In: Hahn, M., & Stoness, S.C., Proceedings of the 21st Annual Meeting of the Cognitive Science Society (pp. 724-729), Mahwah, NJ: LEA.

Conditional Probability and Word Discovery: A Corpus Analysis of Speech to Infants

Daniel Swingley (swingley@bcs.rochester.edu)
Department of Brain and Cognitive Sciences
Meliora Hall; University of Rochester
Rochester, NY 14627 USA

Abstract

Analyses of an idealized corpus of English speech to infants revealed that simple conditional decision rules can separate frequent bisyllabic words from bisyllables not corresponding to words. If infants accurately represent speech in terms of syllables, and compute conditional statistics over these syllables, such computations have the potential to inform infants of likely English words.

Introduction

By the age of 8 months or so, infants know a great deal about the sound structure of the of the language they hear. Infants this age keep track of specific sound sequences or words (Jusczyk & Aslin, 1995; Jusczyk & Hohne, 1997). Also, infants presented with syllables of an artificial language store "words" of that language in accordance with syllables' conditional probabilities (Aslin, Saffran, & Newport, 1998; Goodsitt, Morgan, & Kuhl, 1993; Morgan, 1994; Saffran, Aslin, & Newport, 1996).

Together, these results suggest that infants might implicitly apply conditional probability statistics to analyze not only the speech they hear in laboratory contexts, but also the speech they hear in their everyday life. The purpose of the research presented here is to evaluate the potential linguistic consequences of this memory mechanism. Specifically, we test whether the application of simple conditional-probability decision rules to a corpus of infant-directed speech results in the extraction of words.

Previous research has shown that the segments composing words in speech or in text exhibit distributional regularities that can be computationally exploited for word boundary detection (Aslin, Woodward, LaMendola, & Bever, 1996; Cairns, Shillcock, Chater, & Levy, 1997; Christiansen, Allen, & Seidenberg, 1998; Elman, 1990; see also Brent & Cartwright, 1996; de Marcken, 1996). The current work differs from these previous studies in two respects. First, we have used *syllables* as the units of analysis; and second, we test the potential for statistical information to build a vocabulary, not to parse sentences effectively.

We have used syllables rather than segments because syllables are widely considered to be a unit of speech infants are capable of processing and representing. It is frequently argued in the language acquisition literature that infants do not represent speech in terms of segments, which arise as representational units later in childhood. Several experiments have shown that infants categorize varied sets of words by their number of syllables, but not their number of segments, suggesting that syllables are crucial units in infants' representation of speech (Bertoncini, Floccia, Nazzi, & Mehler, 1995; Bijeljac-Babic, Bertoncini, & Mehler, 1993). Furthermore, Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy, and Mehler (1988) found that when infants were habituated to sets of CV syllables containing a common onset (such as /b/), infants dishabituated equally after the introduction of novel /b/-initial CVs, and after the introduction of novel CVs containing a new C, suggesting that infants did not consider the /b/-initial syllables as similar. Though these results do not necessarily indicate that infants fail to represent segments as units, much evidence now suggests that the syllabic level of representation is significant in infancy, and that syllables may be relevant as units over which statistical computations may be done.

In the present study, a corpus of speech directed to children under 18 months was used to evaluate the utility of transitional probability information for grouping syllables into words. The problem may be stated as follows: given that over 80% of the words children hear are monosyllabic (Aslin et al., 1996; Christiansen et al., 1998), would infants' tendencies to cluster syllables according to conditional probability criteria lead infants to inappropriately conflate monosyllabic words? Or, alternatively, would these clustering mechanisms help lead infants to discover words in speech? Answering these questions requires a statistical analysis of large samples of speech directed to infants. Below, we demonstrate that given certain assumptions about the mechanisms that underlie infants' sensitivities to statistical structure, American English speech does contain statistical regularities that could be used for word discovery.

Methods

Corpora of 15 parents' speech to American infants under 18 months (CHILDES; MacWhinney, 1995; see Bernstein-Ratner, 1984; Bloom, 1973; Hayes & Ahrens, 1988; Higginson, 1986; Sachs, 1983; Warren-Leubecker & Bohannon, 1983) were combined to form a 50,000–word corpus. Spelling of words throughout was regularized by hand, and pronunciations of the resulting words were estimated using the CMU pronouncing dictionary (v. 0.4, 1995). This phonemic corpus was syllabified using an implementation of Kahn's (1980) formalism for slow speech (essentially maximal onset). The syllabification algorithm was run over words, not over utterances; thus, no segments were syllabified across word boundaries (a point we return to below). Over the resulting corpus of syllables, three metrics were calculated

for each consecutive pair of syllables AB (bigrams): predictive transitional probability , or $p(B|A)$; reverse transitional probability, or $p(A|B)$, and mutual information, or $log_2[p(AB)/p(A)p(B)]$. (Mutual information is a measure of how much the occurrence of one syllable is informative about the other syllable; cf. Charniak, 1993).

