COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 35 (2011) 119–155

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2010.01160.x

Learning Diphone-Based Segmentation

Robert Daland,[a] Janet B. Pierrehumbert[b]

[a]*Department of Linguistics, UCLA*
[b]*Department of Linguistics, Northwestern University*

Received 16 August 2009; received in revised form 24 February 2010; accepted 28 April 2010

Abstract

This paper reconsiders the diphone-based word segmentation model of Cairns, Shillcock, Chater, and Levy (1997) and Hockema (2006), previously thought to be unlearnable. A statistically principled learning model is developed using Bayes' theorem and reasonable assumptions about infants' implicit knowledge. The ability to recover phrase-medial word boundaries is tested using phonetic corpora derived from spontaneous interactions with children and adults. The (unsupervised and semi-supervised) learning models are shown to exhibit several crucial properties. First, *only a small amount of language exposure* is required to achieve the model's ceiling performance, equivalent to between 1 day and 1 month of caregiver input. Second, the models are *robust* to variation, both in the free parameter and the input representation. Finally, both the learning and baseline models exhibit *undersegmentation*, argued to have significant ramifications for speech processing as a whole.

Keywords: Language acquisition; Word segmentation; Bayesian; Unsupervised learning; Computational model

1. Introduction

Word learning is fundamental in language development. Aside from communicating lexical meaning in individual utterances, words play a role in acquiring generalizations at multiple levels of linguistic structure, for example, phonology[1] and syntax.[2] Therefore, it is crucially important to understand the factors and processes that shape word learning.

In order to learn a word, the listener must first parse the wordform out as a coherent whole from the context in which it was uttered—word segmentation. Word segmentation is a challenging phenomenon to explain, as word boundaries are not reliably marked in everyday speech with invariant acoustic cues, such as audible pauses (Lehiste, 1960).

Correspondence should be sent to Robert Daland, Department of Linguistics, 3125 Campbell Hall, UCLA, Los Angeles, CA 90095-1543. E-mail: rdaland@humnet.ucla.edu

Therefore, listeners must exploit some kind of language-specific knowledge to determine word boundaries.

In adults, one obvious source for word segmentation is recognition of neighboring words: The end of one word signals the onset of the next, and vice versa. Indeed, a number of computational models such as TRACE (McClelland & Elman, 1986) and Shortlist B (Norris & McQueen, 2008) have explained word segmentation as an epiphenomenon of word recognition in closed-vocabulary tasks, such as an adult might face in a familiar listening environment. Word segmentation in adults is facilitated by recognition of specific words and other ''top-down'' (syntactic/semantic and pragmatic/world) knowledge (Mattys, White, & Melhorn, 2005), which may even override lexical/phonological information (e.g., Levy, 2008). Thus, ''top-down'' knowledge plays a vital role in adult word segmentation.

However, the acquisition facts suggest that word recognition cannot be the only—or even the most important—mechanism for infant word segmentation. This is evident from the fact that infants do not command very much top-down knowledge that might support word segmentation. For example, infants between the ages of 6 and 12 months are reported to know an average of 40–80 word types (Dale & Fenson, 1996), a tiny fraction of the words they encounter. During this same developmental period infants exhibit robust word segmentation, apparently on the basis of low-level cues such as phonotactics and stress (Aslin, Saffran, & Newport, 1998; Jusczyk, Hohne, & Bauman, 1999; Jusczyk, Houston, & Newsome, 1999; Mattys & Jusczyk, 2001; Saffran, Aslin, & Newport, 1996). While word recognition clearly plays some role in infant word segmentation (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005), *word segmentation organizes and supports infant word recognition and learning* (Davis, 2004), rather than being only an epiphenomenon of word recognition.

These facts call for a *phonotactic* account of word segmentation acquisition. Phonotactics refers to tacit knowledge of possible/likely sound sequences, including words, syllables, and stress (Albright, 2009; Chomsky & Halle, 1965; Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999; Hayes & Wilson, 2008; Jusczyk, Luce, & Charles-Luce, 1994). Although phonotactics can refer to a broad array of sound structures, the present paper will focus on segmental sequences and their distribution within and across words. More specifically, this paper explores Diphone-Based Segmentation (DiBS) as previously studied in Cairns, Shillcock, Chater, and Levy (1997) and Hockema (2006).

The underlying idea of DiBS is that many diphones are good cues to the presence or absence of a word boundary. For example, the sequence [pd] occurs in no or almost no English words, so it is a strong cue to the presence of a word boundary between [p] and [d]. Similarly, the sequence [ba] occurs very frequently within English words, but only rarely across word boundaries, so it is a strong cue to the absence of a word boundary. This idea can be formalized as calculating, for every diphone [xy] that occurs in the language, the probability $p(\# \mid xy)$ that a word boundary # falls between [x] and [y].

The DiBS models in previous studies were *supervised* models, meaning that model parameters were estimated from phonetic transcriptions of speech in which the presence/absence of word boundaries was marked. Since this is precisely what infants are trying

to discover, supervised models are not appropriate as models of human acquisition, which is *unsupervised*. Thus, despite the promising segmentation performance of these models, they have attracted little follow-up research, apparently because the model parameters were regarded as unlearnable.

The computational literature shows that when model parameters cannot be directly inferred, they can often be indirectly inferred using Bayes' theorem with reasonable prior assumptions (Manning & Schütze, 1999). The Bayesian approach is especially appropriate for the study of language acquisition because it forces a principled distinction between learner assumptions and the data that the child learns from.

Accordingly, a learning DiBS model that uses Bayes' theorem to estimate parameters is developed here. The approach builds off acquisition literature documenting children's use of phonotactics for word segmentation; specifically DiBS formalizes the finding that children exploit diphones to segment words from unfamiliar sequences (Mattys & Jusczyk, 2001). In estimating parameters, the learning model exploits the fact that phrase boundaries contain distributional information useful for identifying word boundaries (Aslin, Woodward, LaMendola, & Bever, 1996).

The paper is structured as follows. In the background section, we begin with terminology. Next we describe previous computational approaches to word segmentation; then we consider evidence of phonotactic segmentation in infants. Finally, we argue for phonotactic segmentation as a prelexical process in a two-stage (prelexical/lexical) theory of speech processing. In the DiBS section, we begin with our cognitive assumptions and next describe the core learning model; two specific instantiations are introduced: Phrasal-DiBS bootstraps model parameters from the distribution of speech sounds at phrase edges; Lexical-DiBS estimates them from the infant's lexicon. Phrasal-DiBS is an unsupervised algorithm, and lexical-DiBS can be characterized as semi-supervised. The remainder of the paper is devoted to testing the models and discussion of their performance. Simulation 1 uses a phonetic corpus derived from child-directed speech to assess the learning models' ability to recover phrase-medial word boundaries using the supervised model of Cairns et al. (1997) as a baseline. Simulation 2 assesses the models' robustness to variation in the parameter $p(\#)$, the learner's estimate of the global probability of a phrase-internal word boundary. Finally, Simulation 3 assesses robustness to pronunciation variation using a corpus of spontaneous adult speech that represents the phonetic outcome of conversational reduction processes.

2. Background

Word segmentation has been the focus of intensive cross-disciplinary research in recent years, with important contributions from infant experiments (e.g., Saffran et al., 1996; Mattys & Jusczyk, 2001), corpus studies of caregiver speech (e.g., van de Weijer, 1998), computational models (Christiansen, Allen, & Seidenberg, 1998; Fleck, 2008; Goldwater, 2006; Swingley, 2005), or combinations of these methods (Aslin et al., 1996; Brent & Siskind, 2001). Rapid further progress depends on integrating the insights from these

multiple strands of research. In this section, we begin by defining terminology. Next we describe several classes of computational models implementing a variety of theoretical approaches to the acquisition of word segmentation. Then we review evidence of phonotactic segmentation in infants and argue that existing approaches fail to accommodate it. Finally we argue that phonotactic segmentation is a prelexical process.

2.1. Terminology

2.1.1. Units of speech perception

The phonetic categories that infants perceive will be referred to as *phones*. ''Phone'' is used in preference to ''phoneme'' or ''allophone'' because these terms imply to some readers that infants learn the full system of contextual variation and lexical contrast relating cognitive units (phonemes) with their phonetic realization (allophones). For example, alveolar taps, aspirated stops, voiceless unaspirated stops, and unreleased stops are all allophones of the same phoneme /t/. It is nontrivial to learn these relations (Peperkamp, Le Calvez, Nadal, & Dupoux, 2006), and there is no unambiguous evidence that prelexical infants have done so (Pierrehumbert, 2002); hence, the more neutral term ''phone.''

2.1.2. Undersegmentation and oversegmentation errors

An undersegmentation error occurs when there is an underlying word boundary in the input, but the model fails to identify it. An oversegmentation error occurs when there is not an underlying word boundary, but the model identifies one. These terms are used because they refer directly to the perceptual outcome for the infant[3]: In the former case the infant will perceive an unanalyzed whole that underlyingly consists of multiple words; in the latter case the infant will improperly split up a single word into subparts.

2.2. Computational models of segmentation acquisition

Computational models can be regarded as specific instantiations of broader theories, making more specific and sometimes more easily testable predictions than the theories they embody. A variety of modeling frameworks have been proposed for the acquisition of word segmentation, including phonotactic models, connectionist models, and models which treat word segmentation and word learning as a joint-optimization problem. These models differ not only in their internal structure, and in what information they bring to bear, but also in the task they are solving; some are designed to acquire a lexicon as well as segment speech.

2.2.1. Diphone and higher n-phone models

Cairns et al. (1997) used the London-Lund corpus of spoken conversation to test a diphone model. For each diphone [*xy*] in the language, they collected the frequency $f_{\#}(xy)$ with which [*xy*] spans a word boundary, and the frequency $f_{\sim\#}(xy)$ with which [*xy*] occurs word-internally. Then the probability of a word boundary between [*x*] and [*y*] is $p(\# \mid xy) = f_{\#}(xy)/(f_{\#}(xy)+f_{\sim\#}(xy))$.[4] By positing a boundary whenever this probability exceeded

an optimal threshold, the model found 75% of the true word boundaries in the corpus, with only 5% of nonboundaries misidentified as word boundaries. The high level of performance was later explained by Hockema's (2006) finding that English diphones contain a great deal of positional information because most occur within a word, or across word boundaries, but not both.

Cairns et al. (1997) did not regard the diphone model as a suitable model for acquisition, because calculating the model parameters depends on knowing the relative frequency with which word boundaries span different diphones. Observing this information would require knowing when a word boundary has occurred and when it has not, which is precisely the segmentation problem infants are trying to solve. Swingley (2005) followed up with a word-learning model using a related statistic that is observable to infants. Although this model achieved promising results on word learning, it also made a number of ad hoc assumptions that may not be cognitively plausible (for discussion see Goldwater, 2006).

Other studies have revisited the assumption that diphone models are unlearnable. The key insight is that phrase boundaries are always word boundaries (Aslin et al., 1996). Thus, while infants may not observe which diphones span a word boundary phrase-medially, they can observe which phones are likely to occur phrase-initially and -finally. Xanthos (2004) exploited this idea by defining ''utterance-boundary typicality'' as the ratio of the expected probability of a diphone across phrase boundaries to the observed probability within a phrase. This method crucially assumes independence of phonological units *across* word boundaries. Going a step further, Fleck (2008) used Bayes' theorem to derive word-boundary probabilities with the further, counterintuitive, assumption of phonological independence *within* words. Statistical dependencies in this model are represented using all n-phones, $n \leq 5$, that occur more than five times in the corpus, so the model is more powerful than a diphone model and requires correspondingly stronger assumptions about infants' cognitive abilities. Fleck's model also includes a lexical process that repairs morphologically driven segmentation errors, for example, boundaries between stems and suffixes.

