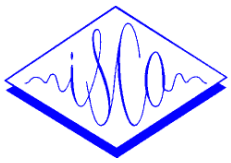

ESCA Workshop on Prosody
Lund, Sweden
September 27-29, 1993

Studies of Intonation and Discourse

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

ABSTRACT

Research on intonation and discourse falls into two major categories: work on the intonational correlates of discourse structure and work on accent and information status. In both categories, problems of specifying an adequate and independently motivated discourse model hinder evaluation of results and generalization across experiments. Also, much work remains to be done on combining these results into general models capturing the mapping between intonational features and discourse features.

INTRODUCTION

Most research on intonation and discourse to date has fallen into one of two categories: investigations of the intonational correlates of topic structure or studies of the relationship between information status and intonational prominence. Much of this work has involved empirical experimentation or corpus-based research. For the latter, the need for large, shareable, prosodically labeled corpora is viewed as increasingly important, given the labor-intensive natural of corpora labeling. To promote the development of such corpora, some efforts have been made to agree upon common labeling standards, such as the current TOBI standard for Standard American English which has recently been proposed (Silverman et al.,1992). While the term 'discourse' might seem to suggest spontaneous conversation, in fact, most work in this area has defined discourse more generally as 'utterances in context'; so, monologues, elicited speech, read speech, and radio speech, have been more frequently examined than natural dialogue.

PROSODIC CUES TO DISCOURSE STRUCTURE

Most researchers who work on discourse accept that it is structured into segments; disagreement arises primarily over the nature of the larger units into which segments are grouped, and the relationship among individual segments. To date, most studies of prosody and discourse structure have focussed on how intonational and acoustic variation signals segment boundaries and conveys larger 'topic' structures.

The notion that discourse structure is signalled by variation in intonational features such as pitch range, timing, and amplitude — and probably variation in some combination of these features — has been widely believed for years. However, there has been surprisingly little empirical testing of this belief. A major problem is noted by Brown et al. (Brown et al.,1980, p. 27) in discussing some production studies designed to elicit acoustic and intonational cues to discourse structure : "... until an independent theory of topic-structure is formulated, much of our argument in this area is in danger of circularity." In fact, most speech-based empirical studies have assumed a particular structure for the discourses they examine or use as stimuli and have then looked for acoustic-intonational indicators of these assumed structures, usually using the experimenter's intuitive notions about changes in topic and topic-subtopic relations. Alternatively, they have used lexcal phenomena believed to be constrained by discourse structure (such as pronominal forms) or supposed

to explicitly indicate structure (such as cue phrases) as indicators of structure, even when these hypotheses themselves remain to be tested.

One of the features most frequently mentioned as important to conveying some kind of 'topic structure' in discourse is PITCH RANGE (the distance between the maximum of the FUNDAMENTAL FREQUENCY (f0) for the vowel portions of accented syllables in the phrase and the speaker's *baseline*, defined for each each speaker as the lowest point reached in normal speech over all . In a study of speakers reading a story, Brown et al. (Brown et al.,1980) found that subjects typically started new topics relatively high in their pitch range and finished topics by compressing their range; they hypothesized that internal structure within a topic was similarly marked. Lehiste (Lehiste,1975) had reported similar results earlier for single paragraphs. Silverman (Silverman,1987) found that manipulation of pitch range alone, or range in conjunction with pausal duration between utterances, could enable subjects to disambiguate utterances that were intuitively potentially structurally ambiguous reliably; for example, he used a small pitch range to signal either continuation or ending of a topic or quotation, and expanded range to indicate topic shift or quotation continuation. Avesani and Vayra (Avesani and Vayra,1988) also found variation in range in productions by a professional speaker which appear to correlate with topic structure, and Ayers (Ayers,1992) found that pitch range appears to correlate more closely with hierarchical topic structure in read speech than in spontaneous speech. Swerts et al (Swerts et al.,1992) also found that f0 scaling was a reliable indicator of discourse structure in spoken instructions, although the structures tested were quite simple. Duration of pause between utterances or phrases has also been identified as an indicator of topic structure in (Lehiste,1979; Chafe,1980; Brown et al.,1980; Silverman,1987; Avesani and Vayra,1988; Swerts et al.,1992; Passoneau and Litman,1993). Brown et al. found that longer, 'topic pauses' (.6-.8 sec.) marked major topic shifts (Brown et al.,1980, 57). Passoneau & Litman (Passoneau and Litman,1993) also found that presence of pause was a good indicator of segment boundaries in Chafe's pear stories, when tested against their own subjects' segmentations of these stories. Another aspect of timing, speaking rate, was found by Lehiste (Lehiste,1980) and by Butterworth (Butterworth,1975) to be associated with perception of text structure: both found that utterances beginning segments exhibited slower rates those completing segments were uttered more rapidly. Amplitude was also noted by Brown et al. (Brown et al.,1980) as a signal of topic shift; they found that amplitude appeared to rise at the start of a new topic and fall at the end. Finally, contour type has been noted (Brown et al.,1980; Swerts et al.,1992) as a potential correlate of topic structure. In particular, (Hirschberg and Pierrehumbert,1986) suggested that so-called 'downstepped' contours (In which one or more pitch accents which follow a complex accent are uttered in a compressed range, producing a 'stairstep' effect.) commonly appear either at the beginning or the ending of topics. Empirical studies reported in (Swerts et al.,1992) showed that 'low' vs. 'not-low' boundary tones were good predictors of topic endings vs. continuations.

