Journal of Experimental Psychology:
Learning, Memory, and Cognition
1996, Vol. 22, No. 3, 714–735

0278-7393/96/$3.00

Reliability of Prosodic Cues for Resolving Syntactic Ambiguity

David W. Allbritton
University of Pittsburgh

Gail McKoon and Roger Ratcliff
Northwestern University

Although previous research has shown that listeners can use prosody to resolve syntactic ambiguities in spoken sentences, it is not clear whether naive, untrained speakers in experimental situations ordinarily produce the prosodic cues necessary for disambiguating such sentences. In a series of experiments, the authors found that neither professional nor untrained speakers consistently produced such prosodic cues when simply reading ambiguous sentences in a disambiguating discourse context. Speakers who were aware of the ambiguities and were told to intentionally pronounce the sentences with one meaning or the other, however, did produce sufficient prosodic cues for listeners to identify the intended meanings.

One of the central issues in theories of parsing has been what kinds of information are considered during sentence processing. One possibility is that syntactic principles are used to guide the initial hypothesis about a sentence's syntactic structure, with nonsyntactic information then being used to evaluate and revise the initial structure (Frazier, 1978, 1991; Frazier & Rayner, 1982; Rayner, Carlson, & Frazier, 1983). Others have argued that nonsyntactic information can also influence the initial decisions of the parser (Altmann & Steedman, 1988; Crain & Steedman, 1985; Marslen-Wilson & Tyler, 1980; McClelland, St. John, & Taraban, 1989; Nagel, Shapiro, & Nawy, 1994; Steedman & Altmann, 1989). Some of the nonsyntactic sources of information that have been proposed include verb control information (Boland & Tanenhaus, 1991; Boland, Tanenhaus, & Garnsey, 1990; Trueswell, Tanenhaus, & Kello, 1993) and information provided by the larger discourse context in which a sentence appears (Altmann & Steedman, 1988; Taraban & McClelland, 1988).

Another possible source of information that the parser might be able to make use of is prosody. Although the question of whether prosody affects syntactic analysis has not been studied extensively, there is a growing body of research indicating a relationship between syntactic and prosodic structure (Cooper & Paccia-Cooper, 1980; Ferreira, 1993; Klatt, 1975; Selkirk, 1984). Comprehension studies have shown that prosodic manipulations such as varying the duration of words or pauses can help cue syntactic structure for sentences heard in isolation (Beach, 1991; Lehiste, Olive, & Streeter, 1976; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Scott, 1982). Production studies have found prosodic correlates of syntactic boundaries (Beach, 1988; Cooper & Sorensen, 1977; Klatt, 1975; Lea, 1973), and a production study by Price et al. (1991) found differences in prosodic structure for two different meanings of a syntactically ambiguous sentence.

In our research, we were interested in whether untrained, naive speakers in an experimental setting would produce prosodic cues sufficient for resolving syntactic ambiguities. The issue of generalizing the role of prosody to naive speakers is an important one in light of the fact that much of the research on prosody and ambiguity resolution (and research examining comprehension of ambiguous sentences in particular) has been based on stimuli that were produced by trained speakers or by speech synthesizers. Studies that have specifically attempted to assess listeners' ability to use prosody to help resolve syntactic ambiguities have used stimuli produced by a single trained speaker (Nespor & Vogel, 1983; Speer, 1995; Speer & Bernstein, 1992), synthesized speech (Beach, 1991), or spoken stimuli that had been modified by a speech synthesizer (Scott, 1982; Streeter, 1978). None of these studies used materials exactly as naive speakers had produced them.

Production studies have also relied on trained speakers. For example, to examine the lengthening of vowels preceding syntactic phrase boundaries, Klatt (1975) himself read aloud the materials to be analyzed. Sorensen and Cooper (1980; Cooper & Sorensen, 1977) analyzed the intonation contours of sentences using a method by which speakers repeatedly practiced pronouncing each sentence, with feedback from the experimenter concerning their performance.

It is surprising that so much research has been based on the productions of trained speakers, in the absence of clear evidence that their productions are representative of speech produced by naive speakers. It is especially surprising in light of hints that have appeared in the empirical literature that there are differences in production between trained and naive speakers. Price, Ostendorf, Shattuck-Hufnagel, and Veilleux (1988), for example, found evidence that prosodic features in the speech of professional broadcasters were more pronounced and consistent than was the case for ordinary speakers. Lehiste (1973) and Wales and Toner (1979) found that listeners were much less successful in identifying the correct meaning of syntactically ambiguous sentences when they were

David W. Allbritton, Psychology Department, University of Pittsburgh; Gail McKoon and Roger Ratcliff, Psychology Department, Northwestern University.

This research was supported by National Institute on Deafness and Other Communication Disorders Grant R01-DC01240, National Science Foundation Grant SBR 9221940, and National Institute of Mental Health Grants MH44640 and MH00871.

Correspondence concerning this article should be addressed to David W. Allbritton, Learning Research and Development Center, University of Pittsburgh, 3939 O'Hara Street, Pittsburgh, Pennsylvania 15260. Electronic mail may be sent via Internet to allbritt+@pitt.edu.

produced by naive speakers than when they were produced by speakers given explicit instructions about the ambiguities. Similarly, Cooper, Paccia, and Lapointe (1978) found preboundary lengthening at appropriate locations for four syntactically ambiguous sentences, but the effects were less consistent when the sentences were read in paragraph contexts than when speakers were informed of the ambiguities and given explicit instructions about which meanings to convey. If, in experimental settings, naive speakers do not consistently produce prosodic cues sufficient for accurate comprehension of ambiguous sentences, whereas trained, nonnaive speakers do, then the issue of generalization becomes problematic. Clearly, it would not be desirable to build a theory of the interaction of syntax and prosody on the basis of experiments with the productions of trained speakers alone, without determining how such a theory might generalize to naive speakers.

Our aim was to investigate and compare the production of prosodic cues by trained and untrained speakers and the use of those cues by listeners. Five of the syntactically ambiguous sentences we selected as stimuli were items used by Price et al. (1991) that had ambiguously attached middle phrases. Another four items had ambiguously grouped triples of noun phrases (NPs), similar to the stimuli used by Scott (1982). These materials were selected because previous research with trained speakers had successfully demonstrated both prosodic differences in production and prosodic effects on comprehension for these items. We also included in our materials a sentence that is not syntactically ambiguous but can be pronounced with two different accent patterns to convey two different pragmatic uses (from Liberman & Pierrehumbert, 1984). We suspected that speakers would consistently place the accent in the appropriate location in this sentence, and this item therefore served as a manipulation check in the production experiments.

To investigate the degree to which prosodic differences in production might be dependent on speakers' training and their knowledge of the goals of the research, we used two different sets of instructions (naive vs. informed) and two separate groups of speakers (untrained vs. trained). With one set of instructions, participants were asked to read the experimental materials without being told the purpose of the research. In this "naive" condition, speakers were told simply that they were reading materials that would be used for future psychological research that required auditory presentation. The critical ambiguous sentences were embedded in short paragraphs. The context of each paragraph demanded one or the other of the two possible meanings for each critical sentence, but no other cues for disambiguating the sentences (such as punctuation) were given. With "informed" instructions, the participants were told the purpose of the experiment, shown paraphrases of the two possible meanings for each critical sentence, and asked to pronounce each sentence once for each meaning, making it clear by the way they pronounced it which meaning was intended.

The speakers in the first set of experiments (Section 1 below) were untrained; they were introductory psychology students participating as part of a course requirement. The productions of these speakers were assumed to be a fairly representative sample of how ordinary, nonprofessional speakers would pronounce the sentences in an appropriate discourse context in an experimental setting. The speakers in our second set of experiments (Section 2) had training that might have made them more likely to produce appropriate prosodic cues. The speakers were amateur and professional actors and broadcasters recruited from the Northwestern School of Drama and the Department of Radio, Television, and Film, and professional actors and broadcasters from the Evanston, Illinois area.

The ambiguous sentences produced by our two groups of speakers were analyzed in two ways. First, knowledgeable judges' ratings of the appropriateness of the prosodic pattern of each utterance from the production experiments, together with the judgments of a group of naive listeners, were used to assess the degree to which speakers were successful in producing prosodically distinguishable pronunciations for the two possible meanings of each sentence. Second, the pairs of productions for each ambiguous sentence that were identified on the basis of the judges' ratings as having meaning-appropriate prosodic patterns were analyzed to see whether they showed evidence of differences in pitch, amplitude, and word duration.

In Section 1, Experiment 1 was a production study using untrained, experimentally naive speakers. Experiments 2 and 3 were judgment studies in which listeners tried to identify the intended meanings for selected acceptable sounding pairs of sentences and selected inappropriate sounding pairs of sentences produced by the speakers in Experiment 1. In Section 2, professional speakers were used for the production studies under naive (Experiment 4a) and informed (Experiment 4b) instruction conditions. In Experiment 5, naive listeners tried to identify the intended meanings of sentences produced in both the naive and the intentional conditions of Experiment 4.

Section 1: Productions From Untrained Speakers

Experiment 1: Production Study, Experimentally Naive Speakers

Method

Participants. The speakers in Experiment 1 were 23 Northwestern University undergraduates (18 men and 5 women) enrolled in introductory psychology who participated for course credit. All were native speakers of American English.

Materials. The materials for Experiment 1 were brief passages (two to six sentences each) that participants were asked to read aloud. Embedded within each of these passages was an ambiguous sentence. Each passage had two versions, providing one context to fit each of two possible interpretations of the ambiguous sentence in the passage. Paragraph contexts were constructed for Items 1–5, and Items 6–10 used the same one-sentence contexts that Price et al. (1991) had used. (See Table 1 for a list of sentences used in the experiment and their two possible interpretations and Appendix A for the passages used for Sentences 1, 2, and 6.) The ambiguous sentences were of three types: ambiguity of background versus new information (Item 1), ambiguity of NP groupings (Items 2–5), and ambiguity of left versus right attachment of a middle phrase (Items 6–10).

Table 1
Ambiguous Sentences and Their Intended Interpretations in Each Context

1. Anna came with Manny.
   (A) It was Manny that Anna came with.
   (B) It was Anna that came with Manny.
2. For our parties, we invite David and Pat or Bob, but not all three.
   (A) We invite David, and we also invite either Pat or Bob.
   (B) We invite both David and Pat, or else we invite Bob.
3. They will use either television or radio and newspapers to announce the sale.
   (A) They will use either television or radio, and they will definitely use newspapers.
   (B) Either they will use television alone, or they will use both radio and newspapers.
4. Automatic seat belts and air bags or antilock brakes are standard.
   (A) You can get both automatic seat belts and air bags, or you can get antilock brakes.
   (B) You get automatic seat belts, and you also get a choice of either air bags or antilock brakes.
5. So, for lunch today he is having either pork or chicken and fries.
   (A) Either he is having pork alone, or else he is having chicken and fries.
   (B) He is having either pork or chicken, and he is definitely having fries.
6. They rose early in May.
   (A) During the month of May, they rose early.
   (B) They rose during the early part of May.
7. Rollo read the review literally learning not an iota.
   (A) He read the review literally, and learned nothing.
   (B) He read the review, and learned literally nothing.
8. As I was eleven only I knew my Dad would be angry.
   (A) I was only eleven, so I knew my dad would be angry.
   (B) I was eleven, so I was the only one who knew my dad would be angry.
9. Although they did run in the woods they were uneasy.
   (A) They ran in the woods, but they were uneasy.
   (B) They ran, but they were uneasy in the woods.
10. When you learn gradually you worry more.
   (A) When you learn slowly, you worry more than when you learn quickly.
   (B) When you learn, you slowly begin to worry more.

Note. Sentence 1 is from Liberman and Pierrehumbert (1984); Sentences 2–5 are similar to sentences used by Scott (1982); and Sentences 6–10 are from Price, Ostendorf, Shattuck-Hufnagel, and Fong (1991). The paraphrases are those that were used in the judgment studies.

