Fundamental Frequency and Other Prosodic Cues to Topic Structure

Margaret Zellers and Brechtje Post

mkz21@cam.ac.uk, bmbp2@cam.ac.uk

Research Centre for English and Applied Linguistics, University of Cambridge

Abstract

Studies of the relationship between prosody and the topic structure of longer discourses have tended to focus on intonational cues, most specifically fundamental frequency (F0). However, this may limit the kinds of topic structure it is possible to identify. A case study of two speakers shows that including speech rate along with F0 factors in an analysis of the prosody of topic structure can be valuable in shaping our understanding of the topic structure of spoken discourses. Furthermore, it demonstrates that F0 cues can alternate not only with each other, but also with other prosodic cues, to effectively signal topic structure in discourse.

I. Introduction

Prosodic studies have generally focused on structure and signals at the level of the individual sentence or utterance, yet it is becoming increasingly apparent that prosody also provides valuable cues for signaling discourse structure in units larger than single utterances. Prosody is used as a way of helping to maintain coherent discourses by indicating, for example, the newness of referents in utterances; it can also be used on the level of whole utterances to signal how those utterances relate to one another in terms of their content, and specifically their discourse topic.

By discourse topic, we here refer to the shared 'aboutness' of a group of utterances; the utterances may be 'about' a referent, a proposition, or some other entity. This is related to, but not necessarily the same as, topic as contrasted with the comment or the focus in a given sentence. Büring (1999) describes D(iscourse)-Topic as constraining the directions which a conversation might take; in his analysis, the D-Topic is relevant to but not the same as the possible values for a S(entence)-Topic, which contrasts with focused or background information in an individual utterance. In his analysis, therefore, the discourse topic can be seen as an active element in the semantic structure of the discourse, helping to shape the discourse by constraining the directions a following utterance may take. All further uses of the term 'topic' in this paper will refer to something like Büring's D-Topic: something that is held in common by multiple utterances within a discourse.

A number of prosodic studies on topic have used the notion of a functional discourse topic, as suggested by Büring, in their analyses. In particular, Nakajima and Allen (1993) and Wichmann (2000) have described prosodic, and particularly intonational, variation in their data on the basis of four topic structure categories, which have to do with the relative newness of the information in different utterances and the semantic relationships between them. In their theories, the relationship to an existing topic, which is normally presented explicitly in the discourse, is the key element in the organization of discourse, and thus in the description of prosodic variation cueing that organization.

However, this is not the only way of seeing discourse topic, and indeed there are some theories of discourse structure which discount the existence of topic as a functional element altogether. Blakemore (1992, 2002) argues that the idea of topic is nothing more than an artifact of the way in which sentences or utterances are connected to each other, with more

closely related sentences appearing in closer physical proximity, and less closely related sentences occurring farther away. From this point of view, discourse might be better described in terms of a simple hierarchy, as presented by Grosz and Sidner (1986). In a hierarchical theory, the relationships between utterances are dominance and subordination relationships. This is not completely in conflict with a categorical view such as that of Nakajima and Allen (1993) or Wichmann (2000), since the categories given by these latter studies also appear in a hierarchy with one another (that is, new topics always dominate the other categories from a hierarchical standpoint). However, other than the basic requirement of coherence that there be identifiable links (or potentially, identifiable disconnects[1]) between subsequent utterances, the semantic content of the utterances and the way in which the informational content relates is not relevant in a hierarchical theory in the same way it is relevant in a categorical theory. In a categorical theory, the relationships are defined on the basis of the informational content; that is, a given utterance can add separate information to a new topic, or it can add more detail on the content of a previous utterance. Both of these relationships, however, could be subsumed under a subordination relationship in a hierarchical system. The different ways in which the new information is added, or perhaps the semantic distance between the two utterances, is less important to the organization of the discourse than the fact that both of the new utterances are subordinate to the new topic.

(1) (New topic) Suddenly there was a huge commotion.

Animals flooded the camp. The captain ordered everyone to remain calm.