Recently, Hirschberg & Grosz (Hirschberg and Grosz,1992; Grosz and Hirschberg,1992) have addressed the problem of acoustic-prosodic correlates of discourse structure, inspired by the need to test potential correlates against an independent notion of discourse structure, as noted by (Brown et al.,1980). We looked at pitch range, aspects of timing and contour, and amplitude to see how well they predicted discourse segmentation decisions made by subjects using instructions based on the Grosz and Sidner 1986 (Grosz and Sidner,1986) model of discourse structure. Our corpus consisted of three AP news stories previously recorded by a professional speaker. Subjects labeled either from text alone (Group T) or from text (with all orthographic markings except sentence-final punctuation removed) and speech (Group S); average inter-labeler agreement for structural elements varied from 74.3% to 95.1% for subject decisions such as where segments began and ended. Decisions subjects all agreed upon were then correlated with variation in the acoustic-prosodic features mentioned above, as well as features such as change in f0 from preceding phrase, subsequent as well as preceding pause, absolute and relative amplitude, and type of nuclear pitch accent. We found statistically significant associations between aspects of pitch range, amplitude, and timing with segment beginnings and segment endings both for labelings from text alone and for labelings from speech.

For phrases labelled as beginning segments (We collapsed this category with phrases identified as SEGMENT MEDIAL POPS, those phrases which immediately followed a segment final phrase.) identified by Group T, we found significant effects for pitch range and subsequent pause; for Group S significant effects were found for pitch range, subsequent pause and preceding pause. So, segment beginnings do appear to be signalled by expanded range and timing, as previous studies had suggested. For phrases ending a segment, for both Group T and Group S, we found a single intonational correlate, subsequent pause; longer subsequent pauses are significantly associated with segment-final phrases. These findings confirm those noted above 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.

We further found that segment beginnings and endings could be reliably identified from the same acoustic and prosodic features with considerable success. For example, automatically generated prediction trees distinguished segment beginnings from other phrases 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 distinguished segment-final phrases from other phrases in 92.5% of cases, using subsequent pause (> 913 msec.), amount of f0 change from prior phrase (< 93%), and overall rate for the story (> 4.76 sps).

While these initial studies were encouraging, they also revealed some problems with our experimental design: First, due to the speech corpus we employed, we had no access to the speaker's own intentions with respect to structure at the time of recording. Inferring these intentions from labelers' performance on text was much too indirect to be satisfying. The subject matter of the recordings, news stories, proved unexpectedly difficult to segment for our subjects. We also felt we had inadequate means to compare inter-labeler segmentation; clearly segment beginnings and endings only capture part of what is going on in a discourse. We will be addressing these problems in the next phase of the study.

