Statistical Cues in Language Acquisition: Word Segmentation by Infants

Jenny R. Saffran
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14627
saffran@bcs.rochester.edu

Richard N. Aslin
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14627
aslin@cvs.rochester.edu

Elissa L. Newport
Department of Brain and Cognitive Sciences
University of Rochester
Rochester, NY 14627
newport@bcs.rochester.edu

Abstract

A critical component of language acquisition is the ability to learn from the information present in the language input. In particular, young language learners would benefit from learning mechanisms capable of utilizing the myriad statistical cues to linguistic structure available in the input. The present study examines eight-month-old infants' use of statistical cues in discovering word boundaries. Computational models suggest that one of the most useful cues in segmenting words out of continuous speech is distributional information: the detection of consistent orderings of sounds. In this paper, we present results suggesting that eight-month-old infants can in fact make use of the order in which sounds occur to discover word-like sequences. The implications of this early ability to detect statistical information in the language input will be discussed with regard to theoretical issues in the field of language acquisition.

Introduction

While it is widely acknowledged that language acquisition is accomplished by an interaction between innate constraints and learning, surprisingly little research has focused on the learning mechanisms which are a critical component of this interaction. Even the richest input imaginable would not allow the child to learn language unless she possessed the mechanisms required to extract pertinent information from this input. Similarly, innate linguistic knowledge would be of no use without mechanisms relating it to linguistic experience. For these reasons, a number of researchers on both sides of the nature/nurture debate have begun to investigate the kinds of learning mechanisms possessed by young language learners.

One class of learning mechanisms which has recently returned to prominence[1] is distributional learning devices, which utilize the statistical properties inherent in linguistic input. The renewed interest in distributional learning in language acquisition results in part from the contributions of recent connectionist models. Importantly, this interest has also been generated by research suggesting that humans extract and remember information about the statistical structure of their native language. Adults possess rich representations of far-flung statistical features of their language, ranging from word-frequency effects to probabilistic prosodic expectancies to frequency-based contingency effects in parsing (e.g., Kelly, 1988; MacDonald, Pearlmutter, & Seidenberg, 1994; Morton, 1969), and can readily learn distributional regularities in laboratory tasks (e.g., Morgan, Meier, & Newport, 1987; Saffran, Newport, & Aslin, in press).

These abilities are not confined to adults. First-grade children, for example, are at least as good as adults at discovering distributional regularities in the lab (Saffran *et al.*, under review). Infants also demonstrate knowledge of some of the statistical regularities of their native language. For example, when nine-month-old infants are presented with phonotactically legal phonetic patterns which are either frequent or infrequent in their native language, they prefer to listen to the frequent patterns (Jusczyk, Luce, & Charles-Luce, 1994). This knowledge must arise through learning from the linguistic environment, suggesting that statistical learning mechanisms exist and, moreover, play a far greater role in language acquisition than most contemporary theories suggest.

The present research seeks to elucidate the nature of the learning mechanisms underlying the acquisition of language. Our strategy in this research is to focus on aspects of language that are undeniably discovered in the language input, rather than potentially an expression of innate knowledge. In particular, we hope to begin to discover how infants' learning mechanisms are structured to make use of the enormous volume of statistical information available in the language input. To do so, we investigated the learning mechanisms underlying word segmentation.

Word Segmentation

One of the earliest and most impressive feats of learning by infants is the discovery of word boundaries. Speech is essentially continuous, without pauses or other consistent acoustic cues present to mark word boundaries. Infants must thus somehow break into the speech stream to discover word boundaries without recourse to silences analogous to the white spaces between printed words. Despite the difficulty of this learning problem (Cole & Jakimik, 1980), experimental evidence indicates that infants can succeed at word segmentation tasks by eight months of age, well before the onset of word production (Jusczyk & Aslin, 1995).