Predictive transitional probability is high when one syllable makes the following one predictable. This metric would be useful if it tended to be higher in words (such as "pretty": $p(B|A) = 1$) than in other sequences (such as "thank you": $p(B|A) = 1$). Reverse transitional probability is high when one syllable makes the previous one predictable. This metric would be useful if sequences like "little" $(p(A|B) = .81)$ were more common than sequences like "the door" $(p(A|B) = .81)$. Mutual information is high when both syllables tend to co-occur, and would be useful if sequences like "daddy" (m.i. 7.7) were more common than sequences like "sit down" (m.i. 7.5).

Because the vast majority of the word types in speech to children are monosyllabic (in the present corpus, about 55%) or bisyllabic (about 42%), the syllable-grouping problem in English reduces largely to a problem of deciding whether two syllables form a bisyllabic word. If this distinction could be made accurately, 97% of words in the corpus would be correctly identified. Thus, the current analyses examined whether conditional-probability metrics could be used to distinguish the bigrams that are words, from those that are not.

In a series of analyses, a threshold value of one of these three conditional probability metrics was set, and the bigrams above that threshold were identified as words. The question asked was whether thresholds for the three metrics could be set to produce a favorable ratio of hits to false alarms (*precision*, or *accuracy*), and of hits to misses (*completeness;* cf. Brent & Cartwright, 1996). Precision indicates whether "yes, it's a word" responses tend to be correct; completeness measures the proportion of words that are identified. At the same time, the frequency of bigrams considered as possible words was varied, to evaluate possible interactions between frequency and conditional-probability information.

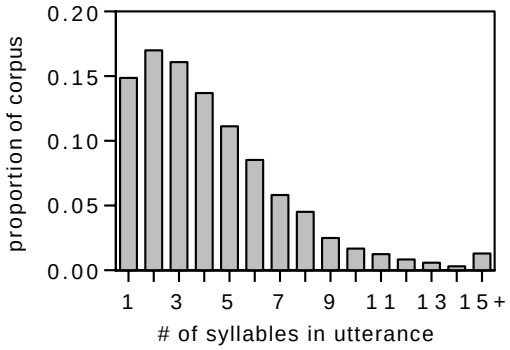

Figure 1: Sentence lengths in syllables.

For computational convenience, analyses included only the utterances containing 2–14 syllables. (The upper bound excluded about 1.3% of the utterances.) The remaining corpus included approximately 41,000 bigram tokens, and about 13,000 bigram types. Among these types were 762 different bisyllabic words, the primary targets of the analyses. As shown in Figure 1, the majority of utterances directed to infants contained only a few syllables. This illustrates a common observation about speech to infants (e.g. Snow, 1972).

Results and Discussion

As a baseline for comparison, we first consider the precision and completeness scores that would be expected from simple guessing, an estimate of "chance" in finding bisyllables. A guessing-based decision rule could say "yes" or "no" equally often (p=0.5), or could say "no" most of the time (say, p=0.8). Considered in Figure 2 are precision and completeness (by types) for a range of guessing rates, from 0 (calling all bigrams words) to 1.0 (calling all bigrams nonwords). The results make clear the fact that bisyllables cannot be located effectively by guessing; they are too rare. Without more information differentiating word and nonword bigrams, the infant would be considerably better off simply assuming that all syllables are monosyllabic words.

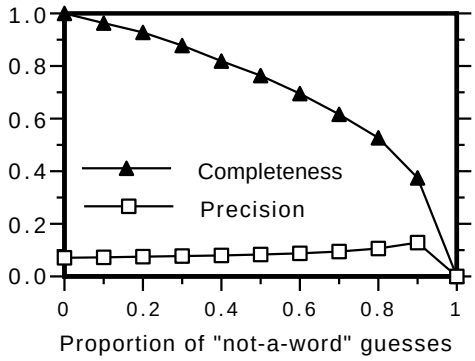

Figure 2: Results for guessing, shown by types (i.e. not frequency weighted).