To anticipate briefly, the present study includes elements from several of these studies. It shares the *core diphone model* from Cairns et al. (1997) and Hockema (2006). From Aslin et al. (1996) and Xanthos (2004) it draws the idea of using *utterance boundary distributions to estimate word boundary distributions*, although it goes beyond these works in offering a principled probabilistic formulation. And in common with Fleck (2008), this work uses *Bayes' theorem to bootstrap model parameters*, although it is a leaner model, because it uses only diphones and does not also attempt to learn words.

2.2.2. Connectionist models

A number of researchers have proposed connectionist models of word segmentation, generally using the Simple Recurrent Network first defined in Elman (1990). Work in this line has illustrated a number of important theoretical points, such as learnability of segmentation from distributional information (Aslin et al., 1996) and the additional leverage gained by combining multiple cues (Christiansen et al., 1998). These results do not directly bear on the nature of the representations or computations that humans bring to bear in word

segmentation, in part because of the well-known difficulty of interpreting connection weights and hidden unit activations (Elman, 1990)—typically it is unclear how the network solved the problem.

2.2.3. Joint-optimization approaches

Some have formulated word segmentation and word learning as joint-optimization problems, in which related problems can be solved jointly by defining a single optimal solution (e.g., Blanchard & Heinz, 2008; Brent & Cartwright, 1996; Goldwater, Griffiths, & Johnson, 2009). As shown in Goldwater (2006), extant approaches have a natural Bayesian formulation in which ''solutions'' are segmentations of the input, and the optimum is defined as the solution with *maximum a posteriori* probability, calculated from a prior on the segmentation-induced lexicon.

To illustrate the core ideas, consider two minimally different orthographic segmentations[5] of the sentence *The dog chased the cat*. Each segmentation induces a lexicon, operationalized as a list of word types and associated frequencies (Table 1).

Crucially, the induced lexicons differ in the number of words and their frequencies. It is these differences which cause joint-optimization models to prefer one solution over another. For example, ''minimum description length'' prefers solution (1a) to (1b) because it uses fewer words to explain the observed corpus (Brent & Cartwright, 1996). The ''Chinese Restaurant Process'' prior of Goldwater (2006) would also prefer (1a) to (1b), because it exhibits a Zipfian frequency distribution in which a few words occur repeatedly (in this case, *the*) and many elements occur only rarely (Baayen, 2001).

While some joint-optimization models adopt an ideal-observer approach, in which the goal is to draw inferences about the cognitive properties of the learner from the optimal solution (e.g., Goldwater, 2006), other models claim to model human cognitive processes (Brent, 1999; Blanchard & Heinz, 2008). The current generation of such models assumes a one-to-one relationship between input segmentation and the learner's lexicon, so positing a word boundary automatically entails incrementing the frequency in the lexicon of the words

Table 1
Segmentation in joint-optimization models

Segmentation																					
(a)	t	h	e	#	d	o	g	#	c	h	a	s	e	d	#	t	h	e	#	c	a	t
(b)	t	h	e	d	#	o	g	#	c	h	a	s	e	d	#	t	h	e	#	c	a	t

Induced lexicon			
(a)		(b)	
Word	Frequency	Word	Frequency
the	2	thed	1
dog	1	og	1
chased	1	chased	1
cat	1	the	1
		cat	1

on either side. These models bear on the crucial assumption of this paper that word segmentation is in part a prelexical process, because they instantiate the alternative hypothesis that word segmentation, word recognition, and word learning are part of the same act and are driven by word frequency distributions.

2.3. Motivation for a phonotactic approach

While the joint-optimization approach is highly illuminating, we argue that current-generation models do not solve the segmentation task in the same way that infants do. Two issues motivate a phonotactic approach: infants' use of phonotactic generalizations in segmentation and the complexity of word learning.

2.3.1. Phonotactic generalizations

A number of studies provide clear evidence that infants make use of phonotactic generalizations (not just lexical knowledge) for word segmentation. As early as 7.5 months of age, English-learning infants treat stressed syllables as word onsets, incorrectly segmenting *TARis* as a word from the sequence *...guiTAR is...* (Jusczyk, Houston, et al., 1999)—a strategy that is highly appropriate for English owing to the fact that most English content words are stress-initial (Cutler & Carter, 1987). By 8 months of age, infants exhibit some familiarity with the segmental phonotactics of their language and use it for word segmentation (Friederici & Wessels, 1993; Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993; Jusczyk et al., 1994; Saffran et al., 1996).

Mattys and colleagues demonstrated that infants exploit diphone phonotactics specifically for word segmentation. Recall that many diphones are contextually restricted, occurring either within a word (e.g., [ba]), or across word boundaries (e.g., [pd]), but not both (Hockema, 2006). Mattys, Jusczyk, Luce, and Morgan (1999) exposed infants to CVC. CVC nonwords, finding that both stress and the medial C.C cluster affected infants' preferences. Then, Mattys and Jusczyk (2001) showed that infants use this diphone cue to segment novel words from an unfamiliar, phrase-medial context.

With the exception of Blanchard and Heinz (2008), current-generation joint-optimization models do not predict segmentation on the basis of phonotactic generalizations (stress, diphone occurrence). Blanchard and Heinz (2008) show that including a phonotactic model yields significantly better performance; however, even this model exhibits word learning from a single exposure, argued below to be cognitively implausible.

2.3.2. Word learning

In current-generation joint-optimization models, positing a word boundary entails incrementing the frequency of the wordforms on either side. If the wordforms are not already present in the lexicon, they are added. This amounts to the assumption that words are always learned from a single presentation. While learning a word from one exposure is clearly possible, even for adults it is not the norm; even after seven presentations adults fail to learn about 20% of novel CVC words (Storkel, Armbruster, & Hogan, 2006), and a number of additional studies suggest that segmenting a word is not sufficient to cause word learning in

infants (Brent & Siskind, 2001; Davis, 2004; Graf Estes, Evans, Alibali, & Saffran, 2007; Swingley, 2005).

Moreover, word learning in infants is apparently subject to many other factors besides word segmentation. Lexical neighbors facilitate word learning in adults and 3- to 4-year-olds (Storkel et al., 2006; Storkel & Maekawa, 2005). Caregiver/infant joint attention also facilitates word learning (Tomasello, Mannle, Kruger, 1986; Tomasello & Farrar, 1986). A comprehensive theory of word learning should include these factors, but they are apparently independent of word segmentation. Thus, while it is fair to ask how a word segmentation model can facilitate word learning, segmentation models should not bear the full explanatory burden for word learning. In short, segmentation makes word forms available to be learned, but word learning is a separate process.

2.4. Segmentation in the cognitive architecture

More precisely, we argue that phonotactic segmentation is a prelexical process (whereas word learning is necessarily a lexical process). For this claim to make sense, it is necessary to accept that there is a distinction between prelexical and lexical processing. This section reviews evidence for a two-stage (prelexical/lexical) account of speech processing (Luce & Pisoni, 1998; McClelland & Elman, 1986). The general principle underlying this distinction is that prelexical processing assigns structure to speech in some way that facilitates lexical access.

The most convincing evidence for a two-stage processing account comes from dissociable effects of phonotactic probability and lexical neighborhood density across a wide range of tasks. The phonotactic probability of a wordform is estimated compositionally from the probabilities of its subparts (e.g., $p([\text{bat}]) = p([\text{b}])p([\text{a}]|[\text{b}])p([\text{t}]|[\text{a}])$). Lexical neighborhood density refers to the number of phonological neighbors of a word (i.e., differing by one phoneme). Bailey and Hahn (2001) and Albright (2009) find unique effects of phonotactics and lexical neighborhood in explaining word acceptability judgements. Luce and Large (2001) found a facilitory effect of phonotactic probability, but an inhibitory effect of lexical neighborhood density on reaction time in a same-different task. While lexical neighbors affect categorization of phonetically ambiguous tokens (Ganong, 1980), experiments on perceptual adaptation (Cutler, McQueen, Butterfield, & Norris, 2008) and phonetic categorization of ambiguous stimuli (Massaro & Cohen, 1983; Moreton, 1997) show there are also nonlexical (phonotactic) effects. Thorn and Frankish (2005) found a facilitory effect of phonotactic probability on nonword recall when neighborhood density was controlled, and a facilitory effect of neighborhood density when phonotactic probability is controlled. Storkel et al. (2006) found a facilitory effect of neighborhood density and an inhibitory effect of phonotactic probability on word learning in adults. These findings can be straightforwardly explained by a theory with distinct sublexical and lexical levels of representation, but they are harder to accommodate under a single-stage approach, such as joint-optimization models appear to take.

That phonotactic word segmentation is attested for novel words in novel contexts (Mattys & Jusczyk, 2001) provides prima facie evidence it must be a prelexical mechanism. By

prelexical, we mean the segmentation mechanism has *no access to specific lexical forms*; instead, it feeds a downstream lexical processor (Fig. 1).

One implication is that segmentation is a *distinct* cognitive process from word learning. As a corollary, we attribute to the downstream lexical processor factors in word learning such as lexical neighborhood density and caregiver/infant joint attention, which exceed the scope of this paper.

3. DiBS

We now present a phonotactic model that learns diphone-based segmentation (DiBS), as described in Cairns et al. (1997) and Hockema (2006). We begin by reviewing our assumptions about the task and what infants bring to it. We follow other computational studies in assuming that *speech input is represented as a sequence of phones and the listener's goal is to recover phrase-medial word boundaries*. In DiBS, the listener recovers word boundaries based on the identity of the surrounding diphone. The core of the model: Given a sequence [*xy*], estimates the probability that a word boundary falls in the middle $p(\# \mid xy)$. In the following section, we outline our assumptions as to what is observable to infants, and the assumptions they must make to estimate model probabilities from these observables.

3.1. Assumptions

We assume the infant knows or can observe the following:

- phonetic categories;
- phonological independence across word boundaries;
- phrase-edge distributions;
- the context-free diphone distribution;
- the context-free probability of a phrase-medial word boundary;
- the lexical frequency distribution.

These assumptions are justified as follows.

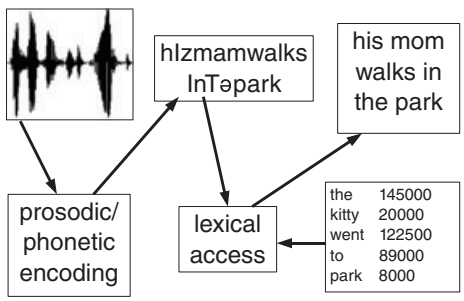

Fig. 1. Cognitive architecture of speech perception.

3.1.1. Phonetic categories

Infant speech perception begins to exhibit hallmark effects of phonetic categorization by 9 months of age.

Phonetic categorization in adults is evident from high sensitivity to meaningful acoustic/phonetic variation, and low sensitivity to meaningless acoustic/phonetic variation. For example, the speech sounds [l] and [r] represent distinct sound categories in English, as evident by minimal pairs such as *leak/reek* and *lay/ray*; the same sounds do not signal a lexical contrast in Japanese, because they represent alternate pronunciations of a single sound category. Thus, Japanese listeners exhibit poor discrimination of the [l]/[r] contrast, whereas English listeners exhibit excellent discrimination (Miyawaki et al., 1975). Adult speech perception is exquisitely tuned to the phonological system of the native language.

The effect of language exposure on phonetic categorization generally becomes apparent between 7 and 11 months of age.[6] Prior to 6 or 7 months of age, infants exhibit similar discrimination regardless of language background (Kuhl et al., 2006; Trehub, 1976; Tsao, Liu, & Kuhl, 2006; Werker & Tees, 1984). Between 7 and 11 months, discrimination of meaningless contrasts decreases (Werker & Tees, 1984), and discrimination of meaningful contrasts improves (Kuhl et al., 2006; Tsao et al., 2006). Thus, infants appear to acquire native language phonetic categories around 9 months of age.