INTONATIONAL PROMINENCE AND INFORMATION STATUS

How speakers decide which words to accent and which to deaccent is an open research question. While syntactic structure was once believed to determine accent placement, it is now generally held that syntactic, semantic, and discourse/pragmatic factors are all involved in accent decisions (Bolinger,1972; Bardovi-Harlig,1983). Word class, grammatical function, syntactic constituency, and surface position are still believed to influence accent location (Ladd,1979b; Erteschik-Shir and Lappin,1983), and there are some recent empirical results supporting this (Altenberg,1987; Terken and Hirschberg,1992). But it has also been found that less easily defined phenomena falling into the broad category of INFORMATION STATUS, including CONTRASTIVENESS (Bolinger,1961; Bing,1983; Bardovi-Harlig,1983; Couper-Kuhlen,1984), FOCUS (Jackendoff,1972; Rooth,1985; Baart,1987; Dirksen,1992; Wilson and Sperber,1979; Enkvist,1979; Gussenhoven,1983; Rochemont and Culicover,1990; Horne,1987; Zacharski,1992; Eady and Cooper,1986), and the GIVEN/NEW distinction (Brown,1983; Fuchs,1984; Kruyt,1985; Fowler and Housum,1987; Terken and Nooteboom,1987; Nooteboom and Kruyt,1987; Koopmans-van Beinum and van Bergem,1989; Horne,1991) influence accent decisions, with most of the empirical studies currently focussing on the last category. All of these types of information status are defined in terms of the structure of the discourse context. At least implicit in the notion of what is 'in focus' or what is 'given' in these accounts is some assumption about how discourses are structured, and what identifies an item as focussed or as given or as contrastive in its context. (Clearly, the cue of accentedness itself cannot be used as such an indicator for the study of accent itself, so some independent notion of discourse structure must be appealed to in order to establish the discourse variables to be tested.) Often, these models are not made explicit, or are greatly simplified for the purposes of the experiment; it is not always clear, thus, how results will generalize. Also, it is not easy to compare results

when researchers have made different assumptions, as, about what defines 'givenness'.

The role of accent in reference resolution has been more speculated upon than studied, although observations such as Lakoff's (Lakoff,1971) classic *'John called Bill a Republican and then HE insulted HIM'* have long been noted. Some empirical work has also been done (Gleitman,1961; Hirschberg and Ward,1991; Horne,1985). However, given the heavy emphasis on this topic in text-based studies of discourse, there would appear to be richer fields to plow here than have yet been touched.

As yet there have been few attempts to combine potential or attested determinants of accent location into unified models of accent assignment, perhaps because the role of individual factors is still an open question. But it is important to start viewing our knowledge of the contribution of individual phenomena, such as 'givenness', within the larger framework of contributions from other discourse features and from syntactic and semantic features. There is of course considerable practical incentive to find solutions to these problems for text-to-speech synthesis. Many current research systems implement algorithms which attempt to make use of discourse-level information for accent assignment (Carlson and Granstrom,1973; Horne,1987; Hirschberg,1990; Monaghan,1991; Quené and Kager,1992); message-to-speech systems have also employed their richer sources of discourse information to improve prominence location (Davis and Hirschberg,1988; House and Youd,1990).

CONCLUSION

So, we have some evidence of some intonational and acoustic features that appear to signal certain aspects of discourse structure, such as topic beginnings and endings. And we have some notion about which discourse-level factors influence the decision to accent an item. In neither case do we know which factors are more important or more reliable than others. Nor do we know what sort of interaction there is among different cues. Nor do we know much about speaker or listener variability. Future work on intonation and discourse must thus address the following questions: What discourse factors influence intonational decisions, and how do these discourse factors interact with other components of the grammar? What sort of individual variation exists in these models? What assumptions are we making about our underlying models of discourse phenomena when we study the mapping between intonation and discourse? Are they justified?

REFERENCES

Bengt Altenberg (1987), *Prosodic Patterns in Spoken English: Studies in the Correlation between Prosody and Grammar for Text-to-Speech Conversion*, Vol. 76 of *Lund Studies in English*. Lund University Press, Lund.

Cinzia Avesani and Mario Vayra (1988), Discorso, segmenti di discorso e un' ipotesi sull' intonazione, In *Att del Convegno Internazionale "Sull'Interpunzione"*, Florence.

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

J. L. G. Baart (1987), *Focus, Syntax and Accent Placement*, Ph.D. thesis, University of Leyden, Leyden.

K. Bardovi-Harlig (1983), Pronouns: When 'given' and 'new' coincide, In *Papers from the 18th Regional Meeting*. Chicago Linguistic Society.

