Prosodic Bootstrapping: A Critical Analysis of the Argument and the Evidence

Anne Fernald and Gerald McRoberts
Stanford University

The hypothesis that young infants rely on prosodic cues in speech to bootstrap their way into syntax has received considerable attention in recent discussions of early language development (e.g., Gleitman, Gleitman, Landau, & Wanner, 1988; Hirsh-Pasek, Kemler Nelson, Jusczyk, Cassidy, Druss, & Kennedy, 1987). The appeal of the prosodic bootstrapping hypothesis is easy to understand. If the boundaries between syntactic constituents in speech were indeed reliably marked by constellations of prosodic features such as pauses, pitch contours, and vowel lengthening, this acoustic punctuation could potentially be useful to the child beginning to learn language. And if this syntax-to-prosody mapping were more distinctive and reliable in infant-directed speech (IDS) than in adult-directed speech (ADS), the prosodic structure of IDS could provide even greater support for the infant’s initial efforts at parsing the speech stream. The prosodic bootstrapping hypothesis has captured the imagination of many researchers in the field on the strength of its apparent plausibility and explanatory promise.

We argue here, however, that the popularity of the prosodic bootstrapping notion has proceeded far in advance of the data necessary to support it. Although a few critical voices have been raised (e.g., Pinker, 1987), there has been insufficient attention either to the logic of the argument or to the limitations of the data. Support for the prosodic bootstrapping hypothesis rests on a selective use of indirect evidence, and some of the central findings cited in its favor need further replication. Because this hypothesis is about how the child uses prosodic cues to induce grammatical rules, it can be directly tested only by manipulating the relation of syntactic and prosodic units in speech to the child and assessing the effects of these manipulations on language acquisition. In the absence of direct evidence, prosodic bootstrapping advocates rely on indirect evidence to argue that prosodic features in language input are operative in the child’s induction of language structure. The force of the argument lies not in any particular finding, but in the broad sweep of what appears to be convergent evidence from diverse sources. Four major categories of indirect evidence are typically cited:
Descriptive studies suggesting that characteristic prosodic features are consistently associated with syntactic boundaries in ADS (e.g., Cooper & Sorensen, 1981);

Experimental studies showing that prosodic cues influence adults’ perception of syntactic boundaries (e.g., Morgan, Meier, & Newport, 1987; Streeter, 1978);

Descriptive studies indicating that prosodic cues to utterance boundaries are more exaggerated in IDS than ADS (e.g., Fernald & Mazzie, 1991; Morgan, 1986);

Developmental research suggesting that young infants are sensitive to syntax–prosody relations, even in unfamiliar languages (e.g., Hirsh-Pasek et al., 1987; Jusczyk, Hirsh-Pasek, Kemler Nelson, Kennedy, Woodward, & Piwotz, 1992).

Our goal is to look closely at each of these frequently cited sources of evidence. We begin with a discussion of cue reliability, a central notion in the prosodic bootstrapping argument that is frequently misconstrued. Our review of the literature on syntax–prosody mappings in ADS leads us to conclude that the reliability of prosodic cues to syntax has been overestimated. We come to a comparably cautious conclusion for IDS, arguing that the cue reliability of prosodic features in speech to infants has not been properly assessed and is relatively low. Finally, we review results of recent experiments attempting to replicate and extend the widely cited finding that young infants prefer to listen to speech segmented at clause boundaries (Hirsh-Pasek et al., 1987; Kemler Nelson, Hirsh-Pasek, Jusczyk, & Cassidy, 1989). These results also suggest that the prosodic bootstrapping hypothesis does not have strong prima facie plausibility.

**CUE RELIABILITY**

The issue of cue reliability is critical to the prosodic bootstrapping story because of the central claim that linguistic novices exploit low-level prosodic cues, discernible to even very young infants, to gain access to higher order linguistic structures that are not at first discernible. Cue reliability is a measure of the extent to which a cue is consistently associated with a particular structure or function and not with others, expressed numerically as “the ratio of the cases in which a cue leads to the correct conclusion, over the number of cases in which it is available” (Bates & MacWhinney, 1987, p. 164). That is, if \( C \) = the number of times when a particular cue is present and leads to the correct conclusion, and \( I \) = the number of times when that same cue is present but leads to an incorrect conclusion, then the reliability of the cue is reflected in the ratio \( C/(C+I) \). In the context of interest here, cue reliability is equivalent to the conditional probability that a particular linguistic structure is present given the occurrence of a particular prosodic cue, \( p(\text{structure} \mid \text{cue}) \). In the debate on prosodic bootstrapping this relation has frequently been confused with a second conditional probability, \( p(\text{cue} \mid \text{structure}) \), which is not an appropriate measure of cue reliability.

