Distributional Regularity and Phonotactics are Useful for Early Lexical Acquisition*

Michael R. Brent, Timothy A. Cartwright, and Adamantios Gafos
Department of Cognitive Science, Johns Hopkins University, Baltimore, MD 21218 USA

Abstract

In the course of language acquisition, infants must segment connected speech into sound sequences that can be stored in the lexicon and eventually paired with meanings. However, there is no known acoustic analog of the blank space that separates printed words, so it is not clear how infants can segment speech into words at the stage where most words are unfamiliar to them. This paper investigates two sources of information that might be useful for speech segmentation at the onset of lexical acquisition: distributional regularity and phonotactics. Informally, distributional regularity *refers to the intuition that sound sequences that occur frequently and in a variety of contexts are better candidates for the lexicon than those that occur rarely and in few contexts. This paper begins by formalizing that intuition. This formalization makes possible the development of algorithms that, starting out without any lexical items, acquire a lexicon from unsegmented utterances. In addition, three types of phonotactic constraints are investigated. One of the three types is exemplified by the fact that* bigdog *cannot be segmented into the two words* bi *and* gdog *because* gdog *is not a possible syllable of English, and therefore it is not a possible word. The other two constraints state that every word and every syllable must have a* syllabic *(vowel-like) segment. By applying computer implementations of lexical acquisition algorithms to phonetic transcripts of child-directed English, we show that both distributional analysis and two of the three phonotactic constraints can be used to significantly improve lexical acquisition. Further, the contributions of these two information sources are not redundant, so using both yields better lexical acquisition than using either one*

Correspondence to: Michael Brent, Dept. of Cognitive Science, Johns Hopkins University, Baltimore, MD 21218 USA. **Internet:** *michael@mail.cog.jhu.edu*

*This work has benefitted greatly from discussions with Lila Gleitman, Peter Jusczyk, Brian MacWhinney, Gary Marcus, Jacques Mehler, Jeff Siskind, and Paul Smolensky. Special thanks are due to Sandor Brent for his careful reading of several drafts and thoughtful comments on them.

alone.

Knowing a language implies, among other things, knowing a lexicon—a set of correspondences between individual sounds, their meanings, and their syntactic privileges. Evidence for this includes the fact that speakers can understand and create an unbounded number of sentences using finitely represented knowledge, as well as the introspective sense that we know individual words and their meanings. Some of the central questions of language acquisition the lexicon. For example, how do children learn the correspondence between meanings and sounds? How do they learn the syntactic privileges of meaningful sounds? In recent years, a seemingly more humble facet of lexical acquisition has become a focus of increasing interest: How do children learn which sounds belong in the lexicon?

At first glance, the problem may seem trivial—words belong in the lexicon, and words can be collected directly from the speech that serves as input to language acquisition. However, no acoustic analog of the blank space that separates printed words has been discovered; certainly, boundaries between spoken words are not generally marked by silent pauses (e.g., Fisher and Tokura, 1994).[1] This suggests that infants may use other strategies for segmenting connected speech into lexical items.

However, it might be thought that children learn words by hearing them in isolation. This theory depends on there being enough isolated words in the input to get children to the stage where they can segment multi-word utterances. It is impossible to determine how many isolated words might be needed without a theory of how the ability to segment is acquired. However, it is known that some mothers do not use isolated words even when consciously striving to teach their infants specific words (Aslin, Woodward, LaMendola, and Bever, 1994). Further, it has been argued that explicit teaching does not occur in some cultures, and even in cultures where explicit teaching does occur, it does not begin until well after the onset of lexical acquisition (Cutler 1994a,b).

Another problem with the isolated-words theory is that children must eventually learn to segment speech in any case. It seems clear that when adults learn new words in sentential contexts, they first segment the sentence and identify all the familiar words in it; whatever remains is the novel word. Thus, the isolated-words theory implies that lexical acquisition begins without the ability to segment speech, then that ability is acquired, and eventually it is integrated into the lexical acquisition procedure. This theory must explain how and why the transition from isolated-word learning to segmentation-based learning comes about.

The isolated-words theory suffers from yet another logical flaw: in order to

[1] However, see (Christophe, Dupoux, Bertoncini, and Mehler, 1994) for evidence suggesting that there are acoustic correlates of word boundaries in French, and that infants can sometimes detect them, at least in a laboratory setting.

enter isolated words into their lexicons, infants would need the ability to distinguish them from multi-word utterances (Christophe, Dupoux, Bertoncini, and Mehler, 1994). There is no reason to believe that this is fundamentally easier than segmentation itself.

Finally, the notion that segmentation develops late is directly contradicted by laboratory evidence. Jusczyk and Aslin (in press) familiarized 7.5-month-old American infants with two different monosyllabic words and then presented them with passages that either did or did not contain the familiar words in sentences. The infants listened significantly longer to the passages containing the target words than to passages without the target words. They also listened longer when the familiarization and test stimuli were reversed, so that the test phase required the infants to recognize isolated words with which they had been familiarized in sentential context. These results indicate that at 7.5 months, well before there is any evidence of lexical acquisition, infants are able segment speech well enough to identify monosyllabic words in sentential context.

These arguments lend substantial support to the theory that speech segmentation plays a central role at the onset of lexical acquisition. But how can infants segment speech before they have a lexicon against which to match utterances? Specifically, what aspects of the input, what innate knowledge, and what acquired knowledge do they use in early segmentation? How do they use it?

Information sources

We begin to address these questions by asking what aspects of the input and what linguistic knowledge are potentially useful for pre-lexical segmentation. Five linguistic regularities that might prove useful are allophonic variation, semantics, metrical stress, distributional regularity, and phonotactics.

Allophonic variation

Church (1987) argues that allophonic variation is useful for finding syllable boundaries, and speculates that future research might show syllable boundaries to be useful for finding word boundaries. However, the syllables of fluent speech often cross word boundaries. For example, /gɪvə/ (*give a*) is syllabified as /gɪ·və/. Neither of these CV syllables constitutes a word, nor does their combination. Using the aspiration rule that Church cites, *tapdance* and *taproom* are syllabified as *tap·dance* and *ta·proom*. The status of the word *tap* is the same in both utterances, but it is split across two syllables in *ta·proom*, while it is left whole in *tap·dance*.

Semantics

Semantics must play some role in lexical acquisition, since a lexical entry is an association among representations of sound, meaning, and syntax. The role of semantics can be understood as either active, passive, or somewhere in between. On the active extreme, sounds are only stored in the lexicon once they can be assigned a tentative meaning—patterns in the sound of the input alone do not lead to long term storage. On the passive extreme, sounds are entered into the lexicon with or without meanings, but they are eventually forgotten if no meaning can be assigned to them.

The traditional view favors the active extreme, but recent evidence suggests that this view may need revision. As described above, Jusczyk and Aslin found that infants can recognize monosyllables in sentential contexts by age 7.5 months, well before the first evidence of comprehension. Hohne, Jusczyk, and Redanz (1994) showed that 8-month-olds who were repeatedly read stories preferred to listen to a list of words that appeared in those stories over a matched control list, even two weeks after the last presentation of the stories. Yet, in tasks that measure word recognition by the infants' orienting toward the referent, subjects do not show word recognition in sentential contexts until 11 to 13 months (Oviatt, 1980; Thomas, Campos, Shucard, Ramsay, and Shucard, 1981). This suggests that attention to meaning makes segmentation *more* difficult rather than less, perhaps as a result of increased cognitive load (Jusczyk and Aslin, in press). On the strength of this evidence, we pursue the passive view of the role of semantics, while acknowledging that the question remains open.

Metrical stress

One of the most frequently discussed cues to segmentation is the fact that the initial syllables of English content words rarely have the reduced vowels /ə/, /ɪ/, and /ɛ/ (Cutler and Carter, 1987, estimate 10%). Conversely, internal syllables of English content words usually have reduced vowels (Cutler and Carter, 1987, estimate 75%). Recently, it has been proposed that these regularities play a role in pre-lexical segmentation and early lexical acquisition (Cutler, 1994a,b). This proposal and related evidence is considered further in the General Discussion.

Distributional Regularity

Sound sequences that belong in the lexicon, such as those of *big* and *dog*, differ from accidental cooccurrences, such as *igdo*, in that the former recur frequently in a variety of contexts, whereas the latter occur less frequently and in fewer contexts. For example, /dɔg/ (*dog*) occurs in *bigdog, smalldog, adog, thedog*, and so on. The string /igdɔ/, on the other hand, almost never occurs except between /b/ and /g/. *Bigdog* also occurs in a variety of contexts, but it is rarer than its components *big* and *dog*. Thus, frequency and context variability must both be considered. We refer to this intuitive relationship among frequency, context variability, and lexical status as *distributional regularity.* One of the aims of this paper is to formalize this intuition.

Phonotactics

Phonotactics refers to the well-formedness rules for syllables. The relation of phonotactics to lexical segmentation is both subtle and, as we argue below, quite important. At its heart is the fact that lexical items are composed of well-formed syllables.[2] For example *gdog*, as an ill-formed syllable, is also an ill-formed word of English.

The syllable consists of an initial sequence of consonants called the *onset*, followed by a *syllabic* (vowel-like) segment called the *nucleus*, followed by a final sequence of consonants, called the *coda.* In /grɔg/ (*grog*), for example, the onset is /gr/, the nucleus is /ɔ/, and the coda is /g/. Every syllable must have a nucleus, but codas and the onsets are optional. Since every syllable must have a syllabic nucleus and every word consists of one or more syllables, it follows that every word must have a syllabic segment. This is the first of the phonotactic constraints investigated below.