Item 1 in our experiments did not contain a syntactic ambiguity. We wanted to compare the degree to which speakers' pronunciations were different for a sentence that differed only in prosodic structure but not syntactic structure, compared with sentences that were syntactically ambiguous. For this reason, we included the sentence *Anna came with Manny,* a sentence for which Liberman and Pierrehumbert (1984) have demonstrated changes in prosodic structure corresponding to changes in which parts of the sentence represent background versus new information. This sentence could be uttered in answer either to the question "Who came with Manny?" or to the question "Who did Anna come with?" The ambiguity is thus between whether *Anna* or *Manny* is the information that answers the question, as opposed to being merely background information. Liberman and Pierrehumbert suggested that the information given in answer to the question is sometimes marked prosodically by a falling boundary tone, whereas background information tends to be deemphasized by a rising boundary tone, and this is the pattern they observed for this sentence.

Four sentences contained ambiguously grouped NPs of the form "*x* and *y* or *z*" or "*x* or *y* and *z*" similar to the materials used by Scott (1982). Scott had speakers attempt to produce two different meanings for sentences such as *Kate or Pat and Tony will come.* She then modified and resynthesized the sentences for a comprehension study and found that listeners could use differences in word and pause durations to determine the intended grouping of the NPs.

The remaining five items were the ambiguous middle-phrase sentences and their context sentences used by Price et al. (1991). These sentences had a phrase that could be attached either to the left or right, such as *literally* in the sentence *Rollo read the review literally learning not an iota.* We made only one change in these materials. For some of their sentences, Price et al. (1991) used punctuation in addition to context to get speakers to pronounce them differently for the two possible meanings (*Rollo read the review, literally learning not an iota,* for example). Because we were interested in whether speakers would spontaneously use prosody to help disambiguate these utterances, we eliminated all punctuation differences between the two versions of the sentences in our experiment.

Design and procedure. The materials for the experiment were printed in booklets and given to participants to read. Two orders were used. Twelve of the participants read Version A of each of the 10 passages first and then were given a second booklet with Version B of each passage. The other 11 participants read Version B of the passages first and Version A second. A cassette tape deck and microphone were used to record the passages as they were read.

Participants were told that the purpose of the recordings was to provide materials for future experiments on language comprehension. They were not told that prosody was the subject of the experiment, nor were they told anything about the possibility of ambiguous sentences appearing in the passages. Participants were instructed to read each passage over carefully to themselves and then read it aloud exactly as it was worded, speaking as naturally and conversationally as possible ("as if you were telling someone a story that you wanted them to understand"). They were instructed that if they mispronounced a word they should reread that passage from the beginning. After giving the instructions, the experimenter gave the participant the first booklet of passages and left the room while they read them aloud.

When all of the passages in the first booklet had been read and recorded, the experimenter gave the participant the second booklet containing the other version of each of the items. Participants were instructed to read the passages the same way they had before, in a conversational manner. They were also told that although the passages in the second booklet were similar to those in the previous one, there were differences between the two, and they should read each passage carefully before saying it aloud.

Results

Judges' ratings. The ambiguous sentences from the passages participants had read were digitized at a sampling rate of 8012 hz and transferred to a NeXT workstation for analysis. Sentences that were mispronounced or had the word order changed (16% of the sentences produced) were not analyzed. Before performing acoustic and pitch-tracking analyses, one of the experimenters, David Allbritton, first rated each sentence on a scale of 1 to 3 to indicate whether the utterance was, in his judgment, a plausible pronunciation for its intended meaning, with 1 meaning that it had the appropriate intonation, 2 meaning that it had neutral intonation, and 3 meaning that it had the intonation appropriate to the opposite version of the sentence. A second judge, who was not one of the experimenters but was informed about the nature of the rating task, also rated approximately 70% of the sentences in order to confirm the judgments of the first rater. The two judges agreed on 85% of the sentences, and all but one of the cases in which they disagreed were resolved through further examination and discussion. In the case where the two judges could not agree, the item was given the neutral rating of 2.

The ratings data from Experiment 1 is given in Table 2 (see Appendix B for individual item data). For each of the 10 items, the number of participants who produced it with appropriate, neutral, or inappropriate intonation was counted, and a proportion labeled *context effect* was calculated to provide a measure of how successful speakers were in producing sentences that distinguished their intended meanings. The context effect for each item was defined as the difference between the number of productions judged to match their contextual meaning minus the number judged to match the inappropriate meaning, divided by the total number of productions for that item (excluding sentences in which words were mispronounced or omitted).

Table 2
Judges' Ratings in Experiments 1, 4a, and 4b

Experiment and item	Context A			Context B			Context effect
	A	N	B	A	N	B	
Experiment 1: 23 naive untrained readers							
Item 1	20	2	1	0	2	20	.87
Items 2–10	135	28	31	95	31	68	.20
Experiment 4a: 9 trained naive readers							
Item 1	8	1	0	1	0	8	.83
Items 2–10	59	9	9	42	8	24	.22
Experiment 4b: 9 trained and informed readers							
Item 1	8	1	0	2	0	7	.72
Items 2–10	74	4	2	10	6	64	.79

Note. Number of productions that were judged to be correct for Meaning A (A), Meaning B (B), or neutral (N) in each presentation context. Context effect is the number of utterances judged to fit their intended meaning minus the number of utterances judged to fit the incorrect meaning, divided by the total number of utterances for that item.

For productions of the sentence containing the nonsyntactic ambiguity between background and new information (Item 1), speakers consistently used a meaning-appropriate prosodic structure. The context effect for this item was .87. The distinction between the two versions of the sentence was a pragmatic one, signaling a difference in conversational implicatures about which information is background and which is new, and speakers distinguished the two versions prosodically nearly without fail.

Performance was not nearly as good on any of the nine syntactically ambiguous sentences, however. For the four sentences containing NP ambiguities (Items 2–5), the mean context effect was .19. For the five middle-phrase sentences, the mean context effect was .21. Clearly, naive speakers did not reliably disambiguate the sentences prosodically. It might have been thought that when participants saw the second version of a passage, they could have deduced the purpose of the experiment and begun pronouncing the ambiguous sentences with more distinctive prosodic features as a result, but the mean context effect for Items 2–10 was .20 for first presentations and .27 for second presentations, only a small increase. Although there was considerable speaker variability for Items 2–10, even the best speaker used appropriate prosody only 78% of the time (context effect = .68), and only 3 speakers were consistent enough to have mean context effects of .50 or better.

Because speakers in Experiment 1 did not reliably produce appropriate prosodic cues for the intended meanings of the syntactically ambiguous sentences, we were concerned that speakers may have simply not understood the intended meanings of the sentences in the contexts in which they read them. To check for this possibility, we conducted a control study in which 9 Northwestern University undergraduates read the passages aloud into a microphone with the same instructions as those used for Experiment 1, and then were asked to identify which of two meanings the ambiguous sentences had in the passages. This was a conservative test of whether they had understood the intended meanings of the sentences because they were not tested immediately after reading a passage, but only after reading all 10 passages. Each participant read one version of each of the 10 passages aloud and then was given a test sheet listing each ambiguous sentence and its two possible meanings. (The paraphrases provided for the two possible meanings were the same ones used in the judgment studies described below.) Participants did appear to have difficulty identifying the correct meaning for Item 9, with 4 of the 9 participants choosing the incorrect alternative. For the other eight syntactically ambiguous items, however, the correct meaning was chosen 82% of the time, indicating that participants were able to correctly identify the intended meaning for most of the items, even though they had to rely on their memory for the passages to make their meaning judgments and they were not expecting a test when they read the passages. The results of this control study, therefore, argue against the possibility that speakers' performance in Experiment 1 was primarily a result of a failure to comprehend the intended meanings of the sentences. Instead, the data support the conclusion that even though the speakers did, in general, understand the intended meanings, they did not produce appropriate prosodic cues.

Pitch-tracking analysis and duration analysis. On the basis of the ratings by the two judges, we selected some of the better productions for pitch-tracking analysis. For each sentence, we analyzed only pairs of productions that were produced by the same speaker and that were both judged to have been pronounced in a way consistent with the context in which they were read. The number of acceptable pairs varied from 15 for Item 1 to only 1 pair for Item 9, and ranged from as few as 1 to as many as 6 pairs per speaker. Over half of the speakers in Experiment 1 produced only 1 or 2 usable pairs of sentences. Of the 230 pairs of sentences produced by the speakers in Experiment 1, 50 pairs were selected for analysis.

We first marked and recorded the word boundaries and word and pause durations for each of the sentences using a sound editor on the NeXT. Word boundaries were marked as points at which amplitude fell to near zero, or as the point at which all of one word could be heard without hearing part of the next word. Pauses were defined as silent intervals between the final boundary of one word and the initial boundary of the following word. For cases in which the end of one word coincided with the beginning of the next with no intervening silence, the pause duration was recorded as zero. After recording the word and pause boundaries, we performed a pitch-tracking analysis. The pitch-tracking application (LPC View v1.0 [linear predictive coding] for the NeXT) provided amplitude and fundamental frequency estimates averaged over windows of about 11 ms each. These pitch and amplitude values for each window were then used for further analyses. First, they were averaged across tokens produced by different speakers to provide means for each version of each item that could be plotted and examined visually. The values for the 15 pairs of spoken sentences for Item 1, for example, were averaged to produce mean amplitude, pitch, and duration plots for Item 1a and Item 1b (see Figure 1). The average pitch and amplitude values displayed in the figure were computed as follows: Each word from each token was divided into a fixed number of segments, with an average segment length of about 10 ms. (The number of segments for a particular word was determined by the mean duration of that word across tokens.) For each token, the value for each segment was then interpolated from the actual values recorded for that token, and the interpolated values were then averaged across tokens. This method of averaging tends to obscure subtle differences in pitch and amplitude patterns, particularly in the case of multisyllable words that contain stop consonants, but it is useful for identifying general trends in pitch and amplitude.

Second, in addition to producing average pitch and amplitude plots for each item, we also used the pitch and amplitude estimates produced by the pitch-tracking application to make peak pitch and amplitude estimates for individual words for each speaker. Paired *t* tests were used to test for differences in pitch, amplitude, and duration in critical words of the sentences, those located near a possible syntactic boundary. The pitch and amplitude values used in the *t* tests were estimates of the peak pitch and amplitude for each word obtained by the following algorithm. Of the pitch values recorded by the pitch-tracking application for a given word, the highest two were discarded, and the mean of the next three highest values was used as the estimated peak pitch value for the word. The purpose of using this method was to reduce the likelihood of outlying values, such as those resulting from noise or pitch-tracking errors, entering into the analysis. The peak amplitude estimates were calculated in the same way using the amplitude values from the pitch-tracking analysis. Means and *t* scores for the peak pitch and amplitude values are reported in Table 3 for only the items that had a sufficient number of observations for tests of significance to be performed.

In the analysis of the productions for Item 1, there were consistent differences in relative pitch, amplitude, and duration between the two versions of the sentence, with *Manny* being accented in Context A and *Anna* in Context B. (See Figure 1 and Table 3). In Version A, when *Manny* was the new information and *Anna* was the background information, the word *Manny* had a longer mean duration, a higher mean peak amplitude, and a higher mean peak pitch than when the given–new relationship was reversed in Version B. Examination of the pitch plots for the individual participants' productions were also consistent with the presence of an accent on *Manny* in Version A, with 11 of the 15 pairs showing a tendency for *Manny* to receive a rising-then-falling pitch contour when it was the new information and a relatively flat pitch contour when it was the background information.

For the four NP-ambiguity sentences, we expected to find indications of prosodic boundaries coinciding with the syntactic constituent phrase boundaries. So, for example, we expected to find phrase-final lengthening preceding the critical NP boundaries (e.g., lengthening of *Pat* in Sentence 2B, "[[David] and [Pat]] or [Bob]" compared with Sentence 2A, "[David] and [[Pat] or [Bob]]"). Qualitative evidence of such lengthening was observed for all four sentences (see Figure 2) and statistical tests confirmed this effect for the two sentences for which more than one pair of acceptable sentences were available (see Table 3, Items 2 and 3). The pitch and amplitude data for these items were less interpretable; there were no reliable differences in peak pitch or peak amplitude.

