Available online at www.sciencedirect.com
SCIENCE DIRECT®
Cognitive Psychology 50 (2005) 86–132
Cognitive Psychology
www.elsevier.com/locate/cogpsych

Statistical clustering and the contents of the infant vocabulary

Daniel Swingley*

Max-Planck-Institute for Psycholinguistics, Nijmegen, The Netherlands

Accepted 30 June 2004
Available online 26 August 2004

Abstract

Infants parse speech into word-sized units according to biases that develop in the first year. One bias, present before the age of 7 months, is to cluster syllables that tend to co-occur. The present computational research demonstrates that this statistical clustering bias could lead to the extraction of speech sequences that are actual words, rather than missegmentations. In English and Dutch, these word-forms exhibit the strong–weak *(trochaic)* pattern that guides lexical segmentation after 8 months, suggesting that the trochaic parsing bias is learned as a generalization from statistically extracted bisyllables, and not via attention to short utterances or to high-frequency bisyllables. Extracted word-forms come from various syntactic classes, and exhibit distributional characteristics enabling rudimentary sorting of words into syntactic categories. The results highlight the importance of infants' first year in language learning: though they may know the meanings of very few words, infants are well on their way to building a vocabulary.

1. Introduction

From the day they are born, infants develop in the presence of language. For the first several months of life, infants hearing speech probably understand little

* Present address: Department of Psychology, University of Pennsylvania, 3401 Walnut Street 302, Philadelphia, PA 19104, USA. Fax: 1 215 746 6848.
E-mail address: swingley@psych.upenn.edu.

0010-0285/$ - see front matter
doi:10.1016/j.cogpsych.2004.06.001

more than the emotions in their parents' expressive intonation contours (e.g. Fernald, 1989). Yet these first months set the stage for later language learning in some surprising ways. Infants begin to learn the phonetic categories of the language around them, revealing effects of the ambient language on speech sound categorization as early as six months of age (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Polka & Werker, 1994). Infants also learn about the frequencies with which sequences of speech sounds occur, distinguishing syllables containing rare elements (e.g., *chedge*) from syllables containing more frequent elements (e.g., *kern*; Jusczyk, Luce, & Charles-Luce, 1994). These results demonstrate that even young infants spontaneously perform statistical analyses of distributions of sounds they hear, thereby learning categories that form the basis of the child's phonology.

The present research examines the possibility that infants also learn the forms of words by using statistical grouping mechanisms. Thus, in addition to learning the general sound structure of the ambient language, infants may learn actual words, or at least words' phonetic forms, through the operation of memory mechanisms that are sensitive to statistical properties such as the frequency with which sound sequences co-occur. If so, eight-month-olds still learning to babble in syllables may already possess a rudimentary lexicon, consisting of words for which meanings will be learned over the next year and a half. Alternatively, the speech that infants store in this period may not be of much use: If the nature of infants' parsing and encoding mechanism does not match the linguistic structure of infant-directed speech, infants may command only a potpourri of half-words and bogus portmanteaux. The primary goal of the current paper is to use computational analyses to estimate the contents of infants' early lexical knowledge: Do the forms that infants store as familiar sequences correspond to actual words?

A second goal is to resolve a puzzling chicken-and-egg problem in early speech processing. Infants have been shown to chunk speech sequences according to the dominant prosodic pattern of words in the language they hear; from about 8 months, prosodic form governs infants' segmentation of speech. Yet while infants use prosody to find words, they must first find words to determine the dominant prosodic form. We will show some proposals intended to account for this problem to be inadequate, and will suggest an alternative account of this developmental process.

2. Word discovery by infants

The idea that words may be identified via distributional analysis is not new (e.g. Brent & Cartwright, 1996; Harris, 1955; Hayes & Clark, 1970; Pinker, 1984) but infants' discovery of words has received less attention than infants' discovery of phonetic categories. It is clear that infants do remember at least some words in the speech they hear; for example, Jusczyk and Hohne (1997) found that 8-month-olds who had been read to at home from storybooks recognized frequent words from those stories when tested several days later in an auditory preference task.

Various factors are known to influence the likelihood that infants will extract a given sequence of speech as a unit. First, infants are reluctant to group together syllables that straddle certain phonetic cues to word boundaries. For example, in one study, 10-month-olds recognized the word *beacon* when presented in the sentence *The very large beacon fused to the ice*, but not when presented in *The very large bee confused all the mice*, evidently because infants detected a phonetic boundary between /bi/ and /kən/ in the latter utterance (Gout, Christophe, Millotte, & Morgan, under review; see also Soderstrom, Seidl, Kemler Nelson, & Jusczyk, 2003). These studies indicate that infants sometimes use phrasal prosody to guide their grouping of parts of utterances into smaller units to be stored in memory (see also Gerken, Jusczyk, & Mandel, 1994; Nazzi, Kemler Nelson, Jusczyk, & Jusczyk, 2000).

A second factor influencing infants' extraction of speech sequences is lexical stress. By 8 or 9 months, infants in English and Dutch language environments tend to group together syllable pairs having a strong–weak (trochaic) stress pattern, and tend not to group syllable pairs with the reverse (iambic) pattern (Echols, Crowhurst, & Childers, 1997; Houston, Jusczyk, Kuijpers, Coolen, & Cutler, 2000; Jusczyk, Houston, & Newsome, 1999; Morgan, 1996b; Thiessen & Saffran, 2003). It has been suggested that this tendency to cluster trochaic sequences is useful for learners of English and Dutch, because the vocabularies of these languages contain far more trochaic than iambic words, whether one considers the adult lexicon (Cutler & Carter, 1987; Vroomen & de Gelder, 1995) or speech to infants (Kelly & Martin, 1994; Morgan, 1996a; van de Weijer, 1998). The reverse clustering pattern has been found for infants learning Canadian French, in which words are usually realized with an iambic pattern (Polka & Sundara, 2003). These results may be viewed as evidence for prosodic *templates* to which children attempt to fit strings of speech, a notion that can be traced to work on children's expressive phonology (e.g. Allen & Hawkins, 1980; Vihman & Velleman, 2000). The results may also be seen as demonstrating precocious application of the Metrical Segmentation Strategy, a parsing bias shown in English- and Dutch-speaking adults whereby strong syllables are interpreted as word onsets (Cutler & Butterfield, 1992; Cutler & Norris, 1988). The fact that infants' clustering tendencies match the ambient language's dominant bisyllable stress pattern indicates that infants are able to discover language-specific prosodic tendencies and use these tendencies to guide lexical parsing.

Third, infants cluster syllables that tend to co-occur. If syllables A and B usually occur together as AB, infants are more likely to consider the sequence AB as a cohesive unit than if each syllable occurs frequently in several environments (Goodsitt, Morgan, & Kuhl, 1993). Infants' clustering of syllables having high co-occurrence probabilities was shown in studies by Aslin, Saffran, and Newport (1998) and Saffran, Aslin, and Newport (1996). Eight-month-olds were familiarized to a 3-min auditory stimulus consisting of a quasirandom ordering of four trisyllabic nonce words strung together to produce a continuous sequence lacking acoustic cues to word boundaries. The only information regarding which syllables should be grouped into words was embedded in the statistical structure of the input stream: the second

syllable of each word was always preceded by the first syllable and followed by the third, and the first and third always flanked the second. Because the order of words in the stream varied, these features of words were absent in trisyllabic sequences not defined as words. For example, the word *pabiku* was followed by three other words; consequently, [ku] was followed by three different syllables, and the transitional probability of any of these syllables following [ku] was therefore lower than the transitional probability of, for example, [ku] given [bi] ($p = 1.0$). Showing that infants extracted these words would provide evidence that infants computed some kind of conditional probability statistic over the syllables.

To demonstrate this extraction process, Saffran et al. used the familiarization preference procedure developed by Jusczyk and Aslin (1995). After familiarization with the continuous stream (pabikutibudodaropi...) for 3 min, a test phase began in which infants' gaze to a flashing light to their right or left triggered presentation of one of two kinds of stimulus: *words* (isolated nonce words) and *part-words* (isolated trisyllables containing two syllables from one nonce word, in the familiar order, plus a third syllable from a different word). Infants' significantly longer gaze (''preference'') on part-word trials showed that infants distinguished the two kinds of stimuli. This result was found in both Saffran et al. (1996) and Aslin et al. (1998). In the former study, the discrimination may have been based on the greater frequency of the words relative to the part-words in the familiarization phase; the latter study controlled for the frequency of the test items across conditions and nevertheless still found that infants listened more to the part-words than the familiar words. These studies showed that infants cluster together statistically associated syllables.

Infants' computation of conditional probability statistics accords with a long history of experimental learning research showing that various organisms, including humans, compute probabilistic contingencies. One classic example is Rescorla's demonstration that rats linked tones with shocks (and as a result made fewer bar-presses for food during tones) only to the extent that tones were *predictive* of shocks, and not in proportion to the absolute number of tone/shock pairs the rats had experienced (Rescorla, 1968). Interpretation of sequential events in terms of contingencies rather than simple co-occurrence frequencies is a basic and general property of cognition. Thus, as Aslin et al. (1998) point out, what is surprising about recent demonstrations of infants' computation of conditional probabilities is not that it happens at all; the surprise is that infants learn a large number of contingencies at once, without extrinsic reinforcement. Infants' facility with such computations over artificial speech materials suggests that infants can perform similar computations over the speech they hear in normal communicative situations.

To summarize, infants (a) make use of some available acoustic cues to avoid conflating distinct words; (b) evidence a ''trochaic bias'' by about 8 months, if exposed to English or Dutch, and an ''iambic bias'' if exposed to Canadian French; and (c) cluster together syllables that tend to co-occur, i.e., syllables that are mutually predictive. The goal of the present research was to assess the linguistic consequences of the operation of the third of these mechanisms, for infants exposed to English or Dutch. Two primary hypotheses were evaluated: that the clusters infants would obtain using this mechanism would usually be words; and that the bisyllabic clusters

would exhibit the trochaic pattern to which infants are sensitive. If the first hypothesis is true, by the end of their first year infants may already possess a stock of word-forms which then form the basis of the early vocabulary. If the second hypothesis is true, it could account for the learning of the trochaic bias.

In principle, answering these questions is simple. One need only characterize the learning algorithm that gives rise to the statistical clustering results, and then apply this algorithm to a corpus of child-directed speech. The output could then be examined to determine the proportion of successfully extracted words (relative to nonwords) and the proportion of extracted trochaic sequences. In practice, however, this procedure necessitates taking an explicit stand on a number of unresolved and often controversial issues. The precise nature of the learning mechanism is underdetermined by the experimental evidence in its favor; the nature of infants' representation of speech sounds is not clear; and finally, because the large child-directed speech corpora that are presently available for analysis are in an orthographic (text) format, we must attempt to estimate the actual phonetic realization of those utterances. Clearly, one premise of this enterprise (and all similar modeling work) is that these estimates can be made in a way that does not distort the analyses in favor of (or against) the tested hypotheses. We have attempted to guard against this problem in a number of ways, as will be described later.

3. Three models of word discovery

To set the present work in context, we will briefly review some of the previous studies that have used corpora to evaluate models of infant word discovery. A more in-depth review of relevant computational models may be found in Batchelder (1997); work not discussed in detail here includes Aslin, Woodward, LaMendola, and Bever (1996), Olivier (1968), Wolff (1977), and de Marcken (1996), among others. We will begin with the DR model of Brent and Cartwright (1996). The core of this model is a method of data compression in which a corpus of phonemes is re-represented as a *lexicon* and a *derivation*. The lexicon is a list of postulated words, each with an index that uniquely identifies that word. The derivation is a list of indices, from which the corpus may be reconstructed by replacing each index with its corresponding word. For a given corpus, there are many possible lexicons and derivations; the task of the algorithm is to assess a large number of such analyses and choose the one that minimizes the number of characters required to re-represent the corpus. Intuitively, the ideal analysis is the one containing the longest and most frequent sequences as words. This is because long sequences reduce the size of the derivation (each index covers more of the corpus) and because storing frequent sequences reduces the size of the lexicon (one word occurring many times also covers more of the corpus). The statistical nature of language corpora is such that this minimization of the lexicon and the derivation will tend to produce a lexicon that contains many actual words of the language. Brent and Cartwright showed that a minimization algorithm of this sort, coupled with an additional search algorithm guiding the selection of analyses evaluated by the minimization algorithm, was quite successful in segmenting a corpus of child-directed speech into words.

However, there is as yet no evidence that infants consider multiple analyses of this sort and select the one that minimizes its description length, thereby bracketing the speech signal into words. Behavioral evidence supporting the DR model is limited, excepting a study of adults showing segmentation preferences comporting with the model's prediction that found words may be used to aid in the discovery of other words (Dahan & Brent, 1999). However, the applicability of this result to infant speech processing is unclear; furthermore, it is not certain that the entire computational apparatus of the DR model is required to account for these results. Thus, as Brent (1996) points out, the model demonstrates the utility of an abstract strategy (roughly, that of using words to find other words, starting from utterances), but the set of algorithms implementing this strategy is not intended to have strong a priori psychological plausibility.

Some other models of speech segmentation evaluate the learning of parsing strategies, rather than the learning of a lexicon per se. In these models, exposure to a training corpus provides probabilistic information about the likelihood of word boundaries in different environments. After training, the models are tested on new sentences, and evaluated according to how many boundaries were inserted in the correct locations. Some models of this type are connectionist simulations in which phonemes (encoded as vectors of binary phonological features) are serially provided as input. The task of the system is to predict the following phoneme, given the present and previous ones. Error in predictions tends to rise at word boundaries (e.g., Elman, 1990). Cairns, Shillcock, Chater, and Levy (1997) used such a model in their analysis of a corpus of English (adult-directed) speech. Their training corpus was derived from the London-Lund Corpus of English Conversation (LLC; Svartvik & Quirk, 1980). To create the input, a one-million-character section of orthographic text was converted to an estimated phonological transcription using a dictionary. This transcription was then passed through a set of "rewrite rules that introduced phonological alternations... such as assimilation and vowel reduction" (p. 124). Phonemes were then converted to binary feature vectors in a "Government Phonology" transcription. Finally, an unspecified amount of noise was introduced by randomly inverting some feature values "in order to encourage the network to rely on sequential information" (p. 124). Two such randomizations were done, resulting in a total corpus of two million characters.

This corpus served as the input to a Simple Recurrent Network (SRN; Elman, 1990). After training, the network was tested with an additional 10,000-character corpus (this time without any added noise). After each character, the prediction error was measured, producing 10,000 error values as output. Because no single error value is a priori the most plausible threshold above which a word boundary should be assumed, Cairns et al. computed hit:false-alarm ratios for all possible error thresholds, making an ROC curve. The curve showed that for all tested thresholds, the network identified more correct boundaries (hits) than incorrect boundaries (false alarms). The authors reported peak performance of 21% of boundaries correctly identified and a hits to false-alarms ratio of 1.5:1. Performance was better when pauses in the corpus were counted as hits on the grounds that gaps in the speech signal are unambiguous segmentation cues.

The network's success in finding word boundaries might be better described as success in finding syllable boundaries. In fact, as Cairns et al. pointed out, the number of word boundaries detected was not significantly different from the number that would be expected if the model had no knowledge of words beyond knowledge of syllable boundaries. Of course, when many of the words in the corpus are monosyllabic, as is the case with the LLC, detecting syllable boundaries via a phonotactic analysis will provide many word boundaries.