An analogy may help to make this distinction clear. Imagine a novice electrician confronted with a cable from which 100 identical wires emerge. Forty of these wires are “hot” because they are connected to switches at the other end of the cable, while the remaining 60 are unconnected and “cold.” The task of the apprentice is to identify the hot wires and connect them to a panel controlling 40 lights. Since the apprentice cannot see the other end of the cable, she initially has no idea where to begin. To help her, a supervisor takes 30 red “cue” labels and tapes each to a different hot wire. Thus the probability that a hot wire is marked with a cue label, \( p(\text{cue} \mid \text{hot wire}) \), is 0.75. Although 25% of the hot wires are not marked, the labels still serve as highly informative cues to structure because they occur only in association with hot wires and never with cold wires. The structural information these labels provide may be incomplete, but it is never misleading. Because \( p(\text{hot wire} \mid \text{cue}) = 1.0 \), every time the apprentice selects a wire with a red cue flag, she makes a correct choice. Now imagine a similar situation, except that the supervisor, out of perversity or ineptitude, places 30 red labels on hot wires and another 30 labels on cold wires. In this case, \( p(\text{hot wire} \mid \text{cue}) \) drops from 1.0 to 0.50. When the apprentice searches for the hot wires and chooses those wires marked with cue labels, she will be wrong on average half the time. However, \( p(\text{cue} \mid \text{hot wire}) \) remains high at 0.75, because the proportion of hot wires marked with cue labels is unchanged. What has changed is that there are now misleading cues as well, resulting in frequent false alarms. Although the probability of a red label given a hot wire is as strong as before, the cue reliability of the labels has been reduced by half. Of course, whether or not cues with a reliability of 0.50 are still useful to the learner depends on how well the learner would do with no cues at all. In this case, choosing wires randomly would yield a hot wire on average only 40% of the time. Thus it would still be marginally more effective to use the cue labels than to ignore them, although the error rate would be high in either case.

The important point here is that from the perspective of a novice who has no initial knowledge of the structure and must rely entirely on cue markers to identify the structure, it is \( p(\text{structure} \mid \text{cue}) \) that is the relevant conditional probability, not \( p(\text{cue} \mid \text{structure}) \). However, proponents of prosodic bootstrapping have relied on the second of these conditional probabilities rather than the first to build their case. As we discuss in the following sections, \( p(\text{cue} \mid \text{structure}) \) can be reassuringly high while \( p(\text{structure} \mid \text{cue}) \), the appropriate measure of cue reliability, is often quite low.
PROSODIC CUES TO SYNTACTIC UNITS IN ADULT-DIRECTED SPEECH

One line of reasoning in the prosodic bootstrapping argument is that in ADS, clause and phrase boundaries are consistently marked by pauses, pitch movement, and vowel elongation. The same few studies are typically cited to support this claim (e.g., Cooper & Paccia-Cooper, 1980), although there is an extensive literature on prosody-syntax mappings in which the findings are far from consistent. In this section, we briefly review research on linguistic and non-linguistic factors that influence the distribution of the three prosodic features associated with syntactic boundaries: pauses, fundamental frequency (F₀), and duration.

Pause Structure as a Cue to Syntax

Systematic empirical research on pauses in fluent speech began in earnest in the 1950s (e.g., Goldman-Eisler, 1951). Reviewing early research on non-linguistic determinants, Rochester (1973) cited 23 studies published from 1956 to 1970 on the effects of cognitive, affective, and social variables on pause production. Factors such as task difficulty (Levin, Silverman, & Ford, 1967), anxiety level (Fope, Blass, Siegman, & Raher, 1970), personality (Ramsey, 1968), and audience (Lay & Paivio, 1969) all affect how often hesitations occur in speech. Psycholinguists also began to explore how pause structure relates to linguistic structure. Maclay and Osgood (1959) found that pauses in spontaneous speech occurred as often within phrases as between phrase or clause boundaries, and that speakers showed striking individual differences in the distribution of pauses in their speech.

As described earlier, the relation between pauses and syntactic constituents in speech can be viewed from two angles. One approach is to consider the probability that a syntactic boundary will be marked by a pause, \( p(\text{pause} \mid \text{boundary}) \). In an analysis of formal academic speech, Goldman-Eisler (1972) reported that 90% of all sentences and 52% of all clauses within sentences were bounded by pauses >250 msec. The combined probability that a pause >250 msec occurred given a clause or a sentence boundary can be estimated at 0.60 from Goldman-Eisler's data. This finding that clauses are frequently not marked by pauses in fluent speech is echoed in other studies as well. Garro and Parker (1982) found that restrictive relative clauses were typically preceded by little or no pause (\( M = 1 \) msec) and followed by very short pauses (\( M = 19 \) msec). Grosjean, Grosjean, and Lane (1979) also observed that NP/VP boundaries are rarely marked by pauses when the NP is short.

In Goldman-Eisler's (1972) data, 40% of the clause boundaries were not marked by pauses. This is not in itself a problem for the prosodic bootstrapping argument, since pauses could still be reliable cues to clause structure if they occurred only at boundaries and never at other locations in the sentence. To see if this condition holds, we need to examine the cue reliability of pauses in this sample, \( p(\text{clause boundary} \mid \text{pause}) \). If \( p(\text{clause boundary} \mid \text{pause}) \) were close to 1.0, a listener with no knowledge of syntax could infer with high accuracy that a stretch of speech bounded by pauses was in fact a clause. If so, even though pauses were present only 60% of the time they were "needed," they would rarely give wrong information when they were present. In reality, however, pauses were as likely to occur within clauses as at clause boundaries in Goldman-Eisler’s data. Looking at the number of pauses 1000 msec or longer, she noted that 119 occurred between words within clauses, while 83 occurred at within-sentence clause boundaries and 116 occurred at sentence boundaries. Even when all pauses >250 msec are taken into account, \( p(\text{clause boundary} \mid \text{pause}) \) is well below 0.50. This estimate is consistent with the earlier findings of Maclay and Osgood (1959) and later studies using instrumental measures that confirmed that only about 50% of pauses occur at sentence, clause, and phrase boundaries (Boomer & Dittmann, 1962; Hawkins, 1971; Henderson, Goldman-Eisler, & Skarbek, 1966). What this means is that the linguistically naive listener who relied on pauses as cues to the beginnings and endings of syntactic constituents would make incorrect inferences about half the time.