For the ambiguous middle-phrase items, we expected that the middle phrase would be lengthened when it was attached to the left and that the word preceding the middle phrase would instead be lengthened when the middle phrase was attached to the right. Qualitatively, the word and pause durations for all five of the middle-phrase items appeared to reflect phrase-final lengthening at the predicted locations (see Figure 3 for word and pause durations for Items 6–10 in Experiment 1, and Table 3 for statistical tests). Only Items 6 and 8, however, had a sufficient number of observations for tests of significance to be conducted for the differences in duration. Of the predicted differences in duration, only that for the word preceding the middle phrase in Item 6 (*rose*) was statistically

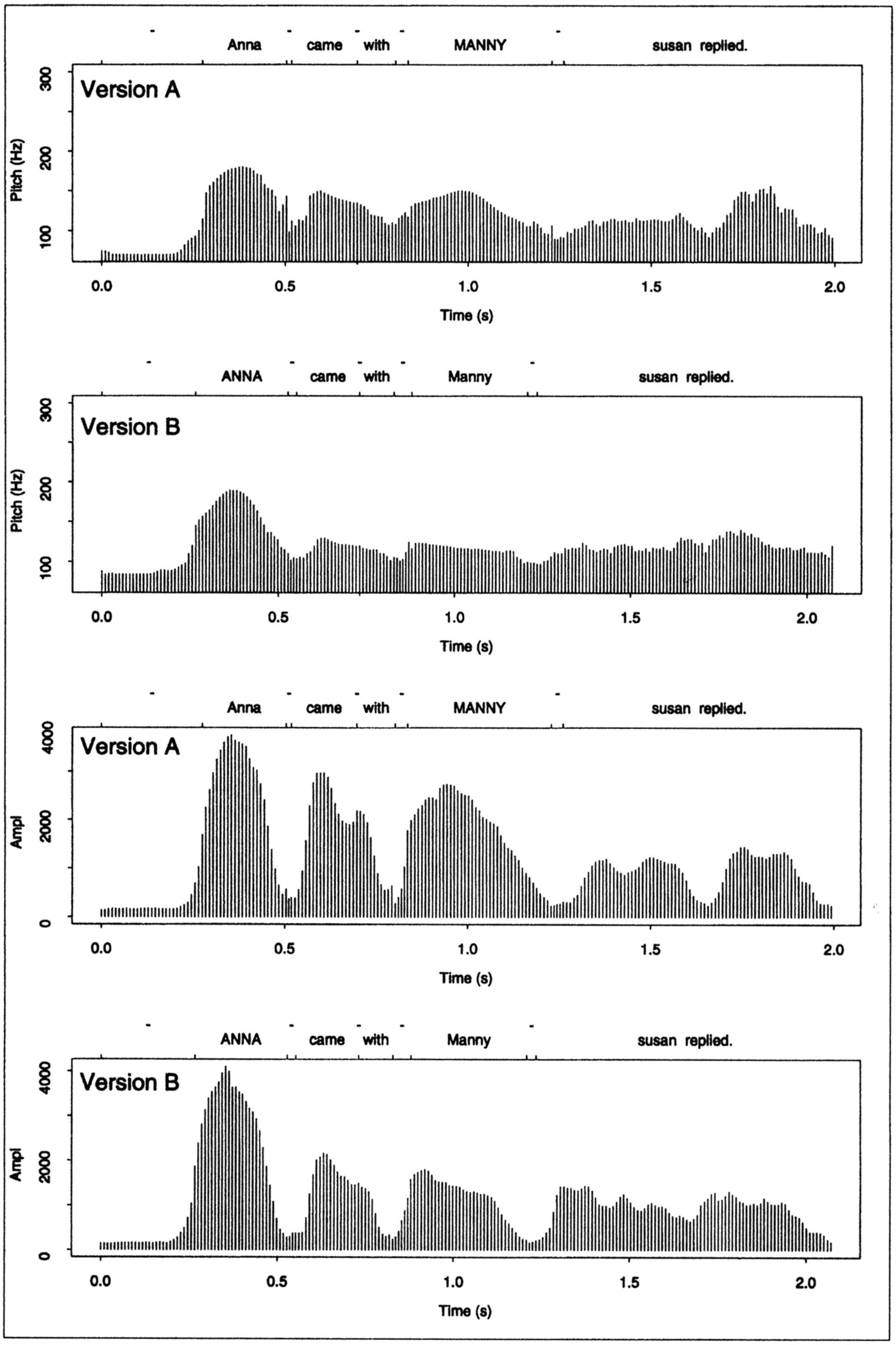

Figure 1. Averaged pitch and amplitude measurements for Item 1 in Experiment 1, averaged across 15 speakers. Each vertical bar represents the mean value across approximately a 10-ms period. Mean word boundaries are indicated by tick marks above each plot. Ampl = amplitude.

significant, although the effect also approached significance for the middle-phrase word *only* in Item 8. There were no reliable differences in pitch or amplitude for the phrase-final words in Items 6 and 8.

Summary. The results of Experiment 1 indicated that naive, untrained speakers do not reliably produce prosodic cues to disambiguate between two possible readings of sentences containing syntactic ambiguities. Only for the one item

Table 3
Mean Duration (in Milliseconds), Peak Pitch (in Hertz), and Peak Amplitude (in Arbitrary Units) for Critical Words and Pauses From Experiment 1

		Duration				Peak pitch				Peak amplitude			
Item	*df*	A	B	*SEM*	*t*	A	B	*SEM*	*t*	A	B	*SEM*	*t*
Item 1													
Manny	14	393	329	12.6	5.10*	152	123	6.4	4.37*	2,902	1,812	277	3.94*
Anna	14	229	262	7.2	−4.56*	185	189	10.0	−0.47	3,813	3,983	421	−0.40
Item 2													
David	5	400	244	38.0	4.11*	131	130	5.8	0.26	4,331	4,355	367	−0.07
Pat	5	221	320	17.0	−5.71*	131	142	8.3	−1.31	2,791	2,720	213	0.33
Item 3													
television	7	510	587	17.0	−4.45*	189	170	22.1	0.84	4,449	4,379	318	0.22
radio	7	454	299	15.0	10.30*	159	169	4.6	−2.20 m	4,205	4,575	409	−0.91
Item 6													
rose	8	252	302	18.0	−2.72*	148	156	4.4	0.06	5,024	4,874	710	0.21
early	8	299	281	19.0	0.81	157	151	10.9	0.55	4,183	3,796	544	0.71
in	8	159	154	19.9	0.26	123	135	6.0	−2.12 m	1,575	2,141	290	−1.95 m
Item 8													
eleven	4	372	427	38.0	−1.42	187	179	7.2	1.13	4,405	3,843	338	1.66
only	4	359	271	37.0	2.39 m	183	161	12.6	1.74	3,572	2,456	757	1.48
I	4	132	186	22.0	−2.51 m	157	175	6.0	−3.25*	2,335	2,705	411	−0.90

Note. For all *t*s, $p > .10$, two-tailed, unless noted as marginal (m, $.05 < p < .10$) or significant ($^*p < .05$). The standard error of the difference between means (*SEM*) is reported for each comparison. A = Context A; B = Context B.

with a pragmatic ambiguity were disambiguating prosodic cues reliably produced. For the sentences for which judges did rate productions as having meaning-appropriate prosodies, acoustical analyses revealed some identifiable prosodic cues at the critical syntactic constituent boundaries, particularly in the analyses of word and pause durations. However, only about 22% of the pairs of utterances for each sentence were included in these analyses (only 50 of the 230 sentences and including sentences from only 20 of the 23 speakers), because the speakers in Experiment 1 so often did not use prosodic patterns appropriate to the intended meanings of the sentences.

Judgment Studies: Experiments 2 and 3

In Experiment 2 we used one of the acceptable pairs of productions for each item from Experiment 1 to elicit meaning judgments from naive listeners, and in Experiment 3 we used unacceptable pairs. These experiments were conducted to serve as a check on the intuitions of the raters concerning whether the sentences had meaning-appropriate prosodic patterns, and also to confirm that when productions were judged by the raters to have meaning-appropriate prosody, listeners could use the prosodic cues to reliably disambiguate the sentences when they were heard out of context.

Participants in the experiments listened to the sentences presented over headphones and decided between two possible meanings for each sentence. Two paraphrases were presented visually on the computer screen, and then a sentence was presented auditorily over the headphones. Listeners were then asked to press a key indicating which meaning they thought the speaker had been trying to convey.

Method

Participants. Thirty-three Northwestern University undergraduates participated in Experiment 2, and 31 participated in Experiment 3. All were enrolled in an introductory psychology class and received course credit for their participation.

Materials. Two paraphrases, indicating the two possible meanings, were written for each sentence for use in the experiments. These paraphrases can be found in Table 1.

Twenty of the productions of ambiguous sentences from Experiment 1 were the stimuli for Experiment 2. We selected one token of each version of each of the 10 items that we judged to be one of the best productions for that item. In all cases, the sentences presented for comprehension in Experiment 2 were among the pairs of sentences used in the pitch-tracking analyses in Experiment 1.

In Experiment 3, the stimuli were 40 productions from Experiment 1 that had been rated as having a prosody that was either neutral between the two meanings or inappropriate for the meaning implied by the context in which it had been read. The only exceptions were Item 1, which had almost no inappropriate pronunciations, and Items 9 and 10, which had been strongly biased toward one pronunciation and were never spoken with the inappropriate prosody in that context. In both of these cases, we included the most neutral sounding sentences that were available. Two tokens for each version of each item were used in the experiment, although results are reported only for tokens that were rated as neutral or inappropriate.

Design and procedure. Participants in the experiments listened to the sentences over headphones and decided which of two possible meanings they thought the speaker had been trying to convey. Presentation of the spoken sentences was controlled by a NeXT workstation, and a real-time microcomputer system was used to present the two possible paraphrases and record participants' choices.

Listeners were told that the experiment was concerned with ambiguous sentences and that they would be asked to decide between two