3.1.2. Phonological independence

Phonological independence across word boundaries means that phonological material at the end of one word exhibits no statistical dependencies with phonological material at the beginning of the next word. While this assumption is not strictly true, it is reasonable to make in the initial stages of acquisition, in the absence of contradictory evidence.

3.1.3. Phrase-edge distributions

We assume that infants know the frequency distribution of phones in the phrase-initial and phrase-final position. This assumption is motivated by the fact that infants are sensitive to phrase boundaries in phonological and syntactic parsing (Christophe, Gout, Peperkamp, & Morgan, 2003; Soderstrom, Kemler-Nelson, & Jusczyk, 2005), and the generalization that they are sensitive to the relative frequency of phonotactic sequences (Jusczyk et al., 1994). Because this study is limited by the coding conventions of the corpora employed, only utterance edges are treated as exemplifying phrase boundaries. The availability of such boundaries to the infant is indisputable. Indeed the works just cited suggest that weaker boundaries, such as utterance-medial intonation phrase boundaries, may also be available to the infant due to pausing and other suprasegmental cues. Including these boundaries would increase the success of the model by increasing the effective sample size for training, and by explicitly providing some boundaries in the test set that our model must estimate (e.g., those word boundaries that coincide with intonational phrase boundaries). Thus, using utterance edge statistics to estimate phrase edge statistics is a very conservative choice.

3.1.4. Context-free diphone distribution

We assume that infants track the context-free distribution of diphones in their input. This assumption, which is shared in some form by all existing models of phonotactic word segmentation, is motivated by evidence that infants attend to local statistical relationships in their input (Mattys & Jusczyk, 2001; Saffran et al., 1996).

3.1.5. Context-free probability of a phrase-medial word boundary

The context-free probability of a word boundary is a free parameter of the model. Because this value is determined by average word length and words/utterance, we assume infants can obtain a reasonable estimate of it. For example, average word length is lower-bounded by the cross-linguistic generalization that content words are minimally bimoraic, a prosodic requirement typically instantiated as CVC or CVCV. Even allowing for the fact that some of the function words are shorter, this implies that the overall probability of a word boundary must be less than about 1/3. Because the assumption that infants can estimate this parameter with adequate reliability is somewhat speculative, Simulation 2 investigates the model's robustness to variation in this parameter.

3.1.6. Lexical frequency distribution

Finally, we assume infants know the relative frequency of the word forms they have learned. This assumption is motivated by the massive body of evidence documenting frequency effects in adults (for a review see Jurafsky, 2003) and findings of frequency sensitivity in closely related levels of representation in infants (Anderson et al., 2003; Jusczyk et al., 1994; Mintz, 2003; Peterson-Hicks, 2006).

3.2. Baseline-DiBS

Diphone-Based Segmentation models necessarily exhibit imperfect segmentation. For every token of a diphone type they make the same decision (boundary/no boundary), whereas at least some diphones occur both word-internally and across word boundaries (e.g., [rn]: *Ernie*, *garner* but *bar none*, *more numbers*). Since a DiBS model must make the same decision in both cases, it will either make errors on the word-internal items, or on the word-spanning items. We define the baseline model as the statistically optimal one, that is, making the smallest possible number of errors—exactly the model described in Cairns et al. (1997) and Hockema (2006). We use this statistically optimal model as the baseline because it establishes *the highest level of segmentation performance that can be achieved* by any DiBS model. Thus, we refer to the baseline model's segmentation as ''ceiling,'' meaning not perfect segmentation, but the best segmentation achievable by DiBS.

3.3. Learning

The core goal of the learner is to derive an estimate of the DiBS statistics $p(\# \mid xy)$ from observable information. Recall that $p(\# \mid xy)$ represents the probability of the presence of a

word boundary in the middle of a sequence, given that the constituent phones of the sequence were [x] and [y].

3.3.1. Bayes' rule

The first step is to apply Bayes' rule, rewriting this conditional probability in terms of the reverse conditional probability:

$$p(\#|xy) = p(xy|\#) \cdot p(\#)/p(xy) \tag{1}$$

where $p(xy)$ is the context-free probability of the diphone [xy] and $p(\#)$ is the context-free probability of a word boundary. Note that these two terms are known by assumption, so the infant now need only worry about estimating $p(xy \mid \#)$.

3.3.2. Phonological independence

The next step is to apply the assumption of phonological independence so as to factor $p(xy \mid \#)$. This represents the joint probability of a word-final [x] followed by a word-initial [y]. Under the assumption of phonological independence, the probability of a word-initial [y] does not depend on the word-final phone of the preceding word. Thus, the joint probability is simply the product of the each event's probability:

$$p(xy|\#) \approx p(x \leftarrow \#) \cdot p(\# \rightarrow y) \tag{2}$$

where $p(x \leftarrow \#)$ represents the probability of observing a word-final [x], $p(\# \rightarrow y)$ represents the probability of observing a word-initial [y], and $\approx$ indicates approximation. The problem of estimating $p(xy \mid \#)$ has been reduced to the problem of estimating the distribution of phones at word edges.

3.3.3. Phrasal-DiBS

The word-edge distribution itself is not observable to infants until after they have begun to solve the segmentation problem. However, infants can get a first-pass approximation by capitalizing on the fact that phrase boundaries are always word boundaries (Aslin et al., 1996), using the phrase-edge distribution as a proxy:

$$\begin{aligned} p(x \leftarrow \#) &\approx p(x \leftarrow \%) \\ p(\# \rightarrow y) &\approx p(\% \rightarrow y) \end{aligned} \tag{3}$$

where $p(x \leftarrow \%)$ and $p(\% \rightarrow y)$ represent the probability of observing [x] phrase-finally and [y] phrase-initially, respectively. The entire model can be written:

$$p_{\text{phrasal}}(\#|xy) = p(x \leftarrow \%) \cdot p(\#) \cdot p(\% \rightarrow y)/p(xy) \tag{4}$$

This first-pass approach is suitable for the very earliest stages of segmentation, when the infant must bootstrap from almost nothing. Recall that utterance boundaries are used here as a conservative proxy for phrase boundaries, due to limitations imposed by the transcripts in the corpora.

3.3.4. Lexical-DiBS

After infants have begun to acquire a lexicon, they have a much better source of data to estimate the distribution of phones at word-edges—namely from the words they know. As discussed in the introduction, by the time infants evince phonotactic segmentation in laboratory studies (about 9 months), they are reported to know an average of 40 words, including familiar names (Dale & Fenson, 1996) and other words which presumably co-occur with them (Bortfeld et al., 2005). However these words are learned, once they are learned, they can be leveraged for phonotactic word segmentation. By estimating edge statistics from known words, infants may avoid errors caused in Phrasal-DiBS by distributional atypicalities at phrase edges.[7]

To use this data, the infant must estimate the probability with which each phone ends/begins a word in running speech. The most accurate method is to use the token probability, that is, weighting word-initial and -final phones according to lexical frequency. (For example, the sound [ð] has a low type frequency but is highly frequent in running speech, because it occurs in a small number of highly frequent words such as *the*, *this* and *that*. Infants need to estimate the token probability in order to make use of this important segmentation cue.) These probabilities can be estimated as follows:

$$\begin{aligned} p_{\Lambda}(x \leftarrow \#) &\approx (\Sigma_{\omega \in \Lambda}(\omega == [\ldots x]) \cdot f(\omega))/(\Sigma_{\omega \in \Lambda} f(\omega)) \\ p_{\Lambda}(\# \leftarrow y) &\approx (\Sigma_{\omega \in \Lambda}(\omega == [y \ldots]) \cdot f(\omega))/(\Sigma_{\omega \in \Lambda} f(\omega)) \end{aligned} \tag{5}$$

where Λ is the listener's lexicon. In these equations, the numerator represents the expected token frequency of words that end/begin with [x]/[y] and the denominator represents the total observed token frequency. The notation (ω==[...x]) is an indicator variable whose value is 1 if word ω ends in [x], and 0 otherwise; (ω==[y...]) similarly indicates whether ω begins with [y]. The full model is given below:

$$p_{\text{lexical}}(\#|xy) = p_{\Lambda}(x \leftarrow \#) \cdot p(\#) \cdot p_{\Lambda}(y \rightarrow \#)/p(xy) \tag{6}$$

The following section discusses Lexical-DiBS in the context of learning theory.

3.3.5. Gradient of supervision

Language acquisition is generally acknowledged to be ''unsupervised.'' In the context of word segmentation, this means that the language input does not include the hidden structure (phrase-medial word boundaries) that the model is supposed to identify at test. While the distinction between unsupervised and supervised models may seem clear, it is not always clear how to apply this distinction to ''bootstrapping'' models, in which some preexisting knowledge is leveraged to solve a different problem. Baseline-DiBS is fully supervised. Phrasal-DiBS is unsupervised because it leverages information that is obviously available in the input. Lexical-DiBS lies somewhere in between.

Lexical-DiBS estimates its parameters from the infant's developing lexicon. It is not unsupervised, because it depends on information (a set of words in the lexicon) whose acquisition has not been modeled here. It is not fully supervised, because the phonological

sequences used in training do not have the word boundaries indicated. In Lexical-DiBS, a number of diphones are identified as boundary-spanning through a nontrivial inductive leap, assuming phonological independence across word boundaries, and estimating word edge distributions from the aggregate statistical properties of the lexicon.

Semi-supervised learning is of interest as a model of human acquisition because infants clearly know some words and learn more during the developmental period modeled here (Bortfeld et al., 2005; Dale & Fenson, 1996). A small early lexicon might be acquired by learning words that have occurred in isolation, through successful application of the segmentation algorithm presented here, or through some other mechanism we have not modeled, such as noticing repeatedly recurring phoneme sequences. Although a full treatment of these factors exceeds the scope of this paper, Lexical-DiBS will reveal how *segmentation can improve if generalizations about the form of words in the lexicon are fed back to the word segmentation task* as soon as such generalizations become possible.

3.3.6. Summary

The core DiBS statistics can be estimated from phrase-edge distributions, and/or word-edge distributions in the listener's lexicon. This section has articulated a learning model for DiBS using Bayes' theorem and the assumption of phonological independence across word boundaries. Two instantiations were proposed: Phrasal-DiBS estimates model parameters from phrase-edge distributions, and Lexical-DiBS estimates them from word-edge distributions. In Simulation 1, the developmental trajectory of these learning models is assessed.

4. Simulation 1

The goal of Simulation 1 is to measure performance of the learning models against the supervised baseline model. Thus, the baseline model should replicate the main findings of Cairns et al. (1997). However, because the focus of the present study is learnability, the methodology differs. The training data are divided into units that represent a ''day'' worth of caregiver input. In accord with contemporary corpus linguistic standards, the model is only tested on unseen data, never on data it has already been trained on. The training and testing data for Simulation 1 are drawn from the CHILDES database (MacWhinney, 2000) of spontaneous interactions between children and caregivers.

4.1. Input

4.1.1. Corpus

The CHILDES database consists of transcriptions of spontaneous interactions between children and caregivers. It contains many subcorpora, collected by a variety of researchers over the past several decades. Ecological validity was the primary motivation for selecting this corpus—it was important to obtain input close to what infants actually hear.

We drew samples from the entire English portion of the database, as very few of the CHILDES corpora target children under 1;5. The motivation for this was to get more data: Acquisition of word segmentation apparently takes several months, so a large data set is required to accurately model the amount and extent of language input that infants hear. Our CHILDES sample contains 1.5 million words; the Bernstein-Ratner corpus used in several other studies of word segmentation acquisition (e.g., Brent & Cartwright, 1996; Goldwater, 2006) contains about 33,000 words, representing about a day of input to a typical child. By sampling from the entire CHILDES corpus, we sacrifice some ecological validity (by including child-directed rather than only infant-directed speech), but we obtain a larger sample more representative of an infant's total language exposure.