J. M. Bing (1983), Contrastive stress, contrastive intonation and contrastive meaning, *Journal of Semantics*, 2:141–156.

Dwight Bolinger (1961), Contrastive accent and contrastive stress, *Language*, 37:83–96.

Dwight Bolinger (1972), Accent is predictable (if you're a mindreader), *Language*, 48:633–644.

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

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

B. Butterworth (1975), Hesitation and semantic planning in speech, *Journal of Psycholinguistic Research*, 4:75–87.
R. Carlson and B. Granstrom (1973), Word accent, emphatic stress & syntax in a synthesis by rule schenme for Swedish, *STL-QPSR*, 2(3):31–35.
W. L. Chafe (1980), The deployment of consciousness in the production of a narrative, In W. L. Chafe, ed., *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Navvative Production*, pp. 9–50. Ablex Publishing Corp, Vol. 3, Advances in Discourse Processes.
Elizabeth Couper-Kuhlen (1984), A new look at contrastive intonation, In Richard J. Watts and Urs Weidmann, eds., *Modes of Interpretation: Essays Presented to Ernst Leisi*, pp. 137–158. Gunter Narr Verlag, Tubingen.
J. R. Davis and J. Hirschberg (1988), Assigning intonational features in synthesized spoken directions, In *Proceedings of the 26th Annual Meeting*, pp. 187–193, Buffalo. Association for Computational Linguistics.
A. Dirksen (1992), Accenting and deaccenting: A declarative approach, In *Proceedings of COLING-92*, pp. 865–869.
S. J. Eady and W. E. Cooper (1986), Speech intonation and focus location in matched statements & questions, *Journal of the Acoustical Society of America*, 80:402–415.
N. Enkvist (1979), Marked focus: Functions and constraints, In S. Greenbaum, G. Leech, and J. Svartvik, eds., *Studies in English Linguistics for Randolph Quirk*, pp. 134–152. Longmans, London.
Nomi Erteschik-Shir and Shalom Lappin (1983), Under stress: A functional explanation of English sentence stress, *Journal of Linguistics*, 19:419–453.
C. A. Fowler and J. Housum (1987), Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction, *Journal of Memory and Language*, 26:489–504.
A. Fuchs (1984), Deaccenting and default accent, In D. Gibbon and H. Richter, eds., *Intonation, Accent and Rhythm*, pp. 134–164. Walter de Gruyter, Berlin.
L. Gleitman (1961), Pronominals and stress in English, *Language Learning*, 11:157–169.
B. Grosz and J. Hirschberg (1992), Some intonational characteristics of discourse structure, In *Proceedings of the International Conference on Spoken Language Processing*, Banff, October. ICSLP.
Barbara J. Grosz and Candace L. Sidner (1986), Attention, intentions, and the structure of discourse, *Computational Linguistics*, 12(3):175–204.
Carlos Gussenhoven (1983), *On the Grammar and Semantics of Sentence Accents*. Foris Publications, Dordrecht.
J. Hirschberg and B. Grosz (1992), Intonational features of local and global discourse structure, In *Proceedings of the Speech and Natural Language Workshop*, pp. 441–446, Harriman NY, February. DARPA, Morgan Kaufmann.
J. Hirschberg and J. Pierrehumbert (1986), The intonational structuring of discourse, In *Proceedings of the 24th Annual Meeting*, pp. 136–144, New York. Association for Computational Linguistics.
J. Hirschberg and G. Ward (1991), Accent and bound anaphora, *Cognitive Linguistics*, 2(2):101–121.
J. Hirschberg (1990), Using discourse context to guide pitch accent decisions in synthetic speech, In *Proceedings of the European Speech Communication Association Workshop on Speech Synthesis*, pp. 181–184, Autrans, France.
M. Horne (1985), English sentence stress, grammatical functions and contextual coreference, *Studia Linguistica*, 39:51–66.
Merle Horne (1987), Towards a discourse-based model of English sentence intonation, Working Papers 32, Lund University Department of Linguistics.
M. Horne (1991), Accentual patterning in 'new' vs 'given' subjects in English, Working Papers 36, Department of Linguistics, Lund University, Lund.
Jill House and Nick Youd (1990), Contextually appropriate intonation in speech synthesis, In *Proceedings of the ESCA Workshop on Speech Synthesis*, pp. 185–188, Autrans. ESCA.
Ray S. Jackendoff (1972), *Semantic Interpretation in Generative Grammar*. MIT Press,