As in the hot wire scenario, the conditional probability relevant to the prosodic bootstrapping argument is \( p(\text{boundary} \mid \text{prosodic cue}) \), although most arguments for the plausibility of prosodic bootstrapping have been formulated the other way around, in terms of \( p(\text{prosodic cue} \mid \text{boundary}) \). Gleitman et al. (1988), for example, support their claim that infants could potentially rely on prosody to identify clause units with the statement that “recent evidence suggests that there are reliable acoustic cues to clause boundaries in speech, including longer pauses, segmental lengthening, declination of fundamental frequency, and stress marking” (p. 163). However, the one study of pauses cited as evidence was by Cooper and Paccia-Cooper (1980), who showed that when speakers read ambiguous sentences to convey alternative interpretations, pause durations vary with the interpretation. Cooper and Paccia-Cooper started with theoretical linguistic units and measured associated pauses, reporting that \( p(\text{pause} \mid \text{boundary}) \) was high. It was not their goal to determine whether equivalent pauses occurred elsewhere in the speech stream, although from the infant’s point of view, this is crucial information. Because the infant must proceed from cue-to-structure rather than from structure-to-cue, these findings tell only part of the story and not the part most relevant to the infant’s task. It is interesting that clauses of different types are bounded by pauses of different durations; however, for the novice with no knowledge of what constitutes a clause, it is more important to know how often pauses serve as reliable cues to clause boundaries and how often they are misleading.

No estimate of actual cue reliability can be derived from the Cooper and Paccia-Cooper (1980) data because unreliable pauses were not assessed. Moreover, the number of unreliable pauses in that study would...
F₀ has been proposed as another cue infants might exploit to gain access to syntactic structure. Although under some conditions sentences read aloud show F₀ fall-rise patterns at certain types of clause boundary (Garro & Parker, 1982), such evidence fails to address the crucial question of cue reliability. In order to demonstrate high cue reliability, it must be shown that F₀ fall-rise occurs primarily at syntactic boundaries and not elsewhere in the sentence, and there is no evidence indicating that this is the case.

Duration as a Cue to Syntax

Several studies show that speech segments are lengthened in phrase-final positions (e.g., Cooper, Paccia, & Lapointe, 1978; Sorenson, Cooper, W. E., & Paccia, 1978), and thus that p(boundary | cue) is relatively high. For the linguistically naive listener, however, what is most important is how often lengthening occurs at syntactic boundaries as compared to other locations. Klatt (1976) provided detailed data on segmental durations that allow us to calculate p(boundary | cue) and assess cue reliability. In an acoustic analysis of 13 sentences of connected discourse, Klatt compared the durations of segments occurring between and within syntactic constituents. To normalize for differences in intrinsic duration associated with phonetic category, the durations of vowels and consonants located at phrase and clause boundaries were compared to the median duration of the same segments averaged across all positions in the sentence. In Klatt’s sample, 61 words contained vowels or consonants greater than 1.2 times their median. Of these words with lengthened segments, 42 occurred at phrase or clause boundaries, and 19 occurred elsewhere within phrasal constituents; thus the cue reliability of segmental lengthening in this sample was 0.65.

That duration is only moderately reliable as a cue to syntactic structure should not be surprising, since many of the non-linguistic factors that influence pause structure and F₀ contour in fluent speech also influence segmental duration. Duration is affected by the speaker’s mood (e.g., Williams & Stevens, 1972), speech rate (e.g., Crystal & House, 1982, 1990), and degree of emphasis (e.g., Cooper, Eady, & Mueller, 1985), as well as by non-syntactic linguistic factors (e.g., House & Fairbanks, 1953). One major source of variability in vowel duration is whether or not the syllable is stressed. Crystal and House (1988) found that stressed vowels in connected discourse were twice as long on average as unstressed vowels. However, the primary source of variability in segmental duration is the phonetic category of the segment. For example, the mean duration for stressed /e/ in Klatt’s (1976) corpus was 155 msec, while the mean duration for stressed /i/ was 70 msec. In a factor analysis, Klatt found that 56% of the variance in stressed vowel duration was accounted for by differences in inherent duration among vowel categories. Although vowels in phrase-final syllables were on
average 40 msec longer than the median for that vowel type, this syntactic factor accounted for only 16.2% of the variance.

There is another reason to be skeptical about the claim that duration is a cue to syntax accessible to very young infants. The problem is that segmental lengthening is a relative measure. Given substantial differences in intrinsic duration associated with different vowel categories, lengthening can only be appreciated in relation to the inherent duration of the vowel in question. For example, a lengthened /i/ is 30 msec shorter in duration than a non-lengthened /e/ (Klatt, 1976). As Klatt pointed out, in order to perceive lengthening as a cue to the end of a syntactic constituent, the listener needs to know the phonetic identity of the lengthened segment. The listener must also be able to take other factors simultaneously into account to compensate for lengthening or shortening effects due to speech rate, stress, emphasis, and phonetic context. It is not clear how adult listeners manage to use durational cues without first working out various other aspects of the sentence, and it is certainly not obvious that young infants should be able to do so.