Next, we consider the possibility that infants cluster syllables in natural speech primarily on the basis of frequency. This has a certain intuitive appeal–one might imagine that, upon hearing a given pair of syllables many times, this pair might come to be considered a linguistic unit. The following decision rule was evaluated: if a bigram appeared *x* or more times in the corpus, it was considered to be a word. Figure 3 shows the results over types (i.e. displayed are the number of different words found, not weighted by their frequency of occurrence). Clearly, many common bigrams are not bisyllabic words; in fact, even if we admit only the bigrams that occur 50 or more times (which is true of 108 bigrams), there are still more than twice as many nonwords as words. Examples of hits include very frequent words like "baby," "doing," and "very." But a frequency criterion false-alarms to sequences like "good girl," "is it," "that's right," and "we put."

The experimental results of Aslin, Saffran, and Newport (1998), however, show that infants' clustering of syllables is not mediated only by frequency. In that study, the frequency of occurrence of words and nonwords in an artificial language was controlled, and differed only in the conditional probabilities with which syllables followed one another. Eight-

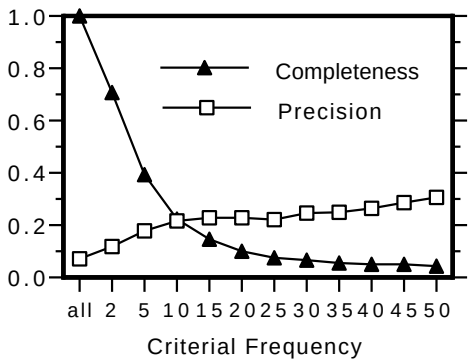

Figure 3: Results for frequency criterion, by types.

month-olds discriminated word and nonword lists, demonstrating sensitivity to conditional probabilities independent of frequency. These results suggest that infants do not simply assume that common bisyllables are words. Given the statistics of the language, this is fortuitous; as shown above, such a mechanism would usually be wrong.

Subsequent analyses consider decision rules based on conditional probability metrics. In Figure 4, precision and completeness scores are shown for decision rules using a mutual information threshold; when the threshold was equaled or surpassed, the bigram was considered a word. As Figure 4 demonstrates, performance using mutual information was better than performance using frequency. However, even the best precision scores were never above about 45%. Similar results were obtained using predictive and reverse transitional probability.

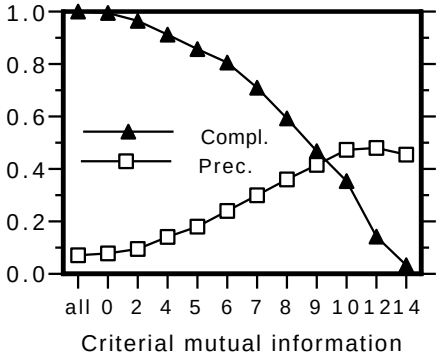

Figure 4: Mutual information criterion with all words counted. Results by types.

However, this analysis included many words of very low frequency. The clustering mechanism that is operating here is likely to be one of implicit memory (Saffran et al., 1997). Implicit learning of sequential stimuli typically involves considerable repetition of the training materials. Thus, it seems reasonable to suppose that the learning mechanism would only make assumptions about bigrams after several exposures to them. Exactly how many exposures we would expect the system to require is not clear; in the subsequent analyses we will examine the degree to which simple decision rules detect words that occurred 5 or more times in the corpus. (The results are similar if we exclude only the bigrams occurring 1–2 times, 1–3 times, etc.) Because (as in any natural corpus) a large proportion of the types occurred infrequently, even a small frequency criterion excludes many types. In the present corpus, a frequency criterion of 5 excludes 32% of the bigram tokens, and excludes 84% of the bigram types; at the same time, this criterion excludes only 11% of the bisyllabic word tokens, and 61% of the bisyllabic word types. Of the 762 bisyllabic words in the corpus, 300 meet the frequency criterion. The following analyses consider whether conditional probability information can help identify the 300 most common bisyllabic words, out of the 1676 most common bigrams. (Note that this is still a nontrivial task, because word types make up less than 18% of the common bigram types.)

