A Prosodic Analysis of Discourse Segments in Direction-Giving Monologues

Julia Hirschberg
AT&T Laboratories, 2C-409
600 Mountain Avenue
Murray Hill, NJ 07974

Christine H. Nakatani*
Harvard University
33 Oxford Street
Cambridge, MA 02138

Abstract

This paper reports on corpus-based research into the relationship between intonational variation and discourse structure. We examine the effects of speaking style (read versus spontaneous) and of discourse segmentation method (text-alone versus text-and-speech) on the nature of this relationship. We also compare the acoustic-prosodic features of initial, medial, and final utterances in a discourse segment.

1 Introduction

This paper presents empirical support for the assumption long held by computational linguists, that intonation can provide valuable cues for discourse processing. The relationship between intonational variation and discourse structure has been explored in a new corpus of direction-giving monologues. We examine the effects of speaking style (read versus spontaneous) and of discourse segmentation method (text-alone versus text-and-speech) on the nature of this relationship. We also compare the acoustic-prosodic features of initial, medial, and final utterances in a discourse segment. A better understanding of the role of intonation in conveying discourse structure will enable improvements in the naturalness of intonational variation in text-to-speech systems as well as in algorithms for recognizing discourse structure in speech-understanding systems.

2 Theoretical and Empirical Foundations

It has long been assumed in computational linguistics that discourse structure plays an important role in Natural Language Understanding tasks such as identifying speaker intentions and resolving anaphoric reference. Previous research has found that discourse structural information can be inferred from orthographic cues in text, such as paragraphing and punctuation; from linguistic cues in text or speech, such as CUE PHRASES[1] (Cohen, 1984; Reichman, 1985; Grosz and Sidner, 1986; Passonneau and Litman, 1993; Passonneau and Litman, to appear) and other lexical cues (Hinkelman and Allen, 1989); from variation in referring expressions (Linde, 1979; Levy, 1984; Grosz and Sidner, 1986; Webber, 1988; Song and Cohen, 1991; Passonneau and Litman, 1993), tense, and aspect (Schubert and Hwang, 1990; Song and Cohen, 1991); from knowledge of the domain, especially for task-oriented discourses (Grosz, 1978); and from speaker intentions (Carberry, 1990; Litman and Hirschberg, 1990; Lochbaum, 1994). Recent methods for automatic recognition of discourse structure from text have incorporated thesaurus-based and other information retrieval techniques to identify changes in topic (Morris and Hirst, 1991; Yarowsky, 1991; Iwańska et al., 1991; Hearst, 1994; Reynar, 1994).

Parallel investigations on prosodic/acoustic cues to discourse structure have investigated the contributions of features such as pitch range, pausal duration, amplitude, speaking rate, and intonational contour to signaling topic change. Variation in pitch range has often been seen as conveying 'topic structure' in discourse. Brown et al. (1980) found that subjects typically started new topics relatively high in their pitch range and finished topics by compressing their range. Silverman (1987) found that manipulation of pitch range alone, or in conjunction with pausal duration between utterances, facilitated the disambiguation of ambiguous topic structures. Avesani and Vayra (1988) also found variation

*The second author was partially supported by NSF Grants No. IRI-90-09018, No. IRI-93-08173, and No. CDA-94-01024 at Harvard University and by AT&T Bell Laboratories.

[1] Also called DISCOURSE MARKERS or DISCOURSE PARTICLES, these are items such as *now*, *first*, and *by the way*, which explicitly mark discourse structure.

in pitch range in professional recordings which appeared to correlate with topic structure, and Ayers (1992) found that pitch range correlates with hierarchical topic structure more closely in read than spontaneous conversational speech. Duration of pause between utterances or phrases has also been identified as an indicator of topic structure, with longer pauses marking major topic shifts (Lehiste, 1979; Brown, Currie, and Kenworthy, 1980; Avesani and Vayra, 1988; Passonneau and Litman, 1993); Woodbury (1987), however, found no such correlation in his data. Amplitude was also found to increase at the start of a new topic and decrease at the end (Brown, Currie, and Kenworthy, 1980). Swerts and colleagues (1992) found that melody and duration can pre-signal the end of a discourse unit, in addition to marking the discourse-unit-final utterance itself. And speaking rate has been found to correlate with structural variation; in several studies (Lehiste, 1980; Brubaker, 1972; Butterworth, 1975) segment-initial utterances exhibited slower rates, and segment-final, faster rates. Swerts and Ostendorf (1995), however, report negative rate results.

