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.

Phonological analyses 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 focused either on relating phonological aspects of prosody to syntax (e.g., [5], [12], [21], [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 of the speech.

In contrast, the more phonological studies have focused either on 'illustrative examples' or on text in which phonetic 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 linguistic analyses of competence for language production, has resulted in neglect of actual language production and neglect of evidence from actual speech (by human or by machine); the mapping from acoustics to meaning. Clearly, speech production 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 phrases, (6) same vs. right attachment of middle phrase, and (7) particles vs. prepositions.

Each pair of ambiguous sentences was preceded by a disambiguating context. The 35 sentence pairs were presented 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 that they thought best matched what they heard. Subjects were 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.

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 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 'intonation phrases'). 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; 1 - boundary within a group, 2 - boundary between groups; 3 - intonational phrase boundary, 4 - intonational phrase boundary, 5 - boundary marking a grouping of intonational phrases, and 6 - sentence boundary.

Break indices of 1, 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 large 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, corresponding to the desired meaning. By and large, smaller break indices tended to mean that syntactic attachment was lower, corresponding to the desired meaning. 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, corresponding to the desired meaning. By and large, smaller break indices tended to mean that syntactic attachment was lower, corresponding to the desired meaning.

The main exception to this picture was the main-internal boundary between sentence types in the sentences. The A versions were typically well-identifiable, whereas the B versions tended to be close to the chance level. This could be the result of a syntactic ambiguity. The difference is interesting since the bracketings differ for the 2 versions of the sentence, and yet they are not apparently separable phonologically. The prosodic transcriptions suggest a rea-
son: both versions have a major prosodic boundary in the same location.

5. PHONETIC ANALYSIS

We have presented perceptual evidence that naive listeners can use prosodic cues to separate structurally ambiguous sentences, and phonological evidence that suggests how listeners might use prosody to select among syntactic alternatives. Although we acknowledge that other cues, such as the application or non-application of phonological rules, contribute to the perception of prosodic boundaries, we hypothesized that such effects are caused 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 was either 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 used to automatically extract and communicated to a parser to reduce ambiguity.

Our results have both theoretical and empirical implications. In speech generation applications, prosodic cues can be used prior to syntactic processing because they are 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 prominence, different 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-8308560 in coordination with DARPA/NSF funding under NSF Grant No. IRI-8508569. 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 gathering the data used in the experiments; the WAT/WAR announcers who recorded the sentences; the subjects who participated in the experiments; N. Velleux and C. Wightman for many hours of assistance during the experiments; and the J. Ac. Soc. Am. 70:4, 955-985.

8. REFERENCES


