Discourse Structure in Spoken Language: Studies on Speech Corpora

Discourse Structure in Spoken Language: Studies on Speech Corpora*

Christine H. Nakatani†
Aiken Computation Laboratory
Division of Applied Sciences
Harvard University
Cambridge MA 02138 USA
chn@das.harvard.edu

Julia Hirschberg
2C-409, AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill NJ 07974 USA
julia@research.att.com

Barbara J. Grosz
Aiken Computation Laboratory
Division of Applied Sciences
Harvard University
Cambridge MA 02138 USA
grosz@das.harvard.edu

Abstract

A better understanding of the intonational characteristics of spoken discourse may lead to new empirical techniques for identifying discourse structure from speech, as well as new algorithms for enhancing the naturalness of synthetic speech. This paper summarizes results of pilot studies that demonstrate reliable correlations of discourse and speech properties, and reports findings on a new corpus of direction-giving monologues, collected in both spontaneous and read speaking styles. Preliminary analyses of the direction-giving corpus show that the availability of speech significantly affects the reliability of discourse segmentation for a set of trained discourse labelers.

Introduction

This paper reports on ongoing corpus-based research on the intonational characteristics of spoken discourse in American English. The scientific goal of this research is to lay the foundations for a bootstrapping process, in which empirical evidence from spoken language informs us of strengths and weaknesses in a discourse theory, and in which our best current understanding of discourse structure suggests more sophisticated interpretations of intonational meaning. The technological goal of this research is to improve the quality of speech synthesis by exploiting the ability of intonation to reliably convey linguistic structure at the discourse level.

Cognitive studies based on linguistic research have shown that the lack of contextually appropriate intonational variation can hinder processing by the human listener (Terken & Nooteboom 1987; Nooteboom & Kruyt 1987). Yet, algorithms for manipulating prosodic variation lag behind even our present understanding of how intonational meaning is conveyed, despite the fact that basic synthesis technologies for producing natural intonation already exist.[1]

Theoretical and Methodological Foundations

Several decades of research have resulted in numerous findings on how discourse level meaning can be conveyed, by acoustic-prosodic properties such as pitch range and pausal duration (Avesani & Vayra 1988; Ayers 1992; Brown, Currie, & Kenworthy 1980; Lehiste 1979; Silverman 1987) (cf. (Woodbury 1987)), amplitude (Brown, Currie, & Kenworthy 1980), speaking rate (Lehiste 1980), and intonational prominence (Brown 1983; Terken 1984). Most of these studies have relied on intuitive analyses of notions such as topic-structure, or operational definitions of discourse-level properties, such as paragraph markings as indicators of discourse segment boundaries.

In contrast to most previous work, two recent studies utilized an independent definition of discourse structure to obtain discourse segmentation data from multiple subjects. In (Hirschberg & Grosz 1992; Grosz & Hirschberg 1992), discourse stuctural elements were determined by trained subjects following (Grosz & Sidner 1986), and were correlated with intonational properties. In (Passoneau & Litman 1993), discourse segmentations were obtained from naive subjects based on an informal notion of speaker intention. For a narrative corpus, pausal duration above a certain threshold predicted segment boundaries with high recall (92%) but low precision (18%). Passoneau and Litman suggest that intonational cues be integrated with text-based cues such as cue phrases (Hirschberg & Litman 1993) and other lexical information (Morris & Hirst 1991; Hearst 1994), in spoken language processing systems using multiple knowledge sources.

The potential contributions of speech cues in such an architecture remain largely unexplored. Intonational variables need to be interrelated in new algorithms,

*The research reported here was partially supported by grants NSF IRI 9009018 and NSF IRI 9308173 from the National Science Foundation.

†Partially supported by a National Science Foundation Graduate Research Fellowship.

[1] That is, input to systems such as DEC-Talk and the AT&T Text-to-Speech System can be *hand* annotated to produce quite natural sounding speech.

and a fuller spectrum of speech properties needs to be correlated with a theoretically motivated yet empirically determined representation of discourse structure. The approach we have taken in our work is to (1) conduct corpus-based empirical work on intonational features of spoken language, (2) analyze discourse properties based on an independently motivated theory of discourse structure, and (3) examine the correlations between the two sources of linguistic structure.