In (1) above, a new topic utterance is followed by two further related utterances. A categorical approach would likely identify the first following utterance (*Animals…*) as giving more detail about the immediately preceding utterance: that is, one specific aspect of the commotion is the presence of many animals. The second following utterance (*The captain…*), on the other hand, would be better identified as the addition of more information about the situation introduced in the topic, although less specifically tied to it. A categorical approach would thus distinguish between these two utterances. A hierarchical approach, on the other hand, would likely identify both of the following utterances as being immediately subordinate to the topic utterance. In other words, they are on the same 'level' of the hierarchy, regardless of the way in which they contribute information to the overall topic at hand. The category approach would therefore predict that the prosodic signals associated with the two following utterances should be different, while a hierarchical approach would predict that the prosodic signalling would be the same (in both cases allowing for an interaction with the sequential order in which the utterances are presented).

Given these two different characterizations of topic structure, the search for prosodic cues to topic structure is difficult to separate from the search for an appropriate theory of topic structure. Studies that have looked at this problem have often either simplified it into the dichotomy of new-topic versus non-new-topic, or they have divided their data into categories which, while reflecting their own data well, may be less applicable in other contexts. These categories in turn may reflect not true variations in the topic structure of utterances, but rather simply variations in one or more prosodic features as evidenced in the data at hand. However,

[1] It is important to note that while Wichmann (2000) in particular describes her topic structure categories as having to do with the relative newness of information contained within the utterances, it has also been suggested that topic structure is to do with the amount of disconnect of utterances from the preceding context (cf. Brazil 1997).

by adding multiple prosodic correlates to an analysis, it may become possible to make more accurate groupings into categories. This was recognized by, for example, Nakajima and Allen (1993), who used several fundamental frequency characteristics to identify four levels of topic structure in their spontaneous speech data. In their data, different height 'settings' for initial highs and final lows, as well as the ratio between consecutive peaks in different utterances, combined to create distinct identifying prosodic patterns for the four topic structure categories. However, no one of these features on its own was sufficient to differentiate all four groups.

In order to gain a better understanding of the prosodic correlates of topic structure, Zellers and colleagues (Zellers 2009; Zellers et al. 2009) conducted a production study on speakers of Standard Southern British English (SSBE). Participants in the study read aloud a written text which had been controlled for segmental factors such as segmental structure of the target word, presence or absence of anacrusis, and position of the target item in a group of utterances, as well as having utterances which were in principle easily classified into four topic structure categories similar to those posited by Nakajima and Allen (1993) and Wichmann (2000). The topic structure categories were used to guide the construction of the original text, and were defined as follows:

(2) *Topic:* the beginning of a new topic
Addition: new information on the same topic
Elaboration: more detail or clarification of a previous utterance
Continuation: completing the speech act begun in the previous utterance

The design of the text meant that different productions of the same target syllable (or in many cases the complete word) by the same speaker in different segmental and topic-related contexts could be compared, in order to gain a better understanding of which kinds of prosodic variation were specifically related to the topic structure versus varying incidentally on the basis of random segmental factors. This was a difficulty encountered by previous studies, which often used spontaneous or semi-spontaneous texts in which it would have been nearly impossible to compare identical lexical items in similar contexts.

Zellers (2009) found that there was a correlation in her data between the size of F0 falls (that is, the distance in semitones between the H peak and a following L valley in a falling pitch accent) and the topic structure categories used, although it was unclear whether the exact categories used were accurate, since it appeared that two of the middle categories did not in fact differ significantly from one another, at least in this measure (see Fig. 1 below).