In addition to the syllabic nucleus constraint, languages impose constraints on which consonant sequences can serve as onsets and which can serve as codas. For example, English specifies that *gr* is a possible onset, but *gd* is not. In many languages the consonant clusters permitted at word boundaries—word-initial onsets and word-final codas—are different from those permitted word-internally. In Greek, for example, onsets consisting of a fricative followed by stop, such /xt/ and /ft/, are permissible word-initially, but not word-internally. Conversely, codas other than /s/ and /n/ are permissible only word-internally, not word-finally. Thus, the constraints on word-initial onsets are more permissive than those on internal onsets, but the constraints on word-final codas are less permissive than those on internal codas. Such examples are not rare. For instance, Arabic (Hooper, 1976), Italian (Chierchia, 1983), and arguably English (Borowsky, 1986) have distinct rules for boundary clusters and internal clusters.

The second constraint investigated here forbids words whose boundary clusters are not permissible in English. For example, a learner who knows that *gd* is not a possible *word-boundary* onset, can rule out the possibility that /bɪgdɔg/ (*bigdog*) consists of the two words /bɪ/ and /gdɔg/. Although boundary-cluster constraints vary from language to language, infants could learn them by assuming that all and only the consonant clusters that occur at *utterance* boundaries and permissible at *word* boundaries. The third constraint requires word boundaries

[2] This does not imply that the syllabification of fluent speech is consistent with word boundaries, only that isolated words can be syllabified.

in consonant clusters that are not permitted word-internally. Since every word is a sequence of syllables, every consonant sequence internal to a word must consist of the coda sequence of one syllable followed by the onset sequence of the next. A consonant cluster that cannot be divided into a permissible internal coda followed by a permissible internal onset cannot be word-internal. That is, such a cluster must contain a word boundary. For example, consider the cluster /zskr/, as in *these scratchy*. The division /·zskr/, where the entire cluster forms an onset, is not permissible.[3] The division /z·skr/ is impossible because /skr/, while it is fine as a word-initial onset, is not a permissible internal onset.[4] The divisions /zs·kr/, /zsk·r/, and /zskr·/ are impossible because /zs/, /zsk/, and /zskr/ are not permissible internal codas. Since there is no possible parse of /zskr/ into a word-internal coda followed by a word-internal onset, it must contain a word boundary. We investigate a version of this constraint that is thought to be cross-linguistically valid.

The central aim of this paper is to assess the utility of the syllabic nucleus constraint, the constraints on boundary clusters, and the constraints on internal clusters, for pre-lexical segmentation and lexical acquisition.

Infants' innate capacities and acquired knowledge

To provide the background for our analysis, this section reviews some evidence about infants' innate capacities and acquired knowledge. This evidence addresses three questions: First, which differences in speech sounds can infants detect? Second, how do infants represent speech sounds? Third, what do infants learn about their native language, and when?

Which differences in speech sounds can infants detect?

Results from this area of research suggest that infants are very good at discriminating stimuli whose differences, when described in linguistic terms, are rather small. For example, they can distinguish pairs of monosyllables or bisyllables that differ in only a single articulatory feature. They can do this whether the distinguishing feature is in a stressed syllable or an unstressed syllable. They can distinguish bisyllables that differ only in stress placement. They can distinguish monosyllables differing by a single feature even when the stimuli include examples from a variety of different speakers. (See Jusczyk, in press, for a review.)

[3] There are a variety of arguments to support this, but the simplest is distributional—if /zskr/ were a permissible internal onset, then it should be able to follow permissible internal codas. However, sequences like /tzskr/, /nzskr/, and /rzskr/ are never observed inside English words, despite the fact that /t/, /n/, and /r/ are perfectly legitimate codas.

[4] In words like *ascribe* and *describe* /skr/ is heterosyllabic.

However, these results do not imply that it is articulatory features or stress per se that the infants are detecting. They could be storing acoustic rather than linguistic representations and simply noting that the two acoustic forms do not match. Thus, these experiments show that infants are able distinguish sounds whose differences can be described in linguistic terms, but they do not show that infants in fact describe those differences in linguistic terms.

How do infants represent speech sounds?

In experiments aimed at determining whether infants describe (or *categorize* or *represent*) speech sounds in terms of phonetic segments, 2 to 3-month-olds were habituated to *sets* of CV syllables that shared either a common vowel or a common consonant (Jusczyk and Derrah 1987; Bertoncini, Bijeljac-Babic, Juszcyk, Kennedy, and Mehler 1988). A new syllable was then added to the stimulus set. In one condition, the new syllables shared the common segment of the original set, while the others it did not. The authors reasoned that if the infants represent the CV syllables as a consonant-vowel sequence rather than an atomic unit of speech they ought to be able to categorize them by their common segment. If they categorize the habituation set by its common segment, then a new stimulus that shares that segment ought to be perceived as less novel than one that doesn't share it. As a result, the degree of dishabituation or the recovery time or both should be reduced. The authors found no statistically significant evidence for such a reduction, and interpreted this as suggesting that infants do not represent speech sounds as sequences of segments.

Although the impact of these experiments on lexical acquisition must be considered carefully (see General Discussion), they are important in that they address the question of which similarities infants detect in speech sounds, rather than which differences they detect.

What do infants learn about their native language, and when?

The two questions discussed so far focus on universal capacities, but it is also possible to investigate what infants have learned about their native language. One approach is to present them with stimuli that either do or do not conform to the regularities of their native language and test whether they prefer to listen to the conforming stimuli. A relevant result from this paradigm is the finding that 9-month-olds listen longer to words that are phonotactically permissible in their language than to those that are not, but 6-month-olds show no preference (Friederici and Wessels, 1993; Jusczyk, Friederici, Wessels, Svenkerud, and Jusczyk, 1993). Among phonotactically permissible words, American 9-month-olds (but not 6-month-olds) prefer to listen to CVC syllables whose segments and segment juxtapositions are common in English, as compared to those that are permissible but rare (Jusczyk, Luce, and Charles-Luce, 1994).

Which information sources are useful for a particular language acquisition task?

The three general questions discussed above can all be addressed by studying infants directly. For example, we have listed five possible sources of information that might aid segmentation: allophonic variation, semantics, metrical stress, distributional regularity, and phonotactics. A natural question is how useful each information source is. We cannot address this question by selectively depriving infants of each information source and comparing the effects on their acquisition of language. Three alternative methods for addressing utility have been used: measuring cue validity by correlation-regression (e.g., Fisher, Gleitman, and Gleitman, 1991; Fisher and Tokura, 1994; Kelly, 1992); measuring learning on an artificial analog of a language acquisition task (e.g., Morgan, Meyer, and Newport, 1987; Valian and Coulson, 1988); and building an artificial learning device for a more natural task (the present approach).

Correlation-regression

An example of the correlation-regression technique is measuring acoustic variables, such as pitch excursion and volume, and regressing them against linguistic variables, such as the locations of clause boundaries (Fisher and Tokura, 1994). This reveals how well a linear function of the acoustic variables can predict the locations of clause boundaries. Some non-linear function might yield better predictions, so the result is only a lower bound on the utility of the acoustic information. While this method can demonstrate the presence of a useful information source, it does not suggest how children could actually use that source.

Analog tasks

An example of the analog-task approach is giving subjects examples from a formally-defined artificial micro-language, and testing the degree to which they can abstract the rules of the language from the examples. This approach has the advantage of taking into account human computational capacities, such as memory and attention. On the other hand, it has the limitation that people generally cannot perform on any but the simplest tasks under laboratory conditions. Further, the relationship between the artificial task and the natural task can be hard to demonstrate.

Simulation

The approach taken in this paper, called *simulation*, involves building a computer program that uses a particular strategy for exploiting the information sources under study. The program can then be selectively deprived of various information sources, and the effect on its learning measured. Simulation has some of the limitations of the other two techniques, but to a lesser degree.

Like correlation-regression, it only provides a lower bound on utility—some other strategy for exploiting the information source might make better use of it. However, simulations can discover a much wider range of relationships among the dependent and independent variables than linear regression. Like analog task studies, simulation studies often abstract away from the natural task. For example, this paper studies the task of extracting a lexicon from unsegmented input represented as strings of phonetic segments. The mental representation of the input from which children must extract their lexicon is not known. Despite these limitations, computer programs have more patience and endurance than human subjects, and hence simulation experiments can generally be applied to a wider range of tasks than analog experiments with human subjects. Another advantage of computer programs over human subjects is that it is possible to determine precisely what they are doing when perform a task.

The interpretation of simulation experiments can vary along a continuum. On one extreme, the simulation serves the same function as the statistics package in a correlation-regression experiment—it simply shows that a certain relationship between an information source and a category of linguistic knowledge exists and can be exploited by some procedure. Just as regression studies make no claim about how child learners exploit the regularities between the dependent and independent variables, this interpretation of simulations makes no claim that the child uses the same algorithm as the program. Simulation differs from regression only in using more powerful methods of predicting the dependent variables from the independent variables. At the other extreme, the simulation program is interpreted as a detailed analog of human processing. The literalist interpretation of neural networks, for example, holds that each processing unit corresponds to a neuron and each connection to a synapse . Under this interpretation, parallels between network behavior and human behavior arise because the dynamics of the artificial network correspond directly to the dynamics of a network in the brain.