In general, these studies have lacked an independently-motivated notion of discourse structure. With few exceptions, they rely on intuitive analyses of topic structure; operational definitions of discourse-level properties (e.g., interpreting paragraph breaks as discourse segment boundaries); or 'theory-neutral' discourse segmentations, where subjects are given instructions to simply mark changes in topic. Recent studies have focused on the question of whether discourse structure itself can be empirically determined in a reliable manner, a prerequisite to investigating linguistic cues to its existence. An intention-based theory of discourse was used in (Hirschberg and Grosz, 1992; Grosz and Hirschberg, 1992) to identify intonational correlates of discourse structure in news stories read by a professional speaker. Discourse structural elements were determined by experts in the Grosz and Sidner (1986) theory of discourse structure, based on either text alone or text and speech. This study revealed strong correlations of aspects of pitch range, amplitude, and timing with features of global and local structure for both segmentation methods. Passonneau and Litman (to appear) analyzed correlations of pause, as well as cue phrases and referential relations, with discourse structure; their segmenters were asked to identify speakers' communicative "actions". The present study addresses issues of speaking style and segmentation method while exploring in more detail than previous studies the prosodic parameters that characterize initial, medial, and final utterances in a discourse segment.

3 Methods

3.1 The Boston Directions Corpus

The current investigation of discourse and intonation is based on analysis of a corpus of spontaneous and read speech, the Boston Directions Corpus.[2] This corpus comprises elicited monologues produced by multiple non-professional speakers, who were given written instructions to perform a series of nine increasingly complex direction-giving tasks. Speakers first explained simple routes such as getting from one station to another on the subway, and progressed gradually to the most complex task of planning a round-trip journey from Harvard Square to several Boston tourist sights. Thus, the tasks were designed to require increasing levels of planning complexity. The speakers were provided with various maps, and could write notes to themselves as well as trace routes on the maps. For the duration of the experiment, the speakers were in face-to-face contact with a silent partner (a confederate) who traced on her map the routes described by the speakers. The speech was subsequently orthographically transcribed, with false starts and other speech errors repaired or omitted; subjects returned several weeks after their first recording to read aloud from transcriptions of their own directions.

3.2 Acoustic-Prosodic Analysis

For this paper, the spontaneous and read recordings for one male speaker were acoustically analyzed; fundamental frequency and energy were calculated using Entropic speech analysis software. The prosodic transcription, a more abstract representation of the intonational prominences, phrasing, and melodic contours, was obtained by hand-labeling. We employed the ToBI standard for prosodic transcription (Pitrelli, Beckman, and Hirschberg, 1994), which is based upon Pierrehumbert's theory of American English intonation (Pierrehumbert, 1980). The ToBI transcription provided us with a breakdown of the speech sample into minor or INTERMEDIATE PHRASES (Pierrehumbert, 1980; Beckman and Pierrehumbert, 1986). This level of prosodic phrase served as our primary unit of analysis for measuring both speech and discourse properties. The portion of the corpus we report on consists of 494 and 552 intermediate phrases for read and spontaneous speech, respectively.

[2] The Boston Directions Corpus was designed and collected in collaboration with Barbara Grosz.

3.3 Discourse Segmentation

In our research, the Grosz and Sidner (1986) theory of discourse structure, hereafter G&S, provides a foundation for segmenting discourses into constituent parts. According to this model, at least three components of discourse structure must be distinguished. The utterances composing the discourse divide into segments that may be embedded relative to one another. These segments and the embedding relationships between them form the LINGUISTIC STRUCTURE. The embedding relationships reflect changes in the ATTENTIONAL STATE, the dynamic record of the entities and attributes that are salient during a particular part of the discourse. Changes in linguistic structure, and hence attentional state, depend on the discourse's INTENTIONAL STRUCTURE; this structure comprises the intentions or DISCOURSE SEGMENT PURPOSES (DSPs) underlying the discourse and relations between DSPs.

