SOME INTONATIONAL CHARACTERISTICS OF DISCOURSE STRUCTURE*

Barbara Grosz
Division of Applied Sciences
Harvard University
Cambridge MA 02138

Julia Hirschberg
2D-450, AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill NJ 07974-0636

ABSTRACT

This paper reports on a study of the relationship between acoustic-prosodic variation and discourse structure, as determined from an independent model of discourse. We present results of two pilot studies. Our corpus consisted of three AP news stories recorded by a professional speaker. Discourse structure was labeled by subjects either from text alone or from text (with all orthographic markings except sentence-final punctuation removed) and speech, following Grosz & Sidner 1986; average inter-labeler agreement for structural elements varied from 74.3%-95.1%, depending upon feature. These elements of global structure, together with elements of local structure such as parentheticals and attributive tags, were correlated with variation in intonational and acoustic features such as pitch range, contour, timing, and amplitude. We found statistically significant associations between aspects of pitch range, amplitude, and timing with features of global and local structure both for labelings from text alone and for labelings from speech. We further found that global and local structures can be reliably identified from acoustic and prosodic features with (cross-validated) success rates of 86-97%. (This paper appears in the Proceedings of the Second International Conference on Spoken Language Processing, Banff, October, 1992, pp. 429–432.)

I. INTRODUCTION

Most computational theories of discourse agree that utterances in a discourse group together into segments, and that the deter-

*This research was partly supported by NSF grant #IRI-9009018.

mination of discourse meaning depends crucially on identifying the ways these segments fit together. However, different theories make different claims about the basis of discourse structure; proposals include coherence relations [8, 10, 11, 14], syntactic features [13], and intentions [6]. In addition, discourse segment boundaries do not always align with paragraph boundaries or other orthographic markers in text. As a result, attempts to apply theories of discourse structure have encountered difficulty with apparent ambiguities in the structure of a particular discourse. Furthermore, there have been no systematic studies of human labeling of discourse segmentation. Thus, one goal of our pilot studies was to devise a set of instructions that would permit consistency in segmentation across different labelers and different texts. A second goal of these studies was to identify intonational features that were strongly correlated with discourse structural elements. Several studies suggest that discourse structure is signalled by intonational variation. Variation in pitch range, timing, and amplitude have all been studied as potential correlates of structural variation [9, 4, 15, 1, 2], with some success. However, a weakness in such studies has been the lack of independent analyses of the structure of the discourses under consideration, using a general theory of discourse structure. We addressed this problem by analyzing intonational and acoustic features of the discourse structures that were identified by the consensus labelings of our subjects, following Grosz & Sidner 1986's model of discourse structure [6] (hereafter, G&S). A third goal was to examine the conjecture that spoken language provides information that enables a listener to identify the structure of a discourse intended by a speaker from among several possible structurings. To this end, we compared discourse labelings by subjects who labeled from text alone with labelings made from both text and speech.

We examined prosodic features of three AP news stories (hereafter, AP1, AP2 and AP5), which had been recorded by a professional newscaster from texts available to us. The texts averaged about 450 words in length and the recordings averaged about three and one-half minutes. AP5 (approximately 550 words and four minutes long) was labeled for discourse features by seven labelers; four labeled from text alone (Group T) and three from text and speech (Group S). Analysis of these labelings with respect to the acoustic-prosodic features observed for the recorded text were presented in [7]; we summarize these results below. AP1 and AP2 were labeled from text alone by three of the labelers who had labeled AP5 from text. In this paper we report additional results for AP5 as well as results for AP1 and AP2.

II. DISCOURSE STRUCTURES

In [7] we described the development of a set of labeling instructions based on G&S's model of discourse structure [6] for guiding labelers in segmenting the news stories and identifying elements of local structure. According to this model, discourse structure includes at least three distinct components. The utterances composing the discourse divide into segments forming the *linguistic structure*. The embedding relationships among segments reflect changes in the *attentional state* component during the discourse; this component represents the entities and attributes that are salient during a particular portion of the discourse. Changes in attentional state, and hence the discourse segment structure, depend on the *intentional structure*, a structure of the purposes or intentions underlying the discourse. The intentional structure thus plays a central role in discourse structure: the determination of discourse segmentation depends on identification of discourse intentions and relationships between them. Hence, we describe it briefly here; additional details may be found in [6, 5].