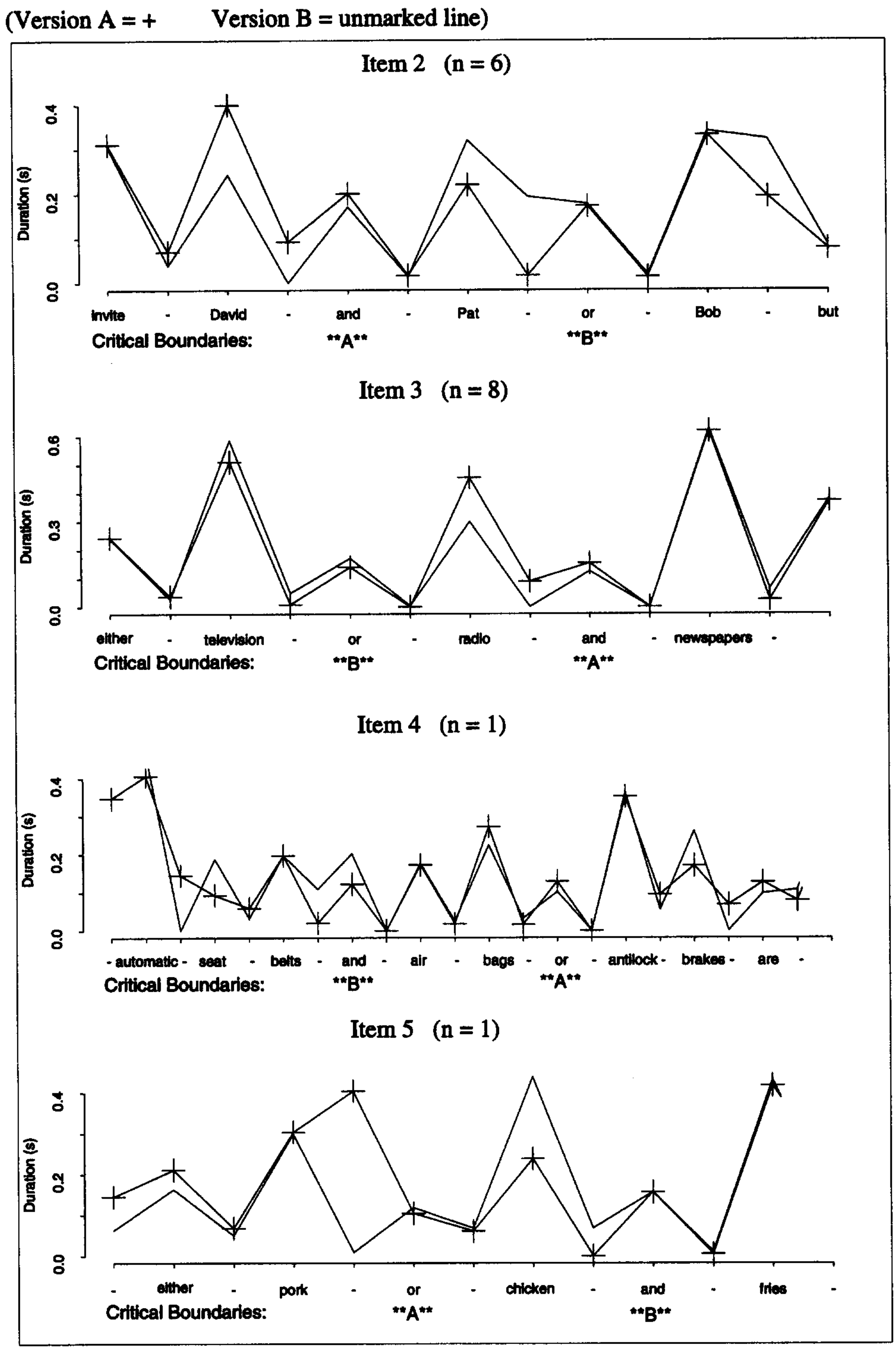

Figure 2. Mean word and pause durations for Items 2–5, Experiment 1. Number of speakers is noted for each item, and the location of the sentence's major syntactic boundary is indicated by **A** or **B** for Versions A and B, respectively.

possible meanings for an ambiguous sentence based solely on how the speaker had pronounced it.

On each trial, the sentence and its two possible meanings (labeled A and B) were first written on the screen. The order of presentation of the two paraphrases was the same for all participants. After 15 s, the sentence was then presented over the headphones, and listeners had 13 s after the onset of the spoken sentence to indicate which meaning they thought it had by pressing either A or B. The sentence and the two

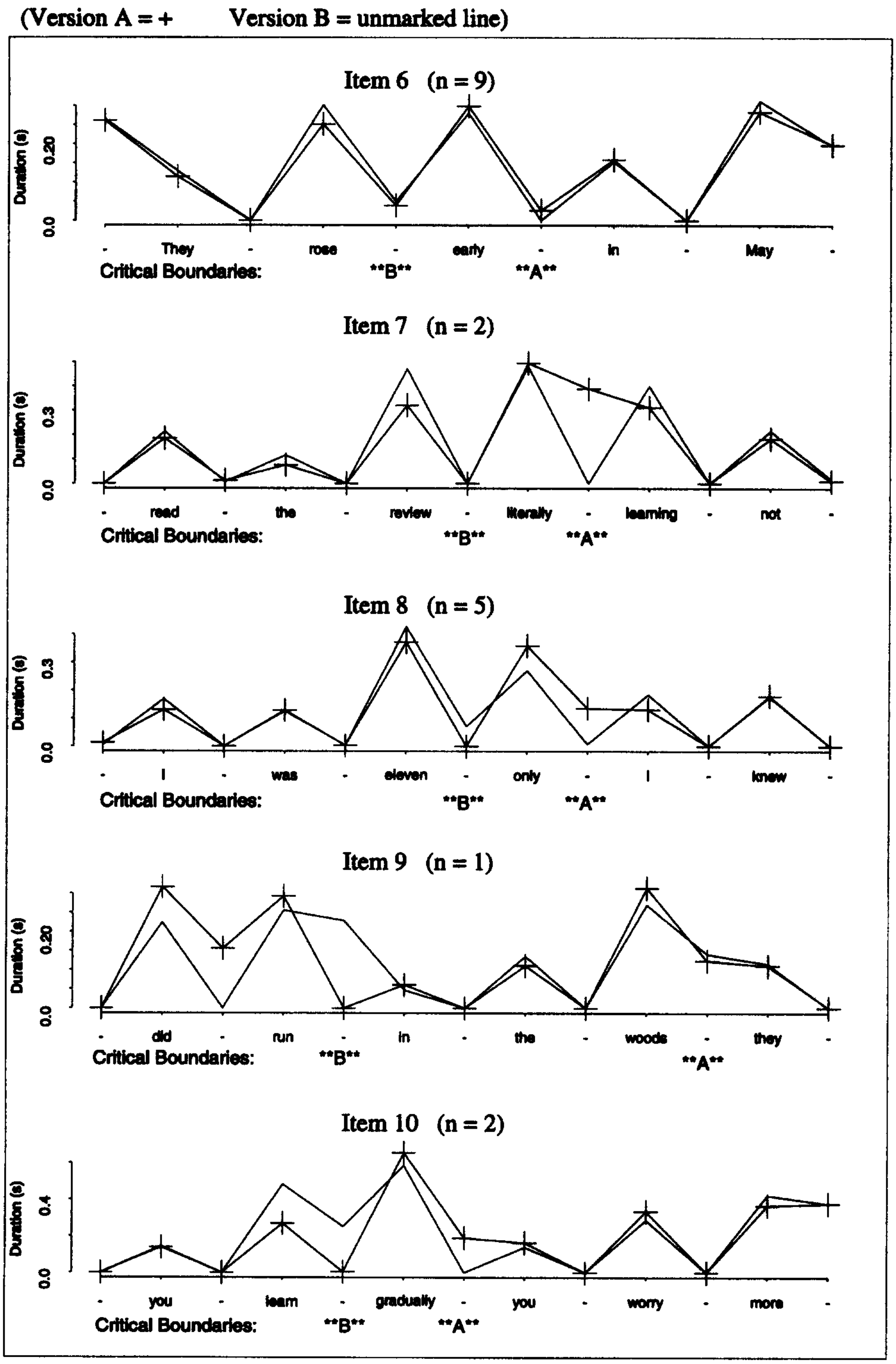

Figure 3. Mean word and pause durations for Items 6–10, Experiment 1. Number of speakers is noted for each item, and the location of the sentence's major syntactic boundary is indicated by **A** or **B** for Versions A and B, respectively.

paraphrases remained on the screen until the listener pressed either the A or B key to indicate which meaning he or she thought the speaker was trying to convey. When the listener pressed a key, the screen was cleared until the presentation of the next item. Sentences were presented one every 30 s, and if listeners did not respond within 13 s following the presentation of the sentence over the headphones the screen was cleared, and it remained blank for 2 s before presentation of the next sentence.

In both of the judgment studies, the spoken sentences (20 for Experiment 2 and 40 for Experiment 3) were presented in 8 different random orders, with approximately the same number of participants for each ordering.

Results

Baseline control study. There were two possible sources of interference with participants' performance in this task that we sought to control for. First, participants might be strongly biased toward one meaning or the other for a given sentence, regardless of how it was said. To provide some measure of this bias, we conducted a baseline study in which 41 Northwestern University undergraduates decided between the two possible meanings for each sentence without hearing the spoken versions of the sentences. On each trial, a sentence was displayed for 5 s on a computer screen, and then it remained on the screen while the two alternative interpretations were displayed below it. (See Table 1 for a list of the paraphrases used.) Participants were instructed to choose the meaning that they thought the sentence was most likely to have and were given up to 30 s to make their response by pressing either A or B on the keyboard. The mean proportion of these participants who chose meaning A is reported as the *Baseline* column in Table 4.

Punctuation control studies. A second possible source of interference we wished to control for was that participants may not always interpret the paraphrases as we expected or may not distinguish clearly between the two possible meanings, leading to an underestimation of listeners' ability to distinguish the intended meaning of the spoken sentences. To get some point of comparison regarding these possibilities, we conducted a second control study with no auditory presentation of the sentences. The sentences were again presented in written form only, but this time with punctuation that removed the ambiguity. In the case of the syntactically ambiguous sentences, a comma was inserted either before or after the middle phrase or the middle NP to indicate where the constituent boundary should occur. For the given–new ambiguity item, the new information was indicated by the use of all capital letters for that word.

Twenty-four Northwestern University undergraduates participated in the punctuation control study for course credit. On each trial, one of the sentences with disambiguating punctuation was first presented on the screen. After 5 s, the two paraphrases for that sentence were displayed below it until the participant responded A or B to indicate which meaning they thought the sentence had.

The mean proportions of A responses in this punctuation control study are shown in parentheses in Table 4 for comparison.

Experiment 2. The dependent measure in this experiment was the proportion of A responses to each item. Trials on which the participant hit a key other than A or B, or on which the participant did not respond within 13 s, were discarded. Table 3 reports three measures for each of the 10 items: the mean proportion of A responses for Version A and for Version B and the context effect for that sentence. The context effect is defined as the proportion of A responses when the sentence was presented with Prosody A minus the proportion of A responses when the sentence was presented in Version B. The same three measures from the punctuation control are reported in parentheses, and the baseline control measure is reported in the column labeled *Baseline.*

For each of the 10 items, there was a sizable context effect, indicating that participants were able to hear the difference between the two versions of the sentences. Two-tailed dependent t tests confirmed that the context effect was significantly greater than zero for all 10 items and in several cases the context effect was nearly at ceiling (see Table 4). This confirms that listeners could use prosodic differences to resolve syntactic ambiguities, and (for Item 1) to determine which was

Table 4
Results for Experiment 2

	Proportion of participants choosing Meaning A					
Item	Baseline [*SE*]	Presented A	Presented B	Context effect	$t(27)$	*SEM*
1	.55 [.079]	.69 (.56)	.17 (.25)	.52 (.31)	4.27	0.117
2	.72 [.065]	.88 (.96)	.03 (.17)	.84 (.79)	10.90	0.072
3	.35 [.078]	.94 (1.00)	.12 (.08)	.82 (.92)	9.88	0.081
4	.41 [.076]	.70 (.96)	.10 (.17)	.60 (.79)	5.42	0.101
5	.48 [.079]	1.00 (.92)	.03 (.08)	.97 (.83)	32.00	0.030
6	.55 [.079]	.78 (.88)	.06 (.25)	.72 (.62)	9.07	0.079
7	.55 [.079]	.97 (.79)	.03 (.17)	.94 (.62)	22.27	0.042
8	.48 [.078]	.97 (.88)	.00 (.12)	.97 (.75)	32.00	0.030
9	.72 [.065]	.84 (.83)	.12 (.17)	.72 (.67)	7.78	0.090
10	.62 [.078]	.97 (.88)	.03 (.00)	.94 (.88)	13.15	0.068

Note. The standard errors reported for the baseline condition represent the standard error of the mean for each item. The context effect is the difference between the proportion of participants choosing Meaning A for Version A and the proportion choosing Meaning A (incorrectly) for Version B. A significant ($p < .05$, two-tailed, for all ts) t value indicates that the context effect for that item was greater than zero, and the reported *SEM* is the standard error of the difference between means for the t test. Results for the punctuation control conditions are in parentheses.

background and which was new information in a sentence that was not syntactically ambiguous. The results of this experiment validate the ratings made in Experiment 1 by showing that other listeners tended to agree with the raters' judgments that the sentences we used had prosodies consistent with their intended meanings.