Two methods of discourse segmentation were employed by subjects who had expertise in the G&S theory. Following Hirschberg and Grosz (1992), three subjects labeled from text alone (group T) and three labeled from text and speech (group S). Other than this difference in input modality, all subjects received identical written instructions. The text for each task was presented with line breaks corresponding to intermediate phrase boundaries (i.e., ToBI BREAK INDICES of level 3 or higher (Pitrelli, Beckman, and Hirschberg, 1994)). In the instructions, subjects were essentially asked to analyze the linguistic and intentional structures by segmenting the discourse, identifying the DSPs, and specifying the hierarchical relationships among segments.

4 Results

4.1 Discourse Segmentation

4.1.1 Raw Agreement

Labels on which all labelers in the group agreed are termed the CONSENSUS LABELS.[3] The consensus labels for segment-initial (SBEG), segment-final (SF), and segment-medial (SCONT, defined as neither SBEG nor SF) phrase labels are given in Table 1.[4]

[3]Use of consensus labels is a conservative measure of labeler agreement. Results in (Passonneau and Litman, 1993) and (Swerts, 1995) show that with a larger number of labelers, notions of BOUNDARY STRENGTH can be employed.

[4]Consensus percentages for the three types in Table 1 do not necessarily sum to the total consensus agreement percentage, since a phrase is both segment-initial and segment-final when it makes up a segment by itself.

Table 1: Percentage of Consensus Labels by Segment Boundary Type

	SBEG	SF	SCONT	Total
READ (N=494)				
Text alone (T)	14%	11%	32%	57%
Text & Speech (S)	18%	14%	49%	80%
SPON (N=552)				
Text alone (T)	13%	10%	40%	61%
Text & Speech (S)	15%	13%	54%	81%

Note that group T and group S segmentations differ significantly, in contrast to earlier findings of Hirschberg and Grosz (1992) on a corpus of read-aloud news stories and in support of informal findings of Swerts (1995). Table 1 shows that group S produced significantly more consensus boundaries for both read ($p<.001$, $\chi=58.8$, df=1) and spontaneous ($p<.001$, $\chi=55.4$, df=1) speech than did group T. When the read and spontaneous data are pooled, group S agreed upon significantly more SBEG boundaries ($p<.05$, $\chi=4.7$, df=1) as well as SF boundaries ($p<.05$, $\chi=4.4$, df=1) than did group T. Further, it is not the case that text-alone segmenters simply chose to place fewer boundaries in the discourse; if this were so, then we would expect a high percentage of SCONT consensus labels where no SBEGs or SFs were identified. Instead, we find that the number of consensus SCONTs was significantly higher for text-and-speech labelings than for text-alone ($p<.001$, $\chi=49.1$, df=1). It appears that the speech signal can help disambiguate among alternate segmentations of the same text. Finally, the data in Table 1 show that spontaneous speech can be segmented as reliably as its read counterpart, contrary to Ayer's results (1992).

4.1.2 Inter-labeler Reliability

Comparisons of inter-labeler reliability, that is, the reproducibility of a coding scheme given multiple labelers, provide another perspective on the segmentation data. How best to measure inter-labeler reliability for discourse segmentation tasks, especially for hierarchical segmentation, is an open research question (Passonneau and Litman, to appear; Carletta, 1995; Flammia and Zue, 1995; Rotondo, 1984; Swerts, 1995). For comparative purposes, we explored several measures proposed in the literature, namely, COCHRAN'S Q and the KAPPA (κ) COEFFICIENT (Siegel and Castellan, 1988). Cochran's Q, originally proposed in (Hirschberg and Grosz, 1992) to measure the likelihood that similarity among labelings was due to chance, was not useful in the current study; all tests of similarity using this metric (pairwise, or comparing all labelers) gave probability near zero. We concluded that this statistic did not serve, for example, to capture the differences observed between labelings from text alone versus labelings from text and speech.