According to G&S, each discourse segment has an underlying purpose intended by the speaker/writer to be recognized by the listener/reader, the *Discourse Segment Purpose* (DSP). Each DSP contributes to the overall *Discourse Purpose* (DP) of the discourse. So, a discourse might have as its DP the intention that the listener be informed that there was a plane accident, and individual segments forming that discourse might have as their DSP's intentions that the listener be informed that the plane lost a piece of its tail (an intention contributing information about the accident) and that the passengers were upset (an intention contributing information about the effect of this event). DSPs may in turn comprise other intentions and relations between them. DSPs a and b may be related to one another in two ways: a *dominates* b if the DSP of a is partially fulfilled by the DSP of b. Segment a *satisfaction-precedes* b if the DSP of a must be achieved in order for the DSP of b to be successful.

At the global level, our analyses focused on phrases identified by subjects as beginning (*SBEG*) or ending (*SF*) discourse segments and on the embedding relationship between such phrases and the phrases that immediately preceded or followed them. At the local level, we examined five types of constituents: parentheticals, direct quotations and their tags, indirect reported speech, and speaker attributions for reported speech. We asked both Group T and Group S labelers to mark parentheticals, because these are not always disambiguated orthographically. In addition, we asked Group S to mark direct quotations. Tags, indirect speech, and speaker attributions for indirect speech were identified independently by the authors from the text.

We found considerable agreement in labelings of global structure for both our groups, although no two segmentations were identical. For AP5, we examined labelings of segment beginnings and endings for all seven labelers: for the binary decision of whether a given phrase began a new segment, there were no statistically significant differences among six of the seven.[1] For the question of whether or not a phrase ended a segment, the seven labelers fell into two groups, with no signifi-

[1] We used Cochran's Q to test for significance.

cant difference within each group; we hypothesize that each may have settled upon a distinct but plausible interpretation of the text's structure. While we had anticipated finding fewer differences among members of Group S than among Group T labelers, consistency across Group S labelers was no greater than consistency among all members of the two groups, for labelings of either segment beginnings or endings.

For each of these decisions, we also examined how often each individual labeler agreed with the majority judgment; means and standard deviations for percentage agreement with majority view are given below. For the segment beginning decision for phrases in AP5, our seven labelers averaged 77.3% (s=4.6) agreement with the majority decision; however, when one labeler is removed, the remaining six improve their mean percent agreement score to 89.5% (s=6.4). The fact that this labeler is a member of Group S is reflected in the observation that the consistency within Group T appears greater than for Group S: for Group T mean percent agreement for each labeler with consensus is 95.1% (s=5.7) while for Group S it is 87.7% (s=9.3). Agreement among the three labelers of AP1 averaged 85.4% agreement with majority (s=15.7); for AP2 it was 91.7% (s=8.3). Agreement for segment endings averaged 75.7% for AP5 (s=7.4), 76.5% for AP1 (s=21.2), and 74.3% (s=4.4) for AP2.

So, while there is considerable labeling agreement for both segment beginnings and segment endings, agreement is higher for segment beginnings, ranging from 85.4-91.7% over the three news stories, with 77.3% agreement for all seven labelers on AP5 and 89.5% for six of the seven. And many of the phrases for which labelers disagreed over the segment beginning decision fell into two categories: (1) utterances that might have initiated (or by themselves formed) small separate segments and were thus classified as segment medial by some labelers and as SBEG by others; (2) utterances classified as SBEG by some labelers but merely as immediately following an SF phrase by others (a segment medial pop, *SMP*). In the latter case, all of the labelers agreed that there was a discourse break of some kind, but they disagreed about the relationship of the utterance in question to the immediately (linearly) preceding segment. Including these all as SBEG would improve agreement scores further.