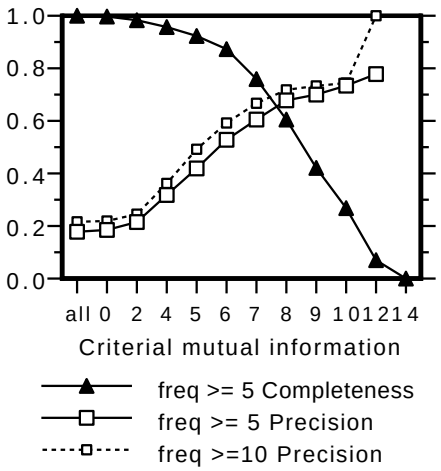

Figure 5: Mutual information criterion with frequency thresholds of 5 and 10.

Figure 5 shows precision and completeness results using mutual information and a frequency criterion of 5 (solid lines). The dashed line shows precision scores when a frequency criterion of 10 is applied. Two results are evident here. First, precision is quite high: in the higher mutual information range, precision varies from about 60% to 80%. Where the lines cross, both precision and completeness are above 60%. Second, precision is only marginally improved by increasing the frequency criterion to 10 occurrences. This small improvement is offset by substantial decreases in completeness (not shown here), reflecting the fact that only 170 bisyllabic words occurred 10 times or more.

Although performance is much better than chance, even in the best case the decision rule makes many errors. An analysis of the false alarms, however, suggests that many of the errors are not pernicious ones: often the “false alarms” are instances in which the decision rule groups together two syllables from a trisyllabic word. Figure 6 shows the proportion of false alarms that are clusters of syllables that form 2/3 of a trisyllabic word, over a range of mutual information criteria.

At the higher mutual information criteria, the proportion of “false alarms” within trisyllabic words is quite high—over a

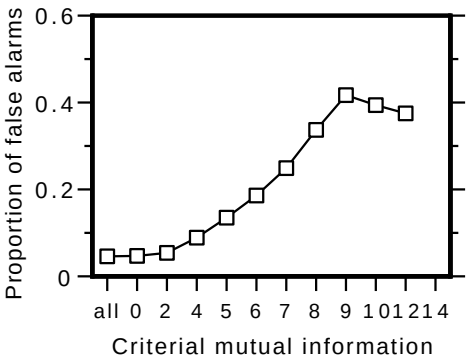

Figure 6: Proportion of false alarm types that form part of a trisyllabic word.

third in some cases. In some of these, there are two "false alarms" within a trisyllable. For example, at a mutual information threshold of 7, the decision rule clusters "kanga" and "garoo." If all false alarms that actually cluster bigrams within trisyllabic words are counted as hits, rather than false alarms, precision at the higher thresholds reaches about 85%. Some persistent errors remain: among them are "I'm so(rry)," "much fun," and "(pe)ter pi(per)." However, in spite of these errors, it is clear that infants computing conditional probabilities are likely to group together syllables that belong within words.

How robust are these results? Changing the frequency threshold does not have much impact on precision, as long as bigrams of very low frequency are not included as possible words. Furthermore, similar results are obtained when using predictive and reverse transitional probability rather than mutual information. Figure 7 shows precision and completeness at probabilities ranging from 0 to 1, using a frequency criterion of 5.

Another sort of decision rule compatible with current experimental results would not use an absolute thresholding mechanism to cluster syllables. Rather, two syllables might be grouped together if their conditional probabilities were higher than those of the neighboring bigrams. Consider the utterance ABCDEF, with each letter a syllable. Suppose the mutual information values between syllables are as follows: $\mathrm{A}_2\mathrm{B}_5\mathrm{C}_1\mathrm{D}_6\mathrm{E}_6\mathrm{F}$. On a "neighbor-comparison" rule, BC would be grouped together, but DE would not, because the value for DE does not exceed the value for EF. Several varieties of this decision rule are possible. For example, a bigram might have to exceed its neighbors by 1, or 2, or 6 (and in this last case, BC would not be considered a word, because the difference between 5 and 2 is less than 6). Rules of this sort amount to attempts to find peaks in the mutual information function across the sentence. Figure 8 shows results from this decision rule. The x-axis represents the number of mutual information units by which a given bigram must be greater than its neighbors, to be considered as a word. As the figure shows, performance using such a rule is comparable to performance using absolute thresholds.