Prosodic Analysis

The methods we use for measuring speech properties such as rate, energy (rms), pauses, and fundamental frequency, are widely used in the speech community. These measures can be obtained automatically given orthographic and prosodic transcriptions of the speech. The prosodic transcription, a more abstract representation of the intonational prominences, phrasing, and melodic contours, is obtained by hand-labeling. We employ the ToBI standard for prosodic transcription (Silverman *et al.* 1992; Pitrelli, Beckamn, & Hirschberg 1994), which is based upon Pierrehumbert's theory of American English intonation (Pierrehumbert 1980).

The ToBI transcription provides us with a breakdown of the speech sample into minor or INTERMEDIATE PHRASES, in Pierrehumbert's terms (Pierrehumbert 1980). This level of prosodic phrase serves as our primary unit of analysis for measuring both speech and discourse properties. For each intermediate phrase, we calculate values for pitch range from the fundamental frequency (f0) maximum occurring within an accented syllable in the phrase; amount of f0 change between phrases, f0(phrase[i])/f0(phrase[i+1]; amplitude and energy (rms) maxima within the vowel of the syllable containing the phrase's f0 peak; contour type and type of nuclear accent, identified in ToBI notation; speaking rate, measured in syllables per second (sps); and pausal duration between intermediate as well as intonational phrases.

Discourse structure analysis

We base our discourse analysis on the theory of discourse structure presented in (Grosz & Sidner 1986) (hereafter G&S), in which discourse structure is comprised of INTENTIONAL STRUCTURE, ATTENTIONAL STATE, and LINGUISTIC STRUCTURE. G&S's model also distinguishes between two levels of discourse processing, GLOBAL and LOCAL (Grosz 1977; Grosz & Sidner 1986). Discourse segments, embedding relations, discourse segment purposes (DSPs) and relations between them are part of the global level. Attentional state at this level is modeled by a stack of focus spaces. The local level of discourse structure concerns features of the utterances within a discourse segment and relations among them. Attentional state at this level is modeled by centering theory (Grosz, Joshi, & Weinstein 1983).

Intonation is an element of the linguistic structure that can provide information important for computing both attentional state and intentional structure. In our research, G&S's model of discourse structure provides both a foundation for segmenting discourses into constituent parts, and a set of theoretical constructs that may serve to mediate our interpretation of the discourse functions of intonational features. Further, intonation provides information about both levels of discourse structure. For example, at the global level, cue phrases that mark segment boundaries (Sidner 1983; Reichman-Adar 1984) exhibit reliable intonational properties (Hirschberg & Litman 1987; Hirschberg 1993). At the local level, intonation may indicate whether a phrase is parenthetical, or may influence the perceived salience of some mentioned entity.

We devised a set of instructions based on G&S for labeling the intentional and linguistic structures at both the local and global levels (Hirschberg & Grosz 1992; Grosz & Hirschberg 1992). While the studies reported here utilize these so-called "expert" instructions, a parallel set of intention-based segmentation instructions suitable for naive subjects is being developed for use in the Boston Directions study, which is described below.

Speech Corpora

We utilize three corpora in our investigations: (1) professionally read AP news stories, (2) non-professional spontaneous narrative, and (3) non-professional elicited task-oriented monologues, both spontaneous and read. Below, we summarize results of two pilot studies utilizing the first two corpora, respectively. The first pilot study investigated intonational correlates of discourse structure, while the second focused on discourse structural constraints on intonational prominence. Although the pilot study results were encouraging, our experiences with the respective corpora revealed ways in which choices of speaking style (e.g. read vs. spontaneous, professional vs. non-professional) and genre generally influence both discourse and speech properties. These single-speaker corpora also did not address the problem of individual variation across speakers. To overcome problems with these corpora, a third corpus of multi-speaker elicited task-oriented monologues was designed.

This corpus, the Boston Directions Corpus, exhibits discourse and speech properties more characteristic of the language used in interactive spoken language systems. After describing the corpus, we present recent initial results on a portion of it that extend findings of our pilot studies.