A concern we had about the procedure used in this experiment was that showing listeners the paraphrases before they heard the sentence might have biased their judgments of the sentence's meaning in some way. The reason for presenting the paraphrases first was so that listeners would not have to try to remember what the sentence sounded like for several seconds while they read the paraphrases. To ensure that reading the paraphrases first did not bias the listeners' judgments, however, we had 28 listeners judge the meanings of the stimuli from Experiment 2 without first reading the paraphrases. The sentence was first presented auditorily over the headphones, with the paraphrases appearing on the screen 1–4 s after the end of the spoken sentence. Listeners then had 13 s to read the paraphrases and indicate which meaning the sentence had. Another 30 undergraduates participated in a punctuation control condition in which a written sentence, disambiguated by punctuation, appeared on the screen for 5 s, followed by a blank screen for 3 s and then the 2 possible paraphrases. For the nine syntactically ambiguous items, the mean context effect when listeners heard the sentence before seeing the paraphrases was .56 (compared with .84 in Experiment 2), and the context effect in the punctuation control was .69 (compared with .76 for the punctuation control for Experiment 2). As expected, the imposition of a memory load did reduce listeners' ability to discriminate between the two versions of each ambiguous sentence but did not change the overall pattern of the results—there was still a significant context effect for each of the nine syntactically ambiguous items.

Experiment 3. The mean proportion of A responses and the context effect for each item for Experiment 3 are reported in Table 5. For Items 2–7 there was a reliable negative context effect, confirming the rater's judgments that these sentences had been pronounced with an inappropriate prosody for their contextually implied meanings. No tokens had been judged inappropriate or neutral for one of the versions of Items 9 and 10 (because speakers almost always pronounced these items the same way regardless of context), and these items are not included in Table 5. Only Item 1, for which no neutral or inappropriate tokens were available, showed any context effect resembling that found in Experiment 2. For Item 1, listeners were able to reliably distinguish between the two versions even in those few cases that the judges had found problematic. Overall, the results of this comprehension experiment supported the judgments made in Experiment 1 concerning the inappropriateness of the sentences' pronunciations for their contexts.

Conclusion: Experiments 1–3

The main conclusion to be drawn from Experiments 1–3 is that, in experimental situations, untrained experimentally naive speakers cannot be relied on to produce consistent prosodic cues that allow listeners to resolve syntactic ambiguities of the type we examined. Speakers did not consistently provide prosodic cues to mark the major syntactic boundaries when reading ambiguous sentences in context. Relatively few of the pairs of utterances produced by speakers in Experiment 1 for the syntactically ambiguous items were rated by the two judges as having prosodic structures that successfully disambiguated the two possible meanings, and the judgments of naive listeners from Experiments 2 and 3 supported the judges' ratings. When speakers did provide appropriate prosodic cues, there were measurable differences in pitch, amplitude, and especially duration corresponding to major syntactic boundaries, and listeners were able to use these cues to identify the intended meaning of a sentence.

Section 2: Productions From Trained Speakers

We were somewhat surprised at how rarely the speakers in Experiment 1 had pronounced the ambiguous sentences with sufficient prosodic cues to allow disambiguation. The obvious question was how badly they compared with trained speakers, so in Experiment 4 we used actors and broadcasters as speakers. The procedure for the experiment was the same as in the first experiment, and the same ambiguous sentences and context passages were used.

Table 5
Meaning Judgments From Experiment 3, Together With the Baseline Control Means

	Proportion of participants choosing Meaning A					
Item	Baseline	Presented A	Presented B	Context effect	*t*(30)	*SEM*
1	.55	.48	.28	.21	2.08	0.108
2	.72	.34	.53	−.18	−2.04	0.095
3	.35	.43	.87	−.44	−4.82	0.094
4	.41	.45	.64	−.19	−2.44	0.086
5	.48	.37	.62	−.24	−2.79	0.092
6	.55	.40	.69	−.29	−3.81	0.076
7	.55	.33	.58	−.25	−3.50	0.069
8	.48	.42	.81	−.39	−4.93	0.082

Note. No speaker pronounced Version A of Items 9 or 10 inappropriately. Therefore there were no data available for Experiment 3 for those items. $p < .05$, two-tailed, for all *t*s.

After our speakers had read the passages in both versions, we then explained the purpose of the experiment to them and asked them to intentionally make the sentences sound like they had one meaning or the other, pronouncing each one in isolation. In the case of items for which we did not find prosodic differences between the two contexts under naive instruction conditions, this allowed us to examine whether our speakers were completely unable to distinguish between the two meanings prosodically or were simply not likely to do so in normal reading without explicit instructions to do so. The experimentally naive condition for Items 1–10 is reported as Experiment 4a, and the session in which speakers were asked to intentionally produce meaning-appropriate prosodies is reported as Experiment 4b.

Experiment 4: Production Study

Method

Participants. The speakers for this experiment were 9 actors and broadcasters. Most were Northwestern University undergraduates or graduate students majoring in either broadcasting or performing arts, and all had amateur or professional experience in acting or broadcasting ranging from 1 to 15 years. Participants were recruited through ads posted on campus at the School of Drama and the Department of Radio, Television, and Film. Participants were paid $12 per hour for their participation in the experiment.

Materials. The 10 sentences and their context passages from Experiment 1 were also used in this experiment. These 10 passages were the materials that speakers read aloud in Experiment 4a.

In Experiment 4b, the materials were the list of sentences in Table 1 along with 2 paraphrases giving the possible meanings for each. These paraphrases can be found in Appendix B.

Design and procedure. In Experiment 4a, participants were given the same instructions as in Experiment 1. They were given all 10 passages in either Version A or B to read aloud and then were given the same passages in the other version to read. Five participants read Version A of each passage first, and 4 read Version B first.

After the first session (Experiment 4a), participants were told that the purpose of the experiment was to examine how differences in pronunciation could affect the meaning of ambiguous sentences, and they were given the list of ambiguous sentences and meanings shown in Appendix B. In Experiment 4b, they were instructed to read aloud each sentence on the list twice, once for each meaning, doing whatever they could to make the sentence sound like it had one meaning or the other.

Results

Judges' ratings. The sentences produced for Items 1–10 were rated for their contextual appropriateness as they had been in Experiment 1. Sentences that were mispronounced or had the word order changed (16% of the sentences produced in Experiment 4a and 11% in Experiment 4b) were discarded. The same two judges as in Experiment 1 rated each spoken sentence in Experiment 4 as having appropriate, inappropriate, or neutral prosody for its intended meaning. The judges agreed on 86% of the cases, and all but two of the disagreements were resolved by discussion between the two judges. The two cases for which agreement could not be reached were assigned a neutral rating of 2. Summary data for these ratings are displayed in Table 2, and the rating data for each item are included in Appendix B. Although the trained speakers in this experiment were able to consistently produce pronunciations that sounded appropriate for the intended meaning of the sentence when they were explicitly instructed to do so (in Experiment 4b), the ratings for the productions from Experiment 4a are not very different from those found in Experiment 1 with untrained speakers. For just the syntactically ambiguous items (2–10), the mean context effects were .22 in Experiment 4a and .79 in Experiment 4b, $t(8) = 7.70$, $SEM = 0.075$. Examination of the mean ratings for each speaker revealed considerable variability, but no clear relationship between performance and amount or type of professional experience (see Table 6). There was no correlation between years of experience in broadcasting or acting and ability to make the sentences prosodically distinguishable, either in Experiment 4a ($r = .17$, *ns*) or in Experiment 4b ($r = .01$, *ns*).

Table 6
Participant Data for Experiment 4

		Experiment 4a: Naive speakers							Experiment 4b: Instructed speakers						
		Context A			Context B				Meaning A			Meaning B			
P	Years of training	A	N	B	A	N	B	Context effect	A	N	B	A	N	B	Context effect
1	2	6	2	2	3	1	5	.32	10	0	0	1	2	7	.80
2	1	7	2	0	5	1	2	.24	8	2	0	3	3	4	.45
3	12	10	0	0	6	2	2	.30	7	3	0	2	0	8	.65
4	15	8	0	1	5	1	2	.24	9	0	0	3	0	6	.67
5	9	10	0	0	5	0	5	.50	10	0	0	0	0	10	1.00
6	4	6	1	3	5	0	5	.15	10	0	0	0	0	10	1.00
7	7	6	1	1	5	1	3	.18	10	0	0	0	1	9	.95
8	5	7	2	1	4	0	5	.37	8	0	2	3	0	7	.50
9	4	7	2	1	5	2	3	.20	10	0	0	0	0	10	1.00
Total		67	10	9	43	8	32	.28	82	5	2	12	6	71	.78

Note. Number of sentences spoken by each participant that was judged to be appropriate for Meaning A (A), Meaning B (B), or neutral (N) in each presentation context. Context effect is the number of utterances judged to fit their intended meaning minus the number of utterances judged to fit the incorrect meaning divided by the total number of utterances for that participant. P = participant.

Pitch-tracking analysis and duration analysis. The sentences were digitized for analysis. As had been the case in Experiment 1, all the pairs of productions for which both versions of the item had been judged to sound appropriate for their intended meanings were included in the analyses. Averaged pitch and amplitude values were calculated for each item, and the peak pitch and peak amplitude values for each word were also analyzed as in Experiment 1.

In Experiment 4a, only Item 1 had a sufficient number of acceptable pairs of productions for statistical tests to be performed, and the results for this item were similar to those found in Experiment 1 (see Table 7). The mean word and pause durations are displayed in Figure 4 for the seven items for which at least 1 speaker produced an acceptable pair of utterances, and the durations were again qualitatively consistent with the existence of phrase-final lengthening preceding the critical phrase boundary in both versions of each item.

In Experiment 4b, there were at least four pairs of acceptable productions for each item. We therefore calculated averaged pitch and amplitude values and analyzed the peak pitch and peak amplitude values for each of the 10 items.

The pitch-tracking analysis for Item 1 replicated the pattern observed in Experiment 1 and Experiment 4a, with relative increases in pitch, amplitude, and duration for *Manny* when it, rather than *Anna,* was marked as the new information in the sentence (see Table 7). Thus, similar results were obtained for Item 1 with both trained and untrained speakers, and both with and without explicit instructions to pronounce the sentence in a way that would disambiguate its meaning.

All four of the NP-ambiguity items once again showed evidence that intonation phrase boundaries occurred before the second NP when the second and third NPs were grouped together, and after the second NP when it was grouped with the first NP. There was evidence of phrase-final lengthening at

Table 7
Mean Duration (in Milliseconds), Peak Pitch (in Hertz), and Peak Amplitude (in Arbitrary Units) for Critical Words and Pauses From Experiment 4

Experiment and item		Duration				Peak pitch				Peak amplitude			
	df	A	B	*SEM*	*t*	A	B	*SEM*	*t*	A	B	*SEM*	*t*
Experiment 4a													
Item 1													
Anna	6	241	255	17	−0.77	175	198	20	−1.14	7,407	7,402	810	0.01
Manny	6	406	349	13	4.53*	138	141	17	−0.16	5,410	3,967	1,123	1.28
Experiment 4b													
Item 1													
Anna	6	250	305	18	−3.10*	205	241	22	−1.62	7,520	8,067	564	−0.97
Manny	6	443	403	19	2.09 m	187	146	10	4.03*	7,788	4,251	571	6.19*
Item 2													
David	6	406	246	43	3.72*	187	184	11	0.21	7,781	8,006	573	−0.39
(pause)	6	215	40	56	3.14*								
Pat	6	240	278	51	−0.73	206	192	8	1.71	5,760	4,670	673	1.62
(pause)	6	26	331	25	−12.10*								
Item 3													
television	6	566	674	32	−3.38*	190	193	8	−0.29	9,348	9,837	909	−0.54
(pause)	6	12	244	36	−6.52*								
radio	6	545	370	21	8.44*	160	158	7	0.22	8,572	9,796	318	−3.85*
(pause)	6	218	11	34	6.16*								
Item 4													
belts	5	280	323	26	−1.66	141	133	10	0.91	6,236	5,560	279	2.42 m
(pause)	5	17	174	51	−3.10*								
bags	5	365	321	29	1.51	152	138	12	1.14	4,721	6,068	447	−3.02*
(pause)	5	155	23	27	4.82*								
or	5	203	139	28	2.24 m	164	136	7	4.16*	8,327	4,820	953	3.68*
Item 5													
pork	7	412	293	15	8.18*	184	166	5	3.59*	8,833	8,281	849	0.65
(pause)	7	263	21	38	6.59*								
chicken	7	311	456	16	−8.87*	174	169	11	0.37	6,066	5,678	638	0.64
(pause)	7	0	206	31	−6.66*								
and	7	150	181	12	−2.60*	148	159	6	−1.69	4,749	4,243	512	0.99
Item 6													
rose	3	271	332	25	−2.39 m	187	172	9	1.61	9,764	10,038	1,273	−0.22
early	3	395	304	36	2.51 m	164	185	18	−1.13	8,639	9,963	1,300	−1.02
Item 7													
review	7	424	530	36	−2.91*	173	168	8	0.54	7,155	8,194	472	−2.20 m
(pause)	7	0	204	73	−2.78*								
literally	7	581	479	29	3.46*	189	211	9	−2.41*	8,695	10,634	499	−3.89*
(pause)	7	201	0	95	2.12 m								

Note. For all *t*s, $p > .10$, two-tailed, unless noted as marginal (m, $.05 < p < .10$) or significant ($^*p < .05$). The standard error of the difference between means (*SEM*) is reported for each comparison. A = Context A; B = Context B.