A similar approach to infant speech segmentation was tested by Christiansen, Allen, and Seidenberg (1998). Like Cairns et al., Christiansen et al. trained an SRN on a corpus that had been converted from orthographic text to a phonemic representation with each phoneme instantiated as a vector of binary feature values. Here, the corpus was Korman's (1984) corpus of infant-directed speech. Christiansen et al. explicitly encoded utterance boundaries as a sort of special phoneme (a *boundary unit*), and trained the network to predict phonemes. In testing, the boundary unit was probed to see if it would tend to be activated at word boundaries. This was indeed the case. The network became sensitive to phonotactic patterns that often appeared at the ends of utterances, and because utterance ends are also word ends, this sensitivity was of use for finding where words ended within utterances. In a second analysis, lexical stress was encoded in the training and testing sets, and predicting stress (as well as phonemes) was added to the training task. Information about stress improved the performance of the network significantly, even though in the corpus stress was not a particularly good cue to word boundaries.

The present research complements previous work evaluating the utility of various algorithms for the detection of words in speech. Here, perhaps to a greater extent than in previous modeling efforts, we have looked to experimental evidence to guide our selection of both the algorithm tested and the nature of the representations over which computations are completed, basing our estimates on what is known about infants of about 8 months of age: the stage in development at which it is clear that infants learn word-length sequences from natural speech. In addition, we have drawn upon somewhat different intuitions regarding the goals of the task. In the next section we describe and attempt to justify these decisions.

4. Infants' representation of speech sounds

Current statistical approaches to word-finding rely on the ability to count the frequency of occurrence of phonetic patterns, either explicitly (as in Brent's model) or implicitly (in the connectionist approaches). This counting requires that utterances be divided into smaller parts, and that at some level tokens be abstracted into types. Without the division into parts, the algorithm could only count utterances, and there would be no sequences to compute. Without the abstraction, each token would count as a new type (because no two naturally produced speech signals are identical) and all frequencies would be equal to one. Thus, one of the premises of research on this topic is that infants are capable of an a priori segmentation of speech into units

smaller than the utterance, and that infants classify different tokens of each of these units as repeated instances of a given type.

This premise need not be true in principle. For example, infants might store utterances as holistic soundscapes of some sort, and compute similarity relations between utterances or parts of utterances according to a metric that does not align closely with the similarity relations that arise from considering various tokens of a phoneme or syllable as equivalent. The fact that each token might count as a new type does not have to be fatal to the word learning process; lexical categories could emerge from matrices of similarity relations. (This view is analogous to the standard view of phoneme category formation, in which distributions of speech sounds are organized into categories. Computational models of this process assume that the infant is capable of isolating tokens of phones from the signal, to serve as input to the distributional analysis, but they do not assume that tokens are composed of categorical features; see e.g., Guenther & Gjaja, 1996.) However, in all current models designed to investigate this task, including that to be presented in this paper, utterances are assumed to be divisible into parts, and each of these parts is assumed to be assigned to a single category type (rather than a series of probabilities over category types, for example). Distributional analyses then proceed over these types.

Nearly all previous statistical analyses of infant word segmentation have used the phoneme as the unit over which infants are assumed to perceive speech and compute probabilities. In the research reviewed above, with the exception of the Cairns et al. work, phonemes were read directly from a dictionary pronunciation; for example, every /t/ in the corpus was assumed to be correctly identified as /t/, whether serving as an onset or coda. In the Cairns et al. study, rewrite rules and added noise altered the phonemic input to some degree, although it is not clear from their report how much. Thus, the standard assumption has been that infants compute statistics over phonological categories.

By contrast, in the present paper the syllable is used as the unit of analysis. This decision was motivated primarily by research suggesting that infants under four months old readily categorize repeated stimuli according to whether they share common syllables, or according to the number of syllables they contain, but not according to common phonemes or number of phonemes. Recognizing that the debate over whether "the unit of representation" is the phoneme or the syllable (or both, or neither) has not been settled, and that it is not necessary to assume that only one level of representation is accessible in infancy, it is argued here that syllables are *plausible* units for statistical computations in infants. Segments (or units at some other level of description) may also participate in related computations (Newport & Aslin, 2004).

Several studies have addressed these questions by testing whether infants note the similarity of syllables that all contain the same consonant or vowel. In one representative study (Jusczyk & Derrah, 1987), 2-month-olds were habituated to a set of syllables such as *bi*, *bo*, *ber*, and *ba*, and then presented with one of two new syllables (among others in conditions not discussed here): *bu*, which maintained the initial /b/; or *du*, which contained a novel consonant and vowel. If infants represented the

familiarization set as "/b/ plus a vowel," one would expect infants to ignore the addition of *ba*, or at least respond to it less often than the change to *du*.[1]

However, this effect was not found. Similar apparent failures of young infants to detect the commonality of syllables sharing phonemes were shown by Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy, and Mehler (1988) and Eimas (1999). At the same time, analogous experiments testing infants' sensitivity to lists of bisyllables containing shared *syllables* rather than shared phonemes have found that infants do detect syllable repetition (Eimas, 1999; Jusczyk, Jusczyk, Kennedy, Schomberg, & Koenig, 1995).

These results were found with infants 3–4 months old or younger. Older infants, in contrast, have been shown to categorize syllable lists according to common consonants or common features. For example, Fodor, Garrett, and Brill (1975), using an operant conditioning procedure, trained 4-month-olds to respond to syllables like /pi/ and /pu/ (but not to /ka/, which they also heard), or to /pi/ and /ka/ (but not to /pu/). After five consecutive days of training, infants were tested without reinforcement for an additional five days. Infants taught the pi–pu category maintained the training better than infants taught the pi–ka category, suggesting that infants perceived some commonality in the /p/-initial syllables and therefore retained the category more easily. Other research showed that by six months, infants trained to make a headturn response to /ba/ and /da/ but not /ma/ can generalize this response to /ga/ without false-alarming to /ŋa/, suggesting infants' sensitivity to some aspect(s) of the similarity of /b/ and /d/ (Hillenbrand, 1983). Still later in development, infants of about 9 months are sensitive to repetition in syllables' onset C and CV (Jusczyk, Goodman, & Baumann, 1999) or rhyme (Hayes, Slater, & Brown, 2000).[2] However, there is no evidence that this developmental increase in categorization according to subsyllabic regularities is accompanied by a decline in the importance of the syllabic level of analysis.

Additional studies have examined infants' categorization of stimuli according to the number of phonemes or syllables they contain. Such studies show that the syllable emerges early in infancy as a "package" of sound that is salient and countable. For example, Bijeljac-Babic, Bertoncini, and Mehler (1993) found that newborns recovered from habituation when a spoken list of various 2-syllable words changed to a list of 3-syllable words (or vice versa), but did not notice analogous changes from 4-phoneme words to 6-phoneme words. Similar results were found by Bertoncini, Floccia, Nazzi, and Mehler (1995; see also Bertoncini & Mehler, 1981). These studies comport with research showing that *explicit* awareness of phonemes is considerably delayed relative to awareness of syllables (e.g. Liberman, Shankweiler, Fischer, & Carter, 1974).

[1] This expectation is grounded in research on visual category formation in infants. In these studies, infants are habituated to various images of objects belonging to a category. If infants have extracted the features common to category members, they may fail to dishabituate upon seeing a novel image that fits in the category, but will dishabituate upon seeing a novel image that does not belong to the category (e.g. Bomba & Siqueland, 1983).

[2] Following common practice, strings of consonants and vowels will be described using C (consonant) and V (vowel); thus, the word *bean* may be described as a CVC.

Thus, perception experiments suggest that young infants do not readily perceive the commonality of phonological units in differing contexts. This is likely to be a consequence of the fact that the acoustic forms of speech sounds vary considerably with the surrounding context (e.g. Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Several researchers have suggested on the basis of these considerations that the syllabic level of analysis is available to infants and is crucial as a unit of segmentation and representation (e.g. Eimas, 1997; Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993, 1997; Mehler, Dupoux, & Segui, 1990). Here, it is assumed that infants compute sequential statistics over syllables. Again, this does not imply that the syllable is the only unit of representation, nor that computations over subsyllabic units cannot occur (cf. Mehler et al., 1990).

5. The goals of the task

Goodsitt et al. (1993) contemplated two sorts of segmentation strategies: *bracketing* strategies, in which infants are assumed to insert boundaries into continuous speech, and *clustering* strategies, in which infants are assumed to group certain speech sequences together into units. The present paper illustrates a clustering strategy. Unlike previous models, the present account does not assume that infants determine a lexical affiliation for every syllable in every utterance. Rather, infants register the frequency with which syllables and syllable sequences occur, and under certain conditions consider syllable sequences to be words. The intuition underlying this system is that infants hear a great deal of speech, but may draw no particular lexical conclusions from much of it. Certain sequences will come to feel more familiar than others as a consequence of repetition and probabilistic clustering. We will call these familiar sequences *postulated words.* Other sequences will not come to feel familiar and will not be considered as words. The notion that familiar sequences emerge from speech as recognizable units seems more consistent with a clustering strategy, which focuses on chunks extracted from the signal, than a bracketing strategy, which focuses on finding dividing points in the speech signal. Similar suggestions have been made by Morgan (1996b) and Perruchet and Vinter (1998).

Infants may not draw a firm boundary between syllable sequences that are or are not possible words. Familiarity, and thus wordlikeness, is assumed to be graded. That said, a categorical distinction between postulated words and nonwords is drawn here for the purpose of analysis, as described below.

6. The base corpora

Two orthographic-text corpora of speech to infants were used: the Korman (1984) corpus, which is English, and the van de Weijer (1998) corpus, which is Dutch. These corpora were selected because they are fairly large, and contain speech to infants under 12 months of age. English and Dutch are both Germanic languages in which words in infant-directed speech tend to be mono- or bisyllabic. Both languages have

relatively simple morphological systems, and as a result the forms of words are fairly fixed across instances. In these respects English and Dutch are languages for which a syllabic clustering model might have a strong chance of success in identifying words.

The Korman corpus consists of speech to six English infants who were each recorded in a series of 24-h sessions ranging from when infants were 6 to 16 weeks of age. Most of the speech was produced by the infants' mothers, though portions of the sample were from other talkers such as the father or the investigator. The corpus contains a total of about 42,000 word tokens and about 1800 word types.

The Van de Weijer corpus consists of speech to a single Dutch infant who was recorded essentially 24 h per day starting from the age of 6 months and extending for 91 days. The text corpus is based on a transcription of 18 of these days, taken from the first week (age 6;0–6;06), a middle week (approximately 7;10–7;16), and a final week (approximately 8;26–9;02). The present analyses considered only the portion of the corpus containing infant-directed speech uttered by adults (usually the mother, father, or a babysitter). Excluded from the corpus were German utterances spoken by the mother (a native German speaker who had been living in the Netherlands for 12 years and who was a fluent Dutch speaker) to the infant. This exclusion was not meant to imply the assumption that this child could distinguish the two languages; the purpose was to obtain a sample of Dutch infant-directed speech and not to model the linguistic situation of a particular child. The corpus consists of a total of about 25,000 word tokens and about 1050 word types.

A number of modifications to each corpus were made, the most important of which was the replacement of each word with an estimated pronunciation taken from the Celex dictionary (Baayen, Piepenbrock, & Gulikers, 1995). In the base corpora, each word took on exactly the same form each time it occurred. This lack of within-type variation was an idealization, of course.[3] Utterances containing untranscribable material (marked *xxx* according to CHILDES conventions; MacWhinney, 1995) were deleted. Utterances containing primarily sound effects (e.g., animal noises, and utterances like *brooooooop?*) were deleted; utterances containing a sound effect, and then words (or vice versa) were retained without the sound effect. For the most part, words were defined as strings of text surrounded by spaces. Exceptions included some words transcribed with "+" such as *upsa* + *daisy* which was considered a single word, and *dear* + *dear* which was considered two words; these decisions were made on a case by case basis for each type. Words not present in the Celex dictionary were given a pronunciation consistent with Celex conventions. Most of these were child-register words such as *drinkies*.

Stress markings and syllable boundaries were taken directly from the dictionary listings, except for monosyllabic words. The syllable boundaries listed in the Celex database were created by a set of rules whose outputs were hand-checked (Baayen et al., 1995). In contrast to (e.g.) Christiansen et al. (1998), who listed all monosyllabic words as stressed, we chose a consistent stress pattern for each monosyllabic

[3] Testing of variant corpora embodying several different assumptions about the severity and nature of infants' misinterpretation of natural variability will be an important extention of this work, lying outside the scope of the present report.

word on a type-by-type basis, by listing the environments in which each word occurred and judging whether it seemed more natural for that word to have been pronounced as stressed or not. For example, across tokens of *the*, in most sentences it seemed natural to pronounce it without stress; the reverse was true for tokens of *car*; thus, *the* was marked as unstressed throughout the corpus, and *car* was marked as stressed.

7. The clustering algorithm

The model worked by computing syllable frequencies and co-occurrence likelihoods over a syllabified corpus. This reflected the assumption that infants are sensitive to how many times they have heard a sequence of speech, and whether a particular syllable tends to appear together with another particular syllable. The statistic used for operationalizing co-occurrence probability was *mutual information*, a standard information-theoretic measure which has also been used in related research (e.g. Redington, Chater, & Finch, 1998; Roy & Pentland, 2002). Mutual information (MI) was computed as follows:

$$MI_{\mathrm{AB}} = \log_2 \frac{p(\mathrm{AB})}{p(\mathrm{A})p(\mathrm{B})},$$

where AB represented two consecutive syllables A and B, and p(AB) referred to the probability of a particular bisyllable AB (that is, the number of times AB appeared in the corpus, divided by the number of bisyllables in the corpus). Taking the logarithm of this fraction is traditional but had no effect on the results here because rank orders (which are not affected by monotonic transformations) were used, as described below. Mutual information is similar to *transitional probability*, which is $p(\mathrm{AB})/p(\mathrm{A})$. Note that transitional probability weights the frequency of a syllable pair by the frequency of its first element; thus, a sequence like *the dog* would not be expected to have a high transitional probability because *the* precedes many different words. Mutual information, by contrast, weights the frequency of the syllable pair by the frequency of both of its elements, and therefore provides a measure of how informative two syllables are about each other.[4]

Modeling the infant's "memory" for heard sequences was two-step process. First, the frequencies of all monosyllables, bisyllables, and trisyllables in the corpus, and the mutual information value of all bisyllables in the corpus, were computed. This procedure resulted in four lists, which were hypothesized to reflect implicitly stored speech data; that is, we assumed that when infants hear speech, these frequency and mutual information counts are updated. The next step was to apply the clustering

[4] The experimental results of Saffran and colleagues are compatible with a number of co-occurrence measures, including MI, as pointed out by Aslin et al. (1998, p. 321). A previous analysis similar to the present one found that MI and transitional probability (both forward, $p(\mathrm{B}|\mathrm{A})$, and backward, $p(\mathrm{A}|\mathrm{B})$) produced roughly equivalent results, although the two varieties of transitional probability often made different false alarms (Swingley, 1999).

algorithm to these data, resulting in a set of speech sequences predicted to be familiar. Familiarity was operationalized as a combination of frequency and mutual information, and multisyllabic sequences high in both were considered words. Rather than choose a fixed definition of "high" (such as "in the top 10% of all syllables"), we will present results for all possible definitions of "high."

To create a common metric for frequency and mutual information, raw scores of monosyllable frequency and bisyllable mutual information were converted into percentiles. For example, a given syllable might be at least as frequent as 85% of all monosyllable types in the corpus, and would therefore receive a percentile score of 85 for frequency.[5] Likewise, a given bisyllable might have a mutual information score as high as 85% of all bisyllable types, and thus receive a mutual information percentile of 85.