III. ACOUSTIC-PROSODIC ANALYSIS

To identify intonational features in the read speech, we labeled the speech for accentuation and phrasing, according to Pierrehumbert's [12] theory of English intonation, using WAVES speech analysis software. We used as our primary unit of analysis Pierrehumbert's intermediate phrase. For each phrase, we calculated values for pitch range, calculated for each minor phrase from the fundamental frequency (f0) maximum occurring within an accented syllable in the phrase;[2] amount of f0

[2]Results from a more conservative measurement

change between phrases, f0(phrase[i])/f0(phrase[i+1]); amplitude, energy maximum within the vowel of the syllable containing the phrase's f0 peak; difference in intensity from prior phrase, measured in decibels (db); contour type and type of nuclear accent, identified in Pierrehumbert's notation; speaking rate, measured in syllables per second (sps); and pausal duration between intermediate phrases and between intonational phrases. Each of these features was then examined as a potential predictor of discourse structure. Results for discourse features labeled from text were compared with those labeled from text and speech, to see whether there were acoustic-prosodic cues which influence Group S labelers to deviate from patterns observed for Group T.

Study I

Results for the analysis of discourse and acoustic-prosodic features in one story, AP5, are summarized in Table 1. For this analysis, we examined consensus labelings (i.e. those on which every member of a group agreed) from Group T with those from Group S for direct quotations, tags, indirect reported speech and attributions, parentheticals, and segment boundaries. In these analyses, we controlled for phrasal position where ANOVAs performed on the data indicated that phrasal position was significantly correlated with the feature under analysis. Significance for the data presented in Table 1 was determined via t-tests, and all results presented were significant at the .05 level or better. A more detailed account appears in [7].

In Table 1, a '+' indicates the row's discourse structural element is characterized by higher values for the column's intonational feature; '-' indicates that the structural element is characterized by lower values for the intonational feature. For example, '-' in the 'Pitch Range' column for parentheticals indicates that parenthetical phrases were uttered in a pitch range that was significantly smaller than the pitch ranges of non-parentheticals.

As shown in Table 1, quoted phrases for Group T were, in general, uttered in a larger pitch range and with less increase in intensity than other phrases; these results are for quote-initial phrases, compared with other phrases in similar utterance position. Also, quote-final phrases were produced with a pronounced drop in intensity compared with other utterance-final phrases. The phrases Group S identified as direct quotations differed significantly only in pitch range from other phrases. Thus, our speaker signaled direct quotes by an expanded pitch and range and hearers apparently perceived this cue.

We found that parentheticals identified by Group T were uttered in a compressed pitch range and that that range represented a smaller than average change in range and in intensity from prior phrase. Those identifed by Group S were even lower in range than those identified by Group T and exhibited an even more pronounced decrease in pitch and intensity, providing some indication that Group S was using such variation, in addition to more semantic cues, to identify parentheticals. Group S

at the amplitude maximum within that syllable were similar.

Table 1: Intonational Correlates of Discourse Features

	Intonational Features						
Discourse Features	Pitch Range	Pitch Range Change	Rms	Db Change	Prec Pause	Subs Pause	Rate
T:Direct quotes	+			-			
S:Direct quotes	+						
T:Parentheticals	-	-		-			
S:Parentheticals	-	-		-			+
T:SF						+	
S:SF						+	
T:SMP	+				+		
S:SMP	+				+		
T:SBEG	+		+			-	-
S:SBEG					+	-	
T:SBEG+SMP	+					-	
S:SBEG+SMP	+				+	-	

parentheticals also were uttered significantly more rapidly than other phrases.