Pilot Study: Intonational Correlates of Discourse Structure

In one set of pilot studies (Hirschberg & Grosz 1992; Grosz & Hirschberg 1992), we analyzed a corpus of three Associated Press (AP) news stories recorded by a

professional speaker. Results confirmed previous findings that pitch range and timing variation are important in signaling topic structure, and further established that these relationships hold when topic structure has been independently determined from consensus subject labeling.

Analysis & Results

Two groups of subjects labeled the stories using the expert intention-based discourse segmentation instructions. One group labeled from text alone (group T), while the other group annotated the text while simultaneously listening to the corresponding speech (group S). We then analyzed intonational and acoustic features for those discourse structural elements agreed upon by all labelers in a given group, the CONSENSUS LABELS.[2] We separately examined group T's and group S's consensus labelings for discourse segment beginnings (SBEGs) and discourse segment endings (SFs) for one news story.

We found statistically significant correlations of aspects of pitch range, amplitude, and timing with features of global and local structure for both group T and S labelings.[3] Further analyses of this and two additional news stories showed that global and local structures could be reliably identified from hand-labeled acoustic and prosodic features with (cross-validated) success rates of 86%-97% (Hirschberg & Grosz 1992; Grosz & Hirschberg 1992). Two central contributions of this pilot study were (a) the discovery of new relationships among intonational features of discourse structure; and (b) the development of a methodology for obtaining discourse segmentations by theoretically motivated empirical methods independent of the acoustic and prosodic factors under investigation.

These results demonstrated the possibility of discovering intonational correlates of discourse structure through corpus analysis. However, the study also revealed certain weaknesses in our corpus and in our methodology. Some aspects of news stories proved difficult for our labelers to segment reliably. Also, radio speech has been claimed to exhibit certain idiosyncracies which might confound our results, since a major goal of radio news writers is to capture and maintain listener attention. This goal of engaging an audience may interact with other discourse purposes. In addition, the normal processes of news editing may alter the originally intended discourse structure of the news story. Issues such as these introduce difficulties into segmentation analysis. We also wanted to see if our results from speech read by a professional speaker would generalize to spontaneous speech and to speech from non-professionals.

[2]Use of consensus labels is a conservative measure of labeler agreement; average inter-labeler agreement for structural elements varied from 74.3%-95.1% within each group. Results in (Passoneau & Litman 1993) show that with a larger number of labelers, more sophisticated criteria such as BOUNDARY STRENGTH can be usefully employed.

[3]Details are reported in (Hirschberg & Grosz 1992; Grosz & Hirschberg 1992).

Pilot Study: Intonational Prominence and Discourse Structure

A second pilot study investigated the relationship between intonational prominence assignment and local and global discourse structure in spontaneous narrative (Nakatani 1993; 1994). Results of (Brown 1983) showed a general tendency for GIVEN information to be unaccented, and NEW information accented (where lexical items referring to referents previously mentioned in the discourse are considered given). Other research on the problem of predicting accented given information has identified a variety of relevant factors (e.g., (Horne 1991) on metrical-phonological constraints, (Selkirk 1993) on syntactic factors, (Terken & Hirschberg 1992) on persistence of grammatical function and surface position, and (Hirschberg 1993) on the interaction of these and other variables in pitch accent assignment algorithms trained on corpora). While Terken studied an additional factor, namely discourse structural constraints, he operationally defined the notion of discourse or topic structure based on task structure (Terken 1984). Referents were identified as given/new relative to a topic segment. We extend Terken's findings by providing an independent discourse analysis for our narrative and by recasting the given/new distinction at two levels of discourse structure, using G&S's attentional state model to identify discourse referents as locally given/new, and globally given/new.

Analysis & Results

A total of 481 animate noun phrase referring expressions in a 20-minute, single-speaker, unrestricted spontaneous narrative were analyzed for lexical form, grammatical function and intonational prominence (i.e. **H*** ("high star") or complex pitch accents in Pierrehumbert's notation (Pierrehumbert 1980)).[4] The linguistic and intentional structures were analyzed according to the expert intention-based instructions. We found statistically significant asymmetries in the interactions of accentuation with grammatical position and lexical form, which we accounted for by noting that the presence or absence of intonational prominence combines with lexical and syntactic information to mark shifts in attention at both the local and global levels of discourse structure.