**THE ROLE OF PROSODY IN ADULT SPEECH PROCESSING**

So far we have reviewed research on the availability of prosodic features as potential cues to syntactic structure. Next we consider research on whether adults actually make use of these prosodic features in speech processing. Numerous studies have examined how prosody influences listeners' perceptions of non-linguistic dimensions of speech such as affective state (e.g., Scherer, 1986), personality (Apple et al., 1979), and intention (e.g., Pierrehumbert & Hirshberg, 1990), as well as perceptions of linguistic organization at levels other than syntax, such as phonology (e.g., Nakatani & Schaffer, 1978) and discourse structure (Fowler & Housum, 1987; Nooteboom & Kruyt, 1987). These studies are not directly relevant to prosodic bootstrapping, except as a reminder that prosody is multiply determined and serves many kinds of communicative functions. We focus here on research on perceptual effects of particular prosodic features, the interaction of prosodic and syntactic cues, and grouping cues in artificial language learning, often cited in discussions of prosodic bootstrapping.

**Prosodic Cues and Boundary Perception in ADS**

When syntactically ambiguous sentences differing in prosodic structure are naturally spoken, listeners can usually identify the intended interpretation (Lehiste, 1973; Price et al., 1991). To discover which prosodic cues are important in resolving ambiguities, one approach has been to manipulate particular prosodic features while holding others constant. Henderson and Nelms (1980) studied the effects of pauses on segmentation, concluding that pauses were not a salient segmentation cue. However, Scott (1982) found that pauses were sufficient as a cue to word boundaries in ambiguous sentences, although a combination of pause plus lengthening was equally effective. The limited data on F0 contour as a cue to boundary perception are also inconsistent. de Rooij (1976) found that F0 fall alone was not effective as a boundary cue, and that duration alone was as powerful as duration plus F0 fall in combination. In contrast, Streeter (1978) claimed that the effects of duration and F0 contour were additive, while Beach (1991) found evidence that pitch and duration cues are perceived as an integrated percept. Although these few parametric studies do not provide a clear picture of the relative influence of different prosodic cues on segmentation, they suggest that duration may be the most consistently effective cue, and that prosodic cues in combination are probably more powerful than in isolation.

**Syntactic Boundary Perception Without Syntactic Cues**

Another approach to the question of how adults use prosody to parse the speech stream has been to use natural speech in which syntactic information is lacking or inaccessible, such as nonsense strings, filtered speech, or speech in an unfamiliar language. Here the question is whether listeners can identify constituent boundaries at all using only prosodic cues, in the absence of syntactic information. Since infants are also initially in the position of having no access to syntactic cues, studies using these methods seem particularly relevant to the prosodic bootstrapping hypothesis.

Wakefield, Doughie, and Yom (1974) and Pilon (1981) investigated sensitivity to constituent boundaries in an unfamiliar language, using Korean sentences with pauses inserted either between or within phrases. When asked which sounded more natural, American subjects tended to choose the sentence in which the pause coincided with the phrase boundary. Performance was far from perfect, however, with means varying from 43% to 80% across groups. Although cited as evidence in favor of the prosodic bootstrapping notion, these results are equivocal. First, the stimuli lacked the prosodic variability typical of natural conversation. Also, although the subjects were unfamiliar with Korean, it was not the case that syntactic cues were completely inaccessible to them. In Korean, particles used as topic markers frequently occur at the end of noun phrases, for example, and these repeated elements could have served as a reliable non-prosodic cue to the NP/VP boundary.

A more convincing approach would be to show that listeners can identify constituent boundaries in conversational speech from which all segmental information has been eliminated. However, adults find it very difficult to identify constituent boundaries using only prosodic information, even in their native language. Lehiste asked listeners to locate sentence boundaries in normal speech and speech that was either spectrally inverted (Lehiste & Wang, 1977) or low-pass filtered (Lehiste, 1979) to remove segmental content. In both studies there was substan-
tial disagreement between listeners' judgments of boundary locations in natural and content-filtered speech. Lehiste found that 50% of the sentences identified in the content-filtered condition were not heard as sentences in the normal speech condition. These included utterances such as Sanskrit at least as a possible and core of courses now of course that were preceded and followed by pauses >250 msec and other prosodic boundary cues. When only prosodic cues were available, listeners identified these fragments as sentences; in the natural speech condition, however, syntactic cues overrode prosody and dominated listeners' judgments. In a similar study using higher quality stimuli, Kreiman (1982) also found that subjects did not hear sentence boundaries in the same locations in normal and content-filtered speech. Because there was a 30% false alarm rate with filtered speech, and 35% of the normal speech sentences were not recognized when filtered, the overall rate of disagreement was high.

These findings illustrate several points relevant to the prosodic bootstrapping debate and the cue reliability issue. Adult listeners do use pauses and other prosodic cues to make decisions about the locations of syntactic boundaries. But to the extent that these prosodic features are uncorrelated with syntactic units (i.e., low in cue reliability), and that no syntactic information is available, listeners make mistakes. Moreover, when syntactic information is available, listeners rely on syntactic cues to override unreliable prosodic cues in identifying constituent boundaries.

Prosodic Cues to Syntactic Structure in Artificial Languages

A third approach to the question of how prosody affects syntactic processing focuses on conditions that facilitate rule induction in artificial linguistic systems. Morgan and Newport (1981) found that subjects learned the syntax of an artificial language much more effectively when information was available about constituent structure. When words in the input sentences were grouped visually into phrasal units, subjects were able to master the constituent structure; when words were not grouped at all, or were grouped arbitrarily with respect to syntactic boundaries, subjects failed to master the grammar of the language. Morgan et al. (1987) extended these findings in studies that included three different kinds of grouping cues: prosody, function words, and concord morphology. Here too they found that when cues to the phrase structure were present in the input, subjects learned the grammar of the artificial language; however, when cues were absent or arbitrarily located, subjects failed to learn the grammar.