Next, a set of decision rules was applied to the four lists. Each decision rule stipulated criteria (in terms of percentiles) a linguistic unit must meet for that unit to be postulated as a word. The decision rules were as follows. Monosyllables were considered wordlike if they exceeded the criterial *frequency* percentile. Bisyllables were considered wordlike if their frequency and their MI both exceeded the criterion percentile values. Trisyllables were considered wordlike if their frequency, and the MIs of their component adjacent bisyllables (i.e. bigrams AB and BC of trigram ABC), all exceeded criterion.[6] To take the 85th percentile example, then, monosyllables exceeding the 85th percentile in frequency were postulated as words; bisyllables whose frequency and mutual information both exceeded the 85th percentile were postulated as words; and trisyllables whose frequency and component bisyllables' mutual information scores all exceded the 85th percentile were postulated as words. Thus, the selectivity of the algorithm (i.e., how miserly it was in granting postulated-word status to speech sequences) was a single parameter varying over a range of zero to 99. A separate analysis was done at each criterion level, resulting in 100 lists of postulated words.

Postulated words embedded in other postulated words were excluded. For example, suppose that the bisyllable *lephone* met the bigram criteria, and the trisyllable *telephone* met the trigram criteria. Only *telephone* would be considered a word. This exclusion was motivated by the finding that infants extracting a bisyllabic word (e.g., *kingdom*) do not appear to treat the stressed syllable of that word (e.g., *king*) as familiar (Jusczyk et al., 1999). Similar biases in favor of groupings that account for more, rather than less, perceptual data are found in other models

[5] To use a consistent frequency metric, frequency for units of all lengths was relativized to monosyllable frequency; thus, a bisyllable was considered frequent if it was as common as (e.g.) 85% of all monosyllables. This use of a single frequency metric seemed more reasonable than assuming that whether sequence feels frequent to a child depends upon how long it is.

[6] Sequences longer than three syllables were not evaluated for wordhood, on the grounds that (1) almost no 4-syllable sequences would exceed the mutual information criteria, and (2) over 96% of Dutch and English word types have fewer than 4 syllables, and the longer ones are nearly always rare. Thus, modeling the acquisition of 1-, 2-, and 3-syllable words provides good coverage of the infant's lexical environment.

as well, though sometimes as outcomes rather than as explicit constraints (e.g. McClelland & Elman, 1986).

The "success" of the algorithms in finding actual words of English and Dutch was evaluated primarily using *accuracy*, which is simply the likelihood that a postulated word was a real word (e.g. Brent & Cartwright, 1996). In signal-detection terms, accuracy may be computed as follows:

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

Of course, infants were not assumed to make this computation, which relies on knowing which sequences were actually words. Another statistic evaluated was the raw number of actual words found by each analysis, to give a sense of the size of the vocabulary being formed. Finally, the proportions of bisyllabic postulated words matching each of four stress patterns (strong–weak, strong–strong, weak–strong, and weak–weak) was computed for each criterion percentile value.

8. Results for the base corpora

Results are presented in the form of three graphs for each language (see Fig. 1). For all graphs, the x-axis represents the criterion percentile. The leftmost portions of the graph show degenerate cases in which hardly any constraint (or no constraint at all, at $x = 0$) was placed on what counted as a word, except for the rule excluding embedded words. Criterion levels between 50 and 100 would seem most likely to reflect infants' representation of what is familiar and what is not; nevertheless the entire range is displayed to provide a better sense of the results.

The first graph shows accuracy as a function of criterion percentile, with one line representing each word length (mono-, bi-, and trisyllables) and a fourth line showing accuracy collapsed over these lengths. For example, in the English graph the line starting at about 0.19 and rising to 1.00 shows that as the criterion for postulating a word was made more stringent (x-axis), accuracy (or likelihood of being correct when postulating a word) for bisyllables increased (y-axis). The overall accuracy of mono-, bi-, and trisyllables also generally increased, as shown by the solid line. Note that the last is not simply an average of the other three plotted lines; rather, it is a count of all true words (regardless of length) divided by the total number of postulated words (also regardless of length).

The English and Dutch accuracy plots have several characteristics in common. Overall accuracy increased from below 5%, at the percentile criterion level of 0, to a plateau of around 80% starting at a percentile criterion level of 75. This indicates that frequency and mutual information were helpful, though not faultless, in separating words from nonwords at moderately high criterion levels. There was also a general tendency for shorter words to be identified with greater accuracy than longer words. With a few exceptions, trisyllable accuracy was consistently lower than bisyllable accuracy, and monosyllable accuracy was consistently high. (In fact, monosyllable accuracy at very low criterion levels reached 100%. This came from the

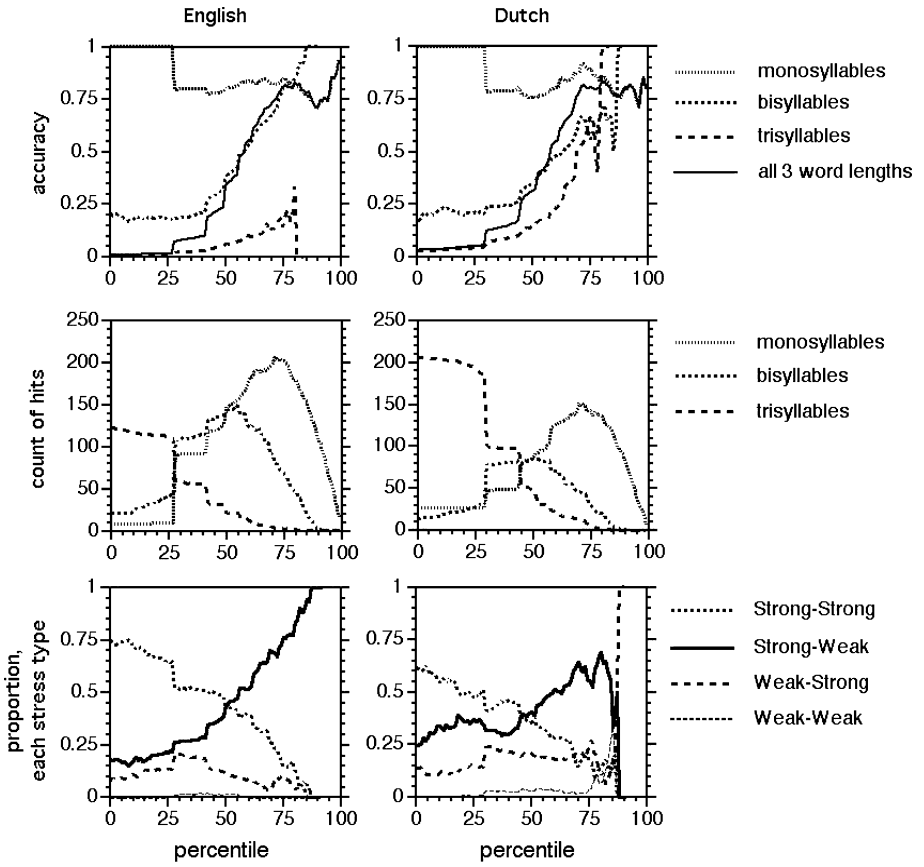

Fig. 1. Accuracy, raw numbers of hits, and proportions of bisyllables having each of four stress patterns, for a range of criterion percentiles, using the base corpora.

embedding constraint. At very low criterion levels, every syllable appearing in a multisyllabic utterance was considered part of a bisyllabic or trisyllabic word and was therefore excluded by the embedding rule. The remaining syllables were those that appeared only in one-syllable utterances; because all such utterances contained no more and no less than an actual word, all such syllables were words, leading to an accuracy level of 100%.)

The second graph shows the raw number of correct identifications (hits) found for each word length and each criterion percentile. Once again, the shapes of the English and Dutch curves were similar. At moderate levels of constraint (e.g., the 75th percentile), the number of monosyllables detected peaked at about 200 (English) and 150 (Dutch). The number of bisyllabic words was smaller, at 66 words (English) and 29 (Dutch). Finally, the number of trisyllabic words identified tended to be quite small. At the 75th percentile, this amounted to 2 English and 6 Dutch words. Considering the number of hits in light of the accuracy results, the optimal level

of constraint lay between about 60 and about 90: below 50, the overall accuracy was very low, and above 90, any high accuracy scores pertained to a few monosyllabic hits, and one or zero multisyllabic hits. (This explains why the accuracy curves are jagged in the upper constraint ranges: when only a few words were postulated, the inclusion or exclusion of a single sequence had large effects on the hit:false-alarm ratio.)

The first and second graphs show a few discontinuities at matching points. For example, in the Dutch analysis, the number of trisyllable hits fell from 186 at the 29th percentile to 100 at the 30th; at the same time, monosyllable accuracy dropped and the number of mono- and bisyllabic hits increased. These effects were cascading consequences of changes from one criterial frequency to the next (e.g., at the 29th percentile, the criterial frequency was one; at the 30th, it was two). For English and Dutch, about half of the trisyllables only occurred once; as a result, the adoption of a criterion of two occurrences halved the number of trisyllables postulated as words. These sequences were then "free" to be considered as shorter words, boosting the number of monosyllabic and bisyllabic hits. Similar effects may be seen elsewhere as the criterial frequency increased (e.g., from 2 to 3 at the 42nd percentile in the English analysis).

The third graph concerns only the bisyllabic postulated words. Each line corresponds to a different stress pattern; the *y*-axis shows the proportion of postulated words conforming to each stress pattern (values of the lines at each percentile sum to one). Generally speaking, as the level of constraint increased, the proportion of trochaic (strong–weak) bisyllables increased, at the expense of the other three patterns. As a result, at the upper constraint ranges, the trochaic pattern clearly dominated, except at the very end of the Dutch range, when the word *hallo*, "hello," was the only postulated word. (*Hallo* was listed in the corpus as iambic, but in fact may be pronounced with various intonational and stress patterns.) For neither language was the dominance of the trochaic pattern merely a strengthening of a trend already present in the corpus; considering all of the bisyllables, the iambic pattern was more frequent by types than the trochaic pattern (English: 26.8% weak–strong, 16.9% strong–weak; Dutch: 29.1% weak–strong, 24.7% strong–weak). Thus, if infants cluster syllables according to conditional probability and frequency, the resulting sequences could indicate to children that the trochaic pattern is characteristic of words in English and Dutch, and could thereby provide the impetus for the trochaic segmentation bias that has been shown in behavioral studies.

Examination of the real words extracted (i.e., the hits) showed that they belonged to various grammatical form classes. To quantify this, all words in each corpus were sorted into the classes *noun*, *action verb*, *adjective/adverb*, *interjection/performative*, *pronoun*, and *other*. The *pronoun* category included pronouns with cliticized verbs, as in *he's*, *I'll,* and *you'd*. Where a word could fit multiple categories, its actual use in the corpus was checked. If there was a clear majority usage, that was taken as the category; if there was not a clear winner (as for English *kiss*), a "miscellaneous" category was assigned (4.3% of English cases, 1.0% Dutch). Table 1 lists the proportions of word types falling into each category (excluding the miscellaneous words) for all words in each corpus; for all words occurring five times or more (which ruled

Table 1
Proportions of words in each of 6 form-class categories in the corpus as a whole, among only the frequent words of the corpus, and among the model's hits at the 70th percentile, for English and Dutch

Form class	English all words	English freq $\geqslant 5$	English 70th %	Dutch all words	Dutch freq $\geqslant 5$	Dutch 70th %
Noun	38.7	29.6	27.3	37.8	28.0	21.4
Action verb	28.1	28.8	23.8	25.0	22.4	17.6
Adj./adv.	17.7	17.5	18.1	15.9	19.3	21.9
Interj.	5.7	7.1	9.2	12.9	12.7	16.1
Pronoun	3.5	7.6	8.2	2.9	6.6	7.1
Other	6.3	9.4	13.4	5.5	11.0	15.9

out slightly over half of the words); and for the hits extracted by the analysis at the 70th percentile.

For the most part, the classes of words extracted by the algorithm did not vary much from the proportions that would be expected based upon their base frequency, except for nouns, which were somewhat less likely to be extracted. As suggested by the "frequency $\geqslant 5$" columns, the lower extraction rate of nouns was a consequence of the model's frequency criterion: many infrequent nouns present in the corpus were not extracted. The fact that the form classes of words were extracted roughly in proportion to their presence in the corpus indicates that the phonological forms of words in different categories, and the contexts in which they occurred, did not markedly favor the extractability of some classes over others. For example, it was not the case that the frequent presence of *the* before English nouns caused nouns to be particularly easy for the algorithm to detect.

Examination of the false alarms showed that the monosyllabic false alarms tended to be syllables that occurred in a promiscuous range of contexts, or which tended to appear together with other promiscuous syllables. (See Appendix A for a list of English false alarms at the 70th percentile.) Monosyllabic false alarms occasionally corresponded to inflections, but these were a small minority. For example, *-ing* was by far the most frequent of the monosyllabic false alarms. But missegmentations of *-ing*, including *-ping* (from *burping*, *keeping,* and *sleeping*, among others), and *-ching* (from *pinching* and *watching*, among others) were also extracted, though they occurred much less often. Apart from the identification of *-ing*, then, the monosyllabic English false alarms did not appear to provide the infant with even a primitive morphological analysis. The monosyllabic Dutch false alarms showed a similar pattern. The prepositional prefix *aan*, which forms the beginning of some infinitive verbs and past participles, was extracted from words like *aangekleed* ("dressed") and *aankomen* ("to arrive"). But most of the errors were syllables having no common meaning across the words from which they originated. In a few cases, words were stripped of the diminutive suffix [jə], resulting in true words that were counted as false alarms because these words had not been used as such in the corpus (e.g., *blok*, "block"; *voet*, "foot"; and *stuk*, "piece"); however, the proportion of such cases was small.

Bisyllabic and trisyllabic false alarms tended to be clusters of words, as opposed to collections of word parts. For example, at the 70th percentile, 40 of the 44 English

bisyllabic false alarms were pairs of frequently co-occurring words. In the Dutch analysis, 82% of the 51 bisyllabic false alarms at the 65th percentile were pairs of monosyllabic words. The trisyllabic false alarms were more heterogenous. At the 70th percentile, about half of the Dutch and English trisyllable false alarms paired a monosyllabic word with a bisyllabic word, and a few more clustered three monosyllabic words together (e.g., *want some more*).

In some cases the false alarms might be thought of as conventional expressions having lexical status. For example, the phrase *come on*, by far the most frequent of the English bisyllabic false alarms, has little apparent semantic relation to the words it nominally comprises, at least in the phrase's occurrences in the corpus (where it usually meant "proceed; do it" but not "move toward me onto *x*"). Several Dutch fixed expressions were counted among the false alarms, including *kijk 's* ("look at this"), *kom maar* ("come along then"), and *zeg maar* ("Say!" or "so to speak"), which often do retain the primary sense of their verbs, but which have distinct conversational functions and cannot be inflected. Word pairs of this sort appeared more common in the Dutch analyses than the English. That said, many of the false alarms did not have this character; although a case could be made for the lexical status of *come on* or *look at*, the same cannot be said convincingly for *you've got*, *give us*, *your bib*, or *play with*. Thus, we conclude that while the false alarms usually did not cluster parts of several words together, the nonword conglomerations that were formed did not appear to consistently offer semantic or syntactic regularities that infants might profit from.

8.1. Differences between Dutch and English results

Although the general pattern of results in the Dutch and English data were similar, a number of differences between the two languages were found. At the higher constraint levels, accuracy in identifying bisyllabic words was greater in the English corpus than the Dutch; the reverse was true for accuracy in identifying trisyllabic words.