Recent discourse annotation studies (Isard and Carletta, 1995; Flammia and Zue, 1995) have measured reliability using the κ coefficient, which factors out chance agreement taking the expected distribution of categories into account. This coefficient is defined as

$$\kappa = \frac{P_O - P_E}{1 - P_E}$$

where P_O represents the observed agreement and P_E represents the expected agreement. Typically, values of .7 or higher for this measure provide evidence of good reliability, while values of .8 or greater indicate high reliability. Isard and Carletta (1995) report pairwise κ scores ranging from .43 to .68 in a study of naive and expert classifications of types of 'moves' in the Map Task dialogues. For theory-neutral discourse segmentations of information-seeking dialogues, Flammia (Flammia and Zue, 1995) reports an average pairwise κ of .45 for five labelers and of .68 for the three most similar labelers.

An important issue in applying the κ coefficient is how one calculates the expected agreement using prior distributions of categories. We first calculated the prior probabilities for our data based simply on the distribution of SBEG versus non-SBEG labels for all labelers on one of the nine direction-giving tasks in this study, with separate calculations for the read and spontaneous versions. This task, which represented about 8% of the data for both speaking styles, was chosen because it was midway in planning complexity and in length among all the tasks. Using these distributions, we calculated κ coefficients for each pair of labelers in each condition for the remaining eight tasks in our corpus. The observed percentage of SBEG labels, prior distribution for SBEG, average of the pairwise κ scores, and standard deviations for those scores are presented in Table 2.

The average κ scores for group T segmenters indicate weak inter-labeler reliability. In contrast, average κ scores for group S are .8 or better, indicating a high degree of inter-labeler reliability. Thus, application of this somewhat stricter reliability metric confirms that the availability of speech critically influences how listeners perceive discourse structure.

The calculation of reliability for SBEG versus non-SBEG labeling in effect tests the similarity of linearized segmentations and does not speak to the issue of how similar the labelings are in terms of hierarchical structure. Flammia has proposed a method for generalizing the use of the κ coefficient for hierarchical segmentation that gives an upper-bound estimate on inter-labeler agreement.[5] We applied this metric to our segmentation data, calculating weighted averages for pairwise κ scores averaged for each task. Results for each condition, together with the lowest and highest average κ scores over the tasks, are presented in Table 3.

Table 2: Comparison of Average κ Coefficients for SBEGs

	% SBEG	P_E	Avg. κ	s.d.
READ				
Text alone	.38	.53	.56	.08
Text & Speech	.35	.55	.81	.01
SPON				
Text alone	.39	.52	.63	.04
Text & Speech	.35	.55	.80	.03

Table 3: Comparison of Weighted Average κ Coefficients and Extra for Each Condition Using Flammia's Metric

	% Weighted Average	Low	High
READ			
Text alone	0.51	.22	.67
Text & Speech	0.70	.48	.87
SPON			
Text alone	0.53	.19	.60
Text & Speech	0.74	.63	1.00

Once again, averaged scores of .7 or better for text-and-speech labelings, for both speaking styles, indicate markedly higher inter-labeler reliability than do scores for text-alone labelings, which averaged .51 and .53.

4.2 Intonational Features of Segments

4.2.1 Phrase Classes and Features

For purposes of intonational analysis, we take advantage of the high degree of agreement among our discourse labelers and include in each segment boundary class (SBEG, SF, and SCONT) only the phrases whose classification all subjects agreed

[5] Flammia uses a flexible definition of segment match to calculate pairwise observed agreement: roughly, a segment in one segmentation is matched if both its SBEG and SF correspond to segment boundary locations in the other segmentation.

Table 4: Acoustic-Prosodic Correlates of Consensus Labelings from Text Alone

	Max F0 (at HighF0)	Avg F0 (phrasal)	Max RMS (at HighF0)	Avg RMS (phrasal)	Rate	Preceding Pause	Subsequent Pause
SBEG							
Read	higher	higher	higher	higher		longer	shorter
Spon	higher	higher	higher	higher		longer	shorter
SCONT							
Read	lower	lower*†	lower*†	lower		shorter†	shorter†
Spon	lower				slower†	shorter†	shorter†
SF							
Read	lower	lower	lower	lower	faster*†	shorter	longer
Spon	lower	lower	lower	lower	faster†	shorter	longer