These findings are frequently cited in discussions of prosodic bootstrapping because they suggest that phrase structure markers such as prosodic cues are crucial for language learning. As Morgan et al. (1987) acknowledged, however, a question of concern is whether research on artificial language learning can be generalized to the acquisition of natural languages. A major problem is that in these two studies, grouping cues were either absent altogether or present but uncorrelated with syntactic structure, or they were perfectly correlated with syntactic structure. That is, in the two conditions in which prosodic cues were available, cue reliability was either 0 (in the Arbitrary Prosody condition) or 1.0 (in the Phrase Prosody condition). When cue reliability was 1.0, subjects were successful at figuring out the rules of the grammar; when cue reliability was 0, subjects failed to learn the grammar, just as when no cues were present at all. In real conversational speech, as we have seen, prosodic cues are often missing at constituent boundaries, and when prosodic cues do occur, they are often misleading. So what can we conclude from these results about learning under more natural conditions in which cue reliability is somewhere midway between 0 and 1.0? Because Morgan et al. did not test subjects in a "noisy cue" condition with less than perfect cue reliability, it is unknown whether listeners would master the syntax in an artificial language more analogous to real speech.

To summarize, in discussions of prosodic bootstrapping, the same few studies of prosodic cues in adult speech production and perception are cited to support broad generalizations about how prosody marks syntactic boundaries and how listeners depend on prosodic cues in decoding speech. A closer reading of the extensive literature in these areas reveals that these generalizations are misleading. In spontaneous ADS, prosodic features are only moderately reliable as cues to syntax. While listeners benefit dramatically from prosodic cues perfectly correlated with syntax, performance is poor under more natural conditions, when listeners attempt to identify syntactic boundaries using only the imperfectly correlated prosodic cues typical of conversational speech.

PROSODY AND SYNTAX IN SPEECH TO CHILDREN

Even if prosody does not map reliably onto syntactic constituents in ADS, bootstrapping may still be an important mechanism during language acquisition. It is crucial to this argument, however, that prosodic and syntactic units should be highly correlated in IDS across languages.

Prosodic Cues in IDS

Although not directly related to the cue reliability issue, several studies show that prosodic features are exaggerated in IDS. Pauses are longer in IDS than in ADS in English and other languages; in fact, the pauses between ID utterances are on average longer than the utterances themselves (e.g., Fernald, Taeschner, Dunn, Papousek, Boysson-Bardies, B., & Fukui, 1989). Lengthening in clause- and phrase-final syllables is also exaggerated in English IDS (e.g., Bernstein Ratner, 1986; Morgan, 1986), although lengthening is not a prominent boundary marker in Japanese IDS (Fisher & Tokura, this volume). Finally, F0 modulation is
greater overall in IDS than in ADS across languages (e.g., Fernald et al., 1989). These exaggerated pitch patterns serve affective and attentional functions (e.g., Fernald, 1993) and may also serve linguistic ends. Because focused nouns typically occur on final pitch peaks in English IDS (Fernald & Mazzie, 1991), utterance boundaries are often marked by dramatic F0 movement. However, rising terminal pitch is much more common in IDS than in ADS in English, given the frequent use of questions and attentionals with infants. Ryan (1991) found that 39% of English ID utterances and 45% of Japanese ID utterances ended with rising pitch. Thus falling pitch is not the only F0 cue to utterance boundaries in IDS.

Very few studies have examined the reliability of prosodic cues in relation to syntax in IDS. However, two of the earliest systematic studies of language input were explicitly motivated by a version of the prosodic bootstrapping hypothesis. Broen (1972) and Dale (1974) were both aware of the findings of Maclay and Osgood (1959) and others that pauses occurred at grammatical boundaries in ADS only about 50% of the time. It was these findings that prompted them to investigate pause location in relation to sentence structure in IDS. In an analysis of pauses >260 msec, Broen found that 93% of sentence boundaries were followed by a pause in IDS, as compared to 29% in ADS. In addition to reporting pause duration, Broen also provided data on cue reliability, or pause (boundary | pause), which was 0.99 in IDS and 0.54 in ADS. Based on similar methods, the data in Dale's study reveal that cue reliability of pauses was 0.96 in speech to 2- to 4-year-olds. In a later study of German mothers' speech to newborns, Fernald and Simon (1984) reported that the reliability of pauses as cues to sentence boundaries was 0.98.

In addition to these studies of isolated prosodic features, Fisher and Tokura (1993, this volume) examined constellations of cues associated with syllables at word, phrase, and utterance boundaries in American and Japanese IDS. In a discriminant function analysis, 43% to 64% of what Fisher and Tokura refer to as "clausal" boundaries were correctly categorized using a combination of pause, segmental lengthening, and pitch movement as predictor variables, although there was no evidence that phrasal boundaries were prosodically marked. In terms of cue reliability, the probability that these correlated cues resulted in the correct classification of "clausal" boundaries was high, averaging 0.88 across the English and Japanese samples. Thus, although approximately 44% of the "clausal" boundaries in IDS were not identified correctly, the constellation of cues most consistently associated with such boundaries led to classification errors only 12% of the time. In other words, very few initial and medial syllables were incorrectly classified as marking "clausal" boundaries.