For global structure, we again found much similarity between intonational features correlated with Group T-identified discourse elements and those correlated with discourse features identified by Group S. However, for global structures we did *not* find that the intonational features associated with discourse features labeled by Group S differed markedly from those labeled by Group T. For SBEG phrases, we found pitch range, amplitude, rate and subsequent pause all were significantly correlated for Group T. However, only preceding and subsequent pause variation distinguished phrases identified as SBEG by Group S. As discussed in [7], this Group S result was unexpected and inconsistent with previous results. However, because of the "topic shifting" similarity of SBEG and SMP utterances, this led us to examine both SMP phrases, and a superordinate category, SBEG+SMP. For SMP, there is a significant effect for pitch range and for preceding pause for Group T, and similar effects for Group S. For SBEG+SMP,[3] identified by Group T, there are significant effects for pitch range and subsequent pause; for Group S significant effects were found for pitch range, subsequent pause and preceding pause. So, SBEG does appear to be signalled by expanded range and timing, as we had anticipated. Finally, for SF phrases for both Group T and Group S, we found a sin-

[3] SBEG phrases significantly outnumber SMP phrases when we collapse categories, so our results do not arise from the latter dominating the former.

gle intonational correlate, subsequent pause. These findings confirm previous work by [4, 9, 15] that pitch range and timing variation are important in signaling topic structure, and demonstrate that these relationships hold when topic structure has been independently determined from consensus subject labeling, which is based upon an independently-motivated theory of discourse.

Study II

Our first experiment identified associations between individual local and global discourse features and, in most cases, *multiple* acoustic-prosodic features. These results support the common observation that a given acoustic-prosodic feature can be associated with variation in multiple, distinct discourse phenomena. However, our initial findings do not tell us how these complex relationships between discourse and acoustic-prosodic features might be modeled. Is variation in one acoustic-prosodic feature either necessary or sufficient to signal a discourse feature? Do all phrases exhibiting a particular discourse characteristic exhibit similar acoustic-prosodic characteristics?

To test these possibilities, we next examined consensus text-based labelings for AP1, AP2, and AP5 from three labelers and also the union of labelers judgments for discourse phenomena. We considered the acoustic-prosodic features discussed above, as well as the mean and, where applicable, initial value for each of these features for each story as a whole (e.g. f0 for initial phrase in story, mean f0 for all phrases in story or all utterance-initial phrases, where appropriate), to control for possible variability of recording situation from story to story. This time we employed Classification and Regression Tree Analysis (CART) [3] to produce decision trees automatically from these feature values; we present results in terms of CART's cross-validated success estimates.[4] CART provided a simple automatic method of identifying reliable associations between acoustic-prosodic features and global and local discourse features for our relative small corpus.

This study confirmed that acoustic-prosodic factors can be used to predict labelers' consensus decisions on both global and local aspects of discourse structure. For example, automatically generated prediction trees distinguish consensus SBEG from other phrases for combined AP1, AP2 and AP5 in 91.5% of cases, using only a simple combination of constraints on duration of preceding pause ($>$ 647 msec.) and pitch range ($<$ 276 Hz.). They distinguish SF from other phrases in 92.5% of cases, from information about subsequent pause ($>$ 913 msec.), amount of f0 change from prior phrase ($<$ 93%), and overall rate for the story ($>$ 4.76 sps). In both cases, all phrases successfully identified by this procedure shared similar acoustic-prosodic features. So, positive nodes can be repre-

[4] In the CART implementation employed here, each decision tree is grown on 90% of the data and tested on the rest five times; results are then averaged to identify a reliable subtree and its likely success rate. Independent testing on other data sets has found that such estimates are reliable to within 1-2 percentage points of the estimated success rate.

sented as a set of simple constraints on a few variables.

Aspects of consensus labeled local structure are identified with similar success with equally simple prediction trees: Attributive tags can be identified with 96.9% success as a set of constraints on rate (< 3.36 sps) and amount of change in f0 from prior phrase (< 79%). Phrases beginning direct quotations are distinguished from other phrases in 86.4% of cases, from preceding pause (> 519 msec.), amplitude (rms <3416), and rate (> 5.10 sps). And phrases beginning indirect quotations were distinguishable from other phrases in 88.5% of cases, in terms of pitch range (> 270 Hz.) and amount of change in range from prior phrase (>62%), and change in intensity from previous phrase (> -4.26 db).