Typically, the most sensible interpretation of simulation experiments lies somewhere in between. For example, the algorithms used in the following experiments abstract away from human memory limitations, in that they use as much memory as necessary, with perfect recall. That limits how literally they can be taken as a models of human processing. Nonetheless, they provide a good starting point for the development of models that are more faithful to the details of human computational capacities. At the same time, they are useful as a sophisticated form of regression which provides information about the predictive power of various information sources.

A formalization of sequence distribution

Informally, *distributional regularity* refers to the intuition that sound sequences that occur frequently and in a variety of contexts are better candidates for the lexicon than those that occur rarely and in few contexts. This intuition is related to Harris's (1951) prescription for identifying morphemes, and to various published suggestions about segmentation at minima of phonetic transition probabilities (e.g., Hayes and Clark, 1970). Olivier (1968) and Wolff (1988) have developed formal algorithms inspired by similar intuitions. Olivier's algorithm is based in part on maximum likelihood segmentation, a theoretically sound notion, but it uses a method of generating new lexical entries that the author himself describes as unprincipled. Wolff's approach seems to be a direct translation of his intuitions, unmediated by a theory of induction. Neither Olivier nor Wolff formalized distributional regularity in a declarative fashion, as distinct from the aspects of their algorithms necessitated by practical considerations like machine time.

The present approach, in contrast to earlier efforts, uses a declarative measure of the distributional regularity of a segmentation and the lexical items it implies. This measure is used to evaluate hypotheses about the correct segmentation of the input. The evaluation function is based on the minimum representation length theory of induction, also called the minimum description length (MDL) principle (Brent, 1993; Ellison, 1992; Li and Vitányi, 1992; Rissanen 1989). We present the evaluation function at an intuitive level here and provide technical details in the Appendix.

Lexical acquisition can be seen as the task of inductively inferring the mental lexicon from which the input was generated. This task can be conceptualized as a search for the lexicon that provides the best explanation for the distributional regularity observed in the input. An induction algorithm, then, can be constructed out of a formal evaluation function that measures the explanatory power of each hypothesis, and a search mechanism that generates plausible hypotheses for evaluation. We focus here on the evaluation function.

Evaluating explanations

The left side of Figure 1 shows one segmentation—the intuitively correct one—for a broad phonetic transcription of three utterances: "Do you see the kitty? See the kitty? Do you like the kitty?" The right hand side of Figure 1 shows the lexical items used in that segmentation (Lexicon) and the order in which they are used (Derivation). The lexicon and derivation together form a hypothesis that provides an explanation of how the input was generated. The explanation says that the input has the regularities it has because the speaker had in his or her mind this lexicon, and chose to use its words in the order specified by this deriva-

Segmentation	Explanation			
	Lexicon			Derivation
du ju si ðə kIti				1 3 5 2 6
si ðə kIti	1 du	2 ðə	3 ju	5 2 6
du ju laIk ðə kIti	4 laIk	5 si	6 kIti	1 3 4 2 6

Figure 1: Left: A segmentation of a phonetic transcription of "Do you see the kitty? See the kitty? Do you like the kitty?" Right: The explanation corresponding to the segmentation shown at left.

tion. There is exactly one segmentation corresponding to each explanation—the one in which the indices in the derivation are replaced by the corresponding items from the lexicon. Conversely, there is exactly one lexicon-derivation pair corresponding to each segmentation. The lexicon consists of all the word types in the segmented text, each paired with an arbitrary unique index. The derivation corresponding to a given segmentation is constructed by replacing each word by its lexical index. Thus, explanations and segmentations are in a one-to-one correspondence. The evaluation function applies to explanations, but because of the one-to-one correspondence, it can just as well be thought of as evaluating segmentations.

There is an enormous number of possible segmentations of each input, including the null segmentation, in which every lexical item is an entire utterance, and the complete segmentation, in which every lexical item is a single phonetic segments. Each segmentation corresponds to one explanation. The learner's task is to find the explanation that is most likely to reflect the long-term regularities of the language, not the accidental regularities of a particular input sample. We propose a method of evaluating competing explanations that uses the minimum representation length principle of induction to implement common intuitions about distributional regularity. This principle can be thought of as formalizing Occam's Razor—the notion that the best explanation of a set of observations (e.g., linguistic inputs) is the least stipulative one. The minimum representation length principle equates the stipulativeness of an explanation with the number of characters in its representation. Thus, shorter explanations are preferred over longer ones. For technical reasons it is better to convert representations of the type shown in Figure 1 into a compact binary representation and then count the number of bits in that representation (see the Appendix). For purposes of gaining intuition, however, it is sufficient to simply count the numbers and letters in the representation shown in Figure 1. Note that this evaluation applies to complete explanations—it cannot be used to evaluate individual lexical items or word boundaries.

The evaluation function formalizes linguistic intuitions

The proposed evaluation seems to faithfully implement intuitions about distributional regularity. Intuitively, a phonetic sequence that appears in a variety of different contexts, such as /ju/ (*you*) in the example, should be given its own lexical entry. On the other hand, if two phonetic sequences always occur together in the same order, like /ðə/ (*the*) and /kIti/ (*kitty*) in the example, then there is no *distributional* reason to split them into two words with separate lexical entries. Figure 2 shows that the proposed evaluation yields the same preferences: the intuitively "correct" explanation (first row) is shorter than one that is identical except for joining *you* and *see* as a single word (second row); however, an explanation that is identical to the correct one except for joining *the* and *kitty* as a single word (third row) is shorter still. (Relevant portions of each alternative explanation in Figure 2 are highlighted with boxes.) Specifically, joining *you* and *see* (second row) saves one index in the derivation, replacing the one occurrence of the sequence 3 5 by 7, the index for *yousee*. However, this savings does not compensate for the four additional letters and one digit required to list *yousee* in the lexicon. The new lexical entry adds to the size of the lexicon because *you* and *see* also appear separately, so *yousee* cannot replace their individual entries. On the other hand, joining *the* and *kitty* (third row) saves three indices in the derivation, replacing every 2, 6 by 2, without increasing the size of the lexicon. Since *the* and *kitty* do not appear separately in this small sample, the combined entry for *thekitty* can replace the separate entries for *the* and *kitty*.

The minimum representation length principle is appealing on several levels. First, it is a formalization of Occam's Razor. Second, people are able to produce and understand an infinite number of utterances using only a finite amount of memory. At an intuitive level, the need to represent an infinite language in finite memory can viewed as a functional pressure to discover regularities that minimize the amount of memory needed to represent the utterances that are known to be in the language. This is precisely what minimum representation length evaluations do.

Experimental Prospectus

In the following experiments, we attempt to quantify the usefulness of distributional and phonotactic regularities for early lexical acquisition. Experiments 1 and 2 focus on phonotactic constraints ruling out words that either (1) lack a syllabic nucleus or (2) have a boundary onset or coda that is not permissible in English. In Experiment 1 we investigate the usefulness of these constraints in combination, the usefulness of distributional analysis, and the interaction between the two. Experiment 2 looks at the effect of imposing the syllabic nucleus

	SEGMENTATION	EXPLANATION	
		LEXICON	DERIVATION
CORRECT	du ju si ðə kɪti si ðə kɪti du ju laɪk ðə kɪti	1 du 2 ðə 3 ju 4 laɪk 5 si 6 kɪti	1 3 5 2 6 5 2 6 1 3 4 2 6
YOUSEE	du $\boxed{\text{jusi}}$ ðə kɪti $\boxed{\text{si}}$ ðə kɪti du $\boxed{\text{ju}}$ laɪk ðə kɪti	1 du 2 ðə 3 $\boxed{\text{ju}}$ 4 laɪk 5 $\boxed{\text{si}}$ 6 kɪti $\boxed{7}$ $\boxed{\text{jusi}}$	1 $\boxed{7}$ 2 6 5 2 6 1 3 4 2 6
THEKITTY	du ju si $\boxed{\text{ðəkɪti}}$ si $\boxed{\text{ðəkɪti}}$ du ju laɪk $\boxed{\text{ðəkɪti}}$	1 du 2 $\boxed{\text{ðəkɪti}}$ 3 ju 4 laɪk 5 si □□	1 3 5 $\boxed{2}$ 5 $\boxed{2}$ 1 3 4 $\boxed{2}$

Figure 2: Three segmentations (left) with the corresponding explanations (right). The explanation in the first row (CORRECT) is the intuitively correct one. The explanation in the second row (YOUSEE) treats adjacent occurrences of *you* and *see* as a single word. The third row shows an explanation that treats adjacent occurrences of *the* and *kitty* as a single word. The explanation in the third row has the fewest characters. Boxes highlight the points at which the second and third explanations differ from the first.

constraint without the boundary-clusters constraint. In Experiment 3 we shift the focus to constraints that require word boundaries in consonant clusters that are not permitted word-internally. Finally, Experiment 4 looks at the effect of input sample size.

Experiment 1: Are distribution and boundary phonotactics useful?

The aim of this experiment is to determine whether distributional analysis is useful for pre-lexical segmentation, whether certain phonotactic constraints are useful, and whether the combination of distribution and phonotactics is more useful than either information source by itself. Logically, it is possible that the intuitions behind distributional analysis are simply wrong. It is also possible that distributional analysis might discover the phonotactic regularities on its own. In that case adding them as explicit constraints would result in little or no improvement.

The experiment is a 2x2 repeated measures design, in which the conditions represent four different segmentation algorithms. The two variables that distinguish the algorithms are whether they use phonotactic constraints and whether they use distributional analysis.