There are many possible cues to word boundaries that might be exploited, including prosodic regularities, as well

[1] Once greatly popular among Bloomfieldian linguists (see, e.g., Harris, 1955), distributional analyses of linguistic structure fell into disfavor with the birth of Chomskian generative syntax.

as the occasional occurrence of utterance-final pauses and words spoken in isolation (see, e.g., Christophe *et al.*, 1994; Jusczyk, Cutler, & Redanz, 1993). While all of these types of information are likely to be helpful, none alone is sufficient to solve the word segmentation problem (see Saffran *et al.*, in press, for discussion). However, one important source of potential information lies in the distributional information offered by the sequences of sounds within and between words (Brent & Cartwright, in press; Hayes & Clark, 1970; Harris, 1955; Saffran *et al.*, in press). A word may be defined as a fixed series of sounds. The learner, however, does not have direct access to this information. Rather, what the learner experiences in the input is complex statistical information over a corpus of utterances resulting from the concatenation of subword units. This information will take the form of relatively strong correlations between sounds found within words, contrasted with weaker correlations across word boundaries (Hayes & Clark, 1970; Saffran *et al.*, in press). On this view, one might discover words in linguistic input in much the same way that one discovers objects in the visual environment via motion: the spatial-temporal correlations between the different parts of the moving object will be stronger than those between the moving object and the surrounding visual environment.

Several recent computational models have illustrated the efficacy of distributional cues in word segmentation. One such model demonstrates that distributional information can provide appropriate segmentations when the algorithm used is the minimum description length principle, an evaluation function which minimizes the amount of memory needed to represent a lexicon derived from a previously unsegmented corpus of speech (Brent & Cartwright, in press). Other models indicate that class-based n-gram and feature-based neural network models can segment speech using transitional probabilities (Cairns *et al.*, 1994); similarly, Elman (1990) describes a simple recurrent network able to discover written words in unsegmented text by computing graded co-occurrence statistics (see Wolff, 1975, for similar findings using a non-connectionist architecture). These corpus-based models, along with many others in the machine speech recognition literature, demonstrate that statistical information is sufficient *in principle* for rudimentary word segmentation[2].

Can human learners make use of statistical cues to word boundaries? If not, then this wealth of information would be of little use to humans confronted with continuous speech in an unfamiliar language. Saffran *et al.* (in press) asked whether adult subjects were able to use differences in the transitional probabilities between sounds to discover word boundaries[3]. Across a language corpus, the transitional probability from one sound to the next will generally be greatest when the two sounds follow one another word-internally; transitional probabilities spanning word boundaries will tend to be relatively low. After only twenty minutes of exposure, adults were able to learn the multisyllabic words of a nonsense language presented as a synthesized speech stream containing no cues to word boundaries except for transitional probabilities (Saffran *et al.*, in press). Moreover, this same result was obtained with first-grade children as well as adults, even when the presentation of the speech stream occurred in the background, while subjects were engaged in another task and neither told to listen nor to learn (Saffran *et al.*, under review). The abilities of human learners to perform such statistical computations implicitly, during mere exposure, are quite impressive. This suggests that this learning mechanism operates automatically, much as one would expect from a learning mechanism hypothesized to underlie learning in children too young to engage in conscious hypothesis testing.

The crucial subjects for such investigations are infants of the age at which rudimentary word segmentation first occurs. Language learning tasks are often seen as too difficult for the limited abilities of infants; indeed, our lack of knowledge regarding infant learning has often led theorists to assume that because a learning task seems difficult, it must be solved innately. However, the sheer volume of information that infants do in fact learn about their native language, much of which could not possibly be encoded innately, suggests that young infants may in fact be far better at extracting statistical regularities from the input than has generally been assumed.

Recent research suggests that infants may in fact be attuned to the kinds of distributional information which serve to cue word boundaries. For example, infants as young as two months of age are able to remember the order of spoken words, as long as the words are spoken with normal sentential prosody (Mandel, Kemler Nelson, & Jusczyk, in press). By eight months of age, infants are able to detect consistently ordered two-syllable units presented in brief repetitive utterances (Goodsitt, Morgan, & Kuhl, 1993). The next step is to determine whether infants are able to keep track of the array of probabilities found in multisyllabic sequences to discover word boundaries, in the absence of any other cues to word boundaries. The present study provides some preliminary indications that infants can in fact use the order of the sounds that they hear to extract word-like units.