4.1.2. Sample

For each target child in the database, a derived corpus was assembled of speech input to the child. Each derived file contained all utterances in the original file *except* those spoken by the child herself. A sample of ''days'' was drawn from this derived corpus, as follows. Based on van de Weijer's (1998) diary study as well as an ecological study of adult production (Mehl, Vazire, Ramirez-Esparza, Slatcher, & Pennebaker, 2007), a ''day'' of input was defined as 25,000 words. (This value is used here as a standardized unit for modeling purposes; it is intended to approximate what a typical English-learning infant hears in a day, but it is not intended as a claim that all infants hear exactly this amount of input.) Files were selected at random from the derived corpus and concatenated to obtain 60 ''days'' of input, each containing approximately 25,000 words. Properties of the corpus and training and test sets for Simulation 1 are given in Table 2, along with comparable figures for Simulation 3.

4.1.3. Phonetic mapping

A phonetic representation was created by mapping spaces to word boundaries and mapping each orthographic word to a phonetic pronunciation using the CELEX pronouncing dictionary (Baayen, Piepenbrock, & Gulikers, 1995) with the graphemic DISC transcription system. Words not listed in the dictionary were simply omitted from the phonetic transcription, for example, ''You want Baba?'' would be transcribed as [ju wQnt], omitting the unrecognized word ''Baba.'' (About 8.75% of tokens were omitted, including untranscribed tokens ''xxx,'' nonspeech vocalizations like ''um'' and ''hm,'' nonstandardly transcribed

Table 2
Corpus properties

Simulation	Type	Corpus	Words	Phrases	Phones	*p*(#)
1	Train	–	750,111	170,709	2,226,561	.2818
1	Test	–	750,125	171,232	2,224,873	.2819
3	Train	Canonical	150,030	25,914	479,741	.2735
3	Train	Reduced	149,998	25,907	442,135	.2981
3	Test	Canonical	16,058	2,490	51,555	.2765
3	Test	Reduced	16,051	2,488	47,516	.3012

speech routines like ''thank+you'' and ''all+right,'' unlisted proper names like ''Ross,'' and phonetically spelled variants like ''goin'' and ''doin.'') Note that the CHILDES standard is to put each sequence of connected speech on its own line, without punctuation or phrase boundaries; thus, an individual ''phrase'' corresponds to something more like an utterance in this corpus.

4.1.4. Training and test sets

The training set consisted of the first 30 ''days'' of input. This length of time is used because the acquisition literature suggests the onset of phonotactic word segmentation occurs shortly after the acquisition of language-specific phonetic categories (cf. Werker & Tees, 1984; Tsao et al., 2006; Kuhl et al., 2006 with Jusczyk, Hohne, et al., 1999; Jusczyk, Houston, et al., 1999; Mattys & Jusczyk, 2001). Thus, a learning model based on categorical phonotactics must be trainable input on the scale of weeks. The test set consisted of the remaining 30 ''days.''

4.2. Models

4.2.1. Phrasal-DiBS

The Phrasal-DiBS model parameters were estimated according to the phrase-edge distributions in the learner's input.

4.2.2. Lexical-DiBS

Lexical-DiBS is based on the learner's lexical knowledge rather than the raw input. In order to properly compare Lexical-DiBS with Phrasal-DiBS, it is necessary to know which words an infant will learn, given what the infant has heard. Unfortunately, no sufficiently predictive theory of word learning exists. As a crude proxy, we use a frequency threshold model: Wordforms are added to the lexicon incrementally as soon as they have occurred n times in the input.

Three frequency thresholds were used: 10, 100, and 1,000. The threshold 10 is used as a lower bound, because it almost certainly overestimates an infant's lexicon size (even 14-month-olds do not learn every word they hear 10 times, e.g., Booth & Waxman, 2003). Similarly, the threshold of 1,000 is used as an upper bound, because only a few words like *dada* and *mama* are actually uttered more than 1,000 times in a typical month of infant input, and all 9-month-olds learn these high-frequency words (Dale & Fenson, 1996). The threshold of 100 is a reasonable compromise between these upper and lower bounds. (NB: Frequency in the learner's lexicon was calculated by subtracting the threshold from the true input frequency.)

4.2.3. Baseline-DiBS

The Baseline-DiBS model parameters were estimated according to the within-word and across-word diphone counts in the training corpus.

In all cases, if the model encountered a previously unseen diphone in the test set, for example, one not expected given the training data, the diphone was treated as signalling a

word boundary. In the context of our analysis, this will cause an oversegmentation error whenever the diphone was actually word-internal.

4.3. Method

Each model was exposed cumulatively to the ''days'' of the training set. After each ''day'' of training, the model was tested on the entire test set.

4.3.1. Hard decisions: Maximum likelihood decision threshold

Formally speaking, a DiBS model estimates the probability of a word boundary given the surrounding diphone, symbolized $p(\# \mid xy)$. Being probabilistic, DiBS does not assign hard decisions as to the presence or absence of a word boundary, but rather a probability. However, the ''correct answer'' is not probabilistic: The speaker intended a particular sequence of words, and the word boundaries are underlyingly there, or not. Thus, for evaluation purposes, the probabilistic output of DiBS models is mapped to hard decisions using the *maximum likelihood decision threshold* $\theta = 0.5$: If $p(\# \mid xy) > .5$, a word boundary is identified, otherwise not. (The value 0.5 is called the maximum likelihood threshold because it results in the minimum number of total errors; that is, it is the threshold with maximum likelihood of yielding a correct decision.) This process is repeated for every diphone in a phrase, as exemplified in Fig. 2.

By scoring in this way, we do not intend to claim that phonotactic segmentation in humans consists of hard decisions, as probabilities are a rich source of information which are likely to be useful in lexical access; hard decisions are used here for simple comparison with other studies.

4.3.2. Segmentation measures

For plotting, the dependent measure used is errors/word, distinguishing both undersegmentation and oversegmentation errors. For example, an oversegmentation error rate of 1/10 means that the listener will incorrectly split up 1 word out of every 10 tokens he or she hears. These measures strike us as highly informative because they indicate the error rate relative to the perceptual object the listener is attempting to identify: the word. In contrast, hearing that boundary precision is 89% does not make it clear how many word tokens the listener will oversegment.

To facilitate comparison with other published studies, we also report *boundary precision and recall*, *token precision and recall*, and *lexical precision and recall*. Boundary precision

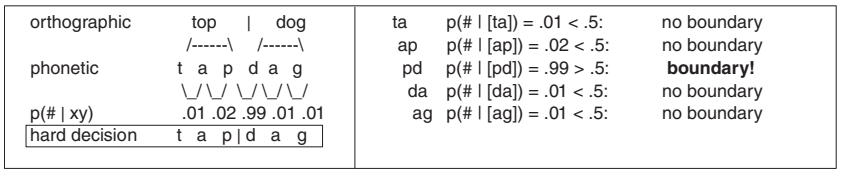

Fig. 2. Segmentation in DiBS.

is the probability of a word boundary given that the model posited one; boundary recall is the probability that the model posits a word boundary given that one occurred. Token precision is the probability that a form is a word token given that the model segmented it (posited boundaries on both sides); token recall is the probability that the model segmented a form, given that it occurred. Lexical precision and recall are analogous, except that wordform types are counted rather than tokens.

Note that because DiBS is intended as a prelexical model, its task is not to identify words per se, but to presegment the speech stream in whatever manner offers maximal support to the downstream lexical process. Since DiBS is intended to get the learner ''off the ground'' in feeding word learning, it is eminently appropriate to assess it in terms of token precision/recall. However, as repeatedly noted above, DiBS is not intended to account for word learning on its own, so it is not appropriate to compare its lexical precision/recall against more complex models that include a lexical module. Type recall is not comparable for another reason: The test set here is about 200 times larger than in comparison studies, so there are a significantly greater number of types.

4.4. Results

Fig. 3 illustrates the undersegmentation and oversegmentation error rates as a function of language exposure.

To facilitate discussion and comparison, the other measures of performance from this Simulation and Simulation 3 are reported in Table 3, as well as reported values from other published studies.

In addition, a small sample of the output (first line of the last test file) is given below in Table 4.

4.4.1. Confidence intervals

Because the undersegmentation and oversegmentation error rates represent probabilities, confidence intervals can be determined by assuming they are Bernoulli-distributed. The half-width of the 95% confidence interval for a Bernoulli distribution is no larger than $.98/\sqrt{n}$, where n is the sample size (Lohr, 1999). As the test set contained 750,111 words, the error rates are accurate to ±0.1%.

4.5. Discussion

4.5.1. Rapid learning

The phrasal and baseline models reach near-ceiling performance within the first ''day'' of training, as evident from the nearly flat trajectory of these models in Fig. 3. That is, while these models do exhibit modest changes in error rates, these changes are small relative to the overall error rate. Only the lexical model continues to exhibit substantial gains with increasing language exposure, and the trajectory of the lexical-10 model suggests that these gains will asymptote eventually as well. For the phrasal and baseline models, most of what can be learned from the training data is learned within

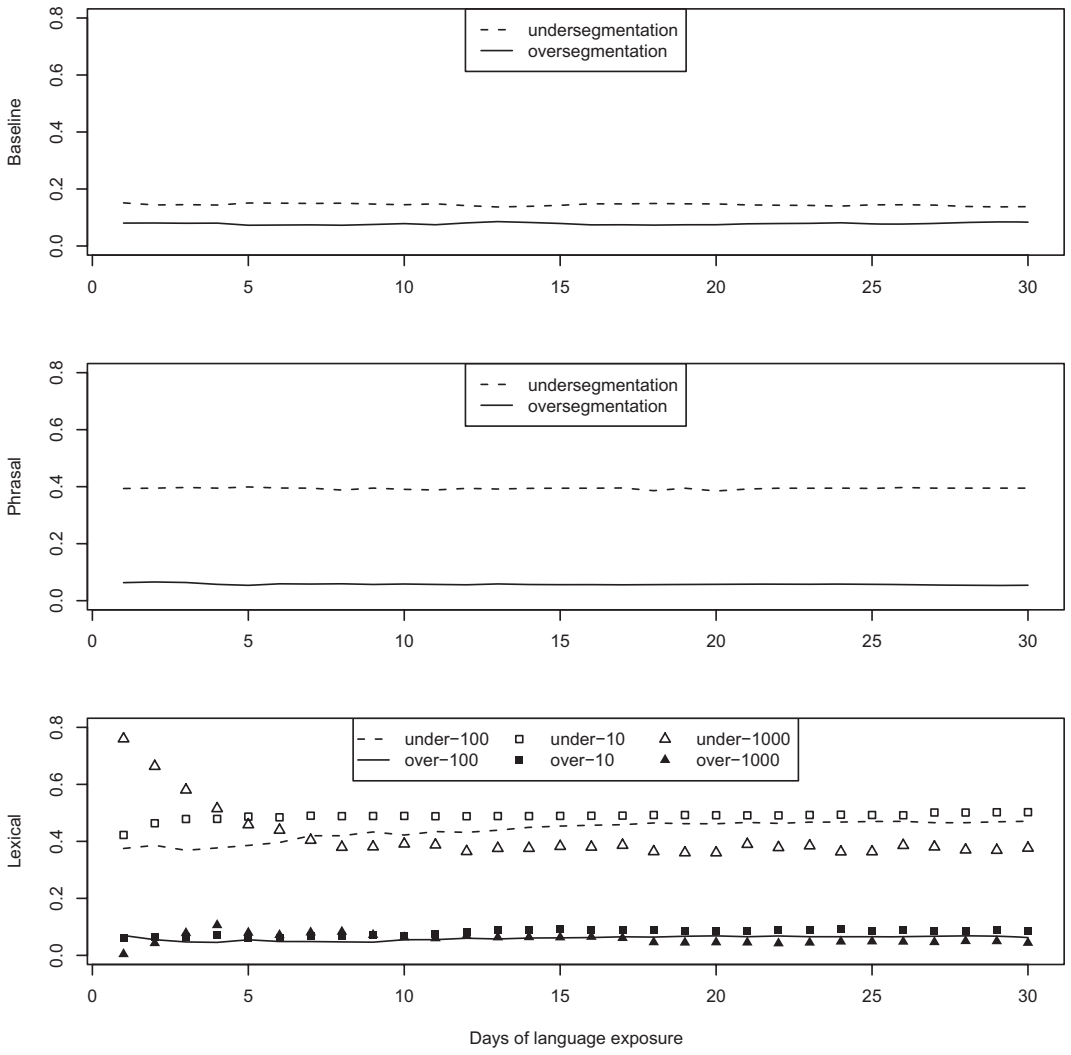

Fig. 3. Undersegmentation and oversegmentation error rates of baseline and learning models. The *x*-axes represent the number of ''days'' of language exposure (approximately 25,000 words/''day''). The *y*-axes represents the probability per word of making an undersegmentation error (dashed/empty) or oversegmentation error (heavy/filled). Panels indicate the baseline model (top), phrasal model (middle), and lexical model (bottom). In the lexical panel, the ''upper bound'' (frequency threshold of 10) is shown with squares; the ''lower bound'' (frequency threshold of 1,000) is shown with triangles.

the first day for these models; for the lexical model, much of what can be learned is learned within a month.