The distributional algorithms are based on the minimum representation length evaluation described above. These algorithms search through many hypothetical segmentations of the text and select the one that yields the shortest explanation. The non-distributional algorithms are a random baseline for comparison, not a plausible approach to lexical acquisition. Given the correct number of word boundaries in the input, they insert that number of boundaries, with equal probability of inserting a boundary between each pair of adjacent phonetic segments. It is important to note that this algorithm overestimates the performance of a system without distributional analysis, since it is given the correct number of boundaries to insert—information that would never be available to a real learner. The distributional algorithms are not given any information about the actual number of word boundaries in the input.

The phonotactic algorithms are provided with a list of all the onsets and codas that English permits at word boundaries. These algorithms do not consider segmentations in which words either lack a syllabic nucleus or have a forbidden boundary cluster. A syllabic nucleus is either a vowel, a syllabic /l/ like the final consonant of *label*, or a syllabic /m/ or /n/. The non-phonotactic algorithms are given no language-specific information of any kind. Note that none of the algorithms explored in this paper start out knowing any words. They must acquire a lexicon from scratch.

Method

Distributional analysis algorithms

The distributional algorithms search through many hypothetical segmentations of the text and select the one that yields the shortest explanation, where the length of an explanation is measured by the formula given in the Appendix. However, there are so many possible segmentations of any input that it is impossible to evaluate them all. The algorithms in this experiment use the following method to decide which segmentations to evaluate. The initial hypothesis is the segmentation that contains no word boundaries except those at utterance boundaries. The following step is repeated until no more boundaries can be inserted: Evaluate all hypotheses that differ from the current hypothesis by the addition of one or two new word boundaries.[5] Among all these, choose the one with the best evaluation to be the new current hypothesis. Without phonotactics, this procedure terminates when word boundaries have been inserted between every pair of adjacent phonemes. With phonotactics, it terminates when inserting an additional word boundary would violate the phonotactic constraints. When the search terminates, the best explanation among all those evaluated is returned.

Input

The four algorithms were applied to broad phonetic transcripts of natural child-directed speech. Orthographic transcripts made by Bernstein-Ratner (1987) were taken from the CHILDES collection (MacWhinney and Snow, 1985) and transcribed phonetically. The speakers were nine mothers speaking freely to their children, whose ages averaged 18 months (range: 13–21).

The philosophy behind our transcription was to preserve all phonemic distinctions while minimizing the number of subjective judgments and the amount of labor required. Accordingly, every word was transcribed the same way every time it occurred, regardless of context. Diphthongs and r-colored vowels were each transcribed as a single character. Syllabic consonants were distinguished from the corresponding non-syllabic consonants. Thus, the first and last consonants in *label* were represented by different characters. Onomatopoeia (e.g., *bang*) and interjections (e.g., *uh* and *oh*) were eliminated for three reasons: they occur in isolation much more frequently than ordinary words, and hence would have skewed the scores toward better performance; their frequency is highly variable from speaker to speaker, so they would have increased the variance in the scores; and there is no standard spelling or pronunciation for many of them, so we could not tell from the orthographic transcript what sound was actually uttered.

Finally, all word boundaries were removed, but utterance boundaries marked in the Bernstein-Ratner transcript were left intact, since utterance boundaries are prosodically marked (Fisher and Tokura, 1994; Jusczyk and Kemler Nelson, 1994). We did not insert any additional utterance boundaries, even at obvious sentence boundaries. (Apparently, the original transcript contains utterance boundaries only between utterances separated by an audible pause.) Each transcript was truncated to about 1350 non-boundary segments. Segments that fall at utterance boundaries were not counted in order to balance the number of points where there is a decision to make about the presence of a word boundary. This choice of length was somewhat arbitrary, although it was constrained from above by both processing time and the length of the shortest transcript.

[5] In pilot work, we tried a faster, narrower search which explored at each step only those hypotheses that differed from the current hypothesis by a single boundary. The performance was markedly worse than that of the current search. Perhaps this is because a pair of boundaries can segment a known word from the middle of an utterance, whereas a single boundary cannot.

Scoring

Each program outputs a segmentation of the text and a lexicon consisting of all the word types used in the segmentation. The results were scored by comparing the lexicon output by each program to the list of word types in a standard segmentation of the corpus.[6] In order to make the segmentation standard as objective as possible, we based it directly on the whitespace and punctuation that would occur in an orthographic representation—all spaces and punctuation marks were taken to be word boundaries, except apostrophes. Thus, *what's*, *its*, and *it's* were all treated as single words in the standard. We did not attempt to use a theoretically motivated notion of *word* because doing so would have involved distinguishing compounds, derivational morphology, inflectional morphology, and historical legacy, as well as estimating productivity.[7] Further, it would have been necessary to establish the psychological reality of any theoretical definition of the lexicon. Determining the contents of the mental lexicon is beyond the scope of this paper.

Results and discussion

Two statistics were computed: accuracy and number of correct lexical discoveries, or *hits*.[8] Accuracy is the proportion of the words in the program's lexicon that are also in the standard lexicon:

$$\text{accuracy} = \frac{\text{hits}}{\text{hits} + \text{false alarms}}$$

Accuracy varies between zero and one. This is a very conservative measure of accuracy—for example, if the input contains *walking* but not *walk*, and the program splits one occurrence of *walking* into *walk* and *ing*, both fragments will count as errors, even though both are reasonable candidates for the lexicon. The number of hits is simply the number of words that were in both the standard lexicon and the lexicon output by the program.

The results, averaged over all 9 subjects, are shown in Table 1. These results suggest that both boundary phonotactics and distributional analysis make a positive contribution to the accuracy of lexical acquisition, and that combining both sources yields a further improvement over either source alone. A 2x2 ANOVA

[6] See Cartwright and Brent (1994) and Brent, Gafos, and Cartwright (1994) for measurements based on scoring the individual word boundaries inserted by these programs.

[7] See Brent (1993) for a study of morpheme discovery in the minimum representation length framework.

[8] The number of hits is not presented as a proportion of the total number possible because discovering every word in the input is not necessary, nor is it desirable from the standpoint of modeling infant learning.

	Accuracy		Number Correct	
	−PHONO	+PHONO	−PHONO	+PHONO
−DIST	15%	43%	48	100
+DIST	35%	65%	30	82

Table 1: Accuracy and number of correct lexical items for four segmentation algorithms. Results are averages over 9 transcripts of child-directed speech. The +PHONO algorithms use both the syllabic nucleus constraint and the constraint the constraints on permissible word-boundary consonant clusters.

results in significant main effects for both phonotactics ($F(1,8) = 702, p < .0001$) and distributional analysis ($F(1,8) = 47.5, p < .0001$). Adding phonotactics improved accuracy on all nine transcripts, both with and without distributional analysis. Likewise, adding distributional analysis improved accuracy on all nine, both with and without phonotactics. There no significant interaction ($F(1,8) <$ 1, NS), which is compatible with the idea that these information sources make roughly independent contributions to accuracy; that is, they are not redundant. In addition to improving accuracy, phonotactics also improves the absolute number of hits significantly, ($F(1,8) = 246, p < .0001$). Adding phonotactics improved the number of hits on all nine transcripts, both with and without distributional analysis. Distributional analysis, on the other hand, yields slightly fewer lexical discoveries than the random baselines ($F(1,8) = 48.0, p < .0001$). For eight of the nine transcripts, distributional analysis reduced the number of hits. There no was significant interaction ($F(1,8) < 1$, NS) between the phonotactic and distributional variables.

Distributional analysis decreases the number of lexical items found by comparison to the random baseline because it is more conservative—it posits fewer lexical items, both correct and incorrect. The algorithm that uses both distribution and phonotactics starts out with no lexicon and discovers 82 new words in less than half an hour of speech, a rate of acquisition that is more than adequate. Further, the random baseline was a deliberate over-estimate of what can be done without distributional analysis, since it relies on knowledge of the number of words in the input—knowledge that is not available to either the child or the distributional algorithms.

In addition to improving accuracy and completeness and number of hits, adding phonotactics to the distributional analysis has an important effect on the nature of the non-word errors. Without phonotactic constraints, most of the non-words consist of single phonemes or other submorphemic fragments. When phonotactic constraints are imposed, an average of 66% of the non-words are concatenations of words in the standard lexicon. These two types of errors are likely to have very different consequences for infants. When a string of words is mistakenly entered into the lexicon, the infant may be able to split it into its component words later, once more input has been considered. Or, if it is a common and syntactically coherent string of words, such as *give me*, the child can fill out its lexical entry with appropriate syntax and semantics. There is, after all, no reason why the mental lexicon cannot store common phrases as well as individual words. None of these considerations apply to sub-morphemic fragments, which have no value and must eventually be purged from the lexicon. Thus, we believe that the measures used in this experiment underestimate both the improvement due to phonotactics and the absolute performance of the algorithm that combines phonotactics and distribution. Given the idealizations involved in this experimental system, however, the differences between conditions are probably more important than the scores in the individual conditions.

Experiment 2: Constraints on nuclear and peripheral segments

The phonotactic constraints in Experiment 1 impose two conditions on hypothesized words: (1) they must contain a syllabic nucleus, and (2) onsets and codas occurring at word boundaries must be permissible in English. It is reasonable to think of (1) as universal (perhaps defined in terms of sonority peaks) and (2) as language-specific, although there are probably some cross-linguistic generalizations about permissible boundary clusters. Experiment 1 showed that the combination of these two constraints improves lexical acquisition. In this experiment, we measure their individual contributions. It does not make sense to impose the boundary cluster constraint without the syllabic nucleus constraint, since onsets and codas are defined with respect to nuclei. Thus, we compared the performance of the distributional analysis algorithm under three conditions: no phonotactic constraint, the syllabic nucleus constraint only, and the syllabic nucleus constraint combined with the boundary clusters constraint.