Examination of the Dutch false alarms suggests two factors that conspired to reduce the accuracy of the Dutch bisyllable analyses. One was the number of fixed expressions consisting of pairs of monosyllabic words. For example, the Dutch false alarm *hou vast* ("hold on") contains two words that hardly ever occurred in other contexts. As noted previously, several bisyllabic false alarms were conventional expressions, particularly in the Dutch analyses. More importantly, Dutch infant-directed speech contains more trisyllabic words than similar English speech; on occasion these words were not detected as trisyllables, but did trigger postulation as bisyllables. Examples include the first two syllables of *boterham* ("sandwich"), *mannetje* ("little man"), *mopperen* ("being grumpy"), *vreetzakje* ("little eating-too-much person"), and *zonnetje* ("little sun"), and the last two syllables of *olifant* ("elephant") and *eventjes* ("just" or "for a little while"). Some of these words are morphologically complex, consisting of a bisyllabic word and the diminutive suffix *-je* or *-tje*. The Dutch diminutive is productive and frequent, making full trisyllables containing the diminutive suffix difficult to extract. Thus, to some

degree the relatively low accuracy of the Dutch analyses can be traced to structural properties of the language.

The Dutch advantage for trisyllables reflected the greater number of frequent trisyllabic words in the Dutch corpus. For example, 9.3% of all Dutch types occurring three times or more were trisyllables; the equivalent English proportion was only 3.4%. But this numerical advantage does not explain the difference in accuracy, which arises not only from the correct detection of Dutch trisyllables but the false postulation of English ones. At the 70th percentile, the Dutch and English analyses postulated a similar number of trisyllabic words (28 and 30, respectively), but half of the Dutch guesses were correct, against 13% of the English. Whether the language difference here was due to general structural properties of the languages is not clear at present. In any event, at the higher constraint ranges the number of trisyllables postulated was dwarfed by the number of monosyllables and bisyllables postulated, so that accuracy over the entire vocabulary was virtually unaffected by the results for trisyllables.

Overall, the results for both languages were similar. Indeed, collapsing over word lengths, word-finding accuracy was nearly identical in Dutch and English. Over the criterion range from the 70th to the 80th percentile, Dutch word-finding accuracy was 80.5%, while English accuracy was 78.7%. The general result, then, was that selecting syllables or syllable sequences having high frequency and high mutual information produced lists of postulated words that were actual words as much as about 80% of the time. At a fairly high level of constraint (e.g., 75%), this selection process resulted in a vocabulary of 270 real words (English corpus) or 178 real words (Dutch corpus) plus about a fourth as many nonwords. The set of selected word and nonword bisyllables contained the strong–weak pattern more often than the other three stress patterns combined, a result that held for both languages. To the extent that the corpus and the clustering algorithm reflect real-world inputs and memory processes, they suggest that the interaction of simple mental mechanisms and the statistical nature of infant-directed language could result in both the foundation of the vocabulary, and the guiding force behind infants' prosodic parsing strategies.

In the remainder of the paper, these results are re-examined after changes to the algorithm representing infants' clustering tendencies, and modifications of the corpora. First, a variant algorithm using only frequency, but not conditional probabilities, is evaluated. Then, the original algorithm is tested on corpora in which syllable boundaries are assumed to be less clear. Finally, a set of analyses examines smaller corpora and very short utterances.

9. Reanalysis using frequency alone

The preceding analyses assumed that infants' syllable-clustering depends on both the frequency and conditional probability of syllable patterns. The relevance of conditional statistics was motivated by the artificial-language experiments of Aslin et al. (1998). However, it may be instructive to consider whether the computation of

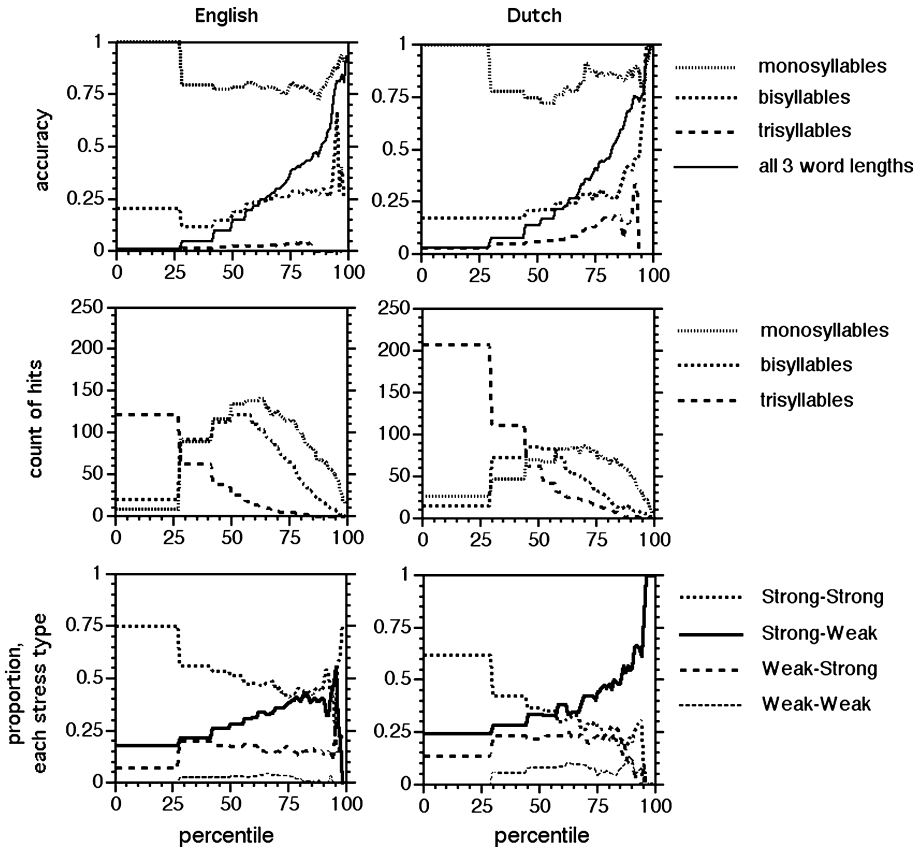

Fig. 2. Accuracy, raw numbers of hits, and proportions of bisyllables having each of four stress patterns, for a range of criterion percentiles, using frequency and not mutual information in criteria.

conditional probabilities is necessary for identifying words. The following analyses addressed this question: are *frequent* syllable sequences words, independently of their conditional probability?[7]

In these analyses, the same corpora were tested using exactly the same procedure, except that here only frequency criteria were applied. Thus, for mono-, bi-, and trisyllabic patterns, only those sequences that exceeded a criterial frequency were postulated as words. Results of these analyses are presented in Fig. 2.

[7] Although mutual information can only be computed if frequencies are counted, it is possible to perform the reverse analysis, in which only mutual information, and not frequency, is criterial for extraction of polysyllabic words. Accuracy levels in such analyses are poor, mainly because there are too many instances of monosyllabic word pairs or triplets that occur once or twice in restricted environments and are postulated as words. In an m.i.-only analysis of the English corpus, overall accuracy remained under 30% for most of the criterion range, and never surpassed 60%. Over the criterion range from 60 to 99%, this analysis showed accuracy levels averaging 51 percentage points lower than the standard analysis.

The first plot reveals a substantial decline in accuracy relative to the analysis that took conditional probabilities into account. Whereas the analysis including frequency and mutual information as relevant criteria yielded accuracy around 80% for much of the upper percentile range, the analysis including only frequency reached 80% only with such strict criteria that just a tiny number of words were postulated. For example, bisyllable accuracy in the Dutch analysis surpassed 75% only at the 97th percentile and above, yielding at most four words. In the English analysis, bisyllable accuracy never surpassed 70%, and exceeded 50% only at the 95th percentile. At the same time, the second plot shows that the raw number of multisyllabic hits detected in the frequency analysis was roughly equivalent to the number of such hits in the analysis that included frequency and mutual information; the number of monosyllabic hits was substantially lower in the frequency-only analysis. (This may seem surprising given that frequency alone provided the exclusion criterion in both analyses; however, in the frequency-only analysis the relatively high number of false alarms resulted in more monosyllabic words being excluded as embedded parts of longer words.) Thus, frequency alone is a misleading indicator of whether syllable groups are words.

This is simply because there are so many common two-word phrases in the speech children hear. The relevance of conditional probabilities may be seen in a comparison of the bisyllabic false alarms containing the word *you* (regularly the most common word in English child-directed corpora, surpassing even *the*). The English frequency-only analysis false-alarmed to 15 *you* phrases at the 90th percentile, including *are you*, *do you*, *you do*, *you are*, *yes you*, and *you want*. By contrast, the conditional analysis never included *you* in bisyllabic or trisyllabic postulated words after the 60th percentile.

Though it is clear that frequency is not a good guide to the lexical status of bisyllables, whether a frequency criterion alone could lead to the trochaic parsing bias depends upon the language and upon the assumptions made about the process. In Dutch, high-frequency bisyllables tend to exhibit the trochaic pattern; in English, high-frequency bisyllables tend to be strong–strong, though the trochaic do outnumber the iambic. The language difference arises in part from a syntactic difference between English and Dutch. Dutch infant-directed speech contains frequent question constructions in which a monosyllabic verb is followed by *je* (''you''), an unstressed word (e.g., *Ga je*, ''Do you go''; *Heb je*, ''Do you have''). English does not have as many of these constructions, and has relatively more constructions like *Do you*, which is not trochaic.

Whether a frequency-based clustering algorithm would be sufficient for producing a trochaic bias depends upon whether infants need only choose between strong–weak and weak–strong (in which case the evidence in favor of strong–weak is clear for both languages, even in the frequency-only analysis), or whether infants assume that the strong–weak pattern holds generally only if it dominates all of the other possibilities. The latter seems more likely; that is, if the strong–strong pattern dominated the child's extracted forms, the child should not then go on to use a strong–weak template in segmenting speech. However, this remains an open question.

10. Results from probabilistically syllabified corpora

In the original corpora, no syllables spanned word boundaries. This was an idealization: in fluent speech, the syllable affiliation of intervocalic consonants may be ambiguous. For example, in the sentence *Who's a pretty girl?* the /z/ of *Who's* may be interpreted as syllabifying with the /ə/ of *a*, yielding the syllable pair /hu.zə/ rather than /huz.ə/. Resyllabification makes the word-finding problem harder, and therefore must be modeled to avoid overestimating the likelihood of correct segmentation. Unfortunately, it is difficult to estimate how often resyllabification occurs (or, more precisely, when infants perceive a consonant as affiliating with the preceding or following vowel). It is likely that in some cases speakers produce acoustic cues that infants can interpret as signaling syllable boundaries, and that in other cases speakers realize adjacent syllables whose consonant affiliations are ambiguous to infants.

To determine the effects of this ambiguity, further analyses examined corpora whose syllable boundaries were established by a probabilistic algorithm that assigned consonants to syllables. The algorithm was biased to split consonants evenly between syllables (e.g., VC.CV was more probable than VCC.V), and was biased toward placing consonants as syllable onsets rather than as offsets whenever the second syllable of the pair was stressed, for VCV and VCCV junctures. The stress bias was motivated by empirical and theoretical arguments that intervocalic consonants generally affiliate with stressed following syllables (e.g. Goslin & Frauenfelder, 2000; Kahn, 1980, p. 41; Treiman & Danis, 1988). The values of the biases are listed in Appendix B.

The algorithm had no knowledge of the speaker's intended word boundaries, embodying the conservative assumption that infants are oblivious to all phonetic cues to syllable boundaries. No specific English or Dutch phonotactic restrictions were applied.[8] However, some syllable assignments were rejected as violations of sonority sequencing (see e.g. Clements, 1990, for a discussion of sonority). For example, the bisyllable *that's me* could in principle be divided as CV.CCCVC (/ðа.tsmi/), or as CVCCC.VC (/ðatsm.i/), but both of these assignments were ruled out on sonority grounds. The sonority restriction was based on a division of all consonants into the classes *glide*, *liquid*, *nasal*, and *obstruent*. Onsets of syllables having falling sonority (such as glide-obstruent) were disallowed, as were offsets having rising sonority (defined over the four classes). Equal-sonority sequences were disallowed for glides, liquids, and nasals. Finally, the obstruent class was subdivided into fricatives and stops. Fricative–fricative and stop–stop sequences were disallowed both as onsets and offsets, reflecting the intuition that infants hearing a talker say *fish food* would not parse it as *fishf ood*.

Because the syllabification algorithm was probabilistic and did not always produce the same output when a corpus was passed through it, each analysis was done

[8] Studies have revealed infants' knowledge of language-specific phonotactics, but thus far only in infants of at least 9 months. Given our goal of modeling the origins of the trochaic bias, which emerges by 8 months, we did not build English- or Dutch-specific sequencing constraints into the syllabifier.

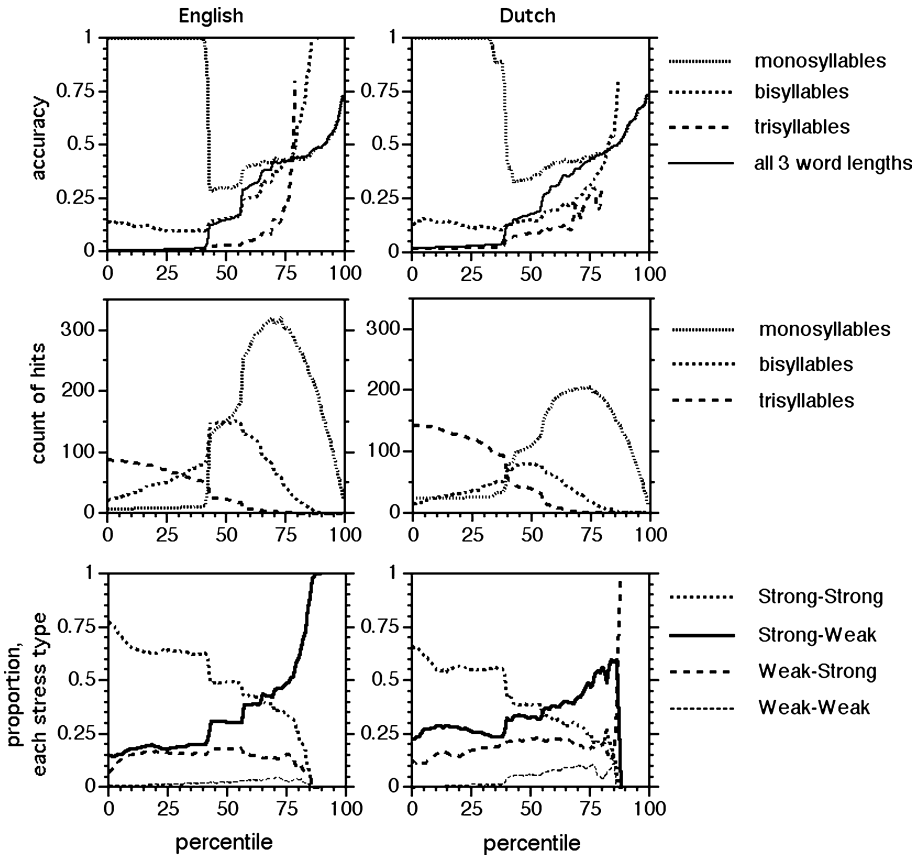

Fig. 3. Accuracy, raw numbers of hits, and proportions of bisyllables having each of four stress patterns, for a range of criterion percentiles, using the resyllabified corpora.

over a set of five randomizations of each base corpus. Results are shown in Fig. 3. Data on each graph represent averages over the five randomizations.[9]

It is immediately clear that accuracy suffered a great deal from the probabilistic assignment of consonants to syllables. Although infants would still have something to gain from the use of frequency and mutual information statistics (the overall accuracy curve increased with higher criteria), the resyllabified corpora yielded accuracy levels that hovered around 40% for much of the upper range. To explain why this was the case, we will first describe the syllable structure patterns in the corpora, and then characterize the false alarms made by the analyses.

