DISAMBIGUATING SENTENCES USING PROSODY

Patti Price, SRI International, Menlo Park, CA USA

Mari Ostendorf, Boston University, Boston, MA USA

Stefanie Shattuck-Hufnagel, MIT, Cambridge, MA USA

ABSTRACT

Prosodic information can provide cues to syntactic structure to help select among competing hypotheses, and thus to disambiguate otherwise ambiguous sentences. We show that some, but not all, syntactic structures can be disambiguated via prosody. The phonological evidence relates the disambiguation primarily to boundary phenomena. Phonetic analyses indicate the importance of both absolute and relative measures. Finally, we describe initial experiments involving the automatic use of this information in parsing.

1. INTRODUCTION

The syntax of spoken utterances is frequently ambiguous, yet communication generally succeeds. This success may arise from a variety of sources; we address here the role of prosody. A clear understanding of the mapping between prosodic and syntactic structure would reveal significant aspects of the cognitive processes of speech production and perception. In addition, it would yield more natural sounding speech synthesis. Further, prosody should be particularly helpful in spoken-language understanding, where lexical and structural ambiguities of written forms are compounded by difficulties in finding word boundaries and in identifying words reliably in automatic speech recognition. Here, we study the mapping between prosody and syntax by minimizing the contribution of other possible cues to the resolution of ambiguity.

With few exceptions (e.g., [7]), previous studies have focussed either on relating phonological aspects of prosody to syntax (e.g., [5], [12], [2], [9]), or on relating phonetic/acoustic evidence to syntax and perceived differences (e.g., [15], [3], [16], [6], [8], [4], [17]). A few studies, e.g., [13], have considered the mapping from phonology to acoustics. The more phonetic/acoustic studies typically used a small number of minimal pairs of utterances in order to facilitate the acoustic measurements and to control parameters more precisely. In contrast, the more phonological studies have focussed either on 'illustrative examples' or on text to which prosodic markers have been assigned on the basis of the syntax of the sentence. These studies have typically ignored the fact that there are several possible prosodic choices for a given syntactic structure. The focus in recent theoretical linguistics on human competence for language production, has resulted in neglect of actual language production and neglect of an area required for speech understanding (by human or by machine): the mapping from acoustics to meaning. Clearly, speech communication involves both production and perception, and it involves performance as well as competence.

The work presented in this paper extends previous work, including the important contribution of [10], in several ways: (1) we investigate the ability of listeners to disambiguate sentences for different types of syntactic structures, using several instances of each type; (2) we consider both production and perception; (3) to increase reliability without assessing a large pool of subjects, we used professional FM radio announcers; (4) we have investigated the possible use of prominence associated with pitch accents, in addition to prosodic phrase boundary cues; (5) to compare durational structures across the various sentences used, and to facilitate generalization beyond the specific sentences used, we present results in terms of relative, rather than absolute, durational patterns; and (6) we consider the automatic use of prosodic information in parsing.

2. CORPUS

We used 35 sentence pairs, ambiguous in that members of each pair contained the same string of phones, and could be associated with contrasting syntactic bracketings. Sentences represented 5 instances each of 7 types of structural ambiguity: (1) parenthetical clauses vs. non-parenthetical subordinate clauses, (2) appositions vs. attached noun or prepositional phrases, (3) main clauses linked by coordinating conjunctions vs. a main clause and a subordinate clause, (4) tag questions vs. attached noun phrases, (5) far vs. near attachment of final phrase, (6) left vs. right attachment of middle phrase, and (7) particles vs. prepositions.

Each pair of ambiguous sentences was preceded by a disambiguating context. For structural categories 1-4, sentence A of the pair involved a larger syntactic break than sentence B. For the attachment ambiguities 5-7, sentence A of the pair had the larger syntactic break later in the sentence than did sentence B. When sentences were recorded by the 4 FM newscasters, contrasting members of a pair did not occur in the same session. Speakers were not told there were target sentences, and recording sessions were separated by a few days to several weeks. Our goal was to create segmentally identical but syntactically different sentence pairs.

3. PERCEPTUAL EXPERIMENT

The target sentences were presented in isolation. The 35 sentence pairs produced by each speaker were presented to listeners in two sessions; only one member of each pair was heard in each session (analogous to the strategy used for recording the sentences). The answer sheet included both disambiguating contexts followed by the target sentence. Listeners marked the context they thought best matched what they heard. Subjects were all native speakers of American English, naive with respect to the purpose of the experiment. The number of listeners who heard both sessions ranged from 12 to 17 for the different speakers.