While this study confirmed general findings that full forms bear accent and reduced forms do not (Brown 1983; Terken 1984), it also went beyond previous work, in suggesting a uniform explanation of both expected

[4]The narrative was collected by Virginia Merlini for the purpose of studying American gay male speech. We thank Mark Liberman at the University of Pennsylvania for making it available to us.

patternings and so-called mismatches between lexical form and accentuation (i.e. cases of accented pronouns and unaccented full forms). So-called mismatches may be reinterpreted as cases in which accent marks the attentional status of a discourse referent where lexical form and grammatical position convey conflicting statuses.

However, the problem of reliable segmentation applied acutely to our narrative, which is over 2,000 words long; the AP news stories averaged 450 words in length, while the task-oriented speech segments studied in (Brown 1983) averaged a few hundred. Also, owing to the subject matter of the narrative, the majority of referring expressions were realized as pronouns and proper names. To further test and refine our hypotheses, we wanted to examine more segmentation data and a fuller variety of referring expressions produced by multiple speakers.

Current Investigations: Boston Directions Corpus

To build upon our preliminary results, we have undertaken a more extensive study using spontaneous and read speech in a direction-giving task. The new corpus is made up of elicited monologues produced by multiple non-professional speakers, who are given written instructions to perform a series of increasingly complex direction-giving tasks. Speakers first explain simple routes such as getting from one station to another on the subway, and progress gradually to the most complex task of planning a round-trip journey from Harvard Square to several Boston tourist sights. The speakers are provided with various maps, and may write notes to themselves as well as trace routes on the maps. For the duration of the experiment, the speakers are in face-to-face contact with a silent experimental partner (a confederate) who traces on her map the routes described by the speakers. The speech is subsequently orthographically transcribed, with false starts and other speech errors repaired or omitted; subjects return several weeks after their first recording to read the transcribed speech. Both sets of recordings are then acoustically and prosodically labeled.[5] Preliminary results described below are available for both the spontaneous and the read speech for one speaker, performing five direction-giving tasks. These tasks resulted in 3.42 minutes or 130 intermediate phrases of read speech, and 4.28 minutes or 145 intermediate phrases of spontaneous speech.

[5]Speech recording and analysis is carried out using the WAVES+ software package (Talkin 1989) and the ToBI labeling convention and tools (Silverman *et al.* 1992; Pitrelli, Beckamn, & Hirschberg 1994) on Silicon Graphics workstations.

Analysis & Results: Discourse Segmentation

Discourse segmentations using the expert instructions were obtained from three subjects labeling from text alone (group T) and three labeling from speech and text (group S). Percentages for consensus labels for segment-initial (SBEG), segment-final (SF), and segment-medial (SCONT, defined as neither SBEG nor SF) are given in Table 1.[6] Two interesting trends emerge. First, in contrast to the AP news story findings, group S segmentations differ significantly from those of group T. Table 1 shows that listening to speech while segmenting produces more consensus boundaries for both read and spontaneous speech than does segmenting from text alone. When the read and spontaneous data are pooled, labelers from speech and text agree upon significantly more SBEG boundaries ($p<.05$, $\chi=4.5$, df=1) as well as SF boundaries ($p<.02$, $\chi=6.3$, df=1) than labelers from text alone. Further, it is not the case that segmenters from text alone simply choose to place fewer boundaries in the discourse; if this were so, then we would expect a high number of SCONT consensus labels where no SBEGs or SFs were identified. Instead, we find that the number of consensus SCONTs is significantly higher for labelings from speech and text, for read ($p<.001$, $\chi=11.7$, df=1) and spontaneous speech ($p< 1.5 \times 10^{-9}$, $\chi=36.6$, df=1). These factors combine to yield significantly higher percentages of consensus labels overall for labelings from speech and text, for both read ($p< 1.8\times10^{-9}$, $\chi=36.1$, df=1) and spontaneous speech ($p< 1.4\times10^{-8}$, $\chi=32.2$, df=1). We conclude that aspects of the speech signal can help disambiguate among alternate segmentations of the same text, and thus the availability of speech critically influences the outcome of discourse structure analysis.