The rapidity with which the models learn is especially important because it demonstrates that diphone-based segmentation is learnable not only in principle, but also in practice. The amount of language exposure required is well within a reasonable timescale of what infants actually hear.

Table 3
Segmentation performance of DiBS and other models

Paper	Corpus	Tokens	BP	BR	BF	TP	TR	TF	LP	LR	LF	Notes
Ba	BR87	33k	–	–	–	67.2	68.2	67.7	–	–	–	
Br	BR87	33k	80.3	84.3	82.3	67	69.4	68.2	53.6	51.3	52.4	rep/GGJ
F	BR87	33k	94.6	73.7	82.9	–	–	70.7	–	–	36.6	
G	BR87	33k	89.2	82.7	85.8	–	–	72.5	–	–	56.2	rep/F
GGJ	BR87	33k	90.3	80.8	85.2	75.2	69.6	72.3	63.5	55.2	59.1	bigram
GGJ	BR87	33k	92.4	62.2	74.3	61.9	47.6	53.8	57	57.5	57.2	$p(\#) = .05$, a = 20
JG	BR87	33k	–	–	–	–	–	88	–	–	–	see JG
V	BR87	33k	81.7	82.5	82.1	68.1	68.6	68.3	54.5	57	55.7	bigram, rep/GGJ
V	BR87	33k	80.6	84.8	82.6	67.7	70.2	68.9	52.9	51.3	52	unigram, rep/GGJ
S	Korman	42k	–	–	–	–	–	–	75	–	–	
D	CHILD	750k	88.3	82.1	85.1	73.7	69.6	71.6	14.6	53.6	23.0	base
D	CHILD	750k	87.4	48.9	62.7	53.4	35.2	42.5	5.6	50.8	10.1	phrasal
D	CHILD	750k	82.8	39.0	53.1	44.8	26.5	33.3	4.5	47.3	8.2	lexical-100
F	Buck	32k	89.7	82.2	85.8	–	–	72.3	–	–	37.4	
F	Buck	32k	71	64.1	67.4	–	–	44.1	–	–	28.6	reduced
G	Buck	32k	74.6	94.8	83.5	–	–	68.1	–	–	26.7	rep/F
G	Buck	32k	49.6	95	65.1	–	–	35.4	–	–	12.8	reduced, rep/F
D	Buck	150k	87.4	76.7	81.7	66.4	59.6	62.8	30.7	53.7	39.1	base
D	Buck	150k	82.5	68.6	74.9	56.5	48.4	52.2	34.7	45.8	39.5	base, reduced
D	Buck	150k	80.5	47.6	59.8	44.1	28.8	34.9	16.5	37.2	22.8	phrasal
D	Buck	150k	76.0	44.1	55.8	39.1	25.2	30.7	21.1	29.4	24.6	phrasal, reduced
F	Switch	34k	90	75.5	82.1	–	–	66.3	–	–	33.7	orthographic
F	Switch	34k	91.3	80.5	85.5	–	–	72	–	–	37.4	
G	Switch	34k	73.9	93.5	82.6	–	–	65.8	–	–	27.8	rep/F
G	Switch	34k	73.1	92.4	81.6	–	–	63.6	–	–	28.4	ortho, rep/F
F	Arab	30k	88.1	68.5	77.1	–	–	56.6	–	–	40.4	
G	Arab	30k	47.5	97.4	63.8	–	–	32.6	–	–	9.5	rep/F
F	Spanish	37k	89.3	48.5	62.9	–	–	38.7	–	–	16.6	
G	Spanish	37k	69.2	92.8	79.3	–	–	57.9	–	–	17	rep/F
S	Weijer	25k	–	–	–	–	–	–	75	–	–	

Note. Column header: B/T/L indicates *boundary/token/lexical*; P/R/F indicates *precision/recall/F-score* (e.g., BR = boundary recall).

Paper key: Ba = Batchelder (2002), Br = Brent (1999), D = DiBS, F = Fleck (2008), G = Goldwater (2006), GGJ = Goldwater et al. (2009), JG = Johnson and Goldwater (2009), S = Swingley (2005), V = Venkataraman (2001); ''rep/X'' indicates the results are reported in paper X.

These results may help to explain a puzzle of the acquisition literature: why the onset of phonotactic segmentation coincides with or shortly follows the acquisition of phonetic categories. The DiBS model crucially assumes that infants possess a categorical representation of speech; however, as long as such a categorical representation is available, only a minuscule amount of language exposure is required to estimate the relevant phonotactic segmentation statistics. Thus, DiBS predicts that phonotactic segmentation should become evident shortly after infants begin to exhibit language-specific phonetic categorization—precisely what occurs. While rapid trainability is presumably not specific to DiBS, to our knowledge

Table 4
Sample output of learning models

ortho	''If you want to eat something give you a cookie but take these''
correct	If ju wQnt tu it sVmTIN gIv ju 1 kUkI bVt t1k D5z
base	If ju wQnt tu itsVmTINgIv ju 1kUkI bVt t1k D5z
phrasal	If ju wQnttuitsVmTINgIv ju 1kUkIbVtt1k D5z
lex-10	IfjuwQnt tuitsVmTIN gIvju1kUkI bVt t1kD5z

Note. Spaces indicate true/posited word boundaries.

we are the first to draw attention to this explanation of why phonotactic segmentation emerges shortly after phonetic categorization.

4.5.2. Undersegmentation

While the baseline and learning models varied considerably in the undersegmentation error rate, they consistently exhibited a low oversegmentation error rate from the beginning of training. In every case, the oversegmentation error rate was below 10%, meaning less than 1 oversegmentation error per 10 words. In fact, the learning models make fewer oversegmentation errors than the baseline model (the baseline model exhibits overall higher accuracy because the undersegmentation error rate is much lower).

This overall pattern, in which some undersegmentation errors are made, but very few oversegmentation errors are made, can be characterized as an overall pattern of undersegmentation. These results show that undersegmentation is the predicted perceptual outcome for all DiBS models considered. We will return to this point in the general discussion.

In summary, Simulation 1 demonstrated three key findings. First, parameters of the Cairns et al. (1997) diphone model can be estimated to some accuracy from information that is plausibly attributable to infants. Second, only a small amount of language exposure is required to make these estimates. Phrasal-DiBS reaches its asymptote with less input than an infant might receive in a typical day; Lexical-DiBS continues to improve with increasing language exposure. Its asymptotic performance is similar to Phrasal-DiBS, indicating that a small lexicon does not supply greatly more information than was already present at utterance boundaries. However, the lex-1000 model already achieves better performance than Phrasal-DiBS within a month, an indication that exploiting the phone statistics of high-frequency words can improve segmentation performance. Finally, all models exhibit undersegmentation, characterized by an error rate of less than 1 oversegmentation error per 10 words.

Of the assumptions required for the learning model to work, the one which is perhaps the most speculative is the assumption that infants know or can learn the context-free probability of a word boundary $p(\#)$. Thus, it is natural to wonder how sensitive the model is to this assumption. In Simulation 2, we investigate the consequences of an error in the infant's estimate of $p(\#)$ by varying this parameter over a reasonable range and assessing the model's performance. The model would not be robust if even small errors in the estimate of $p(\#)$ cause dramatic qualitative shifts in the predicted segmentation pattern. Conversely, the

model is robust if small errors in the estimate of $p(\#)$ cause at most small changes in the segmentation pattern.

The ''reasonable'' range for $p(\#)$ that infants might consider is constrained by the relation between the phrase-medial word boundary probability and average word length. For example, if phrases contain an average of four words, and words contain an average of four phones, there will be three phrase-*medial* word boundaries per 16 phones. Average word length has natural upper and lower bounds, which correspondingly bound $p(\#)$. As discussed above, consideration of the Minimal Prosodic Word (McCarthy & Prince, 1986/1996) generates an upper bound on $p(\#)$ of $\approx$.33. A reasonable upper bound is the longest word that infants are observed to learn (perhaps owing to memory/coding limitations); for English, this is 6–8 phones (8: *breakfast*, 7: *toothbrush*, *telephone*, 6: *grandma*, *peekaboo*, *stroller*, *cheerios*, *outside*; Dale & Fenson, 1996), so $p(\#) > 1/8 \approx .125$. Thus, infants might reasonably consider the range for the context-free probability of a word boundary to be between 1/8 and 1/3. Simulation 2 assesses DiBS' robustness to estimation errors for $p(\#)$.

5. Simulation 2

5.1. Input

The stimuli consisted of the training and test sets of Simulation 1.

5.2. Models

The Phrasal-DiBS and Lexical-DiBS models of Simulation 1 were used.

5.3. Method

Instead of varying language exposure, the free parameter $p(\#)$ was varied in equal steps of .02 from .16 to .40, corresponding to a range of average word lengths from about 2.5 to about 6 phones. Results are reported from the final ''day,'' that is, after exposure to the entire training set.

5.4. Results

The results are shown in Fig. 4, which plots under- and oversegmentation error rates as a function of $p(\#)$. In addition, a sensitivity analysis is presented in Table 5.

Table 5 reports the undersegmentation and oversegmentation error rates with the correct value for $p(\#)$ (columns UE and OE), and beside these columns it reports the range of values for $p(\#)$ that will result in a 5% absolute change to the undersegmentation/oversegmentation error rate. For example, the entries in the far right column indicate that even when the learner estimates $p(\#)$ to be as low/high as .16/.4, the absolute undersegmentation/oversegmentation rate does not decrease/increase by more than 5%.