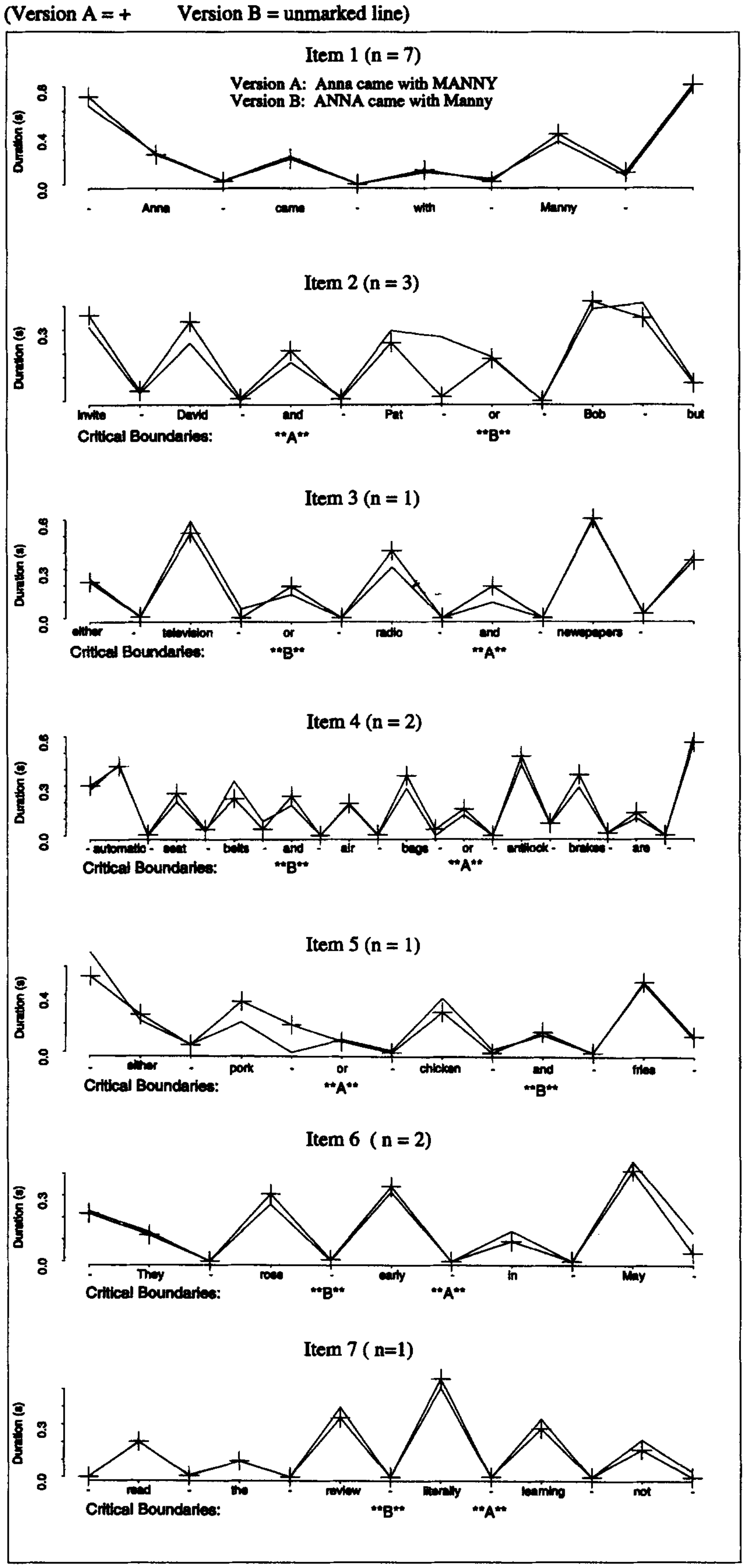

Figure 4. Mean word and pause durations from Experiment 4a, Items 1–7. Items 8–10 are not shown because no speaker produced the sentence correctly in both versions for any of those items. Number of speakers is noted for each item, and the location of the sentence's major syntactic boundary is indicated by **A** or **B** for Versions A and B, respectively.

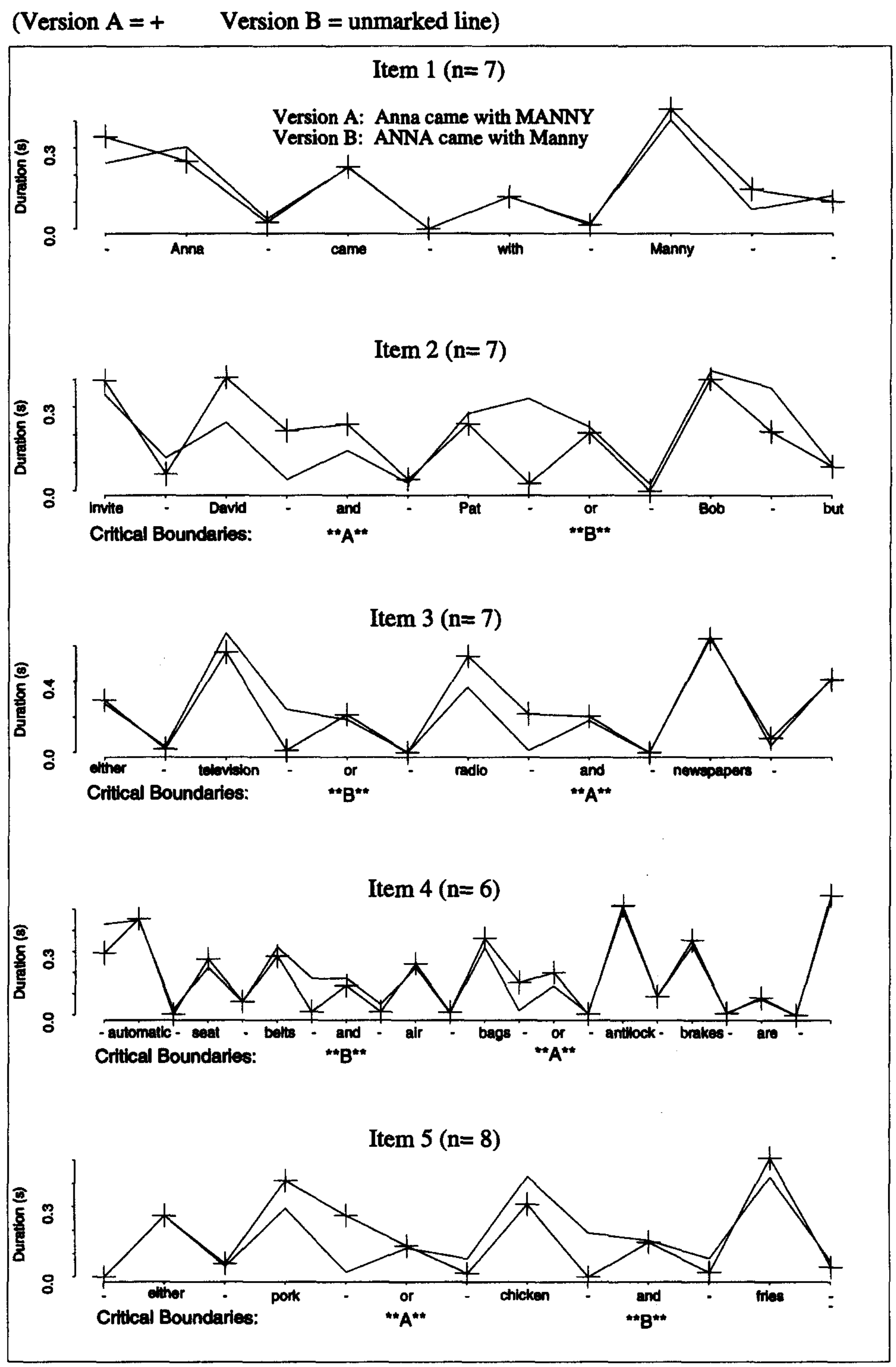

Figure 5. Mean word and pause durations for Items 1–10, Experiment 4b. Number of speakers is noted for each item, and the location of the sentence's major syntactic boundary is indicated by **A** or **B** for Versions A and B, respectively.

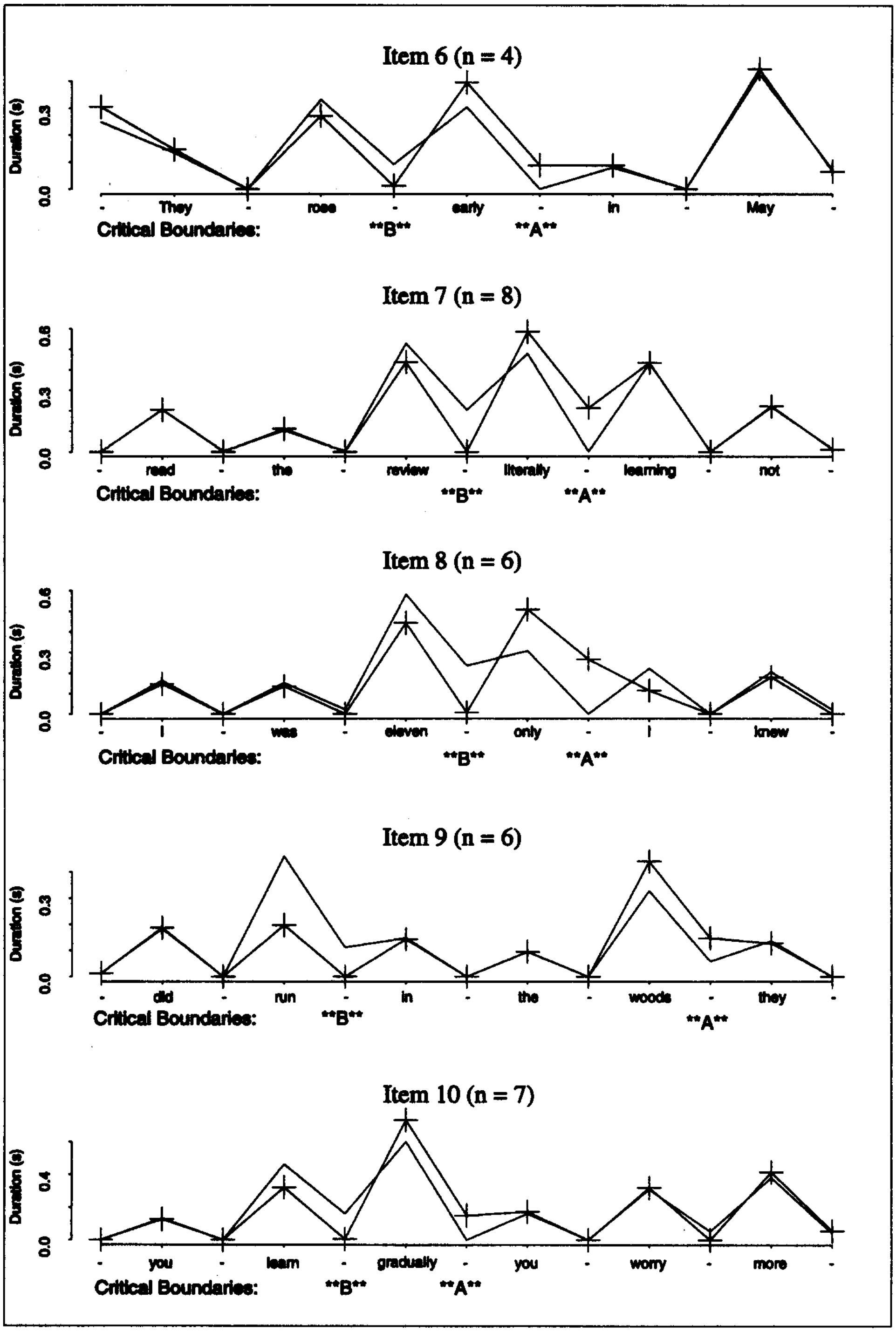

Figure 5 (Continued)

the critical NP boundaries and, to a lesser extent, evidence from differences in pitch and amplitude indicating the presence of intonation phrase boundaries that corresponded to constituent phrase boundaries at the contextually appropriate locations. Words and pauses preceding the major NP boundary tended to have longer durations and higher peak pitch and peak amplitude values, and in Items 4 and 5 the conjunction at the major NP boundary did as well (see Table 7 for mean duration, peak pitch,and peak amplitude values for critical words from Items 2–5, and Figure 5 for word and pause durations).