Method

This study used a brief familiarization period combined with the headturn preference procedure widely used in infancy research (Jusczyk & Aslin, 1995). In this methodology, infants are first exposed to an auditory stimulus which serves as a potential learning experience. Following this exposure, the infant is presented with two types of auditory stimuli: familiar stimuli, like those presented during the familiarization period, and novel stimuli. The infant's

[2]No such algorithm is error-free; neither, however, are young children, who very commonly make segmentation errors (a common undersegmentation error is treating a phrase like "ham'neggs" as a single word). Such errors, however, are not random, but rather reflect the distributional characteristics of the input (e.g., Brown, 1973). Recovery from segmentation errors occurs with more extensive input and the detection of other cues correlated with the correct word boundaries.

[3]The transitional probability of $Y/X = \frac{\text{frequency of XY}}{\text{frequency of X}}$

listening preferences are then assessed. Two possible outcomes suggest that learning has occurred. Infants of this age generally prefer to listen to somewhat familiar items; in this case, the infants should prefer to listen to the items similar to those heard during the familiarization period, if learning did in fact occur. However, the opposite effect would also signal learning: if the infants had learned and habituated to the familiarization stimuli, then a novel stimulus would be more engaging. No preference would, of course, fail to indicate that any learning had occurred.

In the present study, infants were familiarized with an artificial speech stream, consisting of four trisyllabic nonsense words repeated in random order by a speech synthesizer. The synthesizer was given no information regarding word boundaries, and thus spoke the speech stream continuously in a monotone, without any acoustic cues to word boundaries. The only cues to word boundaries were statistical; the transitional probabilities between syllables within words were greater than the transitional probabilities between syllables spanning word boundaries. Following a two-minute exposure, the infants' learning was assessed by determining whether they preferred to listen to 'words' from the nonsense language, or 'nonwords', which consisted of the same syllables that the infants had heard during familiarization but now presented in a novel order. A significant preference for either words or nonwords would, as discussed above, signal that the syllable orders heard during familiarization had been learned to the extent that the infants could distinguish them from novel syllable orders.

Subjects. 16 infants (nine male, seven female), approximately eight months of age, participated in the study. Three additional infants were tested but not included in the analysis for the following reasons: experimenter error (2), and crying (1).

Stimuli. Two counterbalanced stimulus conditions were generated. For each condition, 45 tokens of each of four trisyllabic nonsense words (Condition A: *tupiro, golabu, bidaku, padoti*; Condition B: *dapiku, tilado, burobi, pagotu*) were digitized to create two-minute-long speech streams. The words were spoken in random order, with the stipulation that the same word never occurred twice in a row. A speech synthesizer (MacinTalk) generated the speech stream at a rate of 270 syllables/minute with equivalent levels of coarticulation between all syllables; no pauses or any other acoustic or prosodic cues to word boundaries were present. A sample of the speech stream used in Condition A was analogous to the following orthographic representation: *bidakupadotigolabubidakutupiro...* The only cues to word boundaries were the transitional probabilities between syllable pairs over the language corpus, which were higher within words (all 1.0) than across word boundaries (all .33).

To assess learning, each infant was presented with repetitions of four trisyllabic strings (*tupiro, golabu, dapiku, tilado*) during the test phase. For the infants in Condition A, the first two test strings were 'words' which had been played during familiarization, and the last two test strings were 'nonwords', that is, syllables which they had heard during familiarization but now presented in a novel order (the transitional probabilities between the syllables in the nonwords were all zero relative to the familiarization corpus). For infants in Condition B, the first two test strings were 'nonwords' and the last two test strings were 'words'. This between-subjects counterbalanced design ensured that any observed preferences for words or nonwords across both conditions would not be artifacts of any general preferences for certain syllable strings. A test trial consisted of repetitions of a test string. Each of the four test strings were presented on three different trials, resulting in a total of 12 test trials per infant. Note that the strings used in the test were generated in citation form by the speech synthesizer, and thus had acoustic properties quite different from the same strings presented in the continuous speech stream.

Design. Half of the infants were assigned to each familiarization condition. During the test phase, all infants heard the same 12 test trials, randomized for each subject.

Procedure. During an experimental session, the infant was seated on a parent's lap in a sound attenuated booth. A video camera was placed directly in front of the infant, allowing the experimenter to observe the session via a video monitor outside the booth. Also directly in front of the infant was a blinking red light, used to bring the infant's gaze back to midline between trials. Blinking yellow lights were mounted on the right and left sides of the booth, along with hidden speakers. Both the parent and the experimenter wore headphones playing loud masking music. Because the different test trials were randomly assigned to the right or left speaker, and the experimenter could not hear the stimuli, the experimenter was blind to which stimulus was being presented on any given trial.