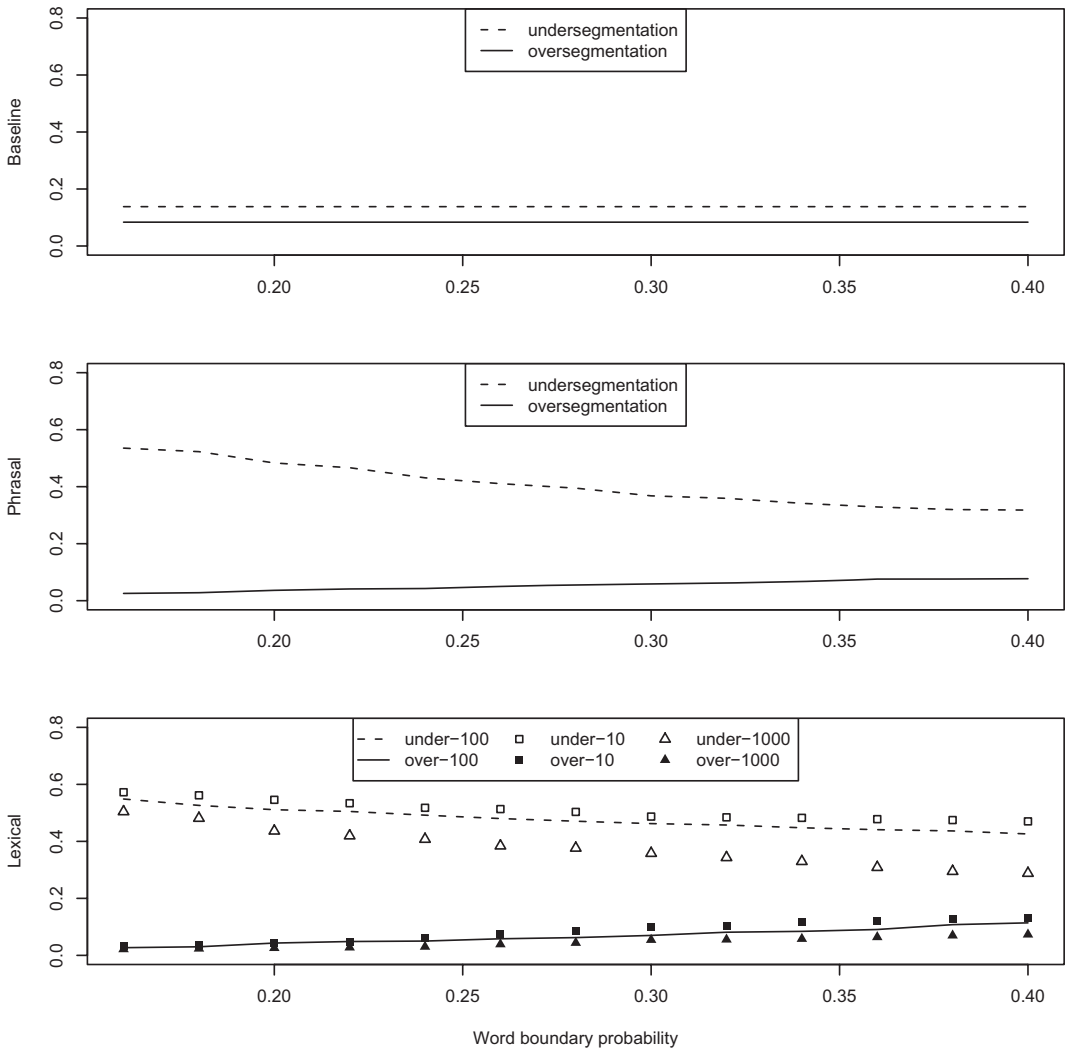

Fig. 4. Undersegmentation and oversegmentation error rates as a function of the probability of a phrase-medial word boundary. The *x*-axes represent *p*(#) and the *y*-axes represent undersegmentation and oversegmentation error rates, as in Fig. 3.

5.5. Discussion

The results of Simulation 2 suggests the learning model is robust to estimation errors. The oversegmentation error rate is particularly robust—varying *p*(#) yields significant changes in the undersegmentation rate, but over a wide range, the oversegmentation rate stays small (near or under 10%).

The phonetic corpus in Simulations 1 and 2 used a canonical, invariant pronunciation for each word. Since pronunciation variation is a general fact of speech (Johnson, 2004;

Table 5
Sensitivity analysis of Simulation 2 results

Model	UE	$-5\% < \Delta UE < +5\%$	OE	$-5\% < \Delta OE < +5\%$
phrasal	39.49	$.22 < p(\#) < .34$	5.41	$.16 \leq p(\#) \leq .40$
lex-10	50.25	$.20 < p(\#) \leq .40$	8.47	$.16 \leq p(\#) \leq .40$
lex-100	47.05	$.18 < p(\#) \leq .40$	6.24	$.16 \leq p(\#) \leq .40$
lex-1000	37.64	$.20 < p(\#) < .34$	4.4	$.16 \leq p(\#) \leq .40$

Mitterer, Yoneyama, & Ernestus, 2008), it is natural to wonder about the ecological validity of these results. The goal of Simulation 3 is to address this question by testing DiBS on a corpus of spontaneous adult speech containing natural pronunciation variation.

The Buckeye corpus (Pitt et al., 2007) consists of interviews with lifetime residents of Columbus, Ohio. The corpus is invaluable for the present purposes because it contains two transcriptions—a ''canonical'' transcript with the citation pronunciation of each word, and a ''reduced'' transcript representing much of the pronunciation variation present in the talkers' speech. Before discussing the details of the experiment, it is worth considering what effects pronunciation variation may have on word segmentation.

While conversational speech is likely to be overall more challenging, it is conceivable that certain reduction processes may actually facilitate segmentation. For example, conversational speech is likely to cause more place assimilation word-internally than across word boundaries $p(Stanford \rightarrow Sta[m]ford) > p(can\ ford \rightarrow ca[m]\ ford)$. This should strengthen the place cue: Heterorganic (different-place) clusters become more reliably associated with a word boundary, and homorganic clusters with no word boundary. Indeed, one appeal of phonotactic segmentation is that it might ameliorate the vulnerability of word recognition processes to pronunciation variability. Conversely, reduction processes might destroy segmentation cues, for example, by deleting phones in the crucial boundary-indicating contexts.

It is difficult to predict the impact of these differences on word segmentation from verbal argumentation alone. The goal of Simulation 3 is to determine the models' predictions for conversational speech by running it on the Buckeye corpus.

6. Simulation 3: Robustness to conversational reduction

6.1. Input

6.1.1. Corpus

The Buckeye corpus (Pitt et al., 2007) contains high-quality recordings from 40 age- and gender-stratified lifelong residents of Columbus, OH. Speech was collected in an interview format; speakers were asked their opinions about a variety of local issues such as sports and politics. The speakers were recorded and their speech was orthographically transcribed. An automatic speech recognition program was used to generate forced alignments between the orthographic transcript and a phonetic transcript using canonical dictionary pronunciations.

The research team inspected and made adjustments to this ''canonical'' transcript so as to represent conversational reduction processes, including foot-medial flapping, vowel nasalization, and segment deletion. Thus, the Buckeye corpus includes two alternate phonetic transcriptions of the same speech, a ''canonical'' transcript, and the ''reduced'' transcript.

6.1.2. Division into ''days''

Each transcript was divided into ''days'' consisting of 25,000 words, as in Simulation 1. Owing to the relatively small size of the corpus, there was only enough transcribed material for 7 ''days.''

6.1.3. Training and test sets

The first 6 days were used as the training set; the remaining (partial) day made up the test set.

6.2. Models

The Baseline-DiBS and Phrasal-DiBS models were used, as in Simulation 1.

The Lexical-DiBS model was not included because of the theoretically problematic status of distinct pronunciation variants of the same word: Are distinct pronunciation variants to be counted as distinct words, equivalent to assuming that infants are unable to recognize them as variants of the same word? An additional concern pertains to the frequency threshold that defines which words are in the model's lexicon. Supposing that distinct pronunciation variants are counted as distinct wordforms, the ''same'' threshold has a different meaning for the ''reduced'' transcript. This is because multiple variants of the same word imply a lower frequency for each one; thus, the same frequency threshold will yield a smaller lexicon on the ''reduced'' transcript. Rather than take a position on these questions here, we defer them to future research.

6.3. Method

The method was identical to Simulation 1, except that the ''canonical'' and ''reduced'' versions of the Buckeye corpus were used. Thus, two versions of each model were trained—one was trained on one part of the ''canonical'' transcript and tested on the remainder; the other was trained on one part of the ''reduced'' transcript (and tested on the remainder). Note that these are different transcriptions of the same speech; the only difference between the transcripts is the extent to which they represent phonetic variation of that speech. Because the central point of interest is the effect of conversational reduction (not the developmental trajectory), performance is shown only at the end of training.

6.4. Results

The undersegmentation and oversegmentation error rates after exposure to the training set are shown in Fig. 5.

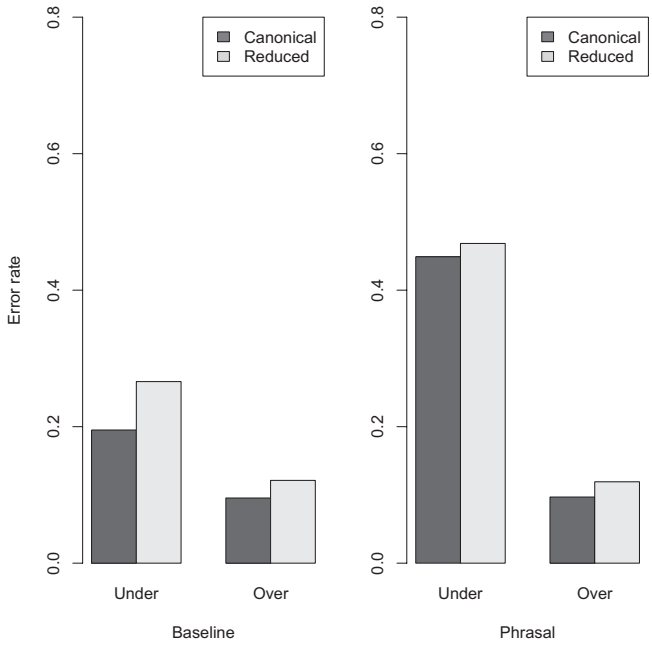

Fig. 5. Undersegmentation and oversegmentation error rates on ''canonical'' (dark gray bars) and ''reduced'' transcripts (light gray bars).

6.5. Discussion

6.5.1. Register

The results of the baseline and phrasal models on the ''canonical'' transcript are in broad agreement with the baseline results from Simulation 1, as evident by comparing Fig. 3 with Fig. 5. The primary difference between the ''canonical'' Buckeye transcript and the child speech is that the Buckeye transcript represents adult-adult conversation.

The qualitative similarity in performance is especially striking in view of the fact that the transcription system differs somewhat between Simulations 1 and 3. Simulation 1 uses the CELEX transcription system, whereas Simulation 3 uses a custom version of the DARPA transcription standard in automatic speech recognition research. While there is considerable overlap in the phone inventories of these two transcription systems, they nonetheless differ in a number of important ways. For example, the CELEX transcription system represents British pronunciation, whereas the Buckeye transcript represents American (Midwestern) pronunciation; one notable difference between these dialects is the phonetic realization of /r/. The fact that a consistent performance profile appears *in spite of* these transcription differences is strong evidence for the robustness of DiBS, as it is exactly the kind of robustness that characterizes human speech processing.

6.5.2. Pronunciation variation

Comparison between the ''canonical'' and ''reduced'' results suggest that pronunciation variation causes only a mild decrement in word segmentation. Consistent with previous experiments, the oversegmentation error rate remains close to 10% in both transcripts (although it goes as high 12% on the ''reduced'' transcript, representing one oversegmentation error per eight words). Again consistent with previous experiments, the undersegmentation error rate is more variable. Thus, both the ''canonical'' and ''reduced'' transcripts exhibit the undersegmentation error pattern found in Simulation 1.

An additional consideration is that pronunciation variability in the Buckeye corpus may present a worst-case scenario for infants. For all languages that have been investigated instrumentally, caregivers use a special speech style when interacting with infants known as ''infant-directed speech,'' characterized by exaggerated pitch excursions and vowel formants (Kuhl et al., 1997). This hyperarticulatory speech style serves to enhance meaningful phonetic contrasts and contains fewer instances of conversational reductions such as segmental deletion. The input that infants get is often much cleaner than the ''reduced'' transcript of the Buckeye corpus.