Adjacent words in both the English and Dutch corpora usually contained one or more consonants at the word boundary. Vowel-final words preceded vowel-initial

[9] Standard errors of the means are not plotted because the error bars are generally too small to see.

words for only 6% of adjacent words (English) or 5% (Dutch). Thus, about 95% of the utterance-internal word boundaries involved consonants whose assignments to syllables were susceptible to change by the probabilistic syllabification algorithm. The most frequently-occurring case was the presence of a single consonant between the nucleus of first word's last syllable and the second word's first syllable (i.e. VCV), as in the pairs *give a* and *to me*. This one-consonant case included 49% of the adjacent English words, and 44% of the Dutch. (In both languages, about 70% of these single consonants were onsets.) Most of the other adjacent-word pairs contained two consonants at the word boundary (43% English, 36% Dutch). Clusters abutting singleton consonants or other clusters were relatively rare (VCCCV cases composing 9% of the English corpus and 7% of the Dutch, and cases with four or more consonants composing about 0.05% of each corpus).

A consequence of these statistical patterns was that the vast majority of word boundaries could be interpreted in exactly two or three ways—VCV sequences as V.CV or VC.V, and VCCV sequences as V.CCV, VC.CV, or VCC.V. As a result, probabilistic resyllabification resulted in a small number of frequent interpretations of each common word pair. This, in turn, yielded many syllable pairs that occurred in limited contexts and which were falsely postulated as words. For example, the English word *it* co-occurred with many other words and was rarely postulated as part of a bisyllable in the analysis of the base corpus; however, in the resyllabified corpus it formed a part of the postulated words *hold it* (as the syllables /hol/ /dɪt/) and *that's it* (as the syllables /ðæ/ /tsɪt/). Whereas *it* had been too promiscuous to be fixed to a postulated bisyllable in the first analysis, resyllabification sometimes converted it into the relatively infrequent /dɪt/, a syllable that surfaced primarily in *hold it*.

Comparison of the false alarms in the base analyses and the resyllabification analyses revealed that the spurious incorporation of very frequent words in postulated words was much more common in the latter analyses. For example, of 15 bisyllabic false alarms in the English base analysis at the 75th percentile, only one contained a top-20 word (*the* in *the ma*, a missegmentation of *the matter*), whereas of the 15 bisyllabic false alarms in a resyllabification analysis (at the 81st percentile), 67% contained a top-20 word (e.g., *a* in *a big*, syllabified as *ab ig*; *you* in *thank you*, syllabified as *than kyou*). Similar effects were present in the Dutch analyses. Thus, the bisyllabic false alarms tended to be pairs of monosyllabic words, sometimes with a stray consonant added to the beginning or end.

The monosyllabic false alarms formed a heterogenous mixture of words missing a consonant (e.g., /lip/ from *sleep*), words with an extra consonant (e.g., /tju/ from *don't you*), and missegmentations of longer words (e.g., /liŋ/ from *telling*). A substantial number of monosyllabic false alarms were syllables that had not appeared in the base corpora, often because they contained consonant clusters that were ill-formed. For example, 24 of the 65 monosyllabic false alarms at the 70th percentile (in the English analysis) had onsets with consonant clusters that were not present in the original corpora, including /dm/, /tθ/, and /vg/.

The fact that the sonority-based parser permitted some unattested consonant clusters suggested that part of the reason for the model's relatively poor performance in the resyllabification corpora was that the range of possible consonant assignments

was too broad, and perhaps broader than should be attributed to even 7-month-old infants. That is, even a relatively naive infant might not parse a sequence like *and mix* as *an dmix*, or *you've got* as *you vgot*, either on phonetic or distributional grounds. Thus, an additional set of corpora was created, in which the syllable parser attempted to assign consonants such that complex syllable onsets were drawn from the set of onsets present at utterance onsets in the corpus, and complex syllable offsets were drawn from the set of codas at utterance offsets. This kind of constraint involves information that is in principle available to infants (e.g. Brent & Cartwright, 1996). The syllabification procedure was otherwise the same as that used in the resyllabification analyses presented above.

Inspection of the results revealed modest improvements in accuracy. To save space, rather than reprint the full set of criterion plots, accuracy scores for each analysis are reported as averages over the criterion percentile range of 70 to 80, a range high enough to effect a reasonable degree of constraint on word postulation, but (at least for the monosyllables and bisyllables) low enough to avoid degenerate cases in which only a few words are postulated.

In the English analysis, monosyllable accuracy increased from 43.3% (the mean over the 70th to 80th percentiles for five randomly syllabified corpora) to 50.7% in the phonotactically constrained corpus; bisyllable accuracy increased from 43.1 to 50.8%.[10] The number of true words found was similar in both analyses, with slightly fewer monosyllables being detected in the phonotactically constrained corpus. Similarly, in the Dutch analysis consonant cluster constraints added about 7% to accuracy, yielding an overall accuracy of 42.6%: monosyllable accuracy increased from 44.4 to 52.2%, bisyllable accuracy from 31.7 to 36.9%. These increases in accuracy are similar in magnitude to those found by Brent and Cartwright (1996) when implementing a similar constraint.

Thus, the assumption of some fairly primitive phonotactic knowledge does make word-finding easier, but does not lead the model to approach the accuracy levels that were obtained when syllable boundaries were assumed to be unambiguous. The reason for this is clear: nearly half of word boundaries involved only one consonant and were therefore unaffected by cluster constraints. Many of the remaining word boundaries involved only two consonants, and often these could be "legally" assigned in two or three different ways.

Grouping syllables based on conditional probability succeeded in the no-resyllabification case because individual words appeared in many contexts, whereas parts of words tended to appear in limited contexts. The errors that did arise in the base corpora were due to recurring groupings of words (which limited some words' range of contexts), and to syllables' appearance in multiple words (if *ing* is attached to many verbs, its appearance in varied contexts makes it wordlike). However, once syllable boundaries were blurred, words no longer circulated freely as units, and this limited the separation of word and nonword sequences.

[10] Trisyllable accuracy appeared to increase substantially (28 to 44%), but this increase concerned less than one word, on average, and hence does not constitute a marked advance.

That said, the relatively low accuracy levels uncovered in the resyllabification analyses do not deny the value of conditional probability in separating words from nonwords; high-frequency sequences in the resyllabification corpora were less likely to be words than sequences high in both frequency and mutual information. For example, in an analysis using frequency alone, mean accuracy over the 70–80th percentile (for all word lengths) was only 21.7% (English) and 23.3% (Dutch), scores about half as large as those in the standard resyllabification analyses. Thus, the use of conditional probability statistics provides an enormous advantage over the use of frequency alone, even when the ultimate level of accuracy is not high.

Resyllabification did not change the tendency of the analysis to extract trochaic units more often than iambic units, as shown in the lower panels of Fig. 3. Beyond about the 60th percentile, trochaic sequences were extracted at least twice as often as iambic ones. The predominance of trochaic sequences was driven by the hits, which were overwhelmingly trochaic at the higher constraint levels. False alarms tended to be strong–strong (about half in the English corpus, and a third in the Dutch), with most of the remainder fairly evenly divided between trochaic and iambic sequences. Thus, even under the assumption that infants have no information about which syllables consonants belong to, statistical clustering biases yielded bisyllables that clearly exemplified the trochaic patterning that apparently guides Dutch- and English-learning infants' lexical parsing by 8 months of age.

The assumption that infants must guess at the syllable affiliation of all consonants could be relaxed in an informed way by inspecting each word boundary in a recorded corpus and estimating the likelihood of infants perceiving a boundary cue that would render syllable affiliations unambiguous. We do not attempt such an analysis here. However, it is possible to approximate the results of such an analysis by assuming that some proportion of the time, syllable boundaries are unambiguous. To do this, a modified version of the syllable parser was implemented. At each boundary, the probabilistic parser was only invoked if a random number exceeded a threshold value; otherwise the boundary was left in its original state, with consonants assigned to the correct syllables. Dutch and English corpora were generated using thresholds of 30, 50, and 70%, representing three estimates of the proportion of syllable boundaries that are unambiguous (and therefore "correct" with respect to word boundaries) to infants.

Here we present averaged data for the criterion percentile range from 70 to 80. For purposes of comparison, equivalent results are also plotted for the previous resyllabification corpora ("none" in the graph legend) and for the base corpora ("100%" in the graph legend). The graphs are presented in Fig. 4.

This finding highlights the importance of research showing infants' increasing sophistication in capitalizing upon various phonetic and probabilistic cues to syllable and word boundaries. Except when parents make syllable boundaries clear (e.g., at intonational phrase boundaries), infants' success in word-finding may depend upon their access to subtle, probabilistic cues to consonant affiliation. Over developmental time, the infant's "database" of syllable statistics should become cleaner and cleaner, resulting in improved accuracy in word discovery, as depicted in Fig. 4.

Determining where along Fig. 4 infants may fall at different ages depends upon future infant perceptual experiments, together with detailed phonetic analyses of

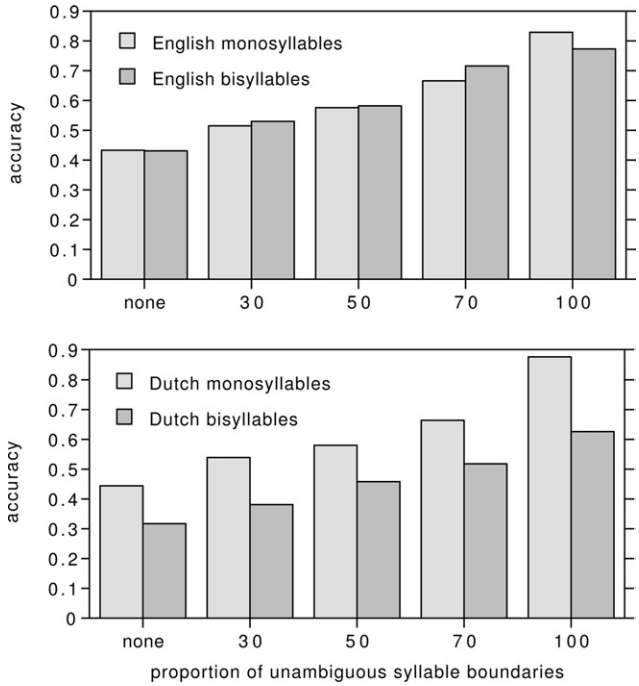

Fig. 4. Monosyllable and bisyllable accuracy scores, given various assumptions about the proportion of syllable boundaries that are unambiguous to infants. At "none," no syllable boundaries are unambiguous; at 100, all syllable boundaries are unambiguous. The top plot shows English results, the bottom plot Dutch. Figures represent averages over the 70th to 80th percentile criterion range. The ascending bars show that the success of the conditional probability analysis in identifying words (and excluding nonwords) depends upon the clarity of syllable boundaries in the input.

auditory (rather than orthographic) infant-directed speech corpora. Some phonetic studies of adult-directed speech have established *articulatory strengthening* at the edges of prosodic domains, a phonetic enhancement of the speech signal that could aid in rendering some word boundaries unambiguous (e.g. Fougeron & Keating, 1997); other studies have show articulatory differences between consonants depending upon whether those consonants are onsets or codas (e.g. Byrd, 1996; Krakow, 1999). The fact that such information is present in the signal suggests that the resyllabification analyses presented above (at left in Fig. 4) may be too pessimistic.

Infants' placement of syllable boundaries might also be aided by statistical computations over phonetic categories. Although the model presented here takes syllables as undecomposed input units for statistical tabulation, infants are undoubtedly capable of detecting subsyllabic regularities (e.g. Jusczyk et al., 1994; Saffran & Thiessen, 2003), and computational models have demonstrated the theoretical feasibility of bootstrapping syllable boundaries based upon distributional regularities in phoneme category sequences (Cairns et al., 1997; Vroomen, van den Bosch, & de Gelder, 1998). At present experimental evidence showing that infants

take phonotactic probabilities into account in word segmentation is limited to infants of 9 months or older, but younger infants too may use some language-specific segmentation heuristics that are founded upon subsyllabic statistics (Mattys, Jusczyk, Luce, & Morgan, 1999). We noted above the utility of disfavoring parses resulting in consonant clusters that were unattested at utterance boundaries, but considerably more information than that is potentially available in the signal. One such source of information is the marked asymmetries in the frequencies of individual consonants at utterance onsets and offsets. For example, in the English corpus, /v/ was ten times more likely to appear as the final consonant in an utterance than as the initial consonant; /g/ was 17 times more likely to appear as an utterance onset than as an utterance offset. These trends were also true of words within the corpus. Thus, in principle, infants' detection of such disparities could lead to improved lexical segmentation.

If syllable boundaries are never clear to infants, their early sound-form vocabularies will include slightly more nonwords than words. These false alarms may present infants with two problems. The first is that if infants and young children are biased to interpret speech in terms of word-forms that sound familiar, missegmentations that have found a niche in the vocabulary may hinder the correct segmentation of future utterances, and lead children to seek meanings for word-forms that are not words. If a child believes that *shall we* is a word, for example, instances of this phrase will not contribute to the child's learning of *shall* and *we*; instead, they may lead to an unsound syntactic analysis.

A second problem is that false alarms distort the database from which infants derive the lexical phonology of their language. We have seen that false alarms do not imperil the trochaic bias; however, less robust phonological tendencies might be obscured. Some examples were mentioned previously, including the consonant clusters appearing in the English resyllabification's false alarms but not attested in the corpus itself. The reverse situation also occurred—patterns shown in the corpus but not in the postulated words, a situation that applied for most of the 3-consonant onsets, such as *skr-* and *str-*. These mismatches between English and the infant's developing vocabulary could lead to persistent segmentation errors, if (e.g.) the infant began to apply the false generalization that *skr* should not be parsed as a syllable onset.

That said, the overall phonotactic probabilities of the input corpus and the postulated words of the resyllabified corpus showed more similarities than differences, particularly among the more frequent onsets and offsets. For example, ranks of the token frequencies of the syllable onsets in the English base corpus and in the postulated words of the resyllabified corpus were correlated at 0.648. Considering only the 20 most frequent (by tokens) syllable onsets, which includes 95.3% of the onset syllables in the corpus, the correlation is 0.965. Similar values were found in calculations of offset consonant cluster frequencies. Thus, resyllabification did alter the likelihoods of some consonantal patterns, but aside from these relatively rare cases, the vocabulary postulated in the resyllabification analysis yielded a good overall picture of the true distribution of consonants and consonant clusters in word onset and offset positions. Knowledge of this sort has been shown to affect infants' parsing of speech (Mattys et al., 1999).

11. Analyses of smaller corpora

The analyses presented here did not assume that infants store the entire corpora in memory; rather, they assumed that frequency data compiled over several thousand utterances contribute to the degree to which infants consider speech sequences as familiar. This is not a radical claim. The infants tested by Jusczyk and Hohne (1997) remembered some previously-heard words for at least two weeks, a period in which they may have heard more intervening speech than that represented by the corpora examined here. Nevertheless, it is informative to consider whether a large number of utterances is required for this simple statistical procedure to detect words. This was tested by repeating the analyses over subsets of the original corpora.

Subset corpora of various lengths were created by selecting continuous blocks of text from the English and Dutch base corpora. The largest subset selected from the English base corpus corresponded to the 2926 utterances from the parent GL. This subcorpus contained 10,037 word tokens, thereby amounting to about 24% of the total English corpus. For purposes of comparison, a Dutch subcorpus of the same length (in word tokens) was created from the first 4039 utterances of the full Dutch base corpus. Additional, smaller corpora were sampled from these blocks and from other portions of the base corpora for both languages.