During the familiarization phase, the two minute speech stream was played continuously through both speakers. Blinking lights were used to help maintain infants' interest; the lights, but not the speech, were contingent upon the infant's looking behavior. Each trial began with the blinking center light. Once the infant had fixated on the center light, the experimenter signaled the Macintosh Quadra 650 running the study to turn off the center light and blink one of the side lights, whereupon the infant would turn to fixate the now blinking side light. The side light would continue to blink until the infant had looked away from it for two seconds. At that point, the center light would begin flashing, and a new trial would begin.

The test phase was similar, except that the number of repetitions of the test stimuli was contingent upon the infants' listening preferences. When the infant turned to look at the blinking side light, one of the four test strings was repeated from the speaker on that side, until the infant looked away for a preset criterion of two seconds (or until the test string had been repeated 15 times). The lookaway criterion signalled a loss of interest in that particular test string. Each infant thus only heard each test string as long as it remained interesting to him/her. Listening times to each type of test stimulus reflected each infant's listening preferences; these were tabulated on-line by the computer.

Results

We first compared the listening patterns of infants in the two counterbalanced conditions with one another to ensure that there were no overall preferences for any particular test

items regardless of familiarization. This was done by computing a difference score between mean listening times for words and nonwords for each infant, and comparing the infants from the two conditions with a t-test. As no differences were found ($t(14) = 1.3$, n.s.), data from the two conditions were combined in the primary analysis.

We then compared listening times to the 'words' versus the 'nonwords'. A matched-pairs t-test revealed that the novel 'nonwords' were listened to significantly longer than the familiar 'words': $t(15) = 2.8$, $p < .02$. Mean listening scores are presented in Figure 1. Twelve of the 16 infants listened longer to the novel stimuli. This novelty preference (or dishabituation effect) indicates that the infants clearly recognized that the novel orderings of test syllables were in fact novel and distinct from the orders that they had learned during the familiarization phase. Moreover, this effect could not have been simply due to memory for the low-level acoustic patterns presented during familiarization, as the acoustic properties of the test 'words' were quite different from the same 'words' present in the speech stream. Rather, the infants appear to have learned and remembered a more abstract representation of the strings of sounds that they heard during familiarization.

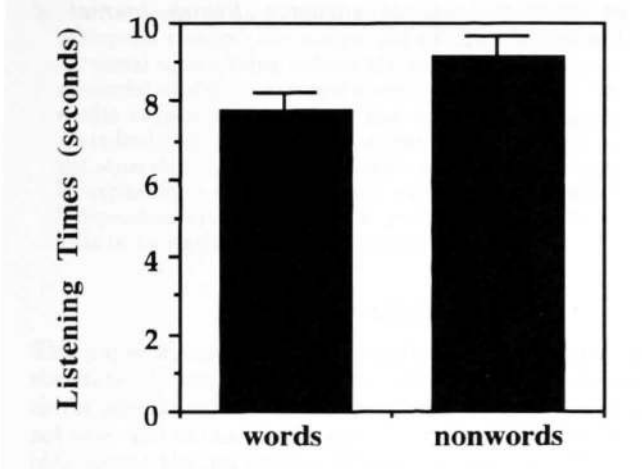

Figure 1: Mean listening times to the familiar (words) and novel (nonwords) stimuli.
Error bars = 1 standard error.

Discussion

Despite the impoverished state of the speech stream used in this study -- a learning stimulus devoid of prosody, pauses, or any other cues to word boundaries save statistical cues -- eight-month-old infants nevertheless succeeded at learning the words of the language to which they were exposed, at least to the extent that they could distinguish them from the same syllables in novel orders. This is by no means a trivial accomplishment. Despite the ubiquity of events which unfold in time, the discovery and representation of serial order is generally considered to be a difficult technical problem (e.g., Elman, 1990). Moreover, infants in this study had no particular incentive to keep track of co-occurrences. Rather, the discovery of the words within the continuous speech stream appears to be a natural outcome of exposure to patterned input. This process is particularly impressive given the brevity of exposure necessary for learning; the infants in this study were familiarized with the speech for a mere two minutes.