In summary, the results of Simulation 3 suggest that the qualitative segmentation performance of DiBS is somewhat robust to pronunciation variation. This constancy of perceptual learning outcomes in the face of variability in the input is a hallmark of human speech perception, and thus an important property to exhibit for a cognitively plausible model of word segmentation. Simulation 3 shows that DiBS exhibits the kind of perceptual robustness that human segmentation does.

7. General discussion

7.1. Summary of key findings

The major theoretical contribution of this paper is a learning model of diphone-based segmentation (DiBS) as discussed in Cairns et al. (1997) and Hockema (2006). Using Bayes' theorem and the assumption of phonological independence across word boundaries, infants can estimate DiBS model parameters from the distribution of speech sounds at the edges of observable lexico-prosodic domains, either phrases (Phrasal-DiBS) or the emerging lexicon (Lexical-DiBS). Empirical assessment of the learning models demonstrates three properties that are crucial from a learning perspective: rapid training, robustness, and undersegmentation.

7.1.1. Rapid training

Diphone-Based Segmentation models require a minuscule amount of training data. As shown in Simulation 1, the Phrasal-DiBS model achieves asymptotic performance with less language input than a typical infant receives in a single day. The Lexical-DiBS model may exceed Phrasal-DiBS within a few weeks and continues to improve with additional language exposure. These results indicate a clear prediction that infants will command good

phonotactic segmentation shortly after they meet the model's assumption, in particular, command of native-language phonetic categories. In other words, the learning model here can explain *why* phonotactic segmentation is evident shortly after the emergence of native-language phonetic categories.

7.1.2. Robustness

Also crucial from a learning perspective is robustness. Simulation 2 demonstrates a consistent performance profile in the face of considerable variation to the free parameter $p(\#)$ that represents context-free probability of a word boundary. Simulation 3 demonstrates the same consistent performance profile in the face of considerable pronunciation variation in the input. This consistency of perceptual outcomes in the face of multiple sources of variation is a hallmark of language acquisition generally, and a necessary characteristic of language models.

7.1.3. Undersegmentation

All DiBS models exhibited a consistent pattern of undersegmentation, that is, an error rate of less than 1 oversegmentation error per 10 words (1/8 with conversational reduction). Thus, when the sublexical system identifies a word boundary, the word boundary can generally be trusted. Since this pattern occurred in both the baseline and learning models, this work illustrates a clear prediction of DiBS: The sublexical/phonotactic segmentation mechanism should undersegment throughout the lifespan. The implications of this prediction are considered below for language acquisition and theories of lexical access.

7.2. Implications of undersegmentation

7.2.1. Undersegmentation efficiently pares the hypothesis space

Explicit computational models of lexical access (e.g., Baayen, Schreuder, & Sproat, 2000; Norris & McQueen, 2008) are generally subject to the problem that longer phrases take longer to process. A substantial processing benefit can be gained by analyzing the input into smaller subparts, even if lexical search in humans is a massively parallel operation. Recall that undersegmentation means sublexically identified word boundaries can generally be trusted. The consequence for lexical access is that many analyses which cross-cut each other at trustable word boundaries can be pared, and errors that occur on one side of the boundary cannot propagate to the other (Pierrehumbert, 2001). Trustable word boundaries will pare away lexical searches which do not correctly match the input, and in a way which can only favor the correct analysis. This reduces the overall level of lexical competition, yielding a faster overall lexical access process.

7.2.2. Word learning

There is a further computational/cognitive benefit of undersegmentation, under standard assumptions about lexical access: If lexical access is left-to-right and recursive, then lexical access failure becomes a nearly unambiguous signal for the occurrence of a novel word. Here, ''left-to-right'' means that lexical searches are initiated earlier for phonological input

that is experienced earlier. ''Recursive'' means that unexplained residues of phonological material trigger additional lexical searches; for example, if the system receives *kiss him* as an input and identifies *kiss* as the initial word of this sequence, it immediately initiates a lexical search for matches to the remaining, unexplained *him*. These assumptions are standard in models of word recognition (cf. Baayen et al., 2000; McClelland & Elman, 1986; Norris & McQueen, 2008). It follows from them that input is steadily decomposed into recognizable lexical chunks. As argued above, undersegmentation speeds this process by reducing the scope of lexical searches, but it does not interfere with it. Thus, lexical access failure is predicted to occur only if a stretch of phonological material *cannot* be matched; that is, if the listener lacks a corresponding lexical entry, because the word is novel.

7.2.3. Syntactic acquisition

Recent work suggests that learners may benefit from undersegmented input in early stages of acquisition. For example, learners make fewer errors on irregulars when they are learned in natural phrasal contexts, for example, *on your feet/*foots* (Arnon & Clark, 2009), apparently because they learn the phrase as an unanalyzed whole. This same effect may explain why children acquire grammatical gender more readily than second-language learners (Arnon & Ramscar, 2009). The apparent dovetail between this work and syntactic acquisition raises the possibility of exciting synergies and merits further research.

7.3. Rapid trainability

One important question that has so far gone unaddressed is why the learning models train up so rapidly. We suspect the answer lies in the fact that diphones exhibit a Zipfian (power-law) distribution: A few diphones occur many times, whereas many diphones occur only a few times. This fact is illustrated in Fig. 6, which shows a scatterplot of the frequency distribution of word-internal and word-spanning diphones from the training set of Simulation 1.

The x-axis represents frequency and the y-axis represents diphone types. As standard for power-law distributions, a log-log scale is used. Thus, a point with the coordinates ($\log_{10} f$, $\log_{10} n$) indicates that there are n diphone types whose frequency is f (frequencies which do not occur are not plotted). For visual clarity, the word-internal and word-spanning diphones are slightly offset from one another vertically; in addition, the counts are jittered to make the density visually apparent.

Note the large clouds on the bottom right of the graph, representing a small number (low y) of very high-frequency (large x) diphones: A few diphones which occur many times, accounting for the bulk of the observed events. Similarly, the points in the extreme upper left portion of the graph represent many diphones (large y) occurring just once or a few times each (low x). Known as *data sparsity*, this condition implies the distribution is unstable because many rare events are undersampled (Baayen, 2001). While data sparsity is generally a problem in language models (Manning & Schütze, 1999), the Zipfian distribution is advantageous in the present case, because it guarantees infants receive the most exposure to precisely the diphones that are most important for word segmentation.

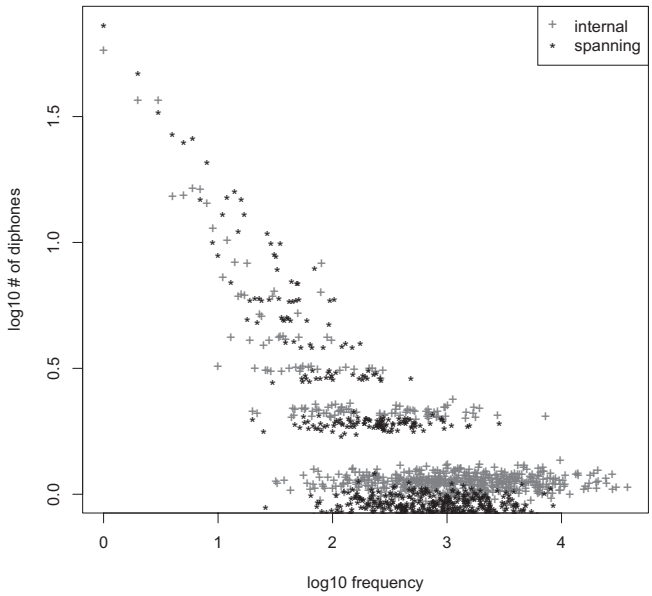

Fig. 6. Zipfian distribution of diphones in the test set. The *x*-axis represents frequency, and the *y*-axis represents the log count of diphone types. Thus, a point at (log f, log n) indicates there are *n* distinct diphone types that have a frequency of *f*. Both word-internal diphones (light gray +'s) and word-spanning diphones (dark gray *'s) are shown. Nonoccurring frequencies are not plotted.

Finally, the overlapping distribution between word-internal and word-spanning diphones indicates that word-internal and word-spanning diphones cannot be distinguished on the basis of frequency alone. Rather, the infant must infer whether a diphone is word-internal or word-spanning based on phonotactics.

7.4. Toward future work

The results of the present study, while promising, are also limited in scope. The present study considered word segmentation using a limited phonotactic domain, and in a single language. It is therefore natural to wonder to what extent these results are specific to diphones and/or English. Understanding this issue provides an entry to the more scientifically important question: Which aspects of DiBS can fruitfully be generalized?

7.4.1. Function/content word distinction

The present study has not carefully addressed the distinction between function and content words owing to lack of space. Inspection of the models' output leads us to speculate that the model is especially good at finding boundaries after content words, and comparatively poor at finding the boundary in a function-content sequence. Framed in terms of the Prosodic Hierarchy (Nespor & Vogel, 1986; Selkirk, 1984), it would appear that DiBS is better suited for segmenting Prosodic Words than orthographic words as prosodically weak

function words are often incorporated with an adjacent content word into a single Prosodic Word. We are testing this hypothesis in follow-up work.

7.4.2. Rapid learning of diphones should generalize

English is relatively unusual among languages for its complex phonotactics, permitting up to three-consonant clusters in onsets (e.g., *strict*) and up to four consonants in codas (e.g., *sixths*). Presumably English allows more diphones than most other languages, and it should therefore take longer to learn. Since English does not take very long, other languages should not either.

7.4.3. Comparative utility of diphones and other cues

English phonotactics exhibits strong contextual constraints on the occurrence of phones. As a result, a considerable amount of positional information is encoded by diphones in English. The high level of segmentation achieved by DiBS stems from the fact that it utilizes this information efficiently. However, diphones do not exhaust the information that is known to be used by English-learning infants in segmenting the speech stream. In an experiment in which diphone statistics were unavailable and phrase boundaries were obscured by ramping the amplitude of stimulus onsets and offsets, Saffran et al. (1996) demonstrated that 8-month-old infants could use transition statistics defined over a two-syllable window to segment a speech stream. A variety of studies (Jusczyk, Houston, et al., 1999; Thiessen & Saffran, 2003) also show that metrical foot structure supports segmentation of the speech stream from 9 months onward.

Anticipating a future model of word segmentation that includes all applicable factors, we observe that the core ideas of DiBS may be of wider utility in modeling other structural domains as well. For example, while the phonotactic possibilities of English are rich enough that unattested syllable types are easily constructed (e.g., *zimp*), Japanese has a closed syllabic inventory that is not much larger than the number of phones in English (Itô & Mester, 1995). This means that the core equations of DiBS could be applied to Japanese syllables rather than phones.

The utility of diphones, in comparison to other phonological units, may prove to be different in different languages. The utility of diphones is likely to be less in languages that have predominately CV (consonant-vowel syllables) than in English. In contrast, the utility of f0 (fundamental frequency) cues might be greater in some other languages than in English. Whereas prosodic words in English have no tonal marking at their boundaries, prosodic words in French exhibit word-final pitch accent. Supposing that infants categorize f0 events in the same general manner as segmental phonetic events, it is possible to collect DiBS-like statistics on syllable-adjacent f0 events rather than segmentally adjacent diphones. The statistics of these events promise to be a good cue for word segmentation in French.

The present study has outlined a learning model for the diphone-based segmentation model discussed in Cairns et al. (1997) and Hockema (2006). The learning model was shown to illustrate three key features: rapid learnability, robustness to free parameter error and conversational reduction processes, and an undersegmentation error profile.