Summary results are shown in Fig. 5. All data are means over the percentile range from 70 to 80. Results for corpora smaller than 10,000 word tokens are means over two analyses from different portions of the Dutch and English corpora. The upper plot shows how accuracy varied with respect to corpus size; the middle plot shows how the number of hits varied with corpus size; and the lower plot shows the number of hits that were found relative to the number of word types in the corpus (a measure called *recall* or *completeness*; e.g. Batchelder, 2002). The most striking result, shown in the top panels, is how little accuracy was affected by reducing the number of word tokens in the analysis. In shrinking the corpus from 42,000 tokens to 500 tokens, English monosyllable accuracy declined only about 6%; Dutch monosyllable accuracy declined by only 13%. Reducing the corpus size actually increased accuracy in identifying bisyllabic words for most of the corpus sizes tested, for the percentile range from 70 to 80%.

However, smaller corpora yielded substantially fewer words (hits), as shown in the middle panels of Fig. 5. Thus, the chief cost of retaining a smaller amount of speech data was that fewer words were extracted. If a corpus of only 1000 tokens constituted the input to the statistical memory mechanism, only a few bisyllables would extracted at moderately high criterion levels (i.e. over the 70th to 80th percentiles).

This was partly a simple consequence of there being fewer words to find in the smaller corpora. As shown in the lower panels of Fig. 5, the proportion of hits relative to the number of types in the corpus (*completeness*) tended to *increase* for the monosyllables, demonstrating that the reduced hit count for smaller corpora was not due to a decline in efficiency, but was an effect of fewer words being available. On the other hand, bisyllable completeness tended to decline with smaller corpora. The opposite effects of corpus size on recall of monosyllables and bisyllables may derive from the fact that monosyllables were identified solely on the basis of frequency,

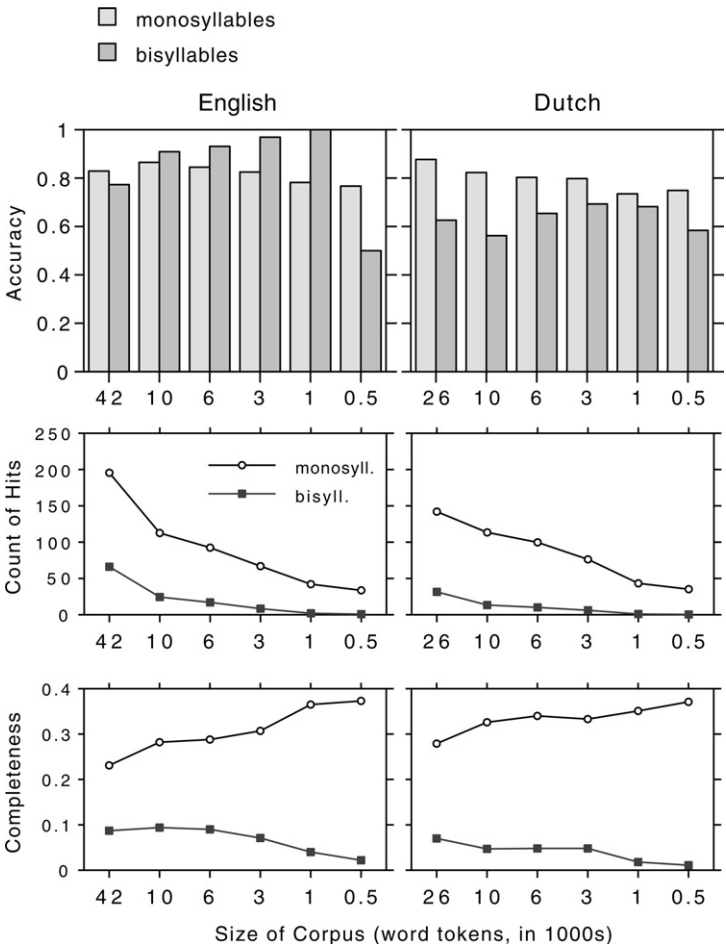

Fig. 5. Accuracy, count of hits, and completeness, for corpora of varying sizes. The leftmost data points for each language correspond to the full corpora.

whereas bisyllables were identified via frequency and conditional probability. Conditional probability is only a cue to wordhood when short words occur in varying environments, thereby deflating the consistency of monosyllable pairs that might be mistaken for bisyllabic words. When the size of the corpus decreases, short words have fewer opportunities to appear in a range of contexts, and as a result, pairs of monosyllabic words become harder to distinguish from bisyllabic words. Nevertheless, decreasing the corpus size had a modest impact on bisyllable accuracy.

Inspection of the stress characteristics of the retrieved bisyllables revealed that the trochaic pattern dominated the iambic and strong–strong patterns by a substantial margin for all corpora except one Dutch 500-word corpus, in which a clear majority did not emerge.

The general conclusion to be drawn from these analyses of smaller corpora is that tabulation of frequency and conditional probability over a small number of utterances can yield a set of sound sequences that are usually words (again, provided that syllable boundaries are not ambiguous), and that the bisyllables among these sound sequences usually exemplify the strong–weak stress pattern.

12. An analysis of very short utterances

Although the present results, together with those of Thiessen and Saffran (2003), suggest that the trochaic parsing bias may emerge as a generalization over statistically extracted bisyllables, the possibility remains that the bias is in fact derived via infants' attention to some other regularity in English or Dutch speech. The most prominent alternative hypothesis is that infants learn the dominant stress pattern of bisyllabic words by hearing these words in isolation (Jusczyk et al., 1999). If parents frequently say "doggie" or "mommy" or, for many children, the diminutive form of the child's name (e.g., "Mikey," "Susie") alone in sentences, children could learn the trochaic bias that way.

However, only a minority of the two-syllable utterances in the English corpus were trochaic. Strong–strong utterances dominated; furthermore, in some counts iambic bisyllables were more frequent than trochaic bisyllables. For example, counting bisyllable *tokens*, 65.5% were strong–strong; 14.2% were trochaic, 19.7% were iambic, and the remainder were weak–weak. Counting all English types, 71.3% were strong–strong, 21.2% were trochaic, and 6.4% were iambic. Similar results were obtained in counts of only the more frequent bisyllable types; for example, considering only the top 10% of bisyllables, 79% were strong–strong, compared to 13.2% trochaic and 7.9% iambic. Thus, while it is true that many English bisyllables were trochaic, the set of two-syllable utterances heard by infants contain too many sentences like *who's this?* and *oh dear* to permit the strong–weak pattern to stand out.

The Dutch corpus contained more trochaic utterances; in fact, the trochaic utterances showed a small lead over the strong–strong utterances in counts by tokens (49.6% versus 43.0%, with only 7.4% iambic). In counts of types, however, strong–strong utterances dominated (strong–strong, 61.3%; trochaic, 27.2%; iambic, 11.5%).

These results show that for English infants, the trochaic bias could be derived from a simple generalization over two-syllable utterances only if (a) the infant had a prior constraint against interpreting the strong–strong pattern as a template, and (b) the proportion of apparent words exemplifying the template could be as low as one-fifth. Dutch infants might arrive at the trochaic bias via the advantage of trochees in token counts, but even Dutch infants would need an a priori constraint against a strong–strong template in order to select the trochaic pattern on the basis of bisyllable types.

In our view, it is unlikely that infants would develop a trochaic parsing bias given only the sort of evidence provided by two-syllable utterances. Another possibility is that infants do rely on bisyllabic utterances to set their parsing preferences, and that infants do not have a trochaic bias at all—rather, infants may operate with a ranking

in which strong–strong parses are favored, followed by strong–weak parses, and then weak–strong parses. That is, infants may incorrectly consider strong–strong bisyllables to be "ideal" words. If this were correct, infants hearing a strong–strong–weak trisyllable would tend to segment it as a strong–strong word followed by a weak monosyllable. This possibility cannot be ruled out with the available experimental data, in which only oppositions between trochees and iambs have been tested systematically. If the strong–strong bias were true, it would put infants at a severe disadvantage in interpreting English and Dutch (in which strong–strong pairs are frequent but are almost never words), and it would contradict the evidence infants get from conditional-probability analyses of syllable sequences, as we have shown.

Thus, the intuitively plausible notion that infants can learn the typical prosodic form of bisyllabic words by attending to bisyllabic utterances is not well supported (see also Morgan, 1996b). By contrast, the numerical supremacy of the trochaic pattern in analyses that take bisyllable frequency and mutual information into account is robust in English and Dutch, even with small corpora and when syllable boundaries are assigned stochastically.

13. Discussion

In the first year of life, infants are exposed to a great deal of speech, the vast majority of which they cannot possibly understand. Yet this talking that infants hear lays the foundation for one-year-olds' enormous progress in learning how sounds are used to convey meaning. Most research on infant speech processing has been devoted to the acquisition of segmental structure: the discrimination of speech sounds and the construction of a language-specific inventory of phonetic categories. Over the past decade, however, a number of experiments have demonstrated that infants extract and retain not only speech sounds, but also word-length sequences of sounds from speech (e.g. Jusczyk & Hohne, 1997; Jusczyk & Aslin, 1995). Infants' biases to extract cohesive sequences (Aslin et al., 1998) and sequences that do not contain prosodic or other word-boundary cues (e.g. Gout et al., under review) could in principle lead to the extraction of a stock of word-forms that would then form the basis for the early vocabulary.

The present results indicate that the word-length sequences that English and Dutch infants retain from speech consist largely of actual words, provided that infants are able to identify many syllable boundaries. If about half of syllable boundaries are ambiguous to infants, then the sequences that infants extract as familiar candidate words may consist of about half actual words and half nonwords. Of the nonwords, the bisyllables tended to be pairs of monosyllabic words that parents often used together, though most of these pairs were not readily interpretable as fixed expressions with distinctive meanings.

Even under the assumption that all syllable boundaries are ambiguous, most of the extracted bisyllables exhibited the trochaic (strong–weak) stress pattern. This was also true in analyses over corpora as small as 1000 word tokens, for both English and Dutch. The dominance of the strong–weak pattern was not found in analyses of

bisyllables selected for high frequency alone, nor in analyses of maternal utterances containing only two syllables. This set of results supports the idea that infants acquire the trochaic parsing bias as a generalization over a ''protolexicon'' of word-forms extracted on the basis of the forms' relatively high conditional probability and frequency.

The present results complement previous modeling efforts that have used segmental input representations and other computational mechanisms (e.g., Batchelder, 2002; Brent & Cartwright, 1996; Cairns et al., 1997; Christiansen et al., 1998.) Taken together, previous studies have shown that phonological descriptions of speech to infants contain statistical regularities that provide word-boundary cues accessible to a range of computational learning systems. The current research does not speak to the plausibility of these previous models. Indeed, there is little in the models' *outcomes* that bears on how reasonably they reflect infants' processing mechanisms, because the available behavioral experiments set only very wide bounds on infants' actual segmentation performance. The merits of these models derive instead from the plausibility of their assumptions and the utility of the results for understanding language learning.

Here, infants were assumed to rely on syllable statistics for identifying words, based on experimental research suggesting that syllables are important units in young infants' judgments of similarity and numerosity. Subsyllabic regularities were proposed as possible aids to identification of syllable boundaries (though perhaps only in older infants) but not of words per se. Word learning was proposed to result from the emergence of some sequences from the surrounding speech as familiar forms, where familiarity results from frequency and from statistical cohesiveness. This notion of familiarity was intended as a simple, formally specified implementation of the empirical conclusion that infants compute conditional probabilities over speech materials (e.g. Saffran et al., 1996; Goodsitt et al., 1993).

Previous analyses of the same problem have usually used established computational techniques (e.g., data compression algorithms, recurrent connectionist networks) rather than the relatively ad hoc set of parameters and rankings used here. This may reflect a difference in the goals of the present project. Rather than serving as a demonstration that certain regularities can in principle be discovered by an efficient statistical mechanism, the current model is an attempt to determine what infants actually know, based on what they have been observed to do. Of course, any attempt at this more ambitious goal is less likely to get it right. Future experimental research may expose limitations of the present implementation and suggest additional constraints on models of infant statistical processing.

The generality of the results described here with respect to other languages remains to be seen. Both English and Dutch permit complex onsets and codas, which results in a large number of syllable types, many of which are words. Languages with fewer syllable types are likely to have more words that share syllables, which would tend to make statistical clustering harder. Also, it stands to reason that languages permitting fewer syllable types will tend to have larger proportions of multisyllabic words, a prediction borne out in a comparison of Spanish and English child-directed speech (Roark & Demuth, 2000). In English and Dutch, performance on trisyllabic

words was relatively poor. On the other hand, languages with simpler syllable types (e.g., only CV) might have syllables that are easier to identify, particularly once infants had discovered the limited range of possible syllable shapes. These considerations point to the importance of testing the model on a range of languages with different phonological properties.

13.1. The benefits of early word-form learning

What might infants have to gain by accumulating a stock of potential word-forms? We suggest that developing this "protolexicon" furthers language acquisition in three ways: by yielding parsing or segmentation heuristics; by obviating phonological encoding difficulties in early word learning; and by allowing the infant to begin tabulating lexical co-occurrence statistics that lay the foundations of syntax. We will discuss each of these in turn.

13.1.1. Building word segmentation strategies

The general principle that within-word sequences should be statistically stable relative to between-word sequences follows from the way that the lexicon is used generatively in sentence formation. Because utterances are often novel combinations of words, one can expect sequences of sounds that compose words to cohere statistically, while sequences crossing word boundaries should not (Harris, 1955).[11] This principle should hold for all languages in which words are markedly more stable than utterances, making the principle a good candidate as an initial parsing strategy in infancy. Of course, infants do not group statistically cohesive syllable sequences because this leads to words; infants group such sequences because statistical clustering is a basic attribute of interpretation, across domains and in both humans and other animals (e.g. Alloy & Tabachnik, 1984; Newport & Aslin, 2004; Rescorla, 1968). Infants hearing some languages may have less to gain from this kind of clustering. For example, highly inflected languages and agglutinative languages allow for considerably more word-form variation than English or Dutch, and the speech heard by infants in such environments might not exhibit the relevant regularities as clearly. Computational analysis of infant-directed speech in such languages would resolve this issue.

Infants who begin parsing speech according to conditional-probability criteria soon acquire a set of familiar sequences that may serve as rough phonological templates. For English and Dutch infants, these templates include information about the predominant stress pattern of bisyllabic words. Once the trochaic pattern is well exemplified in the set of familiar words, infants apparently *use* this pattern in the interpretation of speech (Echols et al., 1997; Johnson & Jusczyk, 2001; Jusczyk et al., 1999; Thiessen & Saffran, 2003). This account implies that each infant should pass through a stage in which he or she relies primarily upon statistical clustering in

[11] Harris actually discussed a procedure for finding *morphemes*, not words, but the principle is the same.

the interpretation of speech, before going on to use cues suggested by statistically derived words. Support for this contention may be found in a study by Thiessen and Saffran (2003), who used an artificial-language learning procedure in which the familiarization stimulus contained statistical cues to wordhood that conflicted with stress cues (specifically, high transitional probabilities between two syllables signaled statistical words, but these words were all iambic). Six- to seven-month-olds' responses were consistent with word extraction based upon conditional probability, whereas 9-month-olds relied upon trochaic stress (see also Johnson & Jusczyk, 2001).