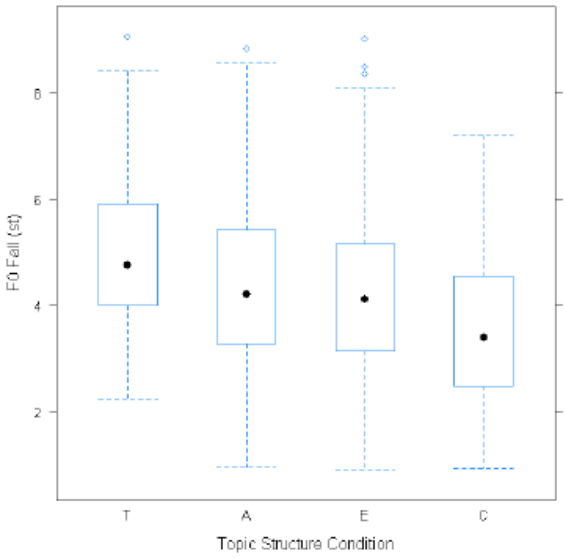

Figure 1: F0 fall range by topic structure category (from Zellers 2009)

However, height of the peak from the speaker's baseline, which had previously been found to correlate with topic structure (e.g. by Wichmann 2000), did not show such a correlation in this data, instead patterning only with the position of an utterance in a group of utterances (beginning, middle, or end). Zellers et al. (2009) found furthermore that F0 peak timing, which had been previously posited to be a prosodic correlate of topic structure, was an unstable cue at best in this data, since it was highly dependent upon the type of phonological theory chosen to describe the pitch accents in the data. Wichmann (2000) had shown that in her data, the F0 peak was delayed relative to the segmental structure of the utterance when a speaker was introducing a new topic. However, Wichmann's study was not able to compare the F0 peak timing across multiple instances of the same word or phrase (that is, the same segmental structure), and it also assumed that all the F0 peaks that she measured fell into the same phonological category. Zellers et al. (2009) found that when the segmental structure was held constant by comparing different instances of the same word, the peak delay effect disappeared almost completely, remaining in only one highly specific segmental context: target words without a consonant onset or an anacrusis. However, we found in contrast that if the pitch accents were divided into two categories, a fall and a rise-fall (following Gussenhoven 2004), an effect of the distribution of the two pitch accents could be seen across the four topic structure categories proposed. Rising-falling pitch accents were most likely in new Topics, and decreasingly so across the other topic structure categories, in an order consistent with the category order in Zellers (2009). We therefore concluded that it is highly likely that peak timing is a cue to the information status of the accented element, which is a factor related to but not equivalent with topic structure. These two studies thus suggest that some of the most popularly recognized intonational cues to topic structure may in fact not be signaling topic structure at all. This leaves us with two possibilities which must be considered. First, prosody may not in fact be in use as a signal of the topic structure of discourses. This could be either because some other, currently unknown, signal is in use, or because the topic structure of spoken discourses is not necessary or useful for discourse understanding. Second, and probably more likely given the studies presented above, it may

be necessary to look beyond the realm of F0 modulation to find prosodic cues to topic structure, for example into pausing, rhythm, or voice quality.

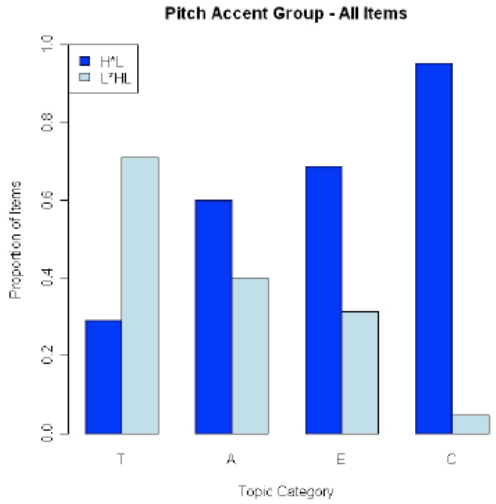

Figure 2: Distribution of pitch accents across topic structure categories (from Zellers et al. 2009)

The current paper addresses the latter possibility by presenting two representative case studies taken from the SSBE production study mentioned above. The two speakers were chosen as being representative of two patterns which emerged in the data. There were sixteen speakers included in the analysis. Speaker F05 is representative of nine of these speakers; F04 is representative of five. The two remaining speakers did not seem to fit either pattern, but were the two speakers with the fewest data points, so it is perhaps not surprising that they did not show a clear tendency to one or the other of these patterns of behavior. Alternatively, they may show variation on other parameters which were not included in this study. Section 2 compares prosodic variation in F0 fall range on the one hand, and speech rate on the other, in two speakers from the study. Section 3 discusses some of the implications of this variation for our understanding of topic structure and its correlates in spoken language.