The second trend to emerge concerns a somewhat surprising effect of speaking style on segmentation, namely that of read versus spontaneous speaking modes. Spontaneous speech is generally claimed to exhibit less reliable prosodic indicators of discourse structure than read speech (cf. (Ayers 1992)). Yet, in our corpus, spontaneous speech actually produced significantly more SCONT consensus labels than did read speech, for groups S and T combined ($p<.004$, $\chi=8.7$, df=1). The higher overall percentages of consensus labels for spontaneous speech are attributable to this difference in SCONT labelings.

[6]Note that the value in Table 1 for "Sum of all types" can be slightly less than the sum of percentages for the three types due to the fact that one phrase may be simultaneously labeled segment-initial and segment-final. This occurs when a single phrase comprises a complete segment.

CONSENSUS LABELS FOR READ SPEECH (N=130)				
	Seg-initial (SBEG)	Seg-final (SF)	Segment-medial (SCONT)	Sum of all types
Text alone (T)	17%	17%	7%	38%
Speech & Text (S)	26%	24%	26%	75%
CONSENSUS LABELS FOR SPONTANEOUS SPEECH (N=145)				
	Seg-initial (SBEG)	Seg-final (SF)	Segment-medial (SCONT)	Sum of all types
Text alone (T)	17%	16%	16%	46%
Speech & Text(S)	25%	23%	33%	78%

Table 1: Percentage of Consensus Labels by Segment Boundary Type

Analysis & Results: Intonational Correlates

We examined the following acoustic and prosodic correlates of consensus labelings of intermediate phrases labeled as SBEGs and SFs: f0 maximum and average f0; rms maximum and average; speaking rate; and duration of preceding and subsequent pauses. We compared segmentation labels not only for group S versus group T, but also for spontaneous versus read speech. As noted, while intonational correlates for segment boundaries *have* been identified in read speech, they have been observed in spontaneous speech rarely and descriptively.

We found strong correlations for consensus SBEG and SF labels for groups S and T in both spontaneous speech and read speech.[7] Results on consensus SBEG labels were as follows: given group T segmentations, we found significantly higher maximum and average f0, and maximum and average rms, and shorter subsequent pause for both spontaneous and read speech; for read speech we also found significant correlations for preceding pauses. Given group S segmentations, we found significantly higher maximum and average f0, higher maximum rms, longer preceding and shorter succeeding pauses for read and spontaneous speech; we found higher average rms as well for read speech. Results on consensus SF labels were as follows: given group T segmentations, we found significantly lower average f0 and rms maximum for both read and spontaneous speech, and lower rms average and subsequent pause in addition for read speech. Given group S segmentations, we found lower average f0, rms maximum and average, shorter preceding pause, and longer subsequent pause for both read and spontaneous speech, and in addition, lower f0 maximum for read speech.

While these results now hold for only a single speaker, they are quite encouraging. We may hypothesize that speakers can convey structural information about a discourse in spontaneous, as well as in read speech. We may also hypothesize that this structural information is in fact signalled at least in part by prosodic and acoustic information, since discourse labelings produced while listening to speech correlated with more acoustic-prosodic features than labelings from text alone. Certain acoustic-prosodic features such as preceding pause, for example, appear to have been made use of in segmentation decisions for group S.

Analysis & Results: Intonational Prominence

A total of 173 noun phrases in the read speech for the five direction-giving discourses were analyzed for lexical form (e.g., proper names, definite/indefinite NPs), grammatical function (e.g., subject, direct object, prepositional object), surface-order position (sentence-initial, medial, final), position in major intonational phrase (phrase-initial, medial, final), and accentuation (unaccented or pitch accent type). Similar to the pilot study findings, 23% of noun phrases were accentually reduced, i.e. bearing fewer pitch accents than the citation-form.[8] Preliminary analysis indicates that lexical form and grammatical function are significant factors in determining accentuation, with names being less reduced than full NPs, and subjects less reduced than objects.[9] Surface-order and intonational phrase position were not significant.