Why should infants use phonological templates if statistical clustering works too? One advantage of using a strategy like trochaic bias (or the Metrical Segmentation Strategy) is that it may be applied to new words on-line. As Thiessen and Saffran (2003) point out, discovering a new word via distributional analysis may require several exposures to that word in different environments, whereas a stress bias can be applied immediately to novel and familiar words. Also, it is not necessary to assume that the precedence of stress cues when they are placed in conflict with statistical information implies that statistical information is no longer relevant to infants. Rather, infants come to balance various sources of information about likely word candidates as infants learn more about the statistical properties of the words that they have detected so far. Such sources of information include allophonic variation (Jusczyk, Hohne, & Baumann, 1999) and the likelihoods with which consonant sequences include word boundaries (Mattys et al., 1999). Counts of strong–weak bisyllables in the base corpora suggest that integrating stress cues and statistical cues would be useful: While only about 60% of trochees are words, 92.6% of the English trochees with frequency and mutual information above the 70th percentile are words, and 86.7% of the Dutch trochees above the 70th percentile are words. At the 75th percentile, these figures are 98.4% (English) and 82.8% (Dutch); at the 80th, 100% (English) and 85.0% (Dutch). Thus, statistical clustering and the consequent formation of an initial lexicon can lead to language-specific heuristics for efficient word finding. Some of these heuristics can be applied on-line, and are used from infancy to adulthood in parsing speech.

13.1.2. One-year-olds' limitations in word learning

Although infants of 6–8 months do learn associations between words and objects (Gogate & Bahrick, 2001; Tincoff & Jusczyk, 1999), word learning is generally assumed to begin within a month or two of the infant's first birthday (e.g. Bloom, 2000). One might sensibly question whether 8-month-olds' acquisition of word-forms is of any consequence if these infants do not yet learn meanings to go along with those word-forms.

We argue that learning word-forms in infancy *is* important, because it gives children a needed head start in assembling a vocabulary. Several recent studies of word learning have shown that when young 1-year-olds are confronted with a new phonetic sequence and a novel object, they may learn the association between the word and the object category, but may not successfully encode phonetic detail in the words. For example, Stager and Werker (1997) found that 14-month-olds who were habituated to the verbal labeling of a novel object with a word like *dih* did not disha-

bituate to a change in this label to *bih*, though they did dishabituate to more obvious changes. However, 8-month-olds did appear to notice the change in pronunciation. Werker and her colleagues argued that at 14 months, the encoding of both the phonetic and the visual features of a word-teaching presentation overwhelmed infants' attentional capacities, resulting in a phonetically underspecified lexical representation. By contrast, the younger infants were not learning the mapping between a word and an object and were therefore capable of encoding the phonetic details of the word (e.g. Werker & Fennell, 2004). Although the nature of the developmental change between 8 and 14 months is obscure, the result is consistent with other data showing 8-month-olds' accurate encoding of words in speech (e.g. Jusczyk & Aslin, 1995; but see Hallé & de Boysson Bardies, 1996).

When children have already learned the sound-form of a word, the word learning process is simplified: only the word's meaning needs to be worked out and linked to the sound-form. In a recent study demonstrating this process, Swingley (2002) showed 18- and 19-month-old children an animated film in which a novel sound-form was used several times, without its referent being presented. After this familiarization phase, the children were shown a novel object, which was labeled using either the familiarized word from the film (for half of the children), or an entirely novel word (for the other half), in ostensive sentences like *This is a tiebie*.[12] Children's learning was tested by showing the taught object and a second novel object on a screen, labeling the taught object (*Where's the tiebie?*), and monitoring children's gaze to it. On some trials, the novel word was mispronounced, as in *Where's the kiebie?*. Previous research has shown that for well-known words like *ball*, children's fixation to a denoted object is greater when the target word is correctly pronounced than when it is mispronounced (e.g. Swingley & Aslin, 2000, 2002). The same result was obtained in this word-teaching study, but only among those children who were taught the meaning for the word they had previously been familiarized with in the film. Thus, children who had had the opportunity to learn (e.g.) *tiebie* as a sound-form before learning its meaning evidently knew that *tiebie* was a good pronunciation and *kiebie* was not; children without prior exposure to the word were indifferent to subtle variations in pronunciation.

Although the generality of this phenomenon has not been tested yet, we suggest that much of the early vocabulary is acquired in such stages. First infants learn the sound-forms of some words using language-general statistical extraction heuristics (and later, language-specific heuristics), and then children figure out what those words mean. Of course, this is not the whole story. For many words, infants may first learn conceptual categories that correspond to the denotations of words; it's possible (and indeed probably normal) for infants to have a fair working knowledge of baby wipes and applesauce before knowing what they're called. Finally, some words may be encountered and learned on the spot (e.g. Woodward, Markman, & Fitzsimmons, 1994), though perhaps without all of their phonetic or conceptual

[12] The study was conducted in Dutch with Dutch-learning children; actual sentences included *Dit is een tiebie.*

details intact. Given our estimate that about 200 word-forms may be familiar to infants via statistical clustering, it seems reasonable to suppose that a large portion of children's "first words" were learned first just as word-forms in infancy.

Thus, even if infants have yet to marshal the cognitive or communicative tools needed for "true" word learning, the word-forms they extract can provide a phonetically complete basis for lexical entries during the first year, when children still have trouble encoding phonetic detail in meaningful new words.

13.1.3. The foundations of syntax

In addition to learning the meanings of words, one-year-old children begin to learn how sequences of words are used by speakers to convey more complex messages, according to the syntactic rules of the language. Because the computational machinery of syntax operates at higher levels of abstraction than the phonological level, syntax cannot be learned simply by refining co-occurrence statistics of syllables or words (e.g. Chomsky, 1957). However, the computation of co-occurrence statistics may lead infants to discover generalizations about syntactic categories or relations. For example, once infants have identified words in the speech stream, knowledge of the co-occurrence privileges of these words could, in principle, permit assignment of words to rudimentary syntactic categories (e.g. Kiss, 1973).

Recent studies evaluating the potential informativeness of lexical co-occurrence probabilities for syntactic category formation have assumed that the input to this distributional analysis consists of properly segmented words (Mintz, Newport, & Bever, 2002; Redington et al., 1998). Given these authors' goal of modeling category formation in children at least two years old, this assumption was not unreasonable. Nevertheless, we might ask whether the local distributional contexts of word-forms familiar to much younger infants also exhibit statistical properties that could in principle lead to grammatical category formation, even if at present little evidence indicates that infants possess the requisite analytical abilities.

We selected the English corpus generated under the assumption that 50% of syllable boundaries are ambiguous to infants, and based the present analysis on the results from the 75th percentile criterion using both frequency and mutual information. At this criterion level, there were 628 postulated words: 376 hits and 252 false alarms. Only the 100 most frequent postulated words were entered into the analysis, amounting to 79 real words and 21 false alarms. Along the lines of Redington et al. (1998) and Mintz et al. (2002), the frequencies with which each of these 100 words preceded or followed one another were tabulated. This yielded, for each word, a vector of 200 values. For example, the *you* vector listed (1) how many times *to* preceded *you* in the corpus; (2) how many times *to* followed *you*; (3) how many times *a* preceded *you*, and so on. False alarms and hits were treated identically; thus, one value in the matrix of vectors indicated how often *you* preceded /ɪŋ/. Sequences of speech that had not been postulated as words were ignored entirely in all computations.

Each of the 100 vectors was then converted to a vector of ranks, and a 100×100 matrix of the Spearman correlations of these rank vectors was computed (the correlation of two vectors provides a measure of how similar those vectors are; when

vectors (and therefore lexical contexts) are similar, they may refer to words having the same grammatical class). This matrix was submitted to an average-link hierarchical clustering analysis using the *hclust* function of the software package *R*, yielding a *dendrogram*, a tree of binary branches splitting at varying heights. The hierarchical clustering technique does not attempt to partition the words into equal-sized categories; rather, where words are similar to one another, they are placed together, and where the categories thus formed are similar, they are linked higher in the tree, until atop the tree all words are in a single category. A set number of categories can be identified by "cutting" the tree at a particular height and disregarding category divisions below that height. To conserve space, rather than present the dendrogram itself, we present a table that preserves some of the information in the dendrogram (see caption).

As Table 2 shows, some form-class similarities could be discerned in the categories extracted in the analysis of contexts, though these categories were not as "pure" as those found by Mintz et al. (2002) or Redington et al. (1998). Contractions that included *is* or *are* formed a cluster, as did pronouns. Most of the verbs were placed together in a cluster also containing a branch full of prepositions. The articles *a* and *the* were paired next to the couple *that* and *this*; the three question words included in the analysis, *what, who,* and *where*, were clustered together. Several other relationships of this sort may be found in the results.

Interestingly, many of the false alarms appeared together (see Table 2). There appeared to be two reasons for this. In some cases the false alarms were similar sorts of missegmentation. For example, /ɪŋ/ and /kɪŋ/ were parts of words containing *-ing* as a suffix; these words had similar following contexts (e.g., *to*, *a*, *on*, *that*,...), making these syllables neighbors. In other cases, missegmentations

Table 2
Results of hierarchical clustering analysis of postulated words' contexts in English corpus allowing 50% of syllable boundaries to undergo probabilistic resyllabification

a	mm, hmm
a	to
a	here, there; then; hey, now; dear; oh, ooh; yes; and, well
b	ya; my; that, this; a, the
b	what's; you're, it's; who's; that's, there's
b	what; who, where; he, she; you've; I; you; we, they
c	tɪs; t/l̩/, /zn̩t/
d	/mɑn/; /mɪ/, /lɪ/; /nə/, /mɪz/
d	girl, /dɪ/; /ləu/; right, /gæn/; /ɪŋ/, /kɪŋ/; it, me
e	don't; an; are, can; got; not, just; do; have, did; is, was; say; tell, see; like; want, get; in; on; up, out; at; for, with
f	/lɪ/; /bɛ/, /tɛ/; good, nice; eh; no, so; smile; one, way
f	go, come; /wɑ/, tickle; /ɪ/; /kʌ/; /mʌ/, /hæ/; /hə/; look, darling

Letters at left (a, b, ..., f) show the 6 categories highest in the tree; each row shows the contents of one of 12 categories formed by cutting the tree at an intermediate height. Words are listed in the order they appeared in the dendrogram; thus, words near each other here tended to be grouped together in the analysis. Pairs of words separated by commas rather than semicolons were classified together near the bottom of the tree.

tended to occur with the same other missegmentations; for example, /tl̩/ (which usually derived from *it'll* or *little*) and /zn̩t/ (which derived from *isn't*) were among the few syllables that followed the missegmented syllable /ɪ/; all the others were also false alarms.

In sum, statistical information relevant to the classification of word-forms into syntactic categories is present in the speech that infants hear, even when word boundaries are not given in the input and when half of the syllable boundaries are subject to probabilistic realignment. The possibility that infants perform similar computations is suggested by research with 12-month-olds showing that under some conditions infants compute the co-occurrence likelihoods of words first extracted using statistical clustering (Saffran & Wilson, 2003). In addition, 17-month-olds have been shown to form categories of words based on distributional evidence (Gerken, Wilson, & Lewis, under review). Although the categories that emerged in the present analysis were not perfect, they might offer infants a valuable head start in acquiring the syntax of their language.

14. Conclusions

Speech perception is a domain in which early experience has profound consequences. A mature speaker of any language cannot be said to perceive speech in a veridical or language-neutral way, because his perceptual system is tuned to match the demands of the language or languages he learned in infancy (e.g. Pallier, Christophe, & Mehler, 1997). This phenomenon has been studied most intensively in research on phonetic category perception in mature learners of second languages (e.g. Best, 1994; Flege, 1995) and in infants learning their mother tongue (e.g. Kuhl et al., 1992; Werker & Tees, 1984). It is generally agreed that these effects reflect the outcome of infants' computation of a distributional analysis of speech sounds (Behnke, 1998; Guenther & Gjaja, 1996), though it is not clear at present how these sounds are individuated (Beckman & Edwards, 2000).

Language-specific perception of speech is not limited to the phoneme, however; listeners tend to interpret speech according to a range of probabilistic criteria, such as the likelihood of two consonants occurring together in particular syllable positions (e.g. Massaro & Cohen, 1983), or the likelihood that a strong syllable serves as the onset or the offset of a word (e.g. Cutler & Butterfield, 1992). The latter case is particularly interesting because the appropriate unit of analysis for forming the relevant generalization (e.g. the generalization that that English or Dutch bisyllabic words are usually strong–weak) is the word. Words, or some reasonably accurate proxy for words, must be available to infants before they begin to show language-specific prosodic parsing biases at 8 months, just as phonemes, or some proxy thereof, must be available for analysis before infants begin to show language-specific categorization biases at around 6 months (e.g. Kuhl et al., 1992). Here we have shown that the selection of syllable sequences high in frequency and mutual information yields bisyllables that largely conform to the trochaic pattern, in both English and Dutch. The robustness of this result may help account for the surprisingly young

age at which infants are guided by the trochaic bias. Given that parsing biases constrain speech segmentation, infants should be expected to employ them only when the evidence in their favor is very strong.

As noted above, infants probably do not make wholesale leaps from one parsing strategy to another in development; rather, infants gradually accrue biases that are supported by the outcome of previous analyses, until a wide range of cues is used. At present, little is known about this process, or about the linguistic data that infants gain. The present analyses suggest that infants of 7 months or so learn a fairly representative sampling of the words they hear, at least in terms of the words' length (mostly monosyllables, plus some bisyllables, and a few trisyllables, reflecting the input) and the words' form classes (for English, about 1/3 nouns, 1/4 action verbs, 1/5 adjectives or adverbs, and so on, as in the input). Infants also make mistakes: even in the most optimistic "base corpus" analyses here, infants were predicted to make at least one false identification for every four correctly identified words. More false alarms were predicted under less idealistic assumptions about the identifiability of syllable boundaries.

One of the mysteries of language acquisition is the rapid pace with which language is learned, given the young child's apparent limitations in various cognitive domains. Part of the solution to this mystery is that children are not as limited as once thought, in domains as diverse as intentional understanding (e.g. Gergely, Bekkering, & Király, 2002), knowledge of the physical world (e.g. Baillargeon & Wang, 2002), and category learning (e.g. Younger, 2003). Another part of the solution is that children get a head start in language learning: When they start to talk at around 12 months, children have at hand the benefit of a great deal of language-specific knowledge that is available from analyses of sound structure. It is clear that infants are a long way toward "solving" the problem of phonetic category learning, and are already making progress in building a lexicon, albeit a bare lexicon of forms largely without meanings. It seems likely that by 12 months, infants have some knowledge of word combinations, and possibly even some intuitions about which words fit into similar sentence positions.

The notion that children's experience in infancy might be so important to the pace of language development owes its plausibility to a large number of recent studies illustrating infants' remarkable capacities in coping with variability in the speech signal and in extracting and remembering word-forms (Jusczyk, 1997; recent examples include Johnson, Jusczyk, Cutler, & Norris, 2003; Soderstrom et al., 2003). At the same time, computational models testing the potential utility of various kinds of learning mechanism for word identification have proclaimed the possible benefits of using multiple phonetic cues (e.g. Brent & Cartwright, 1996; Christiansen et al., 1998; Jusczyk, 1997; Morgan & Saffran, 1995). The message that consistently emerges from this work is that infants improve in their ability to utilize various sources of phonetic information, and that these improvements could provide some benefit to infants in word-finding.

To date, however, it has proven difficult to use these findings to make concrete proposals about the link between knowledge gained in infancy, and the further course of language acquisition. The problem is that knowing only the general

tendency of infants' progress in using language-specific cues more efficiently is not sufficient for modeling the transition from the primarily phonetic processing that occupies infants in the first year, to the central linguistic problems children face thereafter. To understand this transition, reasonable estimates of the *contents* of children's knowledge are required: not only whether infants know word-forms, but how many, and which ones. Better models of the contents of children's knowledge will permit specific tests of children's understanding of constructions predicted to be familiar, shedding light on the obscure period when children speak little, but may know a great deal.