2. Two case studies

Although speakers of the same language must by definition be (generally, at least) mutually intelligible, this does not translate to a requirement that they speak in an identical fashion. Individual variation among speakers of the same dialect is to be expected, even as we see clear overall trends among a group of language speakers. Therefore, although the distribution of a prosodic cue, in this case F0 fall range, may appear as in Figure 1 for a whole group of speakers, it is not surprising to find within this distribution the two very different patterns found in Figure 3a and Figure 4a for individual speakers.

In Figure 3a, we see F0 fall range data for speaker F05, whose speech production pattern basically matches the overall pattern (see Fig. 1) for this data. New Topics have the largest F0 falls, and Continuations the smallest, while Additions and Elaborations fall somewhere in the middle and are difficult to distinguish from one another (ANOVA: $F(3, 41) = 2.602$, $p \approx 0.05$). In Figure 4a (overleaf) the F0 fall range data for speaker F04 is presented. In this case we see a very different picture. Instead of the stepping-down pattern in Figure 1, the

Topic, Addition and Elaboration categories appear to all have more or less the same size fall range, while Continuations are the only category to vary noticeably from the mean of the other groups, having a more compressed fall range ($F(3, 31) = 6.988$, $p<0.01$).

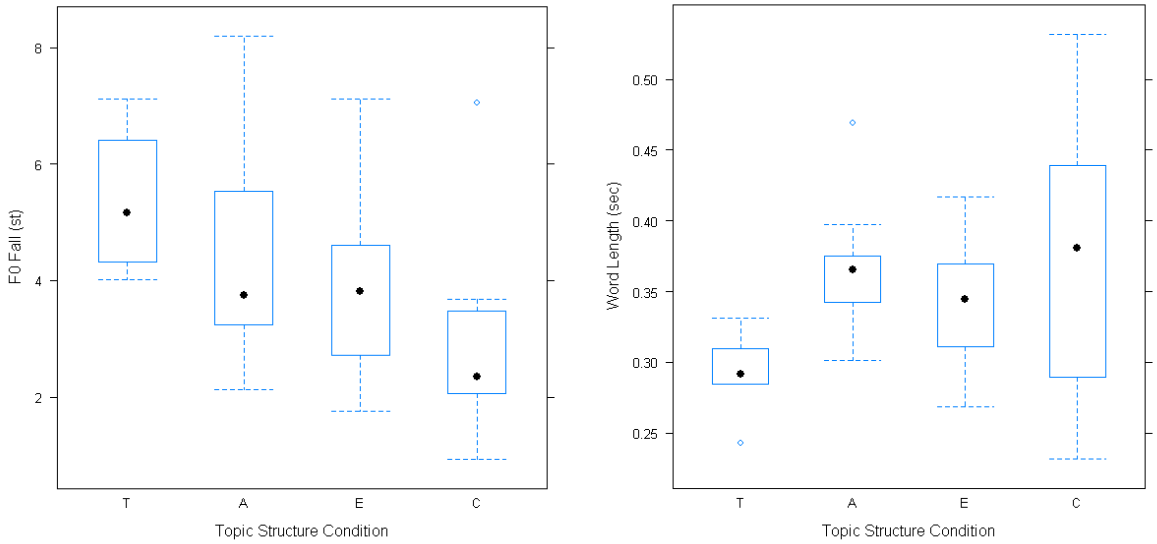

Figure 3: a. F0 fall range (semitones) by topic structure, and b. Speech rate (word length) by topic structure for speaker F05.