The inputs and scoring are identical to those of Experiment 1.

Results and discussion

Table 2 shows that, given a distributional analysis, half of the improvement in accuracy from adding phonotactics is due to the syllabic nucleus constraint and half is due to the boundary-clusters constraint. A t-test shows that adding the syllabic nucleus constraint significantly improved accuracy ($t(16) = 2.46, p < .026$, two-tailed), and adding the boundary clusters constraint improved it significantly again ($t(16) = 2.57, p < .020$, two-tailed). The situation is similar for the number of correct words in the lexicon: three-fifths of the total improvement is due to the

Accuracy			Correct Words		
NONE	NUC	NUC&BC	NONE	NUC	NUC&BC
35%	50%	65%	30	62	82

Table 2: The effects of imposing no phonotactic constraint (NONE), the syllabic nucleus constraint alone (NUC), and the syllabic nucleus constraint with the boundary clusters constraint (NUC&BC).

syllabic nucleus constraint. Adding that constraint alone significantly improved the number of correct lexical items ($t(16) = 4.29, p < 0.0006$), and adding the boundary clusters constraint improved it significantly again ($t(16) = 2.56, p < 0.021$).

These results motivate the theory that infants exploit the universal constraint requiring every word to have a syllabic nucleus. In addition, they motivate the theory that children use language-specific constraints on word boundary clusters. However, the latter theory must address the question of how such constraints could be acquired. We believe that a very plausible candidate is utterance boundaries. Children could assume that all and only the consonant clusters appearing at utterance boundaries are permissible at word boundaries. Although more investigation is certainly needed, that theory appears to be both plausible as well as motivated.

Experiment 3: constraining word-internal consonant clusters

The constraint investigated in this experiment is based on the *sonority hierarchy.* The sonority hierarchy is partially motivated by acoustic and articulatory considerations, but for our purposes it can be viewed as a purely formal device for describing cross-linguistic generalizations about phonotactics. The idea is that all phonological segments are organized into a hierarchy, with the vowels being most sonorous, followed by the glides, the liquids (/l/ and /r/ in English), the nasals (/m/ and /n/ in English), and finally the obstruents—i.e., stops and fricatives (Clements, 1990). This hierarchy serves to demarcate the boundaries of word-internal syllables. Except for the consonant clusters occurring at word boundaries, a syllable consists of a sequence of segments with strictly increasing sonority (the onset and nucleus) optionally followed by a sequence of strictly decreasing sonority (the coda). The most sonorous segment of a syllable is the nucleus, assuming that it is sufficiently sonorous. Nuclei are typically vowels, but sometimes liquids and nasals serve as nuclei (e.g., the last consonant of *bottle*). If some obstruent is left unsyllabified, then the cluster it belongs to must contain a word boundary. For example, all three segments of the sequence /tsð/ in

What's this are obstruents, the lowest level of sonority. Since the sonority of this sequence neither rises nor falls, none of these segments can belong to the same syllable. That leaves the /s/ in a syllable by itself—one that does not have a nucleus. Thus, there must be a word boundary somewhere in /tsð/. We call this the *sonority constraint.* Some version of the sonority hierarchy is thought to be universal.[9]

The sonority constraint indicates sequences in which a word boundary must occur, but not where in those sequences it must occur. This experiment evaluates an algorithm that enforces the sonority constraint in addition to the syllabic nucleus and boundary clusters constraints. In this algorithm, the sonority constraint is enforced at the output stage, not during the search and evaluation. The algorithm outputs the hypothesis with the best evaluation among those that have word boundaries in all the consonant clusters that would otherwise violate the sonority constraint. Recall that the search procedure for the distributional algorithms inserts more word boundaries at each stage until no more can be inserted, but it never deletes boundaries. As more and more boundaries are inserted all the consonant sequences that must be split according to the sonority constraint will eventually be split. Thus, enforcing the constraint is simply a matter of selecting a hypothesis from a sufficiently late stage of the search that all the required boundaries have been inserted.

The inputs and scoring are identical to those of Experiment 1.

Results and discussion

On average, the sonority constraint showed a non-significant trend toward improving accuracy, but had little effect on the total number of correct lexical discoveries. The effect on accuracy was bi-modal, so averaged data are not very useful. On one transcript out of nine, accuracy improved from 57% to 72%. The remaining eight were affected only slightly and showed no particular trend toward improvement or deterioration, since the distributional analysis had already put word boundaries in most of the relevant clusters even without the constraint. This suggests that the current method of exploiting the sonority constraint is redundant with the distributional analysis in most cases.

However, there are other methods of exploiting the sonority constraint. Recall that the current algorithm does not use the sonority constraint during the search, it merely rules out hypotheses that do not split certain consonant clusters. An alternative is to insert word boundaries *before* the distributional analysis at those

[9] Apparently, there many sequences that occur only at word boundaries in English, but are not forbidden by the sonority constraint (Harrington et al., 1989). However, it is not clear how infants could discover them without already having a large sample of individual words.

	Length	Accuracy		Correct Words	
		NUC	NUC&BC	NUC	NUC&BC
Quarters	1279	54%	66%	62	76
Halves	2558	59%	70%	112	130
Whole	5117	61%	70%	188	198

Table 3: Accuracy (center) and number hits (right) as a function of average input size in phonetic segments (left). One algorithm uses only the syllabic nucleus constraint (NUC) and the other uses the constraints on English word-boundary clusters as well (NUC&BC).

locations where: (1) sonority implies the presence of a boundary in a particular cluster, and (2) boundary phonotactics rules out all but one location in that cluster. Whether sonority is more consistently useful with this technique remains to be seen.

Experiment 4: effects of input length

The more frequently a sound sequence occurs, and the more varied the contexts in which it occurs, the more a distributional analysis will tend to create a lexical entry for it. However, sampling error limits frequency and context variability in any particular input sample. In this experiment, we investigate the effect of input sample size on accuracy and number of hits. Two of the longest samples in the Bernstein-Ratner collection (average length 5117 phonetic segments) were each divided into halves and quarters. We ran the two algorithms from Experiment 2 on each of the whole samples, both halves of each sample separately, and all four quarters of each sample separately. If sample size affects the performance of these algorithms, that should be reflected in different average performance on wholes, halves, and quarters.

Results and discussion

Table 3 shows the accuracy (center) and number hits (right) as a function of average input size. The accuracy of both algorithms improves markedly from the quarters (about 1300 segments) to the halves (about 2600 segments), after which it appears to level off. The inputs used in earlier experiments were roughly the size of the quarters, so those experiments underestimate the accuracy of these algorithms on inputs of realistic size. The difference in accuracy between the two algorithms decreases slightly as corpus size increases, but appears to stabilize at about nine or ten points. Thus, the results of Experiment 2, which showed that the boundary clusters constraint is useful, appear to be confirmed for a variety of corpus sizes. The total number hits appears to be increasing roughly as the logarithm of the input size. The algorithm with both constraints discovers an average of 198 correct words from less than an hour's input.

Given the idealizations used in transcribing the input and scoring the output, the absolute performance of the best algorithm is probably not as meaningful as the differences between algorithms. This being said, it is natural to wonder how close the best algorithm comes to solving the problem of pre-lexical segmentation and early lexical acquisition. It would be very surprising if any method of analyzing sound patterns were to achieve 100% accuracy without some way to purge the lexicon of low-frequency items. The reason is that every speech sample contains many words that occur only once, so there is never enough data to attain confidence about every lexical item. The distributional system used here always takes its best guess, rather than abstaining for lack of sufficient evidence. Indeed, the ratio of the number of lexical items to the number of tokens in an input sample (the *type/token ratio*) appears to be a good predictor of accuracy. For example, the 70% accuracy attained by the best algorithm on the largest inputs results from averaging 60% accuracy on one input and 80% on the other. The type/token ratios of these two corpora are 0.21 and 0.15, respectively. Preliminary data indicate that there is a strong positive correlation between the type/token ratio and the age of the listener, and that our algorithms are more accurate on speech directed at younger children (Cartwright and Brent, 1994; Brent, Gafos, and Cartwright, 1994). Since these experiments were conducted on speech directed at children somewhat beyond the age at which lexical acquisition acquisition begins, they may underestimate accuracy on speech to younger infants.

Some mistakes in the output of the algorithms result from a certain inevitable arbitrariness in scoring according to orthographic standards. There is no way for the distributional analysis to guess that *sit down* will be scored as two words but *sitting* will be scored as one.

Finally, it is worth considering the effect of scoring the lexicon as opposed to the word boundaries in the segmentation. Lexicon scoring takes no account of frequency, so an error on a rare word is just as bad as an error on a common word. Since distributional analysis always does better with high frequency items, lexicon scoring produces much lower numbers than segmentation scoring. For example, the algorithm using both constraints produced lexicons in which 70% of the entries were correct from segmentations in which 91% of the word boundaries were correct.

General discussion

The ultimate questions raised by the work presented here are:

1. Do infants in fact rely on phonotactics at the onset of lexical acquisition? If so, which of the following constraint types do they rely on?
   (a) The universal requirement that every syllable must have a nucleus
   (b) Language-specific constraints on the consonant clusters that can occur at word boundaries
   (c) A universal sonority constraint
2. Do infants in fact rely on some kind of distributional analysis—i.e., frequency and context variability? If so, what form does it take? How is the information it provides integrated with any phonotactic constraints that may be used?
3. What other linguistic regularities, either innate or acquired, play a central role at the onset of lexical acquisition?