All five of the middle-phrase items also evidenced phrase-final lengthening at the contextually appropriate syntactic boundaries. In Version A of each of these items, the middle phrase should have been attached to the left according to the context, and in Version B it should have been attached to the right. We therefore expected that the sentences would be pronounced with an intonation phrase boundary preceding the middle phrase in Version B and following the middle phrase in Version A. Consistent with this prediction, the final word, pause, or both of the middle phrase was found to have a greater duration in Version A than in Version B, and the word, pause, or both immediately preceding the middle phrase was longer in Version B for all five items. The evidence from the peak pitch and peak amplitude analyses concerning the location of the intonation phrase boundary was less consistent, although for several of the items there was a significant drop in peak pitch for phrase-final words (see Figure 5 and Table 7).

Experiment 5: Judgment Study

Experiments 5a and 5b were identical in design and procedure to Experiment 2. In Experiment 5a, the sentences presented for meaning judgments were selected randomly from among the sentences produced in Experiment 4a that had been rated as having appropriate prosody for their intended meanings. One token for each version of each sentence was randomly selected from among those that had been rated 1 by both raters. The sentences for Experiment 5b were similarly selected from among the productions from Experiment 4b. Twenty-eight Northwestern University undergraduates participated in Experiment 5a, and another 30 participated in Experiment 5b.

The results of these judgment studies are presented in Table 8. For the sentences produced without explicit instructions (Experiment 5a), most of the items again showed a substantial context effect. The only exceptions were Item 9, for which no appropriate productions had been produced in the B context and no context effect was found, and Item 1, which showed only a small and nonreliable context effect, $t(27) = 1.31, p > .05$. The differences between the two versions of each of the other eight items were all statistically significant (all *t*s > 2.80, $p < .05$).

For the intentionally produced sentences (Experiment 5b), all 10 of the items showed a reliable context effect, all *t*s(29) > 4.0, $p < .05$. Overall, the context effect was larger for Experiment 5b (.83) than for Experiment 5a (.63). This supports the conclusion drawn from the rating data and the pitch-tracking and duration analyses that speakers can produce disambiguating prosodic cues for ambiguous sentences if they are explicitly asked to do so, but are less consistent in the prosodic cues they provide spontaneously.

General Discussion

In three experiments in which participants produced syntactically or pragmatically ambiguous utterances, we found that some speakers, for some sentences, could prosodically disambiguate the sentences. Speakers were sometimes able to produce measurable prosodic cues that allowed listeners in a

Table 8
Comprehension Results From Experiment 5a and 5b, Together With the Baseline Control and Punctuation Control Means (in Parentheses) From Experiment 2

	Proportion of participants choosing Meaning A					
Item	Baseline	Presented A	Presented B	Context effect	*t*	*SEM*
			Experiment 5a			
1	.55	.57 (.56)	.39 (.25)	.18 (.31)	1.31	0.137
2	.72	.89 (.96)	.07 (.17)	.81 (.79)	9.46*	0.079
3	.35	.93 (1.00)	.00 (.08)	.93 (.92)	18.73*	0.050
4	.41	.69 (.96)	.28 (.17)	.41 (.79)	2.77*	0.135
5	.48	.93 (.92)	.04 (.08)	.89 (.83)	15.00*	0.060
6	.55	.86 (.88)	.25 (.25)	.61 (.62)	5.67*	0.107
7	.55	.89 (.79)	.25 (.17)	.64 (.62)	6.09*	0.106
8	.48	.93 (.88)	.04 (.12)	.89 (.75)	15.00*	0.060
9	.72	.86 (.83)	.82 (.17)	.04 (.67)	0.37	0.096
10	.62	.93 (.88)	.04 (.00)	.89 (.88)	15.00*	0.060
			Experiment 5b			
1	.55	.62 (.56)	.13 (.25)	.49 (.31)	4.25*	0.114
2	.72	1.00 (.96)	.03 (.17)	.97 (.79)	25.85*	0.037
3	.35	1.00 (1.00)	.03 (.08)	.97 (.92)	23.55*	0.040
4	.41	.86 (.96)	.00 (.17)	.86 (.79)	12.45*	0.066
5	.48	.93 (.92)	.00 (.08)	.93 (.83)	20.15*	0.046
6	.55	.83 (.88)	.27 (.25)	.57 (.62)	4.01*	0.141
7	.55	1.00 (.79)	.00 (.17)	1.00 (.62)		
8	.48	.97 (.88)	.10 (.12)	.87 (.75)	13.73*	0.063
9	.72	.90 (.83)	.28 (.17)	.62 (.67)	5.17*	0.116
10	.62	1.00 (.88)	.03 (.00)	.97 (.88)	29.00*	0.033

*$p < .05$, two-tailed.

comprehension experiment to identify which of the two possible meanings of the utterance was the one intended by the speaker. Acoustical analyses of the sentences revealed that phrase-final lengthening cued the location of an ambiguous phrase structure boundary. This was true both for the four items we created that had ambiguous NP groupings and for the items with ambiguously attached middle phrases taken from Price et al. (1991). We also found evidence for the use of pitch accents to identify the new information in a sentence that contained a pragmatic ambiguity. The prosodic cues that allowed disambiguation were the same for both the trained and untrained speakers in our experiments. The analyses we report here are consistent with previous research showing that prosodic correlates to syntactic structure can be successfully used to resolve syntactic ambiguities (Beach, 1991; Price et al., 1991; Scott, 1982). In particular, phrase-final lengthening of word and pause durations was observed at the location of constituent boundaries whose correct location could not be determined on the basis of the syntactic structure of the sentences alone, and participants listening to the sentences were able to identify their correct interpretation.

However, most speakers, whether trained or not, did not produce prosodically disambiguated utterances for most sentences. Trained, professional speakers reliably produced appropriate disambiguating prosody only when they were shown the two meanings of a sentence side by side and were explicitly asked to pronounce the sentence twice, once with each meaning. Without the explicit instructions, only the pragmatically ambiguous sentence *Anna came with Manny* (Liberman & Pierrehumbert, 1984) was consistently produced with prosodic cues adequate for identifying the correct meaning, as evidenced by judges' ratings and by judgment studies with naive listeners. Our conclusion that uninstructed speakers in experimental situations do not typically produce disambiguating prosodic cues for syntactic ambiguities is limited by the small number of items we used and the limited number of types of syntactic ambiguities they contained. However, our items were similar to those used in previous research, and the ambiguities they contained were major ones in the sense that they determined important aspects of the meanings of the sentences.

The fact that speakers in experimental situations, even when they have had prior professional training, reliably produce prosodic cues to syntactic structure only when explicitly asked to do so raises serious questions about how prosody and its effects on comprehension processes can be studied. The most serious issue concerns the validity of generalization from experimental setting to real world. The prosody produced by the speakers in our experiments was clearly under their volitional control—they produced reliable cues when we asked them to. What we do not know is whether those cues are the same ones that would be produced for spontaneously generated ambiguous sentences in a natural setting. It directly follows that it cannot be learned from typical results in experimental settings what prosodic cues to syntax (if any) the human comprehension mechanisms use in a natural setting. The seriousness of this problem can be described by analogy to reading: It would be as if participants in reading experiments showed evidence of comprehension only when we explicitly instructed them to comprehend; if we simply instructed them to read (but didn't mention comprehending), they would show no evidence of reliable or correct understanding. If this were the case, we would certainly not want to generalize from our experiments to normal reading in the real world (either in terms of empirical findings or theories to explain those findings), and neither should we generalize from prosodic cues generated with explicit instructions to normal prosodic production. A most extreme conclusion from our results would be that our uninstructed speakers produced the items in our experiments exactly as they would have in a natural setting, that is, with almost none of the standard prosodic cues for disambiguation. Perhaps there were other cues in the course of the spoken passages as a whole, perhaps prosodic cues, perhaps semantic, syntactic, or pragmatic ones (cf. Deese, 1984; Frazier, 1987), that were sufficient to disambiguate the sentences, but these were not cues we could measure in the sentences themselves, and they were not cues that could allow disambiguation of the sentences removed from their passage contexts. Again, the point is that we are not justified in drawing any such conclusions because we have no way to demonstrate generalization from experimental results to natural behavior.

Although our results question whether laboratory studies using read speech can be generalized to the real world, this is not to say that nothing is known about the relationship between prosody and syntactic boundaries in naturally occurring speech. Lea (1973), for example, analyzed hundreds of naturally occurring spoken sentences and found consistent prosodic marking (in the form of a fall–rise intonation pattern) of syntactic boundaries. A number of other studies of spontaneous speech have also found correlations (but not complete correspondence) between prosodic features and syntactic boundaries (Deese, 1984; Goldman-Eisler, 1968; Maclay & Osgood, 1959; Silverman, Blaauw, Spitz, & Pitrelli, 1992; Wichmann, 1991). In one such study, Deese (1984) examined 25 hr of spoken discourse and found that although the prosodic markers of sentence-final pausing and changes in intonation were statistically associated with sentence boundaries, for almost one fourth of the sentences in the corpus neither of these prosodic cues was present to signal the end of the sentence. Studies of pause location in spontaneous speech have found fairly consistent use of pauses at sentence boundaries but not at clause boundaries within a sentence (Beattie, 1983; Brotherton, 1979; Pawley & Hodgetts-Syder, 1983; Stenstrom, 1986). Although some differences have been found between the prosodic structures of read versus spontaneously produced speech (Blaauw, 1994; Howell & Kadi-Hanifi, 1991), similar patterns of phrase-final lengthening (Kloker, 1975; Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992) and fundamental frequency declination (Anderson & Cooper, 1986) have been observed in spontaneous speech and reading aloud, with both types of speech evidencing only a partial correlation of these prosodic markers with syntactic boundaries.

Although it seems clear that there is some relationship between prosodic and syntactic structure in both spontaneous and read speech, the fact that naive speakers in our experiments did not consistently use prosody to mark the location of critical boundaries in syntactically ambiguous sentences (a case in which prosodic cues would seem to be particularly useful) leads us to conclude one of two things. Either the importance of prosodic cues for determining syntactic structure is relatively minimal in comparison with other types of cues available in a discourse context, or what speakers do in a laboratory setting does not generalize to naturally occurring speech. If the question of the generalizability of laboratory findings is to be resolved, the key information for doing so will likely come from future studies of naturally occurring speech.