These results suggest that a variety of decision rules, conditional statistics, and threshold values might lead an infant to

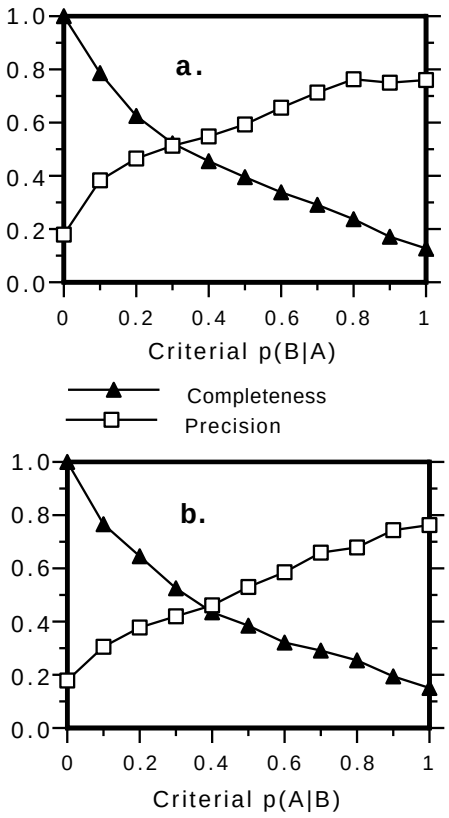

Figure 7: Transitional probability used in decision rule. Graph a, predictive probability; b, reverse probability.

correctly identify frequent bisyllabic words with a high success rate. However, the current results must be qualified by three important considerations. First, the syllabification algorithm did not permit consonants to be resyllabified to adjacent words. This sort of resyllabification does occur in English, although it is not clear how often it occurs in speech to infants, or under what circumstances. It is unlikely that all consonantal codas preceding vowel onsets are resyllabified. For example, upon hearing a sentence like "I put that in," infants probably do not group the /t/ of "that" with "in." In the absence of a model of resyllabification in infant-directed speech, and without experimental data on infants' assignment of consonants to syllables, this remains an open question. However, preliminary analyses based on a "worst-case scenario" in which coda consonants are always resyllabified to the following syllable, if thereby creating a legal onset cluster, show that the use of conditional statistics in decision rules still substantially improves performance. Under these conditions, as expected, precision and completeness scores are lower; but baselines produced by guessing or frequency criteria are also lower. Thus, regardless of our particular assumptions about the transparency of syllable boundaries, information about words is still present in conditional statistics.

Second, the current corpus is idealized in the sense that it assumes a fixed pronunciation for each orthographic word. The truth was certainly messier, although without the recordings themselves we cannot attempt a precise reconstruction.

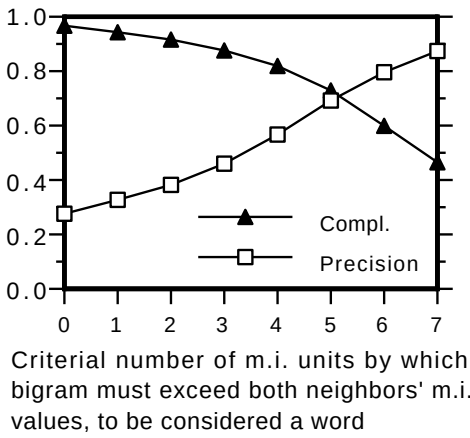

Figure 8: Results with decision rule based on comparison with neighboring bigrams' mutual information values.

In principle, some of the variability in actual pronunciations could be modeled using probabilistic rules (e.g. de Marcken, 1996), yielding a noisier (but more veracious) corpus. This procedure would be particularly useful if more were known about infants' compensation for the processes that lead to variable realizations of words. Our implicit assumption in the current work is that this compensation is perfect, but we recognize that this is an idealization.

Third, it is possible that the structure of English favors the success of decision rules of the sort employed here, and that these decision rules would prove ineffective in the analysis of other languages. This outcome would not necessarily indicate that infants do not perform computations like those we have evaluated. It would indicate that if infants do, it will not lead them to discover words.

The current results are not themselves a model of the infant learner. Presumably infants do not store a corpus of the speech they hear for several months, and then (implicitly) group together cohesive units. Rather, infants' mental computations occur incrementally over time, as more and more speech is heard. A model of the infant learner would take this into account by simultaneously calculating conditional statistics and applying decision rules; such a model is currently under development. The analyses presented here suggest that such a model would show that infants' ability to cluster syllables based on statistical characteristics would result in the identification of words more often than not, perhaps with very high precision.