We begin the discussion by considering each of these questions in light of the results presented above. Next, we discuss the experimental system in the context of uncertainties about the input to infants' lexical acquisition system, the memory capacity of that system, and its output. Finally, we discuss our plans for modifying the experimental system so that it can be used to model memory limitations.

Phonotactics

The results presented here demonstrate that the syllabic nucleus constraint and the English constraints on boundary clusters can improve early lexical acquisition substantially. Combined with a distributional analysis, the syllabic nucleus constraint alone can improve performance substantially. Adding English-specific constraints on boundary clusters yields another improvement of roughly equal magnitude. Adding a constraint that forces word boundaries in sequences of consonants that would otherwise violate the sonority constraint. However, it should not be concluded that the sonority constraint is useless, only that we have not found a way to make use of it.

These results motivate a theory in which infants use the syllabic nucleus and boundary cluster constraints at the onset of lexical acquisition. Further, there is substantial evidence that infants know many phonotactic regularities of their language by 7 to 8 months (Friederici and Wessels, 1993; Jusczyk, et al., 1993a; Jusczyk, et al., 1993b). They may acquire the word boundary constraints by assuming that all and only the consonant clusters occurring at *utterance* boundaries are permitted at *word* boundaries. Taken together, this evidence establishes means and motive for the theory that both constraints play a central role at the onset of lexical acquisition. All that remains is to find a smoking gun. This might be accomplished by manipulating the phonotactic acceptability of targets in infant segmentation and retention tests like those used in (Hohne et al., 1994), (Jusczyk and Aslin, in press), or (Morgan and Saffran, in press).

Distributional analysis

The results of Experiment 1 showed that distributional analysis yields substantially better accuracy in lexical acquisition than a random baseline, even though the random algorithm uses valuable knowledge about the number of words in the input that is not available to either the distributional algorithm or the child. However, the random algorithm is not a plausible alternative to distributional analysis. Indeed, it is difficult to think of any plausible alternative. Those that have been proposed include: using isolated words; detecting some explicit acoustic marking of word boundaries; and using regularities in the pattern of full and reduced vowels, dubbed the metrical segmentation strategy. The isolated words strategy is not plausible for the reasons outlined in the introduction. Infants may exploit acoustic cues for lexical acquisition (Christophe et al., 1994), but the evidence concerning this is limited and indirect. Even if they do, such cues are likely to provide only some of the word boundaries, in which case distributional analysis would be needed to take up the slack. The metrical segmentation proposal is considered in the next section.

An alternative approach: metrical segmentation

What other linguistic regularities play a central role at the onset of lexical acquisition? We focus here on the Metrical Segmentation Strategy (MSS), which was originally proposed for adult lexical access (Cutler and Butterfield, 1992, Cutler and Carter, 1987, Cutler and Norris, 1988), but has recently been presented as a partial explanation of early lexical acquisition (Cutler, 1994b, Cutler, 1994a).

The MSS is based on the assumption that the mental lexicon is divided into two parts, one for function words and one for content words (Cutler and Carter, 1987). The idea is to take advantage of the fact that in English reduced vowels are more likely to be the first vowel of a function word than the first vowel of a content word, whereas the opposite is true of full (unreduced) vowels. If mature speakers could segment consonant clusters properly, then this regularity would suggest which lexicon to probe first with each syllable. This might well reduce the ambiguity that is inherent in the process of inferring word boundaries by lexical lookup. However, recent papers sometimes present the MSS as though utterances consisted entirely of content words, so that having probabilistic cues to the beginnings of content words would help solve the segmentation problem *directly*, rather than by suggesting which lexicon to access. Direct segmentation appears to be the basis for the claim that the MSS would be useful for *pre-lexical* segmentation. However, utterances do not consist entirely, or even primarily of content words. Cutler and Carter's figures suggest that 59% of all tokens in natural speech are function words, and 83% of all tokens are monosyllabic. On average, 82% of the tokens in our samples of child directed speech are monosyllabic. Thus, if the goal were to segment the input into words *without the benefit of a lexicon*, a strategy of segmenting at *every* syllable would compare favorably to the MSS. In summary, lexical access and pre-lexical segmentation are very different problems: while the MSS may be useful for the former, we know of no evidence that it is useful for the latter.

Another potential problem with the MSS is that it locates word boundaries somewhere within a consonant cluster, but does not specify where. For example, the MSS would insert a word boundary somewhere between the reduced vowel /ɪ/ and the full vowel /e/ in /mɪstek/ (*mistake*). But should the boundary be as in /mɪ stek/, /mɪs tek/, or /mɪst ek/? Cutler and Carter assume that allophonic cues provide syllabification, along the lines suggested by Church (1987). However, as we pointed out in the discussion of Church's proposals, allophonic cues frequently indicate syllabifications that cross word boundaries.

Finally, the proposal that the MSS is used for pre-lexical segmentation begs the question of how infants who do not yet know any words could learn the MSS, which is far from universal. To learn the MSS, infants would have to observe the vowel quality patterns of individual content words, then generalize those patterns. But infants do not know any words at the onset of lexical acquisition, much less which are content words and which function words. Since the MSS is based on a probabilistic distinction, learners would need to consider many words to ensure a statistically reliable sample. In an attempt to investigate the plausibility of the MSS for early lexical acquisition, Juscyzk, Cutler, and Redanz (1993) showed that American 9-month-olds prefer to listen to bisyllabic utterances with the full-reduced pattern over those with the reduced-full pattern (Jusczyk, Cutler, and Redanz, 1993). While this experiment is intriguing, it does not necessarily follow that infants are associating the full-reduced pattern with words per se, much less using it for segmentation. Infants might simply prefer *utterances* that begin in full vowels, completely independent of lexical segmentation. A more direct test would be to measure infants' ability to recognize full-reduced words as opposed to reduced-full words in sentential context, using the paradigm of Jusczyk and Aslin (in press).

In summary, there appears to be no evidence that the MSS would be more useful than assuming all words are monosyllabic, the MSS depends on a solution to the problem of segmenting consonant clusters, and it is not clear how infants could learn the MSS. Thus, the available evidence does not support a significant role for the MSS in pre-lexical segmentation.

Comparison of the experimental and natural systems

In order to place the current results in context, it is valuable to compare the experimental system to what is known about the natural one.

Input

The experiments described here were done on input represented as sequences of phonetic segments. However, it uncertain how infants represent speech sounds.

Recent experiments have shown that 2 to 3-month-olds can categorize bisyllabic utterances on the basis of a shared, initial, stressed CV syllable, and can retain the common syllable over a two-minute period of silent visual distraction (Jusczyk, Jusczyk, Kennedy, Schomberg, and Koenig, 1994). The same experiments fail to find categorization on the basis of common phonetic features, phonetic segments, or heterosyllabic VC sequences. These results are reinforced by the failure of earlier studies to find featural or segmental categorization of CV monosyllables (Bertoncini et al., 1988; Jusczyk and Derrah, 1987; Jusczyk and Kemler Nelson, 1994). This pattern of results has been taken to suggest that infants represent speech in syllabic units rather than phonetic units. As Jusczyk et al. point out, however, representation at multiple scales is certainly possible.

Suppose it turns out that 2 to 3 month olds do not analyze CV syllables into segments. It would seem to follow that they must acquire that ability before the onset of lexical acquisition, since syllables often cross word boundaries. Otherwise, /gɪ.və/ (*give a*) would consist of two atomic CV units, so the lexical entry would have to contain either a word fragment or a string of two words that do not form a syntactic or semantic compositional unit.

Even it turns out that infants can somehow begin to acquire words without the ability to analyze CV syllables, this would not change our analysis in any fundamental way. Whatever the size of infants' atomic units of speech representation, most utterances are likely to consist of a sequence of such units. Learners, then, must determine which subsequences should be entered into the lexicon. From the perspective of distributional analysis, this problem is formally identical to the one studied above—the same principles of analysis can be applied. Phonotactic constraints could be adapted as well, although the details would depend on whether all syllables are atomic or only the simple ones. For example, phonotactic constraints could be represented as information about whether each syllable can occur word initially, word internally, and word finally.

Processing: short term and long term memory

A second point of comparison between our system and the infants' lexical acquisition concerns the use of memory. Our programs have perfect memory for the input, while infants presumably do not, although even that is uncertain. As Christophe et al. (1994) put it, "...for all we know, infants might very well function like tape recorders."

If infants do not function like tape recorders, however, then our simulations cannot be interpreted as detailed models of mental processing. Such models would have to process relatively short stretches of the input, store any new items discovered in the lexicon, and then discard the input. In the formalization presented above, no provisions were made for revising a lexicon *formed on the basis of inputs that are no longer available.* On the other hand, questions of memory use have no impact on the interpretation of our experiments as measures of cue validity.

Output: what's in a lexicon?

A third uncertainty about the natural process of lexical acquisition is what it produces. The notion of *word* seems natural, but attempts to define it rapidly run afoul of subtle distinctions among productive, semi-productive, and fossilized morphology, inflectional, derivational, and compounding morphology, and so on. Is *another* a lexical item, or is it formed out of the items *an* and *other*, or even out of *a* and *nother*? For many Americans, *a whole nother* is perfectly grammatical in colloquial speech. Even if a precise, theoretically motivated definition of *word* could be formulated, it would not follow that words are the only items in the mental lexicon. The mental lexicon is not merely a repository of abstract knowledge but a working computational device under functional pressure to respond very quickly. It would make sense for such a device to store common phrases as well as individual words.

