked to other exemplars that occurred in a similar context. For example, if this exemplar was in a Topic Shift context, it would be linked to other exemplars that occurred at Topic Shifts.

(6) Some examples of shared properties which could result in connections being formed among exemplars of the same and different words.



This linking of words occurring in similar contexts leads to a possible scenario for an exemplar model account of the patterns associated with topic transition types. If a word occurs with a lengthened vowel at the end of a sentence preceding a Topic Shift, then other exemplars that are linked to it by virtue of having similarly lengthened vowels, will also be activated. If Topic Shifts and lengthened vowels co-occur again and again, the pattern of activation is reinforced, so that eventually the language user forms a "schema" among exemplars with lengthened vowels occurring in Topic Shift context. According to Bybee (2001:39), schemas are "emergent generalizations over complex representations." That is, the representation stores a great deal of information about the exemplars that the language user has experienced, and the patterns of activation in the links among these exemplars result in the emergence of generalizations about the sets of exemplars connected by these links. These generalizations can then result in the associated properties spreading to new exemplars. Variable patterns such as those due to topic organization can emerge from the process of producing speech, even if they do not always occur, or occur to varying degrees. Variability in the input will be reflected by variability in the output. Both are expected.

As described here, exemplar models are very flexible; essentially any property that is shared by a group of exemplars has the potential to generalize as a schema, if it occurs with sufficient frequency. This is both the strength and weakness of these models. Their flexibility means that they can account for all kinds of patterns of co-occurrence among different kinds of linguistic properties. At the same time, without limits on possible schemas, the model makes no predictions about which are more likely, beyond the fact that more frequently occurring patterns are more strongly activated and more likely to generalize.

Exemplar models predict that generalizations about a specific context will originate with words which the language user has experienced in that context. Only after repeated usage of words associating a property and context can a pattern be generalized as a schema. This process seems like an unlikely path for the generalization of effects such as those associated with topic transitions. These effects depend on the relation between two entire sentences, not specific words, and are very unlikely to occur more than once with the same words. Without repetition, they would not be able to generalize, since the connection between a property and a context is associated with specific exemplars. It is not clear how a schema could develop from a pattern that is not tied to specific linguistic units. The generalization would have to come from repeated co-occurrence of activation in two sets of connections: those connecting exemplars sharing particular topic contexts and those sharing the associated durational properties. Modification or extension of the model would be required in order to allow a schema to develop without repetition of individual exemplars exhibiting a pattern.

The flexibility of the exemplar model means that it could incorporate these effects probably with less modification than the modular model, but this same flexibility means that it could potentially generalize other patterns which co-occur accidentally. The generality of this model is both its strength and its weakness.

4. Summary

In recordings of one American English speaker reading an instructional text aloud, durational differences were observed which appear to depend on the relation between the topics of consecutive sentences. These differences, along with the findings of other similar research, pose a problem for current models of speech production, because they suggest that the structure of the semantic/pragmatic information in a discourse can have measurable effects on the acoustic durations. While it has long been known that durations are influenced by many factors (see, for example, Klatt 1976), neither of the models of speech production as discussed here can immediately incorporate the effects of topic organization. The ideal model would be more constrained than an exemplar model, but more flexible in allowing interaction among different components of the grammar than the modular framework assumed in much current research. Such a model, which remains to be developed, would be valuable in helping to understand how the physical dimensions of speech reflect the different linguistic dimensions of the message.

References

Ayers, Gayle. 1994. Discourse functions of pitch range in spontaneous and read speech. *OSU Working Papers in Linguistics* **44:** 1-49.

Bybee, Joan. 2001. *Phonology and Language Use*. Cambridge: CUP.

Den Ouden, Hanny, Leo Noordman, & Jacques Terken. 2000. Prosodic correlates of text structure. *Proceedings of the 10th Annual Meeting of the Society for Text and Discourse*, 40-41.

Deneba Systems. 1997. *Canvas 5 User's Guide*. Miami: Deneba Systems.

Grosz, Barbara & Julia Hirschberg. 1992. Some intonational characteristics of discourse structure. *Proceedings of the 2nd ICSLP*, Banff, Alberta, 429-432.

Herman, Rebecca. 2000. Phonetic markers of global discourse structures in English. *Journal of Phonetics* **28:** 466-493.

Hirschberg, Julia & Christine Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, 286-293.

Johnson, Keith. 1997. Speech perception without speaker normalization. In K. Johnson & J. Mullennix (eds.) *Talker Variability in Speech Processing*, 145-165. San Diego: Academic Press.

Klatt, Dennis H. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America* **59**: 1208-1221.

Kreiman, Jody. 1982. Perception of sentence and paragraph boundaries in natural conversation. *Journal of Phonetics* **10:** 163-175.

Ladd, D. Robert. 1996. *Intonational Phonology*. Cambridge: CUP.

Lehiste, Ilse. 1975. The phonetic structure of paragraphs. In A. Cohen & S. G. Nooteboom (eds.) *Structure and Process in Speech Perception*. Proceedings of the Symposium on Dynamic Aspects of Speech Perception, 195-203. New York: Springer-Verlag.

Lehiste, Ilse. 1979. Perception of sentence and paragraph boundaries. In B. Lindblom & S. Ohman (eds.) *Frontiers of Speech Communication Research*, 191-201. London: Academic Press.

Nakajima, Shin'ya & James Allen. 1993. A study on prosody and discourse structure in cooperative dialogues. *Phonetica* **50:** 197-210.

Noordman, Leo, Ingrid Dassen, Marc Swerts, & Jacques Terken. 1999. Prosodic markers of text structure. In K. van Hoek, A. Kibrik, & L. Noordman (eds.) *Discourse Studies in Cognitive Linguistics*. Selected Papers from the 5th International Cognitive Linguistics Conference, 133-148. Amsterdam: John Benjamins.

Pierrehumbert, Janet. 2001. Exemplar dynamics: word frequency, lenition and contrast. In J. Bybee & P. Hopper (eds.) *Frequency Effects and Emergent Grammar*, 137-157. Amsterdam: John Benjamins.

Pierrehumbert, Janet. In press. Word-specific phonetics. In C. Gussenhoven & N. Warner (eds.) *Laboratory Phonology VII*. Berlin: Mouton de Gruyter. [from http://www.ling.nwu.edu/~jbp/publications.html]

Rumelhart, David, James McClelland & the PDP Research Group. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press.

Shattuck-Hufnagel, Stephanie & Alice Turk. 1996. A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research* **25:** 193-247.

Smith, Caroline & Lisa Hogan. 2001. Variation in final lengthening as a function of topic structure. *Proceedings of Eurospeech 2001 Scandinavia*, 955-958.

Swerts, Marc & Ronald Geluykens. 1994. Prosody as a marker of information flow in spoken discourse. *Language and Speech* **37:** 21-43.

Van Donzel, Monique. 1999. *Prosodic Aspects of Information Structure in Discourse*. The Hague: Thesus.

Caroline L. Smith
Department of Linguistics
University of New Mexico
Humanities 526
Albuquerque, NM 87131-1196

caroline@unm.edu