Table 5: Acoustic-Prosodic Correlates of Consensus Labelings from Text and Speech

	Max F0 (at HighF0)	Avg F0 (phrasal)	Max RMS (at HighF0)	Avg RMS (phrasal)	Rate	Preceding Pause	Subsequent Pause
SBEG							
Read	higher	higher	higher	higher		longer	shorter
Spon	higher	higher	higher	higher		longer	shorter
SCONT							
Read	lower	lower†		lower		shorter†	shorter†
Spon	lower				slower†	shorter†	shorter†
SF							
Read	lower	lower	lower	lower	faster*	shorter	longer
Spon	lower	lower	lower	lower	faster	shorter	longer

upon. We term these the CONSENSUS-LABELED PHRASES, and compare their features to those of all phrases not in the relevant class (i.e., non-consensus-labeled phrases and consensus-labeled phrases of the other types). Note that there were one-third fewer consensus-labeled phrases for text-alone labelings than for text-and-speech (see Table 1). We examined the following acoustic and prosodic features of SBEG, SCONT, and SF consensus-labeled phrases: f0 maximum and f0 average;[6] rms (energy) maximum and rms average; speaking rate (measured in syllables per second); and duration of preceding and subsequent silent pauses. As for the segmentation analyses, we compared intonational correlates of segment boundary types not only for group S versus group T, but also for spontaneous versus read speech. While correlates *have* been identified in read speech, they have been observed in spontaneous speech only rarely and descriptively.

[6]We calculated f0 maximum in two ways: as simple f0 peak within the intermediate phrase and also as f0 maximum measured at the rms maximum of the sonorant portion of the nuclear-accented syllable in the intermediate phrase (HIGH F0 in the ToBI framework (Pitrelli, Beckman, and Hirschberg, 1994)). The latter measure proved more robust, so we report results based on this metric. The same applies to measurement of rms maximum. Average f0 and rms were calculated over the entire intermediate phrase.

4.2.2 Global Intonational Correlates

We found strong correlations for consensus SBEG, SCONT, and SF phrases for all conditions. Results for group T are given in Table 4, and for group S, in Table 5.[7]

Consensus SBEG phrases in all conditions possess significantly higher maximum and average f0, higher maximum and average rms, shorter subsequent pause, and longer preceding pause. For consensus SCONT phrases, we found some differences between read and spontaneous speech for both labeling methods. Features for group T included significantly lower f0 maximum and average and lower rms maximum and average for read speech, but only lower f0 maximum for the spontaneous condition. Group S features for SCONT were identical to group T except for the absence of a correlation for maximum rms. While SCONT phrases for both speaking styles exhibited significantly shorter preceding and subsequent pauses than other phrases, only the

[7]T-tests were used to test for statistical significance of difference in the means of two classes of phrases. Results reported are significant at the .005 level or better, except where '*' indicates significance at the .03 level or better. Results were calculated using one-tailed t-tests, except where '†' indicates a two-tailed test.

Table 6: Acoustic-Prosodic Change from Preceding Phrase for Consensus Labelings from Text and Speech

	Max F0 Change (at HighF0s)	Max DB Change (at HighF0s)	Rate Change
SBEG versus SCONT			
Read	increase	increase	
Spon	increase	increase	
SCONT versus SF			
Read	increase*	increase*	
Spon		increase	
SBEG versus SF			
Read	increase	increase	
Spon	increase	increase	decrease*†

spontaneous condition showed a significantly slower rate. For consensus SF phrases, we again found similar patterns for both speaking styles and both labeling methods, namely lower f0 maximum and average, lower rms maximum and average, faster speaking rate, shorter preceding pauses, and longer subsequent pauses.

While it may appear somewhat surprising that results for both labeling methods match so closely, in fact, correlations for text-and-speech labels presented in Table 5 were almost invariably statistically stronger than those for text-alone labels presented in Table 4. These more robust results for text-and-speech labelings occur even though the data set of consensus labels is considerably larger than the data set of consensus text-alone labelings.