In summary, there is much yet to learn about the input to lexical acquisition, the memory capacities of infants, and the contents their mental lexica. The results of further studies may well require changes to the input representation and scoring of the experimental system described here, but these changes are likely to be in the details rather than the overall conception. As for memory usage, a conceptual shift may be needed to develop a system that is able to model memory limitations. We outline such a system in the next section.

Future Work

We plan to modify the evaluation algorithms presented here so that they segment relatively short stretches of input, add any newly discovered words to the lexicon, and discard each input stretch before segmenting the next. The length of the input stretches will be adjustable to accommodate various memory models. Alternative segmentations of each input sample will be evaluated by based on length of the sequence of lexical indices in their derivations plus the length of any *new* lexical entries they posit. Lexical items that were used in the segmentation of previous input samples will not be charged against the segmentations of the current sample. As a result, hypotheses that create as few new lexical entries as possible will be preferred, all other things being equal.

The original distributional learning system is based on the idea that the more frequently a sound sequence occurs in the input, and the more varied the contexts in which it occurs, the more it deserves a lexical entry. In the modified algorithm, however, decisions about which sound sequences get lexical entries must be made on the basis of short samples of input, and most words will not occur frequently in short samples. Thus, many sequences will have to be entered in the lexicon on the basis of a single occurrence, and many of these lexical entries will turn out to be chance cooccurrences, like the *igdo* subsequence of *bigdog.* Such mistakes will be especially prevalent early in the lexical acquisition process, when there is little lexical knowledge to go on. This relatively unconstrained process should eventually discover a lot of real words, in addition to the chance cooccurrences. Later in the process, the real words will tend to recur in new utterances, so their presence in the lexicon should improve segmentation. The chance sequences should occur rarely and hence have little negative impact on segmentation. Nonetheless, the lexicon will probably be filled with the low-frequency detritus of the early, error-prone segmentations. What's needed is forgetfulness—some way to purge the lexicon of excess low-frequency items (Siskind, this volume).

One solution would be to impose an arbitrary limit on how long a word can remain in the lexicon without occurring in the input. Such a strict lower limit on frequency would certainly be ad hoc. Worse, it would prevent the learner from ever acquiring very low frequency words. People appear able to acquire words of arbitrarily low frequency, based on the well-known fact that no matter how large a corpus one examines, there are always a great many words of frequency one. A better solution is to fit the frequency *distribution* of the lexical items to the universal lexical frequency distribution proposed by Zipf (1949). If low-frequency detritus is accumulating, it should show up as a long tail in the frequency distribution—longer than that predicted by the Zipf distribution. After each input sample is processed, the learner could purge from the lexicon some of the items that have not occurred for the longest time. The number of items purged could chosen to optimize the fit between the frequency distribution in the learner's lexicon and the universal lexical frequency distribution.[10]

[10] We thank Jacques Mehler for suggesting that the lexical frequency distribution might useful.

Conclusions

The results and arguments presented here suggest that distributional analysis and phonotactics provide the best available explanation of how infants begin to acquire a lexicon. Semantics probably makes a major contribution as well. Other factors may contribute, but their usefulness remains to be demonstrated.

Although many questions about the onset of lexical acquisition remain unanswered, the future appears bright. Currently available techniques are likely to continue providing information about infants' mental representation of speech sounds and the development of their segmentation abilities. Extensions of the work presented here are likely to provide insights into how effective distributional analysis and phonotactic constraint are in other languages, and how sensitive they are to variations in input representation and short term memory. This area of research presents an ideal opportunity for computational and human-subjects research to converge on important psychological questions.

Details of the Formalization

This formalization of the distributional regularity is based on the minimum representation length principle, also known as the Minimum Description Length (MDL) principle. Given a system of representing the hypotheses in some hypothesis space, the MDL says that the shortest one is most likely to be true. Counting characters in representations like those in Figure 2 is a good way to gain intuition about the MDL principle, but the essence of the principle is to choose the hypothesis that stipulates the least *information.* Many details of the representation scheme illustrated in Figure 2 are arbitrary, in the sense that they reflect conventions of rather than information intrinsic to hypotheses. When the lexicon has more than ten entries, for example, some of those entries will have two-digit indices and some will have one-digit indices. Given a hypothesis, the number of characters in the derivation portion of its representation can be increased by assigning the two-digit indices to common words, or decreased by assigning them to rare words. In addition, the number of words that must be assigned multi-character indices in a given hypothesis is affected by the arbitrary choice of using only the ten digits for indices. Since the indices need only be unique identifiers, there is no reason not to use the phonetic symbols in indices too.

In order to eliminate these arbitrary factors, the programs in the experiments evaluate a hypothesis by counting the number of bits (ones and zeros) needed to represent it as a string of ones and zeros, with no spaces, carriage returns, or other delimiters. This binary string contains all the essential information in the representation of Figure 2, in the sense that a computer program could recon-

$$\boxed{blksz}\ \ \boxed{len(w_1)}\dots\ \boxed{len(w_n)}\ \ \boxed{w_1}\ \dots\boxed{w_n} \qquad \text{(a)}$$

$$\boxed{blksz}\ \ \boxed{len([w_1])}\dots\boxed{len([w_n])}\ \ \boxed{[w_1]}\dots\boxed{[w_n]} \qquad \text{(b)}$$

Figure 3: Schematic diagrams for components of the representation

struct the table from the binary string, and vice versa. By eliminating the spaces and arbitrary distinctions between letters and numbers, and by always assigning the shortest indices to the most common words, the binary representation more closely approximates the information intrinsic to the hypothesis.

The representation presented below has two major components, a lexicon and a derivation. We focus first on the lexicon, which itself has two major components, a sequence of lexical items, called the *word inventory*, and a sequence of corresponding indices. The indices are hereafter called *code words*, and the sequence of code words in the lexicon is called the *code-word inventory*, to distinguish it from the sequence of code words in the derivation. The first code word in the code-word inventory is the index for the first lexical item in the word inventory, and so on.

Much of the complexity in representing the sequence of lexical items and code words stems from fact that binary strings contain no spaces or other delimiters. Nonetheless, the information about the location of boundaries between words and code words must be represented somehow. In our treatment, this is done by a sequences of binary integers specifying the length of each lexical item in phonetic segments and the length of each code word in bits. With this length information, the boundaries between items can be reconstructed. The length integers themselves are not separated by delimiters, so their lengths must be specified too. Rather than specify the length of each such integer separately, however, we pad them with leading zeros so that they are all the same length. That length is represented as a unary string ones of terminated by a zero. Figure 3 shows a schematic of the word inventory in the first row, and the code word inventory in the second, each preceded by a corresponding length sequence and by the unary sequence specifying the lengths of the length integers. $w_1 \dots w_n$ stands for the sequence of words in the word inventory, and $[w_1] \dots [w_n]$ stands for the sequence of code words. n is the number of words in the lexicon. It is worthwhile to refer to this schematic throughout the following discussion.

In the word inventory (Figure 3a), the list of lexical items is represented as a continuous string of (binary codes for) phonetic segments, without separators between words. Each phonetic segment is represented by a string of p bits, where

p is the logarithm base 2 of the number of phonemes in the alphabet.[11] If $len(w_i)$ is the number of phonemes in word w_i, then the length of its representation in the lexicon is $p \cdot len(w_i)$. Summing over all words in the lexicon yields:

$$\sum_{i=1}^{n} (p \cdot len(w_i)) = p \sum_{i=1}^{n} len(w_i) \tag{1}$$

Since this representation has no spaces, the information about where each word starts is represented by specifying the number of phonetic segments in each word. These lengths are represented as binary integers between one and the length of the longest word. These integers are padded with leading zeros to make them all the same length as the longest one. Thus, the length of each length integer is:

$$\log_2(\max_{1 \ldots n} \; len(w_i))$$

There are n such integers, one for each word, yielding a total length of

$$n \log_2(\max_{1 \ldots n} \; len(w_i))$$

The length of the length integers is represented once, as a unary string of ones, terminated by a zero. Adding the length of this unary string yields:

$$1 + (n + 1) \log_2(\max_{1 \ldots n} \; len(w_i)) \tag{2}$$

The total length of the word inventory representation is the sum of (1) and (2).