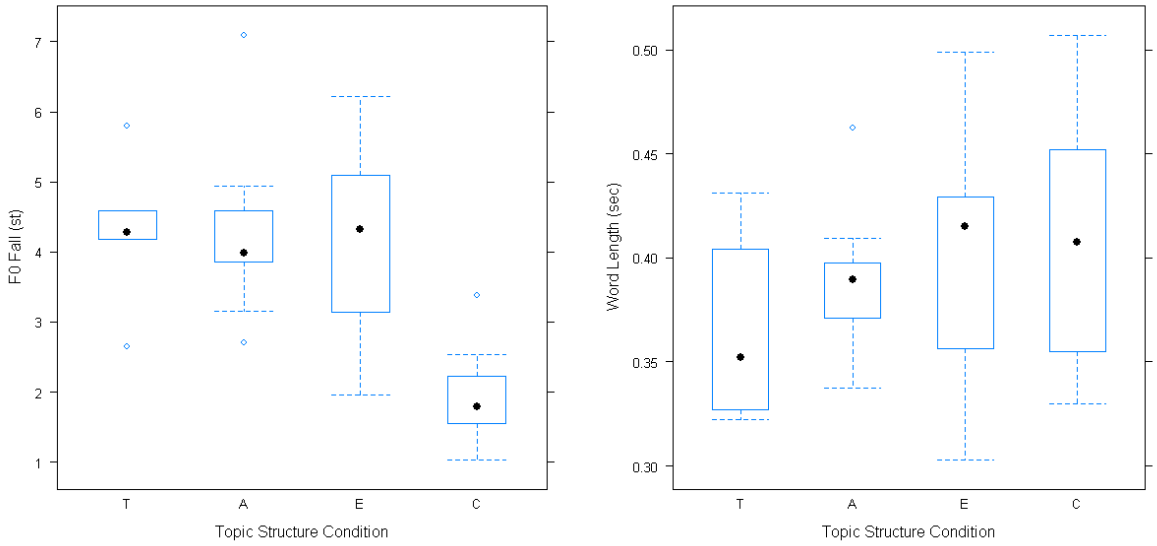

Figure 4: a. F0 fall range (semitones) by topic structure, and b. Speech rate (word length) by topic structure for speaker F04.

If we looked only at the data for speaker F04 (and for the other four speakers in the dataset who pattern with this speaker), we might be led to expect that there are only two topic structure categories, perhaps representing categories involving any kind of new or unpredictable information as opposed to categories with very high predictability. To an extent, this is consistent with Wichmann's (2000) expectation of the content of topic structure categories, but it is much less specific and less detailed than her predictions. Even in a very simple topic structure theory, the key distinction is between utterances beginning new topics

and those not beginning new topics, yet this speaker appears not to make such a distinction. However, the two speakers are speakers of the same dialect of SSBE and read the same text. Even if, as seems likely, there were some small variations in how the individual speakers interpreted the topic structure of the text, these drastic differences are striking

This production behavior becomes less odd when we add other prosodic features into the account of the data. In figure 3b, we see speech rate data for speaker F05, and in figure 4b for speaker F04. For speaker F05, who used F0 fall range in the same way as the overall trend in the data, there is very little meaningful variation in the speech rate, although new Topics are spoken marginally more quickly than the other categories, strengthening this cue ($F(3, 43) = 2.66$, $p \approx 0.06$). Interestingly, for this speaker, Additions and Elaborations may also be differentiable on the basis of speech rate, although this pattern did not hold for all speakers. For speaker F04, however, who made relatively little use of the F0 fall range, the speech rate data are striking. New Topics are spoken at a faster rate than Additions, which in turn are spoken more quickly than Elaborations and Continuations ($F(3, 37)=2.84$, $p \approx 0.05$). The differences between Topics, Additions and Elaborations/Continuations in the speech rate combined with the significant differences between Topics/Additions/Elaborations and Continuations in the F0 fall range mean that by using these two cues in combination it is possible to distinguish between all four categories of topic structure which were posited in the study.