In scoring, we assume speakers produced the intended version, and a *correct* response identifies that version. Accuracy is the percentage of correct listener responses. Table 1 summarizes accuracy for the different structural types. Averages are over the 4 speaker averages, so as not to more heavily weight the utterances that were heard by more listeners.

Type	A	B	Overall
1. Parenthetical or not	77	96*	86
2. Apposition or not	92*	91*	92
3. M-M vs. M-S	88*	54	71
4. Tags or not	95*	81	88
5. Far/near attachment	78	63	71
6. Left/right attachment	94*	95*	95
7. Particle/Preposition	82*	81*	82
Average	87	80	84

Table 1. Perceptual results, averaged over 4 speakers. Version A/B figures are based on 285 observations of each class. An asterisk marks A and B responses with high listener accuracy, where high accuracy is when (average accuracy minus Standard Deviation) > 50%.

Table 1 shows that subjects could reliably disambiguate many, but not all of the ambiguities. On average, subjects did well above chance in assigning sentences to appropriate contexts. Main-subordinate (3B) sentences and near attachments (5B) were close to the chance level; parentheticals (1A), far attachments (5A) and non-tags (4B) were recognized at levels greater than chance but not reliably; all other sentence types were reliably disambiguated.

4. PHONOLOGICAL ANALYSIS

The perceptual experiments show that speakers can encode prosodic cues to structural ambiguities in ways that listeners can use reliably. This section attempts to find a phonological answer to the question: How do they do it? To approach this question, we labeled discrete phenomena that could mark structural contrasts phonologically. We then analyzed the relationship between these labels and patterns in the perceptual study.

We used 7 levels to represent perceptual groupings (or, degrees of separation) between words. These levels appeared adequate for our corpus and also reflected the levels of prosodic constituents described in the literature. We used numbers to express the degree of decoupling between each pair of words as follows: 0 - boundary within a clitic group, 1 - normal word boundary, 2 - boundary marking a grouping of words generally having only one prominence, 3 - intermediate phrase boundary, 4 - intonational phrase boundary, 5 - boundary marking a grouping of intonational phrases, and 6 - sentence boundary.

Break indices of 4, 5, and 6 are *major* prosodic boundaries; constituents defined by these boundaries are marked by a boundary tone and are often referred to as 'intonation phrases'. Boundary tones were labeled using 2 types of falls (final fall and non-final fall), and 2 types of rises (continuation rise and question rise). Prominent syllables were labeled using P1 for major phrasal prominence; P0 for a lesser prominence; and C for contrastive stress (which occurred on fewer than 1% of the total words).The prosodic cues were labeled perceptually by 3 listeners using multiple passes. Correlation across labelers was 0.96.

In general, we found prosodic boundary cues associated with almost all reliably identified sentences. A break index of 4 or 5 was often, but not always, a reliable cue, and was most often observed at embedded or conjoined clause boundaries (often marked by commas in the text). A difference in the relative size of prosodic break indices, or in the location of the *largest* break, was frequently the only disambiguating cue for smaller syntactic constituents (i.e., where fewer brackets would coincide). By and large, larger break indices tended to mean that syntactic attachment was higher rather than lower. Prominence seemed to play a supporting role, and was the sole cue in only a few sentences. Details of these results analyzed by structural types appear in [14].

The main exception to this picture was the main-main (A) vs. main-subordinate (B) sentences. The A versions were typically well-identified, whereas the B versions tended to be close to the chance level. This could be the result of a syntactic response bias. The difference is interesting since the bracketings differ for the 2 versions of the sentence, and yet they are apparently not well separated perceptually. The prosodic transcriptions suggest a reason: both versions have a major prosodic boundary in the same location.

5. PHONETIC ANALYSIS

We have presented perceptual evidence that naive listeners can reliably use prosody to separate structurally ambiguous sentences, and phonological evidence that suggests how listeners might use prosody to select among syntactic hypotheses. In this section we consider phonetic evidence that might be responsible for the prosodic disambiguation. We examine duration and intonation, although we acknowledge that other cues, such as the application or non-application of phonological rules, contribute to the perception of prosodic boundaries. We tried to minimize such effects by asking speakers to reread sentences in which overt segmental cues were produced.