In our view, significant advances in relating infants' learning about speech to later language acquisition are contingent upon progress in three domains: the measurement of phonetic features in infant-directed speech corpora; perceptual experiments evaluating infants' treatment of the phonetic features present in the corpora; and computational modeling integrating what is known about infants and what is known about infant-directed speech.

The present paper, an attempt at the last of these, helps demonstrate the need for refinement in all three domains. For example, the resyllabification analyses showed that the accuracy of word-finding depends strongly upon the availability of syllable boundaries. This is a parameter that could be profitably explored with perceptual experiments (Mattys & Jusczyk, 2001). Further, it is likely that many of the word boundaries that were ambiguous in the corpora used here would in fact be unambiguously marked as prosodic phrase boundaries (Gout et al., under review; Nazzi et al., 2000; Soderstrom et al., 2003). Determining the frequency of these unambiguous boundaries would require detailed analysis of speech recordings. Finally, computational models of early word acquisition, including the present model, have had a number of shortcomings, including the failure to specify why developmental changes in infant's speech processing occur when they do. In our view, progress in each of these three areas separately is entirely feasible with existing techniques. The study of language learning will profit most if these advances are considered together, to produce increasingly accurate models of what infants know about language and how this knowledge changes in development.

Acknowledgments

The author would like to thank Maarten Jansonius for discussion and for help with programming. The portion of this work completed at the MPI was supported by the Max-Planck-Gesellschaft and by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek's Spinoza award to Anne Cutler.

Appendix A. English false alarms

The following nonword sequences were postulated as words in the English base analysis at the 70th percentile. This analysis used mutual information and frequency

criteria to postulate words in a corpus in which no consonants were resyllabified across word boundaries. For ease of interpretation, partial words are filled out here in parentheses. For example, the false alarm *the ma* is listed here as *the ma(tter)*. When a false alarm originated from several lexical sources, a representative set is listed, separated by slashes. False alarms are grouped by length and ordered by frequency (most frequent first). Note that although orthographic spelling is used here, actual computations were done over phonetic spellings.

monosyllable false alarms. (a)gain, be(tter)/be(lly), (a)bout, (fan)cy/(me)ssy, (ha)ppe(ning)/(jum)per, beau(tiful), (a)ble/(trou)ble, pu(tting)/pu(lling), af(ter), (pee)ka(boo), (go)in'/in(sult), wa(ter), (kee)ping/(hel)ping, (remem)ber/(tu)ba, pee(kaboo)/peo(ple), (they)'ll/(tow)el, ta(king)/ta(sty), gi(ving)/gi(ven), (exci)ted/(wan)ted, po(ppit)/po(ppers), (fin)ger/(lon)ger, (e)ven/(gi)ven, (mi)ster/(ye)ster(day), swee(tie)/swee(per), mo(bile)/mo(ment), (peo)ple/(cou)ple, le(ggies)/le(mon), o(kay), ye(llow)/ye(sterday), dri(bbles), (o)pen, su(cking), se(cond), (be)fore/fa(lling), bu(sy)/bi(scuits), (bo)ring/(wonde)ring, (ye)llow/(swa)llow, pe(nny)/pe(tal), do(lly), (ea)ten/(bu)tton, (dan)cing/(dre)ssing, a(llen)/a(pple), na(sty).

bisyllable false alarms. come on, tell me, who's this, good girl, the ma(tter), shall we, you're not, look at, your mum, you've got, I am, play with, smile for, big smile, what's wrong, your fit, big fat, you've had, your mouth, 'tis so, good boy, give us, clap your, we've got, yum yum, you've been, your bib, the sun/sun(shine), see if, your pram, your arms, don't think, out of, smiles for, exci(ted/ting), sit up, no need, nice clean, let's get, let me, cross-patch, you're su(pposed), your dad, bash him.

trisyllable false alarms. (ti)ckle tickle, tickle ti(ckle), mummy's here, looking at, my goodness, (all)right darling, clever boy, got any, allright da(rling), funny face, what-chou loo(king), (what)-chou looking, sleepy head, (mo)mmy's gir(lie), the came(ra), want some more, funny noise, (in)terested in, can't see them, goodness me, the other, quack quack quack, the hiccups, little bit, had enough.

Appendix B. Consonant assignment probabilities of the stochastic syllable parser

The following table indicates the likelihoods set by the probabilistic resyllabifier for each possible placement of a syllable boundary between two adjacent syllables, according to the number of consonants between the vowels of the syllables, and according to the stress pattern of the two syllables. These probabililties were not determined by behavioral experiments; nor were they adjusted to maximize model performance. The resyllabifier operated without information about the correct syllable boundary.

One consonant

	V.CV	VC.V
Strong–strong	.75	.25
Weak–strong	.85	.15
Strong–weak	.50	.50
Weak–weak	.50	.50

Two consonants

	V.CCV	VC.CV	VCC.V
Strong–strong	.15	.80	.05
Weak–strong	.20	.80	.00
Strong–weak	.10	.80	.10
Weak–weak	.10	.80	.10

Three consonants

	V.CCCV	VC.CCV	VCC.CV	VCCC.V
All stress patterns	.10	.40	.40	.10

Four consonants

	V.CCCCV	VC.CCCV	VCC.CCV	VCCC.CV	VCCCC.V
All stress patterns	.05	.10	.0	7.10	.05

References

Allen, G. D., & Hawkins, S. (1980). Phonological rhythm: Definition and development. In G. H. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child phonology: Vol. 1. Production* (pp. 227–256). New York: Academic Press.

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review, 91*, 112–149.

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

Aslin, R. N., Woodward, J. C., LaMendola, N. P., & Bever, T. G. (1996). Models of word segmentation in speech to infants. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 117–134). Hillsdale, NJ: Erlbaum.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database (CD-ROM)*. Philadelphia: Linguistic Data Consortium.

Baillargeon, R., & Wang, S. (2002). Event categorization in infancy. *Trends in Cognitive Sciences, 6*, 85–93.

Batchelder, E. O. (1997). Computational evidence for the use of frequency information in discovery of the infant's first lexicon. Unpublished doctoral dissertation, The City University of New York, New York.

Batchelder, E. O. (2002). Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition, 83*, 167–206.

Beckman, M. E., & Edwards, J. (2000). The ontogeny of phonological categories and the primacy of lexical learning in linguistic development. *Child Development, 71*, 240–249.

Behnke, K. (1998). The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. Unpublished doctoral dissertation, MPI Series in Psycholinguistics 5.

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

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

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development, 4*, 247–260.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception* (pp. 167–224). Cambridge, MA: MIT Press.

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

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Bomba, P. C., & Siqueland, E. R. (1983). The nature and structure of infant categories. *Journal of Experimental Child Psychology, 35*, 294–328.

Brent, M. R. (1996). Advances in the computational study of language acquisition. *Cognition, 61*, 1–38.

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

Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *Journal of Phonetics, 24*, 209–244.

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

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

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

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology I: between the grammar and physics of speech* (pp. 283–325). Cambridge: Cambridge University Press.

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

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

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

Dahan, D., & Brent, M. R. (1999). On the discovery of novel wordlike units from utterances: An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General, 128*, 165–185.

Echols, C. H., Crowhurst, M. J., & Childers, J. B. (1997). The perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language, 36*, 202–225.

Eimas, P. D. (1997). Infant speech perception: Processing characteristics, representational units, and the learning of words. In R. L. Goldstone, P. Schyns, & D. E. Medin (Eds.). *The psychology of learning and motivation* (Vol. 36, pp. 127–169). San Diego: Academic Press.

Eimas, P. D. (1999). Segmental and syllabic representations in the perception of speech by young infants. *Journal of the Acoustical Society of America, 105*, 1901–1911.

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

Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: Is the melody the message?. *Child Development, 60*, 1497–1510.

Flege, J. E. (1995). Second language speech learning. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). Baltimore, MD: York Press.

Fodor, J. A., Garrett, M. F., & Brill, S. L. (1975). Pi ka pu: The perception of speech sounds by prelinguistic infants. *Perception and Psychophysics, 18*, 74–78.

Fougeron, C., & Keating, P. A. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America, 101*, 3728–3740.

Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature, 415*, 755.

Gerken, L. A., Jusczyk, P. W., & Mandel, D. R. (1994). When prosody fails to cue syntactic structure: Nine month olds' sensitivity to phonological versus syntactic phrases. *Cognition, 51*, 257–265.

Gerken, L. A., Wilson, R., & Lewis, W. (under review). 17-month-olds can use distributional cues to form syntactic categories.

Gogate, L. J., & Bahrick, L. E. (2001). Intersensory redundancy and 7-month-old infants' memory for arbitrary syllable-object relations. *Infancy, 2*, 219–231.

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

Goslin, J., & Frauenfelder, U. H. (2000). A comparison of theoretical and human syllabification. *Language and Speech, 44*, 409–436.

Gout, A., Christophe, A., Millotte, S., & Morgan, J. (under review). Phonological phrase boundaries constrain lexical access: II. infant data.

Guenther, F. H., & Gjaja, M. N. (1996). The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America, 100*, 1111–1121.

Hallé, P. A., & de Boysson Bardies, B. (1996). The format of representation of recognized words in the infants' early receptive lexicon. *Infant Behavior and Development, 19*, 463–481.

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

Hayes, J. R., & Clark, H. H. (1970). Experiments in the segmentation of an artificial speech analog. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 221–234). New York: Wiley.

Hayes, R. A., Slater, A., & Brown, E. (2000). Infants' ability to categorise on the basis of rhyme. *Cognitive Development, 15*, 405–419.

Hillenbrand, J. (1983). Perceptual organization of speech sounds by infants. *Journal of Speech and Hearing Research, 26*, 268–282.

Houston, D. M., Jusczyk, P. W., Kuijpers, C., Coolen, R., & Cutler, A. (2000). Cross-language word segmentation by 9-month-olds. *Psychonomic Bulletin and Review, 7*, 504–509.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44*, 548–567.

Johnson, E. K., Jusczyk, P. W., Cutler, A., & Norris, D. (2003). Lexical viability constraints on speech segmentation by infants. *Cognitive Psychology, 46*, 65–97.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

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

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

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

Jusczyk, P. W., Goodman, M. B., & Baumann, A. (1999). Nine-month-olds' attention to sound similarities in syllables. *Journal of Memory and Language, 40*, 62–82.

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

Jusczyk, P. W., Hohne, E. A., & Baumann, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics, 61*, 1465–1476.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39*, 159–207.

Jusczyk, P. W., Jusczyk, A. M., Kennedy, L. J., Schomberg, T., & Koenig, N. (1995). Young infants' retention of information about bisyllabic utterances. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 822–836.

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

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

Kelly, M., & Martin, S. (1994). Domain-general abilities applied to domain-specific tasks: Sensitivity to probabilities in perception, cognition, and language. *Lingua, 92*, 105–140.

Kiss, G. R. (1973). Grammatical word classes: A learning process and its simulation. *Psychology of learning and motivation, 7*, 1–41.

Korman, M. (1984). Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. *First Language, 5*, 44–55.

Krakow, R. A. (1999). Physiological organization of syllables: A review. *Journal of Phonetics, 27*, 23–54.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*, 606–608.

Liberman, A. M., Cooper, F. S., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431–461.

Liberman, I. Y., Shankweiler, D., Fischer, F. W., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology, 18*, 201–212.

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

de Marcken, C. (1996). *The unsupervised acquisition of a lexicon from continuous speech* (Tech. Rep.). MIT AI Laboratory and Center for Biological and Computational Learning.

Massaro, D. W., & Cohen, M. M. (1983). Phonological context in speech perception. *Perception and Psychophysics, 34*, 338–348.

Mattys, S. L., & Jusczyk, P. W. (2001). Do infants segment words or recurring contiguous patterns?. *Journal of Experimental Psychology: Human Perception and Performance, 27*, 644–655.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38*, 465–494.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1–86.

Mehler, J., Dupoux, E., & Segui, J. (1990). Constraining models of lexical access: The onset of word recognition. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 236–262). Cambridge, MA: MIT Press.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*, 393–424.

Morgan, J. L. (1996a). Prosody and the roots of parsing. *Language and Cognitive Processes, 11*, 69–106.

Morgan, J. L. (1996b). A rhythmic bias in preverbal speech segmentation. *Journal of Memory and Language, 35*, 666–688.

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

Nazzi, T., Kemler Nelson, D. G., Jusczyk, P. W., & Jusczyk, A. M. (2000). Six month olds' detection of clauses embedded in continuous speech: Effects of prosodic well-formedness. *Infancy, 1*, 123–147.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48*, 127–162.

Olivier, D. C. (1968). Stochastic grammars and language acquisition mechanisms. Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Pallier, C., Christophe, A., & Mehler, J. (1997). Language-specific listening. *Trends in Cognitive Sciences, 1*, 129–132.

Perruchet, P., & Vinter, A. (1998). Parser: A model for word segmentation. *Journal of Memory and Language, 39*, 246–263.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: MIT Press.

Polka, L., & Sundara, M. (2003). Word segmentation in monolingual and bilingual infant learners of English and French. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1021–1024).

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 421–435.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*, 425–469.

Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology, 66*, 1–5.

Roark, B., & Demuth, K. (2000). Prosodic constraints and the learner's environment: A corpus study. In S. C. Howell, S. A. Fish, & T. Keith-Lucas (Eds.), *Proceedings of the 24th annual conference on language development* (pp. 597–608). Cascadilla Press.

Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science, 26*, 113–146.

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

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology, 39*, 484–494.

Saffran, J. R., & Wilson, D. P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy, 4*, 273–284.

Soderstrom, M., Seidl, A., Kemler Nelson, D., & Jusczyk, P. W. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language, 49*, 249–267.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388*, 381–382.

Svartvik, J., & Quirk, R. (1980). *A corpus of English conversation*. Lund: London Studies in English.

Swingley, D. (1999). Conditional probability and word discovery: A corpus analysis of speech to infants. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the 21st annual conference of the cognitive science society* (pp. 724–729). Mahwah, NJ: LEA.

Swingley, D. (2002). *On the phonological encoding of novel words by one-year-olds.* Paper presented at the 27th Annual Boston University Conference on Language Development. Boston.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition, 76*, 147–166.

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science, 13*, 480–484.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39*, 706–716.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science, 10*, 172–175.

Treiman, R., & Danis, C. (1988). Syllabification of intervocalic consonants. *Journal of Memory and Language, 27*, 87–104.

Vihman, M. M., & Velleman, S. L. (2000). The construction of a first phonology. *Phonetica, 57*, 255–266.

Vroomen, J., & de Gelder, B. (1995). Metrical segmentation and lexical inhibition in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 98–108.

Vroomen, J., van den Bosch, A., & de Gelder, B. (1998). A connectionist model for bootstrap learning of syllable structure. *Language and Cognitive Processes, 13*, 193–220.

van de Weijer, J. (1998). *Language input for word discovery.* Unpublished doctoral dissertation, MPI Series in Psycholinguistics 9.

Werker, J. F., & Fennell, C. T. (2004). Listening to sounds versus listening to words: Early steps in word learning. In D. G. Hall & S. Waxman (Eds.), *Weaving a lexicon* (pp. 79–109). Cambridge, MA: MIT Press.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49–63.

Wolff, J. G. (1977). The discovery of segments in natural language. *British Journal of Psychology, 68*, 97–106.

Woodward, A. L., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning in 13- and 18-month-olds. *Developmental Psychology, 30*, 553–566.

Younger, B. A. (2003). Parsing objects into categories: Infants' perception and use of correlated attributes. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development: Making sense of the blooming, buzzing confusion* (pp. 77–102). London: Oxford University Press.