In summary, we have seen that by using a combination of prosodic cues, it becomes possible to identify topic category structure that would have remained hidden when using only a small subset of prosodic signals. These two speakers are generally representative of the patterns in the data, and it is possible to divide the other speakers into two groups, although not all speakers showed the exact patterns or had them attain statistical significance. The 9 speakers who produced a similar pattern to speaker F05 (henceforth the "Fall Range Group") varied the size of F0 falls in relation to the topic structure of the discourse but not speech rate; the 5 speakers showing a similar pattern to speaker F04 (henceforth the "Speech Rate Group") varied speech rate but not the size of F0 falls. The fact that individual speakers' patterns vary in terms of which distinctions attain statistical significance can in part be attributed to the fact that not all speakers provided an equal number of data points. Since the data analysis was limited to non-nuclear falling intonational contours, and different speakers phrased their utterances differently, the number of target items which could be included in the analysis for each speaker varied. Speakers with more data points were more likely to conform to the patterns presented here, and speakers with fewer data points were more likely to deviate from them, suggesting that more consistent patterns would be observable if more data were available.

	Fall Range Group (like speaker F05)	**Speech Rate Group (like speaker F04)**
Topic – fall range (st)	5.564019 *	4.315985
Addition – fall range	4.716109	4.175885
Elaboration – fall range	4.232264	3.784505
Continuation – fall range	3.431596 *	3.30309 ·
Topic – word length (sec)	0.327549 ·	0.329359 *
Addition – word length	0.343265	0.378102
Elaboration – word length	0.353144	0.381215
Continuation – word length	0.363623	0.410123 ·

Table 1: Mean fall range and word length by category for each group of speakers (shown with ID of representative speaker). Comparisons were carried out by ANOVA. For fall range, F(3, 492)=16.8; for word length, F(3, 544)=9.74. '' indicates that a category differs significantly from the other categories (within the speaker group) to the p<0.5 level; '·' indicates a marginally significant difference, 0.05<p<0.08 in all cases.*

Table 1 shows the mean fall ranges and word lengths for the Fall Range Group and the Speech Rate Group, respectively. The two groups differ from each other in their production of both fall range variation ($F(15, 480)=7.2648$, $p<0.001$) and word length variation ($F(15, 532)=8.5334$, $p<0.001$). The patterns of variation in fall range and speech rate shown in the case studies are still apparent, although in many cases to a slightly different degree. However, the overall pattern as shown in the case studies is still apparent. It is interesting to note that when the data from different speakers are combined, the difference between Additions and Elaborations once again becomes obscured. It is possible that although some speakers, such as F04 above, make a distinction between these categories, others do not.

3. Discussion

Although previous prosodic studies of topic structure have not looked only at F0 modulation, that and pausing have certainly been the main foci of such studies. However, it is clear on the basis of the data presented here (as well as in other studies; cf. Ní Chasaide and Gobl 2004) that F0 variation by itself is insufficient to describe the prosodic characteristics of topic structure, even among speakers of the same dialect. Furthermore, F0 variation does not only combine with other prosodic signals, it also alternates with them. It appears, therefore, that at least as regards the signalling of topic structure, we must consider prosodic signalling as a whole, rather than creating a separation between intonation (as the main element related to F0) and other areas of prosody, since other phonetic parameters are commonly used to enhance F0 cues, and this kind of cue trading can complement cue trading within the F0 domain, and have perceptual effects on intonational meaning (Post et al. 2007).

Up to this point we have been making the assumption that there is a specific prosodic cue, or a consistent set of specific prosodic cues, that are associated with signalling topic structure in SSBE. It might instead be more valuable to think of these kinds of variation as a whole as having to do with prominence marking, rather than trying to separate different elements of the production. In other words, rather than asserting that variation in F0 fall range and speech rate signal topic structure, it could be argued that topic structure is signaled by a variation in prominence of the items in question. While SSBE speakers might be more inclined to use the F0 fall range or the speech rate in this context to make elements more or less prominent, this

would not necessarily lead to the conclusion that varying the F0 fall ranges or the speech rate would be obligatory in topic structure signalling.