It is tempting to conclude from these studies that prosodic cues to syntactic structure are much more reliable in IDS than in ADS. However, there is a serious problem with this interpretation: In none of these studies were "sentences" or "clauses" rigorously defined using a linguistic criterion, as clauses are in studies of ADS. Broen (1972) and Fernald and Simon (1984) classified vocatives and exclamations as complete sentences; Fisher and Tokura (1993) categorized exclamations and elliptical utterances such as Oh and Very pretty as complete clauses, and Japanese sentence-final particles such as ne were also classified as "one syllable clauses." Although there is linguistic debate as to what exactly constitutes a clause—whether He wants to go consists of one clause or two, for example—Oh and ne would not qualify as complete clauses by any reasonable syntactic criterion. It would have been more appropriate to refer to these units of IDS as "utterances," and then to distinguish between clausal and non-clausal utterances, as defined by linguistic criteria. Using such an analysis, these investigators would probably have found that prosodic markers were no more reliably associated with clausal than with non-clausal utterance boundaries in their samples. Such data are needed to address an empirical question of central relevance to the prosodic bootstrapping argument: Are prosodic features indeed uniquely diagnostic of particular syntactic constituents in IDS? As we discuss in the next section, declaring all ID utterances to be "sentences" or "clauses" by fiat begs this important question.

The Syntactic Structure of ID Utterances

If early language input came prepackaged in clausal units, each syntactically complete and acoustically demarcated from the next, then the prosodic bootstrapping hypothesis would seem quite reasonable. The problem with this optimistic premise is that although prosody does provide robust cues to utterance boundaries in IDS, many of these utterances are not in fact clauses, as Fisher and Tokura (this volume) point out. In Newport's (1977) sample of maternal speech, 40% of the utterances consisted of sub-clausal fragments or stock expressions. We found a comparable distribution in an analysis of 100 American English ID utterances from each of five mothers in the Fernald et al. (1989) sample, in which fewer than half the utterances were complete simple clauses. More than 40% of the utterances were sub-clausal, consisting of single or multiword fragments (Doggy. Big black nose), attentionals (Hey there!), stock expressions (Thanks very much), and the like. Although a small proportion of these sub-clausal utterances consisted of complete noun or verb phrases, most were sub-phrasal as well.

The fact that a substantial proportion of ID utterances are sub-clausal fragments is problematic for the prosodic bootstrapping argument. A cornerstone of this argument is the claim that if infants can exploit bottom-up cues to identify utterance boundaries, they have access to a rudimentary phrase structure description of the sentence. As Gleitman et al. put it: "if we are correct in our reading of the evidence, bracketed (and partially labeled) clausal representations are available at the initial stage of language learning, and serve as the primary linguistic data to
be paired with meanings" (1988, p. 154). Using prosody to identify clausal and phrasal units, even without knowing the words, could help infants discover the rules governing the distributional patterns of the ambient language. After all, fundamental rules of English syntax such as those governing subject-verb agreement prevail within the domain of the clause, and other rules such as those specifying the distribution of determiners relative to nouns hold up only to the phrase boundary. If through prosody the infant has access to a phrase structure parsing of heard sentences, the job of discovering these constraints should be easier, since the units within which the constraints prevail would be delimited in advance. However, to the extent that the "units" demarcated by prosodic cues are highly variable in constituent structure, it is less clear how this form of prosodic prepackaging would facilitate the discovery of syntax. In the Fernald et al. (1989) sample, for example, a prosodically bounded multiword utterance might be a complete clause, or it might be a noun phrase, a verb phrase, or something else (Up and over! Hey good job!). An infant biased to assume that each of these prosodically isolated utterances had clausal structure would obviously run into frequent analytical difficulties.

The assumption that IDS makes syntax transparent for the infant has another problem: Even when clauses are complete, they are frequently not in canonical form. Fewer than 20% of the clausal utterances in our sample were simple declaratives, whereas the great majority were imperatives and interrogatives, a distribution similar to that reported by Newport, Gleitman, and Gleitman (1977). Newport et al. have eloquently argued that although ID utterances are typically short and focused on the here and now, it is a mistake to equate these features with syntactic "simplicity." In fact, imperatives with their missing subjects, and yes/no questions with their fronted auxiliary verbs, are in some sense syntactically more complex than the longer declarative sentences predominant in ADS. Thus even when prosodically bounded utterances in IDS do contain complete clauses, the task of discovering the phrase structure of the utterance is hardly straightforward for the infant.

How Useful Are Prosodic Cues That Are Low in Cue Reliability?

In the Fernald et al. (1989) study, utterances were defined using an acoustic criterion—as stretches of speech bounded by pauses >300 msec. The mean pause duration in American English IDS was 1312 msec. Since the mean utterance duration was only 1345 msec, such substantial pauses provided prominent boundary markers often accompanied by other prosodic cues as well. But what kinds of linguistic units were these acoustic cues actually marking? When the infant heard a pause, the probability that the pause marked a single complete clause was on average less than 0.50, and the probability that the pause marked a clause in canonical form was less than 0.10. We can conclude that prosody buys the infant utterance boundaries in IDS, but only sometimes
but that the 6-month-old American infants showed a preference only in English, no longer discriminating between the coincident and noncoincident stimuli in Polish. Finally, in a series of studies by the same research group, infants were presented with English speech stimuli edited so that pauses were coincident or non-coincident with phrase boundaries rather than clause boundaries (Jusczyk et al., 1992). Nine-month-old but not 6-month-old infants preferred the coincident stimuli, even when the speech was filtered to remove all segmental content.