It was argued that undersegmentation offers clear cognitive benefits over alternative error patterns allowing high rates of both undersegmentation and oversegmentation. Undersegmentation efficiently pares the hypothesis space for lexical access by supplying the lexicon with only trustable word boundaries. Thus, a lexical access failure becomes a reliable indicator for the occurrence of a new word—an obvious benefit for word learning. Finally, undersegmentation offers the potential to explain certain puzzles of syntactic acquisition. While the present study has focused on diphones and English specifically, it is to be hoped that the overall approach may generalize to other languages and structural domains.

Notes

1. Albright & Hayes (2003); Beckman, Munson, & Edwards (2007); Edwards, Fox, & Rogers (2002); Hayes & Wilson (2008); Munson (2001).
2. Akhtar & Tomasello (1997); Gathercole, Sebastian, & Soto (1999); Lieven, Pine, & Baldwin (1997); Rubino & Pine (1998); Tomasello (2004); Tomasello & Brooks (1998).
3. These error types are referred to in various ways across different disciplines. An oversegmentation error might be referred to as a false positive in machine learning, or as a Type I error in the social sciences; an undersegmentation error might be called a miss or false negative in machine learning, or a Type II error in the social sciences.
4. Cairns et al. (1997) used the odds ratio, which contains the same information as probability. Probability is used here for formal unity with the rest of the paper.
5. The sentence is represented orthographically for ease of reading. Of course the actual input is speech sounds rather than English letters.
6. This generalization is not absolute. Important questions have been raised owing to an exception (Zulu clicks—Best, McRoberts, & Sithole, 1988); a case in which the shift is evident earlier (<6 months—Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992); a case in which it is evident later (>11 months—Polka, Colantonio, & Sundara, 2001), and a frequency-driven asymmetry (Anderson, Morgan, & White, 2003). These studies do not fundamentally challenge the claim that infants have a categorical representation of speech by 9 months.
7. Function words will generally be more likely at one or both phrase edges for syntactic and pragmatic reasons. For example, the determiners *a* and *the* are likely to begin English sentences because of its default SVO order.

Acknowledgments

We wish to acknowledge support from Northwestern University in the form of a dissertation year fellowship, and from the James S. McDonnell foundation for grant no. 21002061 to the second author. We also wish to thank Matt Goldrick, Jessica Maye, and an anonymous reviewer for helpful comments.

References

Akhtar, N., & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. *Developmental Psychology*, *33*, 952–965.

Albright, A. (2009). Feature-based generalization as a source of gradient acceptability. *Phonology*, *26*, 9–41.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*, 119–161.

Anderson, J., Morgan, J., & White, K. (2003). A statistical basis for speech sound discrimination. *Language and Speech*, *46*, 155–182.

Arnon, I., & Clark, V. E. (2009). Words in frames: Why on your feet is better than feet. Presented at the 83rd Annual Meeting of the Linguistic Society of America, San Francisco, CA.

Arnon, I., & Ramscar, M. (2009). How order-of-acquisition shapes learning: The case of grammatical gender. Presented at the 33rd Boston University Conference on Language Development, Boston, MA.

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

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht, The Netherlands: Kluwer.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (Release 2)*. Philadelphia, PA: Linguistic Data Consortium.

Baayen, R. H., Schreuder, R., & Sproat, R. (2000). Morphology in the mental lexicon: A computational model for visual word recognition. In F. Van Eynde & D. Gibbon (Eds.), *Lexicon development for speech and language processing* (pp. 267–293). Dordrecht, The Netherlands: Kluwer.

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *44*, 568–591.

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

Beckman, M. E., Munson, B., & Edwards, J. (2007). Vocabulary growth and developmental expansion of types of phonological knowledge. In J. Cole & J. I. Hualde (Eds.), *Laboratory Phonology 9* (pp. 241–264). Berlin: Mouton de Gruyter.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 345–360.

Blanchard, D., & Heinz, J. (2008). Improving word segmentation by simultaneously learning phonotactics. In A. Clark & K. Toutanova (Eds.), *Proceedings of the Conference on Natural Language Learning (CoNLL)* (pp. 65–72). Stroudsberg, PA: Association for Computational Linguistics.

Booth, A. E., & Waxman, S. R. (2003). Mapping words to the world in infancy: Infants' expectations for count nouns and adjectives. *Journal of Cognition and Development*, *4*, 357–381.

Bortfeld, H., Morgan, J., Golinkoff, R., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech stream segmentation. *Psychological Science*, *16*, 298–304.

Brent, M. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*, 71–105.

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

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, B33–B44.

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

Chomsky, N., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics*, *1*, 97–138.

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.

Christophe, A., Gout, A., Peperkamp, S., & Morgan, J. L. (2003). Discovering words in the continuous speech stream: The role of prosody. *Journal of Phonetics*, *31*, 585–598.

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., McQueen, J. M., Butterfield, S., & Norris, D. (2008). Prelexically-driven perceptual retuning of phoneme boundaries. In *Proceedings of Interspeech 2008* (p. 2056). Brisbane, Australia: 9th Annual Conference of the International Speech Communication Association.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, *28*, 125–127.

Davis, M. H. (2004). Connectionist modelling of lexical segmentation and vocabulary acquisition. In P. Quinlan (Ed.), *Connectionist models of development: Developmental processes in real and artificial neural networks* (pp. 151–187). Hove, UK: Psychology Press.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 1568–1578.

Edwards, J., Fox, R. A., & Rogers, C. (2002). Final consonant discrimination in children: Effects of phonological disorder, vocabulary size, and articulatory accuracy. *Journal of Speech, Language, and Hearing Research*, *45*, 231–242.

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

Fleck, M. M. (2008). Lexicalized phonotactic word segmentation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 130–138). Madison, WI: Omnipress.

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

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 110–125.

Gathercole, V. C. M., Sebastian, E., & Soto, P. (1999). The early acquisition of Spanish verbal morphology: Across-the-board or piecemeal knowledge? *International Journal of Bilingualism*, *3*, 138–182.

Goldwater, S. (2006). *Nonparametric Bayesian models of lexical acquisition*. Unpublished dissertation, Brown University.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21–54.

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, *18*, 254–260.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*, 379–440.

Hockema, S. A. (2006). Finding words in speech: An investigation of American English. *Language Learning and Development*, *2*, 119–146.

Itô, J., & Mester, R. A. (1995). Japanese phonology. In John. A. Goldsmith (Ed.), *The Handbook of Phonological Theory* (pp. 817–838). Cambridge, MA: Blackwell Handbooks in Linguistics, Blackwell Publishers.

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (Eds.), *Spontaneous speech: Data and analysis. Proceedings of the 1st Session of the 10th International Symposium* (pp. 29–54). Tokyo, Japan: The National International Institute for Japanese Language.

Johnson, M., & Goldwater, S. (2009). Improving nonparametric Bayesian inference: Experiments on unsupervised word segmentation with adaptor grammars. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, pp. 317–325. Boulder, Colorado, June 2009.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 39–96). Cambridge, MA: MIT Press.

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

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

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., & Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to Infants. *Science*, *277*, 684–686.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*, F13–F21.

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.

Lehiste, I. (1960). An acoustic-phonetic study of open juncture. *Phonetica, Supplementum ad*, *5*, 1–54.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 234–243). Madison, WI: Omnipress.

Lieven, E. V. M., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, *24*, 187–219.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.

Luce, P., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes*, *16*, 565–581.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing*, *19*, 1–36.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT.

Massaro, D. W., & Cohen, M. M. (1983). Integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 753–771.

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, *78*, 91–121.

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.

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, *134*, 477–500.

McCarthy, J., & Prince, A. (1986/1996). *Prosodic Morphology 1986* (Technical Report No. 32). New Brunswick, NJ: Rutgers University Center for Cognitive Science.

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

Mehl, M. R., Vazire, S., Ramirez-Esparza, N., Slatcher, R. B., & Pennebaker, J. W. (2007). Are women really more talkative than men? *Science*, *317*, 82.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91–117.

Mitterer, H., Yoneyama, K., & Ernestus, M. (2008). How we hear what is hardly there: Mechanisms underlying compensation for /t/-reduction in speech comprehension. *Journal of Memory and Language*, *59*, 133–152.

Miyawaki, K., Strange, W., Verbrugge, R. R., Liberman, A. M., Jenkins, J. J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception and Psychophysics*, *18*, 331–340.

Moreton, E. (1997). Phonotactic rules in speech perception. *Journal of the Acoustical Society of America*, *102*, 3091–3092.

Munson, B. (2001). Relationships between vocabulary size and spoken word recognition in children aged 3 to 7. *Contemporary Issues in Communication Science and Disorders*, *28*, 20–29.

Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Dordrecht: Foris.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*, 357–395.

Peperkamp, S., Le Calvez, R., Nadal, J.-P., & Dupoux, E. (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, *101*, B31–B41.

Peterson-Hicks, J. (2006). *The impact of function words on the processing and acquisition of syntax*. Unpublished dissertation, Northwestern University.

Pierrehumbert, J. B. (2001). Why phonological constraints are so coarse-grained. *Language and Cognitive Processes*, *16*, 691–698.

Pierrehumbert, J. B. (2002). Word-specific phonetics. In C. Gussenhoven & N. Warner (Eds.), *Laboratory phonology 7* (pp. 101–139). Berlin: Mouton de Gruyter.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). *Buckeye corpus of conversational speech (2nd release)* [www.buckeyecorpus.osu.edu]. Columbus, OH: Department of Psychology, Ohio State University (Distributor).

Polka, L., Colantonio, C., & Sundara, M. (2001). Cross-language perception of/d – ð/: Evidence for a new developmental pattern. *Journal of the Acoustical Society of America*, *109*, 2190–2200.

Rubino, R., & Pine, J. (1998). Subject–verb agreement in Brazilian Portuguese: What low error rates hide. *Journal of Child Language*, *25*, 35–60.

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

Selkirk, E. O. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.

Soderstrom, M., Kemler-Nelson, D. G., & Jusczyk, P. W. (2005). Six-month-olds recognize clauses embedded in different passages of fluent speech. *Infant Behavior and Development*, *28*, 87–94.

Storkel, H. L., Armbruster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*, *49*, 1175–1192.

Storkel, H. L., & Maekawa, J. (2005). A comparison of homonym and novel word learning: The role of phonotactic probability and word frequency. *Journal of Child Language*, *32* (4), 827–853.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*, 86–132.

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

Thorn, A. S. C., & Frankish, C. R. (2005). Long-term knowledge effects on serial recall of nonwords are not exclusively lexical. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 729–735.

Tomasello, M. (2004). The item-based nature of children's early syntactic development. *Trends in Cognitive Science*, *4*, 156–163.

Tomasello, M., & Brooks, P. (1998). Young children's earliest transitive and intransitive constructions. *Cognitive Linguistics*, *9*, 379–395.

Tomasello, M., & Farrar, J. M. (1986). Joint attention and early language. *Child Development*, *57*, 1454–1463.

Tomasello, M., Mannle, S., & Kruger, A. (1986). Linguistic environment of 1- to 2-year-old twins. *Developmental Psychology*, *22*, 169–176.

Trehub, S. E. (1976). The discrimination of foreign speech contrasts by infants and adults. *Child Development*, *47*, 466–472.

Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2006). Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants. *Journal of the Acoustical Society of America*, *120*, 2285–2294.

van de Weijer, J. (1998). *Language input for word discovery*. Unpublished dissertation, Nijmegen: Max Planck Institute for Psycholinguistics.

Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech *Computational Linguistics*, *27*(3), 351–372.

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

Xanthos, A. (2004). Combining utterance-boundary and predictability approaches to speech segmentation. In W. G. Sakas (Ed.), *Proceedings of the First Workshop on Psycho-computational Models of Language Acquisition at COLING 2004* (pp. 93–100). Geneva, Switzerland.