4.2.3 Local Intonational Correlates

With a view toward automatically segmenting a spoken discourse, we would like to directly classify phrases of all three discourse categories. But SCONT and SF phrases exhibit similar prominence features and appear distinct from each other only in terms of timing differences. A second issue is whether such classification can be done 'on-line.' To address both of these issues, we made *pairwise* comparisons of consensus-labeled phrase groups using measures of *relative change* in acoustic-prosodic parameters over a local window of two consecutive phrases. Table 6 presents significant findings on relative changes in f0, loudness (measured in decibels), and speaking rate, from prior to current intermediate phrase.[8]

First, note that SBEG is distinguished from both SCONT and SF in terms of f0 change and db change from prior phrase; that is, while SBEG phrases are distinguished on a variety of measures from all other phrases (including non-consensus-labeled phrases) in Table 5, this table shows that SBEGs are also distinguishable directly from each of the other consensus-labeled categories. Second, while SCONT and SF appear to share prominence features in Table 5, Table 6 reveals differences between SCONT and SF in amount of f0 and db change. Thus, in addition to lending themselves to on-line processing, local measures may also capture valuable prominence cues to distinguish between segment-medial and segment-final phrases.

5 Conclusion

Although this paper reports results from only a single speaker, the findings are promising. We have demonstrated that a theory-based method for discourse analysis can provide reliable segmentations of spontaneous as well as read speech. In addition, the availability of speech in the text-and-speech labeling method led to significantly higher reliability scores. The stronger correlations found for intonational features of the text-and-speech labelings suggest not only that discourse labelers make use of prosody in their analyses, but also that obtaining such data can lead to more robust modeling of the relationship between intonation and discourse structure.

The following preliminary results can be considered for incorporation in such a model. First, segment-initial utterances differ from medial and final utterances in both prominence and rhythmic properties. Segment-medial and segment-final utterances are distinguished more clearly by rhythmic features, primarily pause. Finally, all correlations found for global parameters can also be computed

[8] We present results only for text-and-speech labelings; results for text-alone were quite similar. Note that 'increase' means that there is a significantly greater increase in f0, rms, or rate from prior to current phrase for category 1 than for category 2 of the comparison, and 'decrease' means that there is a significantly greater decrease. T-tests again were one-tailed unless marked by †, and significance levels were .002 or better except those marked by *, which were at .01 or better.

based on relative change in acoustic-prosodic parameters in a window of two phrases.

Ongoing research is addressing the development of automatic classification algorithms for discourse boundary type; the role of prosody in conveying hierarchical relationships among discourse segments; individual speaker differences; and discourse segmentation methods that can be used by naive subjects.

References

Avesani, Cinzia and Mario Vayra. 1988. Discorso, segmenti di discorso e un' ipotesi sull' intonazione. In *Corso di stampa negli Atti del Convegno Internazionale "Sull'Interpunzione"*, Vallecchi, Firenze, Maggio, pages 8–53.

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

Beckman, Mary and Janet Pierrehumbert. 1986. Intonational structure in Japanese and English. *Phonology Yearbook*, 3:15–70.

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

Brubaker, R. S. 1972. Rate and pause characteristics of oral reading. *Journal of Psycholinguistic Research*, 1(2):141–147.

Butterworth, B. 1975. Hesitation and semantic planning in speech. *Journal of Psycholinguistic Research*, 4:75–87.

Carberry, Sandra. 1990. *Plan Recognition in Natural Language Dialogue*. MIT Press, Cambridge MA.

Carletta, Jean C. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2), To appear.

Cohen, Robin. 1984. A computational theory of the function of clue words in argument understanding. In *Proceedings of the 10th International Conference on Computational Linguistics*, pages 251–255, Stanford.

Flammia, Giovanni and Victor Zue. 1995. Empirical evaluation of human performance and agreement in parsing discourse constituents in spoken dialogue. In *Proceedings of EUROSPEECH-95*, Volume 3, pages 1965–1968. Madrid.

Grosz, B. and J. Hirschberg. 1992. Some intonational characteristics of discourse structure. In *Proceedings of the 2nd International Conference on Spoken Language Processing*, pages 429-432, Banff, October.

Grosz, Barbara. 1978. Discourse analysis. In D. Walker, editor, *Understanding Spoken Language*. Elsevier, pages 235–268.