These results have been interpreted as evidence that “clauses are perceptual units” for preverbal infants (e.g., Hirsh-Pasek et al., 1987). If even very young infants listen longer to natural prosody in an unfamiliar language, they must need little or no experience to establish this bias. The implication is that infants are born sensitive to clause-typical prosodic patterns, and that these clause-typical contours are universal across languages. However, exposure to a particular language begins to show its effects midway through the first year. Infants’ preference for coincident over non-coincident stimuli in Polish disappears by 6 months, although the preference for coincident stimuli in the ambient language is maintained. The story for sensitivity to phrasal units is different, since a preference for stimuli with appropriately segmented phrases, even in filtered speech, is shown by 9-month-old but not by 6-month-old infants. The explanation offered is that the larger clausal units are detected first, presumably by means of a language-independent awareness of the sound shape of clausal prosody, while sensitivity to the smaller subunits develops gradually with exposure to a particular language. Jusczyk et al. (1992) were careful to point out that infant listening preferences for speech with natural syntax—prosody mappings tell us nothing about whether infants actually use prosody in segmenting speech and figuring out syntax. However, these findings are often cited as if they pointed firmly in that direction. After all, if months before speaking a word, infants can recognize where clauses and phrases begin and end using only prosodic cues, this capability could potentially be powerful in enabling infants to induce the syntactic rules of their language.

There are several reasons to resist this interpretative leap. First, the research reviewed in previous sections indicates that neither in ADS nor in IDS are clauses consistently marked by prosodic cues. Thus even if infants can discriminate between natural and interrupted prosodic contours, this finding would not justify the inference that “clauses are perceptual units” for infants. Second, because the non-coincident speech stimuli used in these experiments consisted of artificially truncated utterances, they were “unnatural” in two quite different respects. The unnatural feature that was intentionally manipulated was the disruption of the normal relation of prosodic cues to utterance boundaries, since pauses were inserted midway through utterances without the correlated cues of falling pitch and final lengthening. However, since these truncated stimuli were created by splicing into continuous speech waveforms, they were also unnatural in the sense that human vocal tracts cannot produce such sounds. Even if carefully prepared using a waveform editor, the non-coincident stimuli stopped abruptly at points in the utterance where the vocal tract had not closed down. Thus the offset and onset characteristics of the non-coincident stimuli were different from those of naturally produced vocalizations. Because of this confound between prosodic and physiological unnaturalness, an alternative explanation for the findings is not ruled out: If infants indeed listen less to the non-coincident stimuli, it could be because they recognize what kinds of sounds human vocal tracts can and cannot produce and prefer physiologically possible vocalizations. This finding would also be interesting, but it would not provide evidence that infants are sensitive to the relation of prosodic and syntactic units in speech.

In addition to concerns about how an infant listening preference for continuous as compared to truncated utterances should be interpreted, the developmental findings need further replication. Jusczyk (1989) claimed that English-learning infants as young as 4 months show this listening preference, even in an unfamiliar language such as Polish, but that by 6 months the preference for coincident stimuli is limited to English. It is these results that are cited most often to support the inference that infants’ sensitivity to clause-typical prosody is perhaps innate and is initially language-independent, and that exposure to the ambient language in the early months reduces sensitivity to clause-typical prosody in languages to which the infant is not exposed. Because these findings have broad implications for theories of language development, it is essential that their replicability and generality be confirmed by other researchers working independently.

With the goal of extending these findings to other languages, we first undertook a replication of the Hirsh-Pasek et al. (1987) experiment in which 7-10-month-old infants were presented with coincident and non-coincident stimuli in English. Our speech stimuli were based on a simple children’s story called Bye Bye Baby (Ahlberg & Ahlberg, 1989) spoken with ID intonation. As in the original study, 1000 msec pauses were inserted either at clause boundaries or within clauses. Extreme care was taken to interrupt the waveform only at zero crossings, to avoid splicing artifacts. As in the original study, infants were tested using a modified version of the auditory preference procedure developed by Fernald (1985) in which stimuli were presented laterally, contingent on a criterion headturn. The only significant procedural difference was that in the Hirsh-Pasek et al. study the observer who controlled trial onsets and offsets stood directly behind the test booth listening to music or speech over headphones, while the observer in our study was located in a soundproof control room and viewed the infant’s eyes on a video monitor. This precaution was necessary in order to eliminate any possibility of observer bias, since we have found that even when music and speech maskers are presented at very high intensities, they are not...
completely effective at masking acoustic stimuli if the observer is located close to the loudspeakers (Fernald, Pinto, Cole, & McRoberts, in preparation). In this first study we found no significant difference between infants’ mean looking/listening times to stimuli with coincident (M = 4.8 sec) and non-coincident (M = 4.2 sec) pauses.

We decided to try again with subjects at 4, 7, and 10 months of age, using a potentially more sensitive auditory preference procedure. Developed by Cooper and Aslin (1990) for use with newborns, this procedure is also effective with older infants (Asgari, Pinto, & Fernald, 1993). Because sound presentation is contingent on the infant’s fixation of a central checkerboard display, headturns are not required and task demands are reduced compared to the Fernald (1985) procedure. This study included German as well as English stimuli, a comparison language chosen because of the strong similarities between German and English prosodic structure. Moreover, because the German stimuli were based on a direct translation of the Bye Bye Baby story, as told to an infant by a native speaker of German, we could make the coincident and non-coincident stimulus sets structurally parallel across languages. Subjects were 144 English-learning infants, 48 in each age group; half heard English, and half heard German. We found that English-learning infants did not discriminate at any age between coincident and non-coincident speech stimuli in German. Nor did 4-month-old infants discriminate between coincident and non-coincident stimuli in English. However, 7- and 10-month-old infants did respond differentially to the two stimulus types: 7-month-olds looked significantly longer during non-coincident speech trials, while 10-month-olds looked significantly longer during coincident speech trials.