The representation of the code word inventory (Figure 3b) is nearly identical to that of word inventory—a sequence of binary code words preceded by a sequence of integers specifying the length (in bits) of each code word. The specific code words are kept as short as possible, subject to the following constraints: Each code word must be unique; no code word may be a prefix of another code word, so that once the the complete code-word inventory has been deciphered, there is no ambiguity about where each code word starts in the representation of the derivation; and the shortest possible code words must be assigned to every word, with priority going to the high frequency words. A basic theorem of information theory says that these constraints can be satisfied by assigning each word w a code word whose length, $len([w])$, is given by:

$$len([w]) = \left\lceil \log_2 \frac{\sum_{i=1}^{n} f(w)}{f(w)} \right\rceil = \left\lceil \log_2 \frac{m}{f(w)} \right\rceil$$

[11] In an actual binary string there is no such thing as a fraction of a bit, so all length calculations must be rounded up to the nearest bit. Since we are only using the representation as a means to measure information, we do not round up. This yields a smooth evaluation function whose ranking of hypotheses is independent of the radix (base) of the representation.

where $f(w)$ is the frequency of the word w in the derivation portion of the explanation, and m is the total frequency of all words (see, e.g. Rissanen, 1989, or Li and Vitányi, 1993). [check publisher]. Dropping the ceiling (round-up) notation and summing over all words yields:

$$\sum_{i=1}^{n} len([w]) = \sum_{i=1}^{n} \log_2 \frac{m}{f(w_i)} \tag{3}$$

As in the word inventory, the length of each code word is represented by a fixed-length binary string. According to the formula for the lengths of code words, the longest codes word are assigned to words of frequency one, and these code words $\log_2(m)$ bit long. The length integers for code words are therefore between one and $\log_2(m)$, so they can each be represented in $\log_2(\log_2(m))$ bits. (No word has frequency zero, so we actually represent the lengths minus one.) The total number of bits needed to represent the lengths of all n code words is therefore $n\log_2(\log_2(m))$. Finally, the number of digits used to represent each length is itself represented as a unary string of $\log_2(\log_2(m))$ ones followed by a zero, yielding a total of

$$1 + (n+1)\log_2(\log_2 m) \tag{4}$$

bits of information about code word lengths. The total length of the code word inventory is the sum of the terms in (3) and (4).

The derivation portion of the hypothesis is represented by a sequence code words. As mentioned above, the set of code words is selected so that no code word begins with a sequence that is also a code word. As a result, there is no ambiguity about where each code word ends, once the complete set of code words is known. Thus, there is no need to provide length information for each code word in the derivation.

If a code word $[w_i]$ occurs $f(w_i)$ times in the derivation, the number of bits contributed by those occurrences is $f(w_i)len([w_i])$. Summing over all words yields

$$\sum_{i=1}^{n} f(w_i) \cdot len([w]) = \sum_{i=1}^{n} \left[f(w_i) \cdot \log_2 \left(\frac{m}{f(w_i)} \right) \right] \tag{5}$$

Finally, note that the binary representations of the word inventory, code word inventory, and derivation are all strung together into one long binary string. The information about where each component begins and ends is represented by specifying n, the number of words in the lexicon, and m, the number of code words in the derivation. Integers can be represented in a self delimiting fashion, using a unary based scheme similar to one described above, in approximately $\ell^{(2)}(x)$ bits, where

$$\ell^{(2)}(x) = 1.5 + \log_2(x+1) + 2\log_2(\log_2(x+2) + 0.5)$$

(Rissanen, 1989; Li and Vitányi, 1993). The total length of the representation of the entire hypothesis is the sum of $\ell^{(2)}(n)$, $\ell^{(2)}(m)$, and the terms in (1), (2), (3), (4), and (5).

References

Aslin, R. N., Woodward, J. C., LaMendola, N. P., and Bever, T. G. (1994). Models of word segmentation in fluent maternal speech to infants. In (Morgan and Demuth, 1994). In Press.

Bernstein-Ratner, N. (1987). The phonology of parent child speech. In Nelson, K. and van Kleeck, A., editors, *Children's Language: Vol. 6*. Erlbaum, Hillsdale, NJ.

Bertoncini, J., Bijeljac-Babic, R., Juszcyk, 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.

Borowsky, T. (1986). *Topics in the Lexical Phonology of ENglish*. PhD thesis, University of Massachusetts.

Brent, M. R. (1993). Minimal generative models: A middle ground between neurons and triggers. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, pages 28–36, Hillsdale, NJ. Erlbaum.

Brent, M. R., Gafos, A., and Cartwright, T. A. (1994). Phonotactics and the lexicon: Beyond bootstrapping. In Clark, E., editor, *Proceedings of the 1994 Stanford Child Language Research Forum*. CLSI Press, Stanford, CA.

Cartwright, T. A. and Brent, M. R. (1994). Segmenting speech without a lexicon: The roles of phonotactics and speech source. In *Proceedings of the First Meeting of the ACL Special Interest Group in Computational Phonology*, pages 83–90. Association for Computational Linguistics.

Chierchia, G. (1983). Length, syllabification, and the phonological cycle in Italian. Unpublished ms. Brown University, Providence, RI.

Christophe, A., Dupoux, E., Bertoncini, J., and Mehler, J. (1994). Do infants perceive word boundaries? an empirical study of the bootstrapping of lexical acquisition. *Journal of the Acoustical Society of America*, 95:3.

Church, K. (1987). Phonological parsing and lexical retrieval. *Cognition*, 25:53–69.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In Kingston, J. and Beckman, M. E., editors, *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*. Cambridge University Press, Cambridge.

Cutler, A. (1994a). Prosody and the word boundary problem. In (Morgan and Demuth, 1994). In Press.

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

Cutler, A. and Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperceptions. *Journal of Memory and Language*, 31:218–236.

Cutler, A. and Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2:133–142.

Cutler, A. and Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology*, 14:113–121.

Ellison, T. M. (1992). Learning vowel harmony. In Daelmans, W. and Powers, D., editors, *Background and Experiments in Machine Learning of Natural Language: Proceedings of the 1st* SHOE *Workshop*. Institute for Language Technology and Artificial Intelligence, Katholieke Universiteit, Brabant, Holland.

Fisher, C., Gleitman, H., and Gleitman, L. R. (1991). On the semantic content of subcategorization frames. *Journal of Cognitive Psychology*, 23(3):331–392.

Fisher, C. and Tokura, H. (1994). Prosody in speech to infants: Direct and indirect acoustic cues to syntactic structure. In (Morgan and Demuth, 1994). In Press.

Friederici, A. D. and Wessels, J. M. I. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception and Psychophysics*, 54:287–295.

Harrington, J., Watson, G., and Cooper, M. (1989). Word boundary detection in broad class and phoneme strings. *Computer Speech and Language*, 3:367–382.

Harris, Z. S. (1951). *Methods in Structural Linguistics.* University of Chicago Press, Chicago.

Hayes, J. R. and Clark, H. H. (1970). Experiments in the segmentation of an artificial speech analog. In Hayes, J. R., editor, *Cognition and the Development of Language*, pages 221–234. Wiley, New York.

Hohne, E. A., Jusczyk, A. M., and Redanz, N. J. (1994). Do infants remember words from stories? Poster presented at the 127th meeting of the Acoustical Society of America, June 1994, Boston, MA.

Hooper, J. B. (1976). *An Introduction to Natural Generative Phonology.* Academic Press.

Jusczyk, P. W. (in press). Language acquisition: Speech sounds and the beginnings of phonology. In Miller, J. L. and Eimas, P. D., editors, *Speech, Language, and Communication*, volume 11 of *Handbook of Perception and Cognition.* Academic, Orlando, FL.

Jusczyk, P. W. and Aslin, R. N. (in press). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology.*

Jusczyk, P. W., Cutler, A., and Redanz, N. J. (1993a). Infants' preference for the predominant stress patterns of English words. *Child Development*, 64:675–687.

Jusczyk, P. W. and Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology*, 23:648–654.

Jusczyk, P. W., Friederici, A. D., Wessels, J., Svenkerud, V. Y., and Jusczyk, A. M. (1993b). Infants' sensitivity to the sound patters of native language words. *Journal of Memory and Language*, 32:402–420.

Jusczyk, P. W., Jusczyk, A. M., Kennedy, L. J., Schomberg, T., and Koenig, N. (1994a). Young infants' retention of information about bisylabic utterances. Manuscript, Dept. of Psychology, State University of New York at Buffalo, Buffalo, NY 14260-4110, USA.

Jusczyk, P. W. and Kemler-Nelson, D. G. (1994). Syntactic units, prosody, and psychological reality during infancy. In (Morgan and Demuth, 1994). In Press.

Jusczyk, P. W., Kennedy, L. J., and Jusczyk, A. M. (in press). Young infants' retention of information about syllables. *Infant Behavior and Development.*

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

Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignment. *Psychological Review*, 99(2):349–364.

Li, M. and Vitányi, P. M. B. (1992). Inductive reasoning and kolmogorov complexity. *Journal of Computer and System Sciences*, 44:342–384.

Li, M. and Vitányi, P. M. B. (1993). *An Introduction to Kolmogorov Complexity and Its Applications.* Springer Verlag, New York.

MacWhinney, B. and Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, 12:271–296.

Morgan, J. L. and Demuth, K., editors (1994). *Prosody and the Word Boundary Problem.* Erlbaum, Hillsdale, NJ. In Press.

Morgan, J. L., Meier, R. P., and 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. and Saffran, J. R. (in press). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Development.*

Olivier, D. C. (1968). *Stochastic Grammars and Language Acquisition Mechanisms.* PhD thesis, Harvard, Cambridge, MA, USA.

Oviatt, S. L. (1980). The emerging ability to comprehend language: An experimental approach. *Child Development*, 51:97–106.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* World Scientific Publishing, Teaneck, NJ.

Thomas, D. G., Campos, J. J., Shucard, D. W., Ramsay, D. S., and Shucard, J. (1981). Semantic comprehension in infancy: A signal detection analysis. *Child Development*, 52:798–803.

Valian, V. and Coulson, S. (1988). Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language*, 27:71–86.

Wolff, J. G. (1988). Learning syntax and meanings through optimization and distributional analysis. In Levy, Y., Schlesinger, I. M., and Braine, M. D. S., editors, *Categories and Processes in Language Acquisition*, pages 179–215. Erlbaum, Hillsdale, NJ.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort.* Addison-Wesley, New York, NY.