Segment duration was determined automatically using an HMM-based speech recognition system, the SRI Decipher system [18]. Each phone duration was normalized according to speaker- and phone-dependent means as described in [11]. We observed longer normalized durations for phones preceding major phrase boundaries and for phones bearing major prominences compared to other contexts. We measured average normalized duration in the rhyme of word-final syllables and found that higher break indices are generally associated with greater normalized duration. Pauses are also associated with major prosodic boundaries, occurring at 48/212 (23%) boundaries marked with 4 and 17/25 (67%) boundaries marked with 5. No pauses were found after a 0, 1, or 2, and only one pause occurred after a 3.

Analysis of normalized duration of vowel nuclei revealed: (1) major prominences (P1, C) tend to be longer than unmarked or minor (P0) prominences, although the effect is small before major prosodic breaks (where duration is already lengthened); (2) word-final syllables tend to be longer than non-word-final syllables; (3) syllables are longer in words before major breaks than in words before smaller breaks, especially for word-final syllables; and (4) the effects seem to be somewhat independent: the longest vowels are those with a major prominence, in word-final position, before a major break.

Intonational cues included boundary tones, pitch range changes and pitch accents. Boundary tones are involved for break indices 4 - 6. Sentence-final boundary tones are typically either final falls or question rises; level 5 boundary tones were usually labeled non-final falls; and level 4 boundary tones were most often continuation rises, but occasionally non-final falls. Another intonational cue was a drop in pitch baseline and range in a parenthetical phrase, relative to the context. This pitch range change was not always apparent for appositives. Though intonation is an important cue, duration and pauses alone provide enough information to automatically label break indices with a high correlation (greater than 0.86) to hand-labeled break indices [11].

6. AUTOMATING DISAMBIGUATION

We have shown that listeners can pay attention to prosodic information, and we have shown phonological and phonetic evidence bearing on how this might be done. The next step is to be explicit enough about the use of the phonetic evidence that it could be used automatically to select the appropriate parse. In our initial attempt, since there was a good correlation between normalized rhyme duration and the hand-labeled break indices, we used a 7-state Gaussian HMM to convert automatically estimated duration values to break indices [11], and passed to the parser a break index between every pair of words. This procedure required modification of the existing grammar to handle the new break index category and to allow for empty nodes and their interaction with the break indices. The grammar before and after these changes yields the same number of parses for a given sentence.

In order to make use of the prosodic information an additional important change is required: how does the grammar use this information? This is a vast area of research. In this initial endeavor, we focussed on prepositional phrases, and made very conservative changes. We changed the rule N -> N link PP so that the value of the link (break-index) must be less than 3 for the rule to apply. We made essentially the same change to VP -> V link PP, except that the value of the link must be less than 2.

After setting these two parameters we parsed each of the sentences in the 14 sentences in our corpus containing prepositional phrase attachment ambiguities or particle-preposition ambiguities, and compared the number of parses to the number of parses obtained without benefit of prosodic information. For half of the sentences, i.e., for one member of each of the sentence pairs, the number of parses remained the same. For the other member of the pairs, the number of parses was reduced, on average to half the previous number. Thus, the incorporation of the prosodic information resulted in a net reduction of about 25% in the number of parses, without ruling out any correct parses. In many cases the use of prosodic information allowed the parser to identify a unique parse. More details on these procedures and results appear in[1].

7. DISCUSSION

We have confirmed that, for a variety of syntactic classes, but not all, naive listeners can reliably separate meanings on the basis of differences in prosodic information. We have further shown phonological and phonetic evidence bearing on how they might do this: by the tendency to associate relatively larger prosodic breaks with larger syntactic breaks. Though evidence relating to boundary phenomena appeared to be most important, there were some structures for which phrasal prominence either was the only cue or played a supporting role in distinguishing between the 2 versions. Further, we have presented initial evidence showing how extracted phonetic information (normalized duration) can be automatically extracted and communicated to a parser to reduce ambiguity.

Our results have both theoretical and empirical implications. In speech generation applications, the relation between syntax and prosody is important since different prosodic markers will affect the interpretation of a sentence. Prosodic cues are particularly important in computer speech understanding applications, where the semantic rules available to the system are limited relative to the capabilities of human listeners. In addition, in these applications, prosodic cues can be used prior to semantic analysis, to reduce the number of syntactically acceptable parses by eliminating those inconsistent with the prosody [1].