Cambridge MA.
F. J. Koopmans-van Beinum and D. R. van Bergem (1989), The role of 'given' and 'new' in the production and perception of vowel contrasts in read text and in spontaneous speech, In J. P. Tubach and J. J. Mariani, eds., *Proceedings of the European Conference on Speech Communication and Technology*, pp. 113–116, Edinburgh. Eurospeech, CEP, Vol. 2.
J. G. Kruyt (1985), *Accents from Speakers to Listeners: An Experimental Study of the Production and Perception of Accent Patterns in Dutch*, Ph.D. thesis, University of Leyden.
D. R. Ladd (1979b), Light and shadow: A study of the syntax and semantics of sentence accents in English, In L. Waugh and F. van Coetsem, eds., *Contributions to Grammatical Studies: Semantics and Syntax*, pp. 93–131. University Park Press, Baltimore.
George Lakoff (1971), Presupposition and relative well-formedness, In *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics, and Psychology*, pp. 329–340. Cambridge University Press, Cambridge UK.
I. Lehiste (1975), The phonetic structure of paragraphs, In A. Cohen and S. G. Nooteboom, eds., *Structure and Process in Speech Perception*, pp. 195–203. Springer, Heidelberg.
I. Lehiste (1979), Perception of sentence and paragraph boundaries, In B. Lindblom and S. Oehman, eds., *Frontiers of Speech Research*, pp. 191–201. Academic Press, London.
I. Lehiste (1980), Phonetic characteristics of discourse, Paper presented at the Meeting of the Committee on Speech Research, Acoustical Society of Japan.
A. Monaghan (1991), *Intonation in a Text-to-Speech Conversion System*, Ph.D. thesis, University of Edinburgh, Edinburgh.
S. G. Nooteboom and J. G. Kruyt (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.
R. Passoneau and D. Litman (1993), Feasibility of automated discourse segmentation, In *Proceedings of ACL-93*, Ohio State University. Association for Computational Linguistics.
Hugo Quené and René Kager (1992), The derivation of prosody for text-to-speech from prosodic sentence structure, *Computer Speech and Language*, 6:77–98.
Michael S. Rochemont and Peter W. Culicover (1990), *English Focus Constructions and the Theory of Grammar*. Cambridge University Press, Cambridge UK.
Mats Rooth (1985), *Association with Focus*, Ph.D. thesis, University of Massachusetts, Amherst MA.
K. Silverman, M. Beckman, J. Pierrehumbert, M. Ostendorf, C. Wightman, P. Price, and J. Hirschberg (1992), TOBI: A standard scheme for labeling prosody, In *Proceedings of the Second International Conference on Spoken Language Processing*, Banff, October. ICSLP.
K. Silverman (1987), *The Structure and Processing of Fundamental Frequency Contours*, Ph.D. thesis, Cambridge University, Cambridge UK.
M. Swerts, R. Gelyukens, and J. Terken (1992), Prosodic correlates of discourse units in spontaneous speech, In *Proceedings*, pp. 421–428, Banff, October. International Conference on Spoken Language Processing.
J. Terken and J. Hirschberg (1992). Deaccentuation and persistence of grammatical function and surface position. Ms.
J. Terken and S. G. Nooteboom (1987), Opposite effects of accentuation and deaccentuation on verification latencies for given and new information, *Language and Cognitive Processes*, 2(3/4):145–163.
Dierdre Wilson and Dan Sperber (1979), Ordered entailments: An alternative to presuppositional theories, In C.-K. Oh and D. A. Dinneen, eds., *Syntax and Semantics*, Vol. 11, pp. 229–324. Academic Press, New York.
Ron Zacharski (1992), Generation of accent in nominally premodified noun phrases, In *Papers Presented to the 15th International Conference on Computational Linguistics*, pp. 253–259, Nantes. International Conference on Computational Linguistics.