The results of the experiments presented here, like many experiments before, are consistent with the claim that it is possible to use prosody to help inform parsing decisions. However, the experiments also make clear that it is not yet possible to say how much naive speakers in natural settings rely on prosody to disambiguate their utterances, nor is it possible to say how much comprehension processes typically rely on prosody to parse spoken sentences.

References

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition, 30,* 191–238.

Anderson, S. W., & Cooper, W. E. (1986). Fundamental frequency patterns during spontaneous picture description. *Journal of the Acoustical Society of America, 79,* 1172–1174.

Beach, C. M. (1988). The influence of higher level linguistic information on production of duration and pitch patterns at syntactic boundaries. *Journal of the Acoustical Society of America, 84* (Suppl. 1), S99.

Beach, C. M. (1991). The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of Memory and Language, 30,* 644–663.

Beattie, G. (1983). *Talk: An analysis of speech and nonverbal behavior in conversation.* Bristol, PA: Taylor & Francis.

Blaauw, E. (1994). The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. *Speech Communication, 14,* 359–375.

Boland, J. E., & Tanenhaus, M. K. (1991). The role of lexical representations in sentence processing. In G. B. Simpson (Ed.), *Understanding word and sentence* (pp. 331–366). Amsterdam: North-Holland.

Boland, J. E., Tanenhaus, M. K., & Garnsey, S. M. (1990). Evidence for the immediate use of verb control information in sentence processing. *Journal of Memory and Language, 29,* 413–432.

Brotherton, P. (1979). Speaking and not speaking: Process for translating ideas into speech. In A. W. Siegman & S. Feldstein (Eds.), *Of speech and time: Temporal speech patterns in interpersonal contexts* (pp. 179–209). Hillsdale, NJ: Erlbaum.

Cooper, W. E., Paccia, J. M., & Lapointe, S. G. (1978). Hierarchical coding in speech timing. *Cognitive Psychology, 10,* 154–177.

Cooper, W., & Paccia-Cooper, J. (1980). *Syntax and speech.* Cambridge, MA: Harvard University Press.

Cooper, W. E., & Sorensen, J. M. (1977). Fundamental frequency contours at syntactic boundaries. *Journal of the Acoustical Society of America, 62,* 683–692.

Crain, S., & Steedman, M. (1985). On not being led up the garden path: The use of context by the psychological syntax processor. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 320–358). Cambridge, England: Cambridge University Press.

Deese, J. (1984). *Thought into speech: The psychology of a language.* Englewood Cliffs, NJ: Prentice-Hall.

Ferreira, F. (1993). The creation of prosody during sentence production. *Psychological Review, 100,* 233–253.

Frazier, L. (1978). *On comprehending sentences: Syntactic parsing strategies.* Bloomington: Indiana University Linguistics Club.

Frazier, L. (1987). Structure in auditory word recognition. *Cognition, 25,* 157–187.

Frazier, L. (1991). Exploring the architecture of the language processing system. In G. T. Altmann (Ed.), *Cognitive models of speech processing* (pp. 409–433). Cambridge, MA: MIT Press.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in analysis of structurally ambiguous sentences. *Cognitive Psychology, 14,* 178–210.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech.* New York: Academic Press.

Howell, P., & Kadi-Hanifi, K. (1991). Comparison of prosodic properties between read and spontaneous speech material. *Speech Communication, 10,* 163–169.

Klatt, D. (1975). Vowel lengthening is syntactically determined in connected discourse. *Journal of Phonetics, 3,* 129–140.

Kloker, D. (1975). Vowel and sonorant lengthening as cues to phonological phrase boundaries [Abstract]. *Journal of the Acoustical Society of America, 57,* S33.

Lea, W. A. (1973). An approach to syntactic recognition without phonemics. *IEEE Transactions on Audio and Electroacoustics, AU-21,* 249–258.

Lehiste, I. (1973). Phonetic disambiguation of syntactic ambiguity. *Glossa, 7,* 197–222.

Lehiste, I., Olive, J. P., & Streeter, L. (1976). Role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustical Society of America, 60,* 1199–1202.

Liberman, M., & Pierrehumbert, J. (1984). Intonational invariants under changes in pitch range and length. In M. Aronoff & R. Oehrle (Eds.), *Language sound structure* (pp. 157–233). Cambridge, MA: MIT Press.

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word, 15,* 19–44.

Marslen-Wilson, W. D., & Tyler, L. (1980). The temporal structure of spoken language understanding. *Cognition, 8,* 1–71.

McClelland, J. L., St. John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes, 4,* 287–336.

Nagel, H. N., Shapiro, L. P., & Nawy, R. (1994). Prosody and the processing of filler-gap sentences. *Journal of Psycholinguistic Research, 23,* 473–485.

Nespor, M., & Vogel, I. (1983). Prosodic structure above the word. In A. Cutler & D. R. Ladd (Eds.), *Prosody: Models and measurements* (pp. 123–140). New York: Springer-Verlag.

Pawley, A., & Hodgetts-Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–226). London: Longman.

Price, P., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America, 90,* 2956–2970.

Price, P., Ostendorf, M., Shattuck-Hufnagel, S., & Veilleux, N. (1988). A methodology for analyzing prosody. *Journal of the Acoustical Society of America, 84* (Suppl. 1), S99.

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior, 22,* 358–374.

Scott, D. R. (1982). Duration as a cue to the perception of a phrase boundary. *Journal of the Acoustical Society of America, 71,* 996–1007.

Selkirk, E. (1984). *Phonology and syntax.* Cambridge, MA: MIT Press.

Silverman, K., Blaauw, E., Spitz, J., & Pitrelli, J. F. (1992). A prosodic comparison of spontaneous speech and read speech. *Proceedings of the International Conference on Spoken Language Processing, Banff, Canada, Vol. 2,* 1299–1302.

Sorensen, J. M., & Cooper, W. E. (1980). Syntactic coding of fundamental frequency in speech production. In R. A. Cole (Ed.), *Perception and production of fluent speech* (pp. 399–440). Hillsdale, NJ: Erlbaum.

Speer, S. R. (1995, March). *The influence of prosodic structure on the resolution of temporary syntactic closure ambiguities.* Paper presented at the Eighth Annual CUNY Conference on Sentence Processing, Tucson, Arizona.

Speer, S. R., & Bernstein, M. K. (1992, July). *Prosodic resolution of temporary syntactic ambiguity.* Paper presented at the 25th International Congress of Psychology, Brussels, Belgium.

Steedman, M., & Altmann, G. (1989). Ambiguity in context: A reply. *Language and Cognitive Processes, 4,* 105–122.

Stenstrom, A. (1986). A study of pauses as demarcators in discourse and syntax. In J. Aarts & W. Meijs (Eds.), *Corpus linguistics II: New studies in the analysis and exploitation of computer corpora* (pp. 203–218). Amsterdam: Rodopi.

Streeter, L. A. (1978). Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America, 64,* 1582–1592.

Taraban, R., & McClelland, J. (1988). Constituent attachment and thematic role assignment in sentence processing: Influences of content-based expectations. *Journal of Memory and Language, 27,* 597–632.

Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 528–553.

Wales, R., & Toner, H. (1979). Intonation and ambiguity. In W. E. Cooper & E. C. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 135–158). Hillsdale, NJ: Erlbaum.

Wichmann, A. (1991). A study of up-arrows in the Lancaster/IBM spoken English corpus. In S. Johansson & A. Stenstrom (Eds.),*English computer corpora: Selected papers and research guide* (pp. 165–178). New York: Mouton de Gruyter.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America, 91,* 1707–1717.

Appendix A

Sample Materials for Experiments 1 and 4a

1. Anna came with Manny

A.

Phil and Susan were gossiping about who they had seen together at last night's party.
"Did you see who Ben was with?" Phil asked.
"Yeah, I can't believe he and Laura are back together," Susan said.
Then Phil asked, "What about Anna? Who did she come with?"
"Anna came with Manny," Susan replied.
"I think they make a nice couple."

B.

Phil and Susan were gossiping about who they had seen together at last night's party.
"Did you see who Ben was with?" Phil asked.
"Yeah, I can't believe he and Laura are back together," Susan said.
Then Phil asked, "What about Manny? Who came with him?"
"Anna came with Manny," Susan replied.
"I think they make a nice couple."

2. We invite David and Pat or Bob.

A.

David's roommates, Pat and Bob, really don't get along.
In fact, they usually try to avoid each other, as much as that's possible for roommates.
Whenever there's a party in the frat house, David will come, and Pat or Bob will come, but you won't see them all together.
For our parties, we invite David and Pat or Bob, but not all three.

B.

David and Pat really don't get along with their roommate Bob.
In fact, they usually try to avoid him, as much as that's possible for roommates.
Whenever there's a party in the frat house, David and Pat will come, or Bob will come, but you won't see them all together.
For our parties, we invite David and Pat or Bob, but not all three.

6. They rose early in May.

A.

In spring there was always more work to do on the farm.
May was the hardest month.
They rose early in May.

B.

Bears sleep all winter long, usually coming out of hibernation in late April, but this year they were a little slow.
They rose early in May.

(*Appendix B follows on next page*)

Appendix B

Ambiguous Sentences as Presented in Experiment 4b (Intentional Production Condition) and Mean Context Effect From the Judges' Ratings of the Ambiguous Sentences Produced by Speakers in Experiments 1, 4a, and 4b

Experiment	Context effect
1. Anna came with Manny.	
a. (In response to: "What about Anna? Who did she come with?")	
b. (In response to: "What about Manny? Who came with him?")	
Exp. 1	.87
Exp. 4a	.83
Exp. 4b	.78
2. For our parties, we invite David and Pat or Bob, but not all three.	
a. ("David and [either Pat or Bob]")	
b. ("Either [David and Pat] or else Bob")	
Exp. 1	.40
Exp. 4a	.50
Exp. 4b	.78
3. They will use either television or radio and newspapers to announce the sale.	
a. ("[Either television or radio] and definitely newspapers")	
b. ("Either television alone or [both radio and newspapers]")	
Exp. 1	.40
Exp. 4a	.18
Exp. 4b	.83
4. Automatic seat belts and air bags or antilock brakes are standard.	
a. ("Either [both automatic seat belts and air bags) or [antilock brakes alone]")	
b. ("[Automatic seat belts] and [either air bags or antilock brakes]")	
Exp. 1	.05
Exp. 4a	.22
Exp. 4b	.83
5. So, for lunch today he is having either pork or chicken and fries.	
a. ("Either [pork alone] or else [chicken and fries]")	
b. ("[Either pork or chicken] and [fries]")	
Exp. 1	−.12
Exp. 4a	.38
Exp. 4b	1.00
6. They rose early in May.	
a. ("They rose early in the morning during the month of May")	
b. ("They rose during the early part of May" or "on May 1st")	
Exp. 1	.52
Exp. 4a	.39
Exp. 4b	.56
7. Rollo read the review literally learning not an iota.	
a. ("Rollo read it literally rather than figuratively")	
b. ("Rollo read it learning literally nothing")	
Exp. 1	.05
Exp. 4a	.14
Exp. 4b	.89
8. As I was eleven only I knew my Dad would be angry.	
a. ("I was only eleven")	
b. ("I was the only one that knew")	
Exp. 1	.29
Exp. 4a	.25
Exp. 4b	.78

Appendix B *(continued)*

Experiment	Context effect
9. Although they did run in the woods they were uneasy.	
a. ("They ran in the woods and they were uneasy")	
b. ("They ran but when they were in the woods they were uneasy")	
Exp. 1	.04
Exp. 4a	.00
Exp. 4b	.67
10. When you learn gradually you worry more.	
a. ("You worry more when you learn gradually, compared to if you learned quickly")	
b. ("You gradually begin to worry more when you learn")	
Exp. 1	.13
Exp. 4a	−.11
Exp. 4b	.78

Received December 2, 1994
Revision received August 21, 1995
Accepted August 31, 1995 ■