One way to implement the idea of prominence variation in topic structure might be to suggest that there is a baseline neutral level for a variety of cues (Xu 2005), and all variation away from this baseline level contributes to the impression of prominence. In this case, a larger F0 fall span could contribute to the higher prominence of a given element, but equally, an increase in speech rate could do so in the same context. This kind of description would allow cues to alternate but also to combine: one could imagine an 'additive' effect of a slightly larger F0 fall range plus a slightly increased speech rate, which could cumulatively become equivalent to a larger increase in one cue or the other independently. This is consistent with what we observe in the data above, where more than one cue is necessary to both identify topic structure in all of the speakers, and potentially to identify a fourth topic structure category.

One question that still remains is whether or not, despite the combinatory effect of the cues, there is a single cue that is a 'default' or majority choice for signalling topic structure. In the current data, F0 fall span was used by more speakers than speech rate variation, but the number of speakers is relatively small and therefore it is difficult to draw a useful conclusion from this distribution. There may be a meaningful difference in this case, for example between the signals used by more proficient versus less proficient readers; but alternatively it may simply be representative of variation within the population on the basis of some as yet unknown factor. Such a factor may potentially even be external to the structure of the language, for example a sociolinguistic marker. Perceptual testing as to the relative prominence of these cues may shed some light on this issue, but for the moment the source of these different behavioral patterns remains an open question.

Acknowledgements

The authors would like to thank Mariapaola D'Imperio for valuable discussion on this topic, the three anonymous reviewers for their helpful comments and suggestions on the original abstract, and attendees of the IDP workshop for their comments on the poster version. This research was supported by the EC Marie Curie Research Training Network/*Sound to Sense*/(MRTN-CT-2006-035561), and by the ESRC grant 'Categories and gradience in intonation' (RES-061-25-0347) held by the second author.

References

Blakemore, D. (1992) *Understanding Utterances: An Introduction to Pragmatics*. Oxford: Blackwell.

Blakemore, D. (2002) *Relevance and Linguistic Meaning: The Semantics and Pragmatics of Discourse Markers*. Cambridge: Cambridge University Press.

Brazil, D. (1997) *The Communicative Value of Intonation in English*. Cambridge: Cambridge University Press.

Büring, D. (1999) Topic. In: Bosch, Peter & Rob van der Sandt (eds.) *Focus -- Linguistic, Cognitive, and Computational Perspectives*. Cambridge: Cambridge University Press, 142-165.

Grosz, B. & Sidner, C. (1986) Attention, intentions, and the structure of discourse. *Computational Linguistics* 12(3):175-204.

Gussenhoven, C. (2004) *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.

Nakajima, S. & Allen, J.F. (1993) A study on prosody and discourse structure in cooperative dialogues. *Phonetica* 50:197-210.

Ní Chasaide, A. & Gobl, C. (2004) Voice quality and f0 in prosody: towards a holistic account. *Proceedings of 2nd International Conference on Speech Prosody*, Nara, Japan, 189-196.

Post, B., D'Imperio, M. & Gussenhoven, C. (2007) Fine phonetic detail and intonational meaning. *Proceedings of ICPhS XVI*, Saarbrücken, Germany, 191-196.

Wichmann, A. (2000) *Intonation in Text and Discourse*. Harlow: Longman.

Xu, Y. (2005) Speech melody as articulatorily implemented communicative functions. *Speech Communication* 46:220-251.

Zellers, M. (2009) Fundamental frequency and discourse meaning in SSBE. Presentation given at Phonetics and Phonology in Iberia, Las Palmas de Gran Canaria, 17-18 June 2009.
Zellers, M., Post, B. & D'Imperio, M. (2009) Modelling the intonation of topic structure: two approaches. *Proceedings of 10th Interspeech*, Brighton, UK, 2463-2466.