Julia Hirschberg

AT&T Bell Laboratories

Murray Hill, New Jersey 07974

Comparing the questions proposed for this discourse panel with those identified for the TINLAP-2 panel eight years ago, it becomes clear that some progress has been made in discourse studies. While TINLAP-2's questions seem concerned with identifying an ideal model of discourse (e.g., what constitutes an appropriate model of discourse, how domain-dependent must such models be, what makes discourse coherent, and how can relationships between utterances best be characterized), TINLAP-3's questions reflect the debates which have attended individual attempts to define such models (e.g., are current discourse theories testable, are plan-based theories really feasible, how are processing theories defined and what have their authors failed to take into account, and what aspects of discourse study are really subsumed by more general problems of general communication or knowledge representation). However, in one area -- the role of speech in discourse research -- the questions posed at TINLAP-3 appear remarkably similar to those posed at TINLAP-2: essentially, do we need to study speech as well as text to study discourse? While a simple 'yes' is probably no more controversial now than it would have been in 1978, I think we can make an even stronger claim today: that while one *may* still study text in silence, one *cannot* study discourse without taking speech into account.

Before providing support for this claim, I should admit to some bias. For the past two and one-half years I have been interested in intonation and its contribution to utterance and discourse interpretation. It is significant that, for much of this time, this was more of a hobby than a professional interest: both in Linguistics and in Natural Language Processing, intonation has commonly been seen as 'the edge of language'. More recently, however, this attitude seems to be changing -- witness the speech session at ACL86, for example. I think much of the reason for this change in NLP arises from the fact that the study of spoken language promises to help overcome critical obstacles to progress in better-established areas of the field.

1. Speaker Intentions

First, prosody provides help in identifying speaker intentions. It is now widely accepted that the contour or *tune* a speaker employs in an utterance can communicate semantic or pragmatic information. However, since there are few particular tune types for which we can specify with any confidence just what the meaning might be, it has been difficult to generalize about what type of information tunes in general can convey. But from those whose 'meaning' is better understood -- namely, *declarative*, *yes-ñō ˊquestion*, *surprise/redundancy*, *rise-fall-rise*, and *continuation rise* contours -- it seems (Hirschberg & Pierrehumbert 1986) that tunes convey two sorts of inforation: first, propositional attitudes the speaker wishes to associate with an utterance and, second, speaker commitment to some structural relationship holding between utterances. In the first case, the speaker may convey via choice of tune that s/he knows **x**, believes **x**, does *not* believe **x**, is uncertain about **x**, or is ignorant of **x**. Consider 'I'm the Easter Bunny.' intoned declaratively, interrogatively, incredulously, or with rise-fall-rise, for example. In the second case, tunes can convey speaker commitment to some intentional relationship holding between utterances. We have proposed that continuation rise, for example, can convey that a subordination relationship holds between the (interpretation of a) phrase uttered with continuation rise (a) and the succeeding utterance (b), which we gloss informally as 'b completes a'.

Intonational contours also provide information about speaker intentions by helping to distinguish between direct and indirect speech acts (Sag & Liberman 1975). 'Can you read the bottom line?' when uttered with a declarative pattern is more likely to be taken as an indirect request than when it is uttered with yes-no question intonation. Furthermore, when 'can' and 'you' are deaccented, the indirect request interpretation appears even more favored.

2. Semantico-Syntactic Disambiguation

At an even more basic level, accent and phrasing can affect utterance interpretation by helping hearers disambiguate among potential syntactic parses (Marcus & Hindle 1985) or logical forms.

For example, when 'only' associates with intonational focus in an utterance, varying the focussed (stressed) element can affect the truth conditions of the underlying sentence (Rooth 1985). Consider: 'Bill only INTRODUCED Mary to John.' (He didn't do anything else); 'Bill only introduced MARY to John' (He didn't introduce anyone else to John); and 'Bill only introduced Mary to JOHN' (He didn't introduce Mary to anyone else). Phrasing can similarly influence hearers' interpretation of the scope of negation. Compare 'John doesn't drink because he's unhappy.' (John drinks for some other reason) with 'John doesn't drink, because he's unhappy.' (John is too unhappy to drink). It also appears that some form of intonational boundary in otherwise ambiguous conjoined structures can influence disambiguation: If there is a boundary between 'old' and 'men', for example, in 'old men and women', it appears more likely that 'old' will modify the entire conjunct; if there is a boundary between 'men' and the conjunction, this interpretation is less likely. Finally, phrasing, accent, and pauses may distinguish metalinguistic disjunction from true disjunction (Ball 1986). 'Nominate George or the person you most admire.' may propose a choice of nominees or convey that George is admired by the hearer, depending upon intonational cues. While speakers do not always provide intonational cues in cases such as these, and while such cues, when provided, frequently do not **force** one interpretation over another, theories of syntax and semantics cannot afford to ignore them.