Of course, in actual language learning, other cues are likely to be present and used by infants discovering the word boundaries of their native language. Such cues are likely to be particularly effective when used in tandem with distributional cues. For example, Brent & Cartwright (in press) demonstrated that phonotactic information makes an additive contribution to distributional information in their computational model of word segmentation. Allen and Christiansen (1996) argue that the integration of such cues allows for an interaction which in itself is a powerful catalyst for learning. These modeling results are supported by behavioral data which suggest that 9-month-old infants are sensitive to mismatches of distributional and prosodic regularities (e.g., Morgan & Saffran, 1995).

The results presented here indicate that infants possess at least the minimal computational machinery needed to discover the regularities of their language: the ability to detect and represent serial order information. This in itself is not sufficient for word segmentation, which requires the extraction of relative frequencies of ordered strings to compute transitional probabilities, but it is a necessary prerequisite for this process. In fact, recent research in our laboratory has demonstrated that eight-month-old infants can use the relative frequencies of co-occurrence of sound pairs to detect word boundaries (Saffran, Aslin, & Newport, under review), lending further support to the present results.

More generally, future research must continue to investigate the means by which young learners make use of the wealth of statistical information available to them in the language input. The combination of innately constrained learning mechanisms and statistically rich input is potentially immensely powerful, and it is imperative that we gain a greater understanding of the ways in which this interaction renders young humans such superb language learners. The present experiment is one of a few recent studies which have begun to document the rapidity and extent of infant learning using carefully controlled exposures in the laboratory (see also Goodsitt *et al*, 1993; Jusczyk & Aslin, 1995; Morgan & Saffran, 1995). It may therefore be premature to assume, as many researchers have, that the prodigious abilities of young infants necessarily reflect innately specified knowledge. Rather, what may be innate is the human capacity to learn and reorganize the regularities which structure our environment, thereby allowing infants to make sense of what may initially be a "blooming, buzzing, confusion".

Acknowledgements

This research was supported by an NSF predoctoral fellowship to J.R.S., NSF grant SBR-9421064 to R.N.A., and NIH grant DC00167 to E.L.N. We thank J. Gallipeau, J. Hooker, P. Jusczyk, A. Jusczyk, K. Ruppert, J. Sawusch, and especially T. Mintz for their help with various aspects of this research.

References

Allen, J., & Christiansen, M. H. (1996). Integrating multiple cues in word segmentation: A connectionist model using hints. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Earlbaum Associates.

Brent, M. R., & Cartwright, T. A. (in press). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition.*

Brown, R. (1973). *A first language.* Cambridge, MA: Harvard University Press.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1994). Lexical segmentation: The role of sequential statistics in supervised and un-supervised models. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Earlbaum Associates.

Christophe, A., Dupoux, E., Bertoncini, J., & Mehler, J. (1994). Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. *Journal of the Acoustical Society of America,* **95,** 1570-1580.

Cole, R., & Jakimik, J. (1980). *A model of speech perception.* Hillsdale, NJ: Lawrence Earlbaum Associates.

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

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

Harris, Z. S. (1955). From phoneme to morpheme. *Language,* **31,** 190-222.

Hayes, J. R., & Clark, H. H. (1970). Experiments in the segmentation of an artificial speech analog. In J. R. Hayes (Ed.), *Cognition and the development of language.* New York: Wiley.

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., Cutler, A., & Redanz, L. (1993). Infants' sensitivity to predominant stress patterns in English. *Child Development,* **64,** 675-687.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language,* **33,** 630-645.

Kelly, M. H. (1988). Phonological biases in grammatical category shifts. *Journal of Memory and Language,* **27,** 343-358.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review,* **101,** 676-703.

Mandel, D., Kemler Nelson, D., & Jusczyk, P. W. (in press). Infants remember the order of words in a spoken sentence. *Cognitive Development.*

Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology,* **19,** 498-550.

Morgan, J. L., & Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development,* **66,** 911-936.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review,* **76,** 165-178.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (under review). Learning of sequential statistics by 8-month-old infants.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (in press). Word segmentation: The role of distributional cues. *Journal of Memory and Language.*

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (under review). Incidental language learning: Listening (and learning) out of the corner of your ear.

Wolff, J. G. (1975). An algorithm for the segmentation of an artificial language analogue. *British Journal of Psychology,* **66,** 79-90.