However, phrases marked by our labelers as parentheticals provide a more complex picture of the relationship between discourse and intonation: a larger number of features are involved in distinguishing parenthetical from non-parenthetical phrases in our corpus, and not all parentheticals share common acoustic-prosodic characteristics. In general, parentheticals can be distinguished from other phrases in 89.2% of cases, on the basis of absolute f0, amount of f0 change from previous phrase, type of nuclear accent, rate, and preceding pause. However, while all parenthetical phrases successfully identified in this corpus are characterized as uttered in a relatively low pitch range (peak <235 Hz), exhibiting relatively little f0 change from prior phrase (.86<f0change<1.04), and sharing the same set of three nuclear accent types, one group of parentheticals are uttered very rapidly (>6.08 sps) while those uttered at slower rates are much lower in pitch range (<196 Hz.). So, while most discourse features appear to be predicted in this corpus from simple constraints on two or three acoustic-prosodic features, these data suggest that there may be trade-offs in variation of different features, and that more complex models may be necessary to model these.

CONCLUSIONS

Our pilot studies provide evidence that discourses can be segmented reliably by labelers given instructions based on Grosz & Sidner 1986's theory of discourse structure. For segment beginnings and endings, mean percent agreement among our labelers was better than 74% in all conditions. Agreement with majority on segment beginnings ranged from a low of 77.3% for all labelers on one story to 89.5% for six of the seven on that story and 91.7% for three labelers on another.

Our experiments also support the hypothesis that discourse structure is marked intonationally; we found statistically significant correlations between specific discourse structures determined independently of linguistic form and acoustic-prosodic features. In a pilot study that examined seven discourse labelings of one news story in conjunction with a recorded version of that story, we found the following both for those who labeled from text and those who labeled from speech: Phrases beginning discourse segments (including SBEG and SMP) were uttered in a larger pitch range and followed by shorter pauses than other utterance initial phrases. Phrases ending segments were followed by longer pauses than other utterance-final phrases. Phrases initiating direct quotations were uttered in a larger pitch range than other utterance-initial phrases. Parentheticals were uttered in a compressed range, and exhibited less change in f0 and in intensity from the prior phrase than non-parenthetical phrases.

Our results also demonstrate that aspects of discourse structure can be predicted reliably on the basis of acoustic-prosodic features, although the relationship between structure and intonational features is sometimes a complex one — a given discourse structural feature may be signaled by several intonational variables, which may or may not be independent; thus, the values of acoustic-prosodic features associated with a given discourse feature may vary, depending upon the values of other features. From our pilot study of three labelers' consensus labels for three news stories, we found that phrases beginning and ending discourse segments could be predicted with better than 90% (cross-validated) success, and that aspects of local structure could also be predicted with considerable success. Tags are predicted in this data with 96.9% success, phrases beginning direct quotations with 86.4% success, those beginning indirect quotations with 88.5% success, and parentheticals with 89.2% success. While in most cases those phrases that were reliably predicted could be identified by a few simple constraints on acoustic-prosodic features, predictions for parentheticals as well as the data not correctly predicted suggest that more complex models may be needed to model the relationships between intonational and discourse features.

To confirm these preliminary results, we are currently expanding our analysis to include more texts, additional speakers, and additional labelers. We also will examine the difference between professional and non-professional read speech and will examine acoustic-prosodic characteristics of additional discourse features, such aslevel of embedding of discourse segments.

References

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

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

[3] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth & Brooks, Pacific Grove CA, 1984.

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

[5] B. J. Grosz and C. L. Sidner. Plans for discourse. In P. R. Cohen, J. L. Morgan, and M. E. Pollack, editors, *Intentions in Communication*. MIT Press, 1990.

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

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

[8] J. Hobbs. Coherence and coreference. *Cognitive Science*, 3(1):67–90, 1979.

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

[10] W.C. Mann and S.A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[11] Johanna D. Moore and Cécile L. Paris. Planning text for advisory dialogues. In *Proceedings of the 27th Annual Meeting*, pages 203–211, Vancouver, June 1989. Association for Computational Linguistics.

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

[13] L. Polanyi. A formal model of the structure of discourse. *Journal of Pragmatics*, 12, 1988.

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

[15] K. Silverman. *The Structure and Processing of Fundamental Frequency Contours*. PhD thesis, Cambridge University, Cambridge UK, 1987.