The results reported here provide evidence for systematic relationships between prosody and syntax that should be explored further in several ways. First, a larger number of syntactic structures must be examined in order to make the prosody/syntax relationship more explicit. Second, we note that some sentences were successfully disambiguated with cues that were not represented in our labeling scheme. Since prominences were not differentiated as to type of pitch accent, a more detailed classification of intonation in such contexts could yield more information. Finally, for computer speech understanding applications, it will be important to investigate the extension of these results to spontaneous speech by non-professional speakers, where hesitation phenomena and speech errors will affect the prosodic structure.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IRI-8805680 in coordination with DARPA/NSF funding under NSF Grant No. IRI-8905249. The government has certain rights in this material. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the government funding agencies. We thank A. Levitt and L. Larkey for help in generating the ambiguous sentences; the WBUR announcers who recorded the sentences; the subjects who participated in the experiments; N. Veilleux and C. Wightman for many hours of prosodic labeling. We thank J. Bear for providing syntactic bracketings and for grammar modifications and parsing results, and we thank C. Wightman for the automatic labeling of break indices.

8. REFERENCES

[1] Bear, J. and Price, P. (1990). "Prosody, syntax and parsing," *Proc. ACL*.

[2] Bing, J. (1984). "A discourse domain identified by intonation," pp. 11-19 in *Intonation, Accent and Rhythm: Studies in Discourse Phonology*, New York, de Gruyter.

[3] Cooper, W. and Sorensen, J. (1977). "Fundamental frequency contours at syntactic boundaries," *J. Acoust. Soc. Am. 62:3*, 683-692.

[4] Duez, D. (1985). "Perception of silent pauses in continuous speech," *Language and Speech 28:4*, 377-389.

[5] Gee, J. P. and Grosjean, F. (1983). "Performance structures: A psycholinguistic and linguistic appraisal," *Cognitive Psychology 15*, 411-458.

[6] Garro, L. and Parker, F. (1982). "Some suprasegmental characteristics of relative clauses in English," *J. Phonetics 10*, 149-161.

[7] Geers, A. (1978). "Intonation contour and syntactic structure as predictors of apparent segmentation," *J. Exp. Psych: Hum. Perc. and Perf. 4:3*, 273-283.

[8] Kutik, E., Cooper, W., and Boyce, S. (1983). "Declination of fundamental frequency in speakers' production of parenthetical and main clauses," *J. Ac. Soc. Am. 73:5*, 1731-1738.

[9] Ladd, D. R. (1986). "Intonational phrasing: the case for recursive prosodic structure," *Phonology Yearbook 3*, 311-340.

[10] Lehiste, I. (1973). "Phonetic disambiguation of syntactic ambiguity," *Glossa 7:2*, 107-121.

[11] Ostendorf, M., Price, P., Bear, J. and Wightman, C. (1990). "The use of relative duration in syntactic disambiguation," *Proceedings of the 3rd DARPA Workshop on Speech and Natural Language*.

[12] Nespor, M. and Vogel, I. (1983) "Prosodic structure above the word," *Prosody: Models and Measurements*, Cutler and Ladd, eds., Springer-Verlag, pp. 123-140.

[13] Pierrehumbert, J. (1981). "Synthesizing intonation," *J. Ac. Soc. Am. 70:4*, 985-995.

[14] Price, P., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C. (1991). "The use of prosody in syntactic disambiguation," *Proc. 4th Darpa Workshop on Speech and Natural Language*, ed. P. Price, Morgan Kaufman. A longer version of this paper has been submitted for journal publication.

[15] Scholes, R. (1971). "On the spoken disambiguation of superficially ambiguous sentences," *Language and Speech 14*, 1-11.

[16] Thorsen, N. (1980). "A study of the perception of sentence intonation -- Evidence from Danish," *J. Ac. Soc. Am. 67:3*, 1014-1030.

[17] Thorsen, N. (1985). "Intonation and text in standard Danish," *J. Ac. Soc. Am. 77:3*, 1205-1216.

[18] Weintraub, M. et al. (1989), "Linguistic constraints in hidden Markov model based speech recognition," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 699-702.