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

Hearst, Marti A. 1994. *Context and Structure in Automated Full-Text Information Access*. Ph.D. thesis, University of California at Berkeley. Available as Report No. UCB/CSD-94/836.

Hinkelman, E. and J. Allen. 1989. Two constraints on speech act ambiguity. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 212–219, Vancouver.

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

Isard, Amy and Jean Carletta. 1995. Transaction and action coding in the map task corpus. Research paper HCRC/RP–65, March, Human Communication Research Centre, University of Edinburgh, Edinburgh.

Iwańska, Łucja, Douglas Appelt, Damaris Ayuso, Kathy Dahlgren, Bonnie Glover Stalls, Ralph Grishman, George Krupka, Christine Montgomery, and Ellen Riloff. 1991. Computational aspects of discourse in the context of Muc-3. In *Proceedings of the Third Message Understanding Conference (Muc-3)*, pages 256–282, San Mateo, CA, Morgan Kaufmann, May.

Lehiste, I. 1979. Perception of sentence and paragraph boundaries. In B. Lindblom and S. Oehman, editors, *Frontiers of Speech Research*. Academic Press, London, pages 191–201.

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

Levy, Elena. 1984. *Communicating Thematic Structure in Narrative Discourse: The Use of Referring Terms and Gestures*. Ph.D. thesis, The University of Chicago, Chicago, June.

Linde, C. 1979. Focus of attention and the choice of pronouns in discourse. In T. Givon, editor, *Syntax and Semantics, Vol. 12: Discourse and Syntax*. The Academic Press, New York, pages 337–354.

Litman, Diane and Julia Hirschberg. 1990. Disambiguating cue phrases in text and speech. In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 251–256, Helsinki, August.

Lochbaum, Karen. 1994. *Using Collaborative Plans to Model the Intentional Structure of Discourse*. Ph.D. thesis, Harvard University. Available as Tech Report TR-25-94.

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

Passonneau, R. and D. Litman. 1993. Feasibility of automated discourse segmentation. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 148-155, Columbus.

Passonneau, Rebecca J. and Diane J. Litman. To Appear. Empirical analysis of three dimensions of spoken discourse: Segmentation, coherence and linguistic devices. In E. Hovy and D. Scott, editors, *Burning Issues in Discourse*. Springer Verlag.

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

Pitrelli, John, Mary Beckman, and Julia Hirschberg. 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the 3rd International Conference on Spoken Language Processing*, volume 2, pages 123–126, Yokohama.

Reichman, R. 1985. *Getting Computers to Talk Like You and Me: Discourse Context, Focus, and Semantics*. Bradford. The Massachusetts Institute of Technology, Cambridge.

Reynar, Jeffrey C. 1994. An automatic method of finding topic boundaries. In *Proceedings of the Student Session of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 331–333. Las Cruces, NM.

Rotondo, John A. 1984. Clustering analysis of subjective partitions of text. *Discourse Processes*, 7:69–88.

Schubert, L. K. and C. H. Hwang. 1990. Picking reference events from tense trees. In *Proceedings of the Speech and Natural Language Workshop*, pages 34–41, Hidden Valley PA. DARPA.

Siegel, S. and Jr. Castellan, N. John. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York, second edition.

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

Song, F. and R. Cohen. 1991. Tense interpretation in the context of narrative. In *Proceedings of the 9th National Conference of the American Association for Artificial Intelligence*, pages 131–136.

Swerts, M., R. Gelyukens, and J. Terken. 1992. Prosodic correlates of discourse units in spontaneous speech. In *Proceedings of the International Conference on Spoken Language Processing*, pages 421–428, Banff, Canada, October.

Swerts, Marc. 1995. Combining statistical and phonetic analyses of spontaneous discourse segmentation. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, volume 4, pages 208–211, Stockholm, August.

Swerts, Marc and Mari Ostendorf. 1995. Discourse prosody in human-machine interactions. In *Proceedings ESCA Workshop on Spoken Dialogue Systems: Theories and Applications*, pages 205–208, Visgo, Denmark, May/June.

Webber, B. 1988. Discourse deixis: Reference to discourse segments. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 113–122, Buffalo.

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

Yarowsky, David. 1991. Personal communication, December.