As for the pilot study, the simple notion that references to given entities in the discourse should be accentually reduced fails to predict accentuation, since reintroductions of discourse-old entities were often accented. Interestingly, it was not the case either that repetitions of the same referring expressions were accentually reduced any more than were alternate lexical expressions referring to discourse-old entities. In the case of repeated referring expressions, the second

[7] T-tests were used to test for statistical significance of difference in the means of phrases, e.g. beginning and not beginning segments. Results reported are significant at the .025 level or better.

[8] The AT&T Text-to-Speech System (TTS) was used to determine the majority of citation-form accent assignments. Two native speaker judgments were used for items not in the TTS lexicon, such as street and restaurant names.

[9] Chi-square tests were used to test significance. Results reported are at the .02 level or better.

mention was usually unreduced in intonational prominence when a discourse segment boundary intervenes between it and the first mention. Thus, accentual reduction cannot be considered an epiphenomenon of lexical givenness; if it were so, then lexical repetitions should simply be reduced. Rather, discourse structure interacts with lexical and other factors to constrain the deaccentuation of given information.

Finally, certain cases of accentual reduction did not arise previously in the narrative pilot study. In the Boston Directions Corpus, head nouns were frequently deaccented in full NPs with accented adjectival modifiers. This phenomenon typically occurs when the speaker contrasted two referential tokens of the same type (e.g. RED line SUBWAY vs. GREEN line subway, RIGHT TURN vs. ANOTHER right turn). However, the two tokens were not confined to the same discourse segment. This poses a problem for the global focusing mechanism in (Grosz & Sidner 1986), which assumes that entities in sister segments cannot be simultaneously in global focus. We will explore whether limited relaxations of this assumption, such as considering the most recently popped focus space to be in non-immediate global focus, can account for our cases of accentual reduction. A similar relaxation was necessary to account for the deaccentuation of object proper names in the narrative study.

Conclusion

Our studies of intonation and discourse provide empirical evidence that discourses can be segmented reliably, that intonation is used by speakers to convey linguistic structure at the discourse level, and that the relationship among intonational features and discourse elements is more complex than previous studies have suggested. Preliminary analysis of the Boston Directions Corpus has supported these hypotheses, and has also uncovered important effects of speaking style and segmentation methodology on the ability to obtain reliable analyses of discourse structure. Contrary to expectation, we found that discourse structure analysis is most robust for spontaneous speech labeled from speech and text together. Our continuing analysis of this corpus will test the generality of these trends against more data, including speech from multiple speakers and discourse segmentations produced by naive subjects. Findings of these corpus-based studies in sum suggest that looking at spoken language can lead to improvements in the descriptive and computational adequacy of theories about discourse structure as well as theories of intonational meaning.

Acknowledgements

The authors thank Nancy Chang, Andy Kehler, Candy Sidner and Gregory Ward for their expert participation in this research.

References

Avesani, C., and Vayra, M. 1988. Discorso, segmenti di discorso e un' ipotesi sull' intonazione. In *Att del Convegno Internazionale "Sull'Interpunzione"*.

Ayers, G. M. 1992. Discourse functions of pitch range in spontaneous and read speech. Presented at the Linguistic Society of America Annual Meeting.

Brown, G.; Currie, K.; and Kenworthy, J. 1980. *Questions of Intonation*. Baltimore: University Park Press.

Brown, G. 1983. Prosodic structure and the given/new distinction. In Ladd, D. R., and Cutler, A., eds., *Prosody: Models and Measurements*. Berlin: Springer Verlag. 67–78.

Grosz, B., and Hirschberg, J. 1992. Some intonational characteristics of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing*. Banff: ICSLP.

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

Grosz, B.; Joshi, A.; and Weinstein, S. 1983. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting*, 44–50. Cambridge MA: Association for Computational Linguistics.

Grosz, B. J. 1977. The representation and use of focus in dialogue understanding. Technical Report 151, SRI International, Menlo Park Ca. University of California at Berkeley PhD Thesis.