Several researchers have proposed that infants might use prosodic information, such as lexical stress, to help identify words in speech (e.g. Cutler, 1994; Gleitman, Gleitman, Landau, & Wanner, 1988). In fact, evidence from infant experiments suggests that English-learning infants tend to extract (from continuous speech) words with strong-weak stress patterns more readily than words with weak-strong patterns (Newsome & Jusczyk, 1995). Because English content words tend to begin with strong syllables, this tendency may well help English-learning infants to discover words. It is not clear at present whether these tendencies hold for all infants (in which case infants learning some other languages will be disadvantaged by this decision rule), or only for infants in certain language environments (in which case an account of how this prosodic knowledge is acquired will be necessary). In either case, however, there is no reason to suppose that a prosodic strategy and a statistical-learning strategy are incompatible. Although prosodic cues to word boundaries vary with different languages, it may be that statistical cues of the sort examined here are true of most languages. If so, statistical cues might help "bootstrap" a prosodic (or any other) strategy. This is obviously an important area for future cross-linguistic research.

Acknowledgment

This research was supported by NIH grant F32–HD08307.

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8–month–old infants. *Psychological Science, 9*, 321–324.

Aslin, R. N., Woodward, J. C., LaMendola, N. P., & Bever, T. G. (1996) Models of word segmentation in fluent maternal speech to infants. In Morgan & Demuth (Eds). *Signal to syntax: Bootstrapping from speech to grammar in early acquisition.* Mahwah, NJ: LEA.

Bernstein Ratner, N. (1984). Patterns of vowel modification in mother-child speech. *Journal of Child Language, 11*, 557–578.

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., and Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General, 117*, 21–33.

Bertoncini, J., Floccia, C., Nazzi, T., & Mehler, J. (1995). Morae and syllables: Rhythmical basis of speech representations in neonates. *Language & Speech, 38*, 311–329.

Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do 4–day–old infants categorize multisyllabic utterances? *Developmental Psychology, 29*, 711–721.

Bloom, L. (1973). *One word at a time: the use of single word utterances before syntax.* The Hague: Mouton.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61*, 93–125.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. P. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology, 33*, 111–153.

Charniak, E. (1993). *Statistical language learning.* Cambridge, MA: MIT Press.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language & Cognitive Processes, 13*, 221–268.

Cutler, A. (1994). Segmentation problems, rhythmic solutions. *Lingua, 92*, 81–104.

Elman, J. L. (1990). Finding structure in time. Cognitive *Science, 14*, 179–211.

Gleitman, L., Gleitman, H., Landau, B., & Wanner, E. (1988). Where learning begins: Initial representations for language learning. In Newmeyer, et al. (Eds). *Language: Psychological and biological aspects.* Cambridge, UK: CUP.

Goodsitt, J., Morgan, J. L., & Kuhl, P. K. (1993). Perceptual strategies in prelingual speech segmentation. *Journal of Child Language, 20,* 229–252.

Hayes, D. P., & Ahrens, M. G. (1988). Vocabulary simplification for children: A special case of "motherese?" *Journal of Child Language, 15,* 395–410.

Higginson, R. P. (1986). *Fixing: Assimilation in language acquisition.* Doctoral dissertation, Washington State University.

Jusczyk, P. W. & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology, 29,* 1–23.

Jusczyk, P. W. & Hohne, E. A. (1997). Infants' memory for spoken words. *Science, 277,* 1984–1986.

Kahn, D. (1980). *Syllable-based generalizations in English phonology.* New York: Garland.

MacWhinney, B. (1995) *The CHILDES project: Tools for analyzing talk (2nd ed.).* Hillsdale, NJ: LEA.

de Marcken, C. (1996). *Unsupervised acquisition of a lexicon from continuous speech.* MIT AI Lab Memo 1558; CBCL memo 129.

Morgan, J. L. (1994). Converging measures of speech segmentation in preverbal infants. *Infant Behavior & Development, 17,* 389–403.

Newsome, M. R., & Jusczyk, P. W. (1995). Do infants use stress as a cue in segmenting fluent speech? In *Proceedings of the 19th Boston University Conference on Language Development* (pp. 415–426). Boston, MA: Cascadilla.

Sachs, J. (1983). Talking about the There and Then: The emergence of displaced reference in parent-child discourse. In K. E. Nelson (ed.) *Children's Language*. Hillsdale, NJ: LEA.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8–month–old infants. *Science, 274,* 1926–1928.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8,* 101–105.

Snow, C. (1972). Mothers' speech to children learning language. *Child Development, 43,* 549–565.

Warren-Leubecker, A., & Bohannon, J. (1983). The effects of verbal feedback and listener type on the speech of preschool children. *Journal of Experimental Child Psychology, 35,* 540–548.