The results of these studies provide only partial support for the claim that infants prefer to listen to speech in which pauses coincide with utterance boundaries, even in an unfamiliar language. Four-month-old infants failed to discriminate coincident and non-coincident stimuli in either English or German, contrary to the results of Jusczyk (1989). Only when 10-month-old infants were presented with English did we find the significant listening preference for coincident speech reported by Hirsh-Pasek et al. (1987) and Kemler Nelson et al. (1989). When 7-month-old infants listened to English, they showed a significant preference for non-coincident speech. Although this preference was in the opposite direction from that reported in previous studies, it nevertheless indicates that 7-month-old infants were able to discriminate the two types of speech stimuli.

Given these inconsistent findings across studies, what general conclusions can be drawn about infants’ sensitivity to “clause-typical” prosodic units? First, the notion that very young infants show a listening preference even in unfamiliar languages is not supported by the evidence. We were unable to replicate either the finding that 4-month-old English-learning infants prefer coincident speech in English, or that they show a preference in a foreign language prosodically similar to English. Since Jusczyk, Mazuka, Mandel, Kiriatzi, and Hayashi (1993) also recently reported that 4 1/2-month-old English-learning infants fail to discriminate coincident and non-coincident speech in Japanese, the evidence overall seems to suggest that a listening preference for coincident speech does not emerge until later in the first year, and may be limited to speech in the ambient language.

Second, regardless of when this listening preference emerges, it is a misinterpretation of the evidence to claim that such a preference reveals infants’ sensitivity to “clausal units.” Although the stimuli in all these experiments were chosen because they contained complete clauses, in this respect they were not representative of spontaneous IDS, in which almost half the utterances are not clausal units. A more limited and appropriate interpretation of the findings of Hirsh-Pasek et al. (1987) and Kemler Nelson et al. (1989), as well as the findings presented here, is that by 7 months infants have learned to discriminate continuous vocalizations in their own language from those that are artificially interrupted. The premise that continuous vocalizations in IDS are isomorphic with clausal constituents is false (except in experiments). Thus the broad conclusion that clauses are perceptual units for infants does not follow from these findings.

**CONCLUSIONS**

In the 1970s, the first systematic studies of early language input agreed that ID utterances are generally short, often fragmentary, and often non-canonical in sentence form. When Snow (1972) and others suggested that IDS might provide an ideal language lesson for the infant, Newport et al. (1977) objected strenuously and convincingly that, even though short, these often fragmentary, often non-canonical utterances are not syntactically simple, and that we could not look to the structure of the input for easy answers to the question of how children induce syntax. This argument was so convincing, in fact, that the view that simplified speech is crucial for language learning was soon eclipsed by the view that clever infants rather than accommodating mothers are the “prime movers of the acquisition process” (Gleitman, Newport, & Gleitman, 1984, p. 70).

This more nativist position assumes that infants come equipped with perceptual filters and attentional biases that enable them to exploit acoustic correlates of linguistic units in speech. However, this new vision of processing biases as critical to language acquisition does not shift the burden entirely from the input to the infant, for the following reason: Processing biases require appropriately structured input to be useful. The frog’s specialized sensitivity to moving dark spots is functional only because, in the frog’s world, moving dark spots correspond to prey; if moving dark spots instead mapped randomly onto objects irrelevant to frogs, such a perceptual bias would have no utility. The
prosodic bootstrapping argument also requires that speech input be structured to accommodate the hypothesized processing biases of the human infant. If infants are predisposed to attend selectively to a prosodic feature such as pauses in speech, this bias would be useful in language acquisition only to the extent that pauses are actually linguistically informative. That is, implicit in the new view of the perceptually prepared infant are complementary assumptions about the nature of the input that these infant processing biases require to operate effectively.

Unfortunately, the idealized input assumed in the prosodic bootstrapping argument is at odds with the empirical evidence. More recent descriptions of early language input have characterized ID utterances in terms of prosodic boundary markers. But have these detailed acoustic analyses actually increased our understanding of key features of IDS that might facilitate syntax acquisition? Not really. If individual clauses within multiclause utterances were prosodically segmented, this would be revealing—but multiclause utterances are extremely rare in spontaneous IDS. If phrasal boundaries within single clauses were prosodically segmented, this would also be revealing, but research on ADS and IDS has shown that NP/VP boundaries are not reliably marked by prosodic cues. Thus we have not learned much new about the prosodic fine structure of ID utterances that is relevant to the discovery of syntax. Moreover, declaring all ID utterances to be “clauses” or “sentences” regardless of their actual syntactic structure does not change the fact that almost half are sub-clausal fragments. So we are essentially back where we started: When ID utterances are respecified in terms of prosodic boundary markers, they still consist of the short, often fragmentary, often non-canonical sentences described in the 1970s. And the central argument of Newport et al. (1977) is as relevant today as it was then: Whatever their pragmatic merits, the modifications in maternal speech of IDS that might facilitate syntax acquisition? Not really. Whatever their pragmatic merits, the modifications in maternal speech and over time enable the induction of syntactic rules.

REFERENCES