3. Information Status

It is commonly accepted in linguistic pragmatics that so-called 'contrastive stress', or *accent*, can indicate an item's information status (Chafe 1976, Schmerling 1976). In 'John hit BILL', 'Bill' probably represents 'new' information; whereas in 'JOHN hit Bill', John does. In addition, the accenting of prowords can alter reference resolution: Consider 'John hit Bill and then HE hit HIM.' or compare 'John called Bill a Republican and then he insulted him' with 'John called Bill a Republican and then HE insulted HIM'. While, in the first and third sentences, referents to entities clearly not new information are in fact accented, the simple association of given information with deaccented items and new with accented can be salvaged by postulating that position in

predicate-argument structure is also information subject to the given/new distinction. The pronouns in the first sentence, for example, can be accented to indicate that they represent new instantiations of the arguments to 'hit', distinguishing 'hit(HE[Bill],HIM[John])' from 'hit(John,Bill)'. However, if one is trying to construct a general theory of appropriate accenting -- say, for spoken text generation -- the simple association of old or given information with deaccented items and new with accented is not so easily retained. In particular, choice of overall tune appears to influence felicity of accenting of old information.

4. Discourse Structure

Finally, there are several ways in which intonational features can indicate the structure of a discourse. Diane Litman and I have been looking at the relationship between so-called 'clue' words and phrases[1] and their 'non-clue' uses. Recently, we have been studying the intonational distinction between clue and non-clue uses. From a pilot study of 'now' in natural speech, we have found that clue uses can often be distinguished by the following characteristics: they usually represent separate intonational phrases, they are generally accented, and they frequently are given low accents. Non-clue uses are frequently deaccented and rarely represent full intonational phrases. We are also comparing the examining the interaction between the use of clue words to signal discourse structure and other cues, such as pitch range manipulation.

Janet Pierrehumbert and I have proposed (Hirschberg & Pierrehumbert 1986) that speakers can manipulate *pitch range* and *final lowering ratios* to signal the topic structure of a discourse. When speakers increase their pitch range, they can signal various degrees of topic change; they can return to a prior topic level by return to a roughly similar pitch range. Similarly, degree of final lowering in an utterance can be used to signal the 'level' of topic which that utterance concludes; maximum final lowering signals the conclusion of major topics, for example. In both cases, of course, it is

1. Words and phrase such as 'now', 'anyway', 'by the way', which can signal the structure of a discourse. See (Reichman 1985, Cohen 1983).

the relationship of pitch ranges and final lowering ratios employed, rather than any absolute values, that is at issue. While we developed these hypotheses in the course of synthesizing prepared text,[2] we are currently testing them empirically by presenting subjects with ambiguous anaphora resolution tasks which can be disambiguated according to the perceived structure of the discourse. Other pilot studies under way[3] also report encouraging results on the role of pitch range, final lowering, speech rate, and pausal duration in perceived discourse structure.

5. Some Conclusions

So, intonational studies provides a fertile field for researchers in many areas of NLP by suggesting potentially simpler answers to questions such as how hearers recognize and communicate discourse structure, relational predication, goals, and speech acts. To date we have only begun to explore what prosody can convey. Since better tools are currently available for speech analysis and synthesis, it is to be hoped that more researchers will join in the task. Speech synthesizers like Dectalk, Prose, and the Bell Labs Text-to-Speech system allow us to vary intonational features in a principled way to examine what sounds bad as well as what sounds good. The commercial availability of such synthesizers makes it quite possible that the end result of a text generation system can be spoken. Software for the analysis of natural speech (pitch trackers and wave form editors, for example) is faster and more precise, allowing us to examine intonational contours and measure durations of natural data and of speech elicited during empirical studies.

Despite this increased accessibility of tools for speech research, it does seem likely that studies of intonation and discourse will proceed more rapidly if they are interdisciplinary in nature. We need not become phonologists or psychologists to work in this field, but we have a golden opportunity now to collaborate toward common ends. Furthermore, the empirical tradition in speech studies makes it imperative that our hypotheses be tested rigorously before they are incorporated into

2. The text of TNT, a talking tutor for the Unix screen-oriented text editor *vi*.

3. Conducted by Kim Silverman.

spoken text generation systems or, someday, into recognition systems. In conclusion, it is relatively easy now to take the position that prosody presents information crucial to NL generation and understanding. However, the consequences of taking this position, the need for interdisciplinary collaboration and for empirical testing, may be more controversial -- and will increase the magnitude of the task.

REFERENCES

Ball, C. N. 1986. Metalinguistic disjunction. Papers of the Penn Linguistics Colloquium, University of Pennsylvania: Penn Linguistics Colloquium.

Chafe, W. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. Subject and topic, ed. by C. Li. New York: Academic Press.

Cohen, R. 1983. A computational model for the analysis of arguments. dissertation, University of Toronto.

Hirschberg, J., and J. Pierrehumbert. 1986. The Intonational structuring of discourse. Proceedings, New York: Association for Computational Linguistics.

Marcus, M., and D. Hindle. March 1985. A Computational account of extracategorial elements in Japanese. La Jolla CA. Paper presented at the First SDF Workshop in Japanese Syntax.

Reichman, Rachel. 1985. Getting computers to talk like you and me. Cambridge MA: MIT Press.

Rooth, Mats E. 1985. Association with focus. dissertation, University of Massachusetts, Amherst.

Sag, I. A., and M. Liberman. 1975. The intonational disambiguation of indirect speech acts. Papers from the Eleventh Regional Meeting, 487-498.

Schmerling, S. 1976. Aspects of English sentence stress. Austin: University of Texas Press.