Hearst, M. 1994. Multi-paragraph segmentation of expository discourse. In *Proceedings of the 32nd Annual Meeting*. Las Cruces, NM: Association for Computational Linguistics.

Hirschberg, J., and Grosz, B. 1992. Intonational features of local and global discourse structure. In *Proceedings of the Speech and Natural Language Workshop*, 441–446. Harriman NY: DARPA.

Hirschberg, J., and Litman, D. 1987. Now let's talk about *now*: Identifying cue phrases intonationally. In *Proceedings of the 25th Annual Meeting*, 163–171. Stanford University: Association for Computational Linguistics.

Hirschberg, J., and Litman, D. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*.

Hirschberg, J. 1993. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence* 63.

Horne, M. 1991. Why do speakers accent 'given' information. In *Proceedings of the Second European Conference on Speech Communication and Technology*. Genova: Eurospeech-91.

Lehiste, I. 1979. Perception of sentence and paragraph boundaries. In Lindblom, B., and Oehman, S.,

eds., *Frontiers of Speech Research*. London: Academic Press. 191–201.

Lehiste, I. 1980. Phonetic characteristics of discourse. Paper presented at the Meeting of the Committee on Speech Research, Acoustical Society of Japan.

Morris, J., and Hirst, G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics* 17:21–48.

Nakatani, C. H. 1993. Accenting on pronouns and proper names in spontaneous narrative. In *Proceedings of the European Speech Communication Association Workshop on Prosody*. Lund, Sweden: ESCA.

Nakatani, C. H. 1994. Discourse structural constraints on accent in spontaneous narrative. In *Proceedings of the European Speech Communication Association/IEEE Workshop on Speech Synthesis*. New Paltz, NY: ESCA/IEEE.

Nooteboom, S. G., and Kruyt, J. G. 1987. Accent, focus distribution and the perceived distribution of given and new information: An experiment. *Journal of the Acoustical Society of America* 82(5):1512–1524.

Passoneau, R., and Litman, D. 1993. Feasibility of automated discourse segmentation. In *Proceedings of ACL-93*. Ohio State University: Association for Computational Linguistics.

Pierrehumbert, J. B. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. Dissertation, Massachusetts Institute of Technology. Distributed by the Indiana University Linguistics Club.

Pitrelli, J. F.; Beckamn, M.; and Hirschberg, J. 1994. Evaluation of prosodic transcription labeling reliability in the tobi framework. In *Proceedings of ICSLP*. Yokohama: International Conference on Spoken Language Processing.

Reichman-Adar, R. 1984. Extended person-machine interface. *AI Journal* 22(2):157–218.

Selkirk, E. O. 1993. *Sentence Prosody: Intonation, Stress and Phrasing*. Basil Blackwell.

Sidner, C. L. 1983. Focusing in the comprehension of definite anaphora. In Brady, M., and Berwick, R., eds., *Computational Models of Discourse*. Cambridge MA: MIT Press. 267–330.

Silverman, K.; Beckamn, M.; Pierrehumbert, J.; Ostendorf, M.; Wightman, C.; Price, P.; and Hirschberg, J. 1992. Tobi: A standard scheme for labeling prosody. In *Proceedings of ICSLP*. Banff: International Conference on Spoken Language Processing.

Silverman, K. 1987. *The Structure and Processing of Fundamental Frequency Contours*. Ph.D. Dissertation, Cambridge University, Cambridge UK.

Talkin, D. 1989. Looking at speech. *Speech Technology* 4:74–77.

Terken, J., and Hirschberg, J. 1992. Deaccentuation and persistence of grammatical function and surface position. Ms.

Terken, J., and Nooteboom, S. G. 1987. Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. *Language and Cognitive Processes* 2(3/4):145–163.

Terken, J. 1984. The distribution of pitch accents in instructions as a function of discourse structure. *Language and Speech* 27:269–289.

Woodbury, A. C. 1987. Rhetorical structure in a central Alaskan Yupik Eskimo traditional narrative. In Sherzer, J., and Woodbury, A., eds., *Native American Discourse: Poetics and Rhetoric*. Cambridge UK: Cambridge University Press. 176–239.