J. Child Lang., Page 1 of 25. © Cambridge University Press 2010
doi:10.1017/S030500090999050X

Modeling the contribution of phonotactic cues to the problem of word segmentation*

DANIEL BLANCHARD, JEFFREY HEINZ
AND ROBERTA GOLINKOFF
University of Delaware

(*Received* 19 *December* 2008 – *Revised* 27 *July* 2009 – *Accepted* 5 *December* 2009)

ABSTRACT

How do infants find the words in the speech stream? Computational models help us understand this feat by revealing the advantages and disadvantages of different strategies that infants might use. Here, we outline a computational model of word segmentation that aims both to incorporate cues proposed by language acquisition researchers and to establish the contributions different cues can make to word segmentation. We present experimental results from modified versions of Venkataraman's (2001) segmentation model that examine the utility of: (1) language-universal phonotactic cues; (2) language-specific phonotactic cues which must be learned while segmenting utterances; and (3) their combination. We show that the language-specific cue improves segmentation performance overall, but the language-universal phonotactic cue does not, and that their combination results in the most improvement. Not only does this suggest that language-specific constraints can be learned simultaneously with speech segmentation, but it is also consistent with experimental research that shows that there are multiple phonotactic cues helpful to segmentation (e.g. Mattys, Jusczyk, Luce & Morgan, 1999; Mattys & Jusczyk, 2001). This result also compares favorably to other segmentation models (e.g. Brent, 1999; Fleck, 2008; Goldwater, 2007; Johnson & Goldwater, 2009; Venkataraman, 2001) and has implications for how infants learn to segment.

[*] This work was supported by a University of Delaware Research Foundation grant to the second author, and by NIH (5R01HD050199) and NSF grants (BCS-0642529) to the third author. We thank Vijay Shanker for valuable discussions, and Regine Lai and Aimee Stahl for feedback on the manuscript. Address for correspondence: Daniel Blanchard, University of Delaware – Computer & Information Sciences, 101 Smith Hall, Newark, Delaware 19716, United States. e-mail: dsblanch@udel.edu

INTRODUCTION

How do infants come to identify words in the speech stream? Adults break up speech into words automatically and effortlessly, without realizing that there are no pauses between words in the same sentence. Unlike many written languages, speech does not generally have reliable markers for word boundaries (Cole & Jakimik, 1980). When such markers can be found, they vary across languages (Cutler & Carter, 1987). These facts make the task of isolating the cues used for picking out words from a speech signal especially difficult. The task facing the human infant is more daunting. Adults have a lexicon they can use to recognize familiar words in the speech stream, but when infants are born, they have no pre-existing lexicon to consult. In spite of these challenges, by the age of six months, infants are already segmenting some words from speech (Bortfeld, Morgan, Golinkoff & Rathbun, 2005).

Here we present an efficient word segmentation system called PHOCUS, for PHonotactic CUe Segmenter, aimed to model how infants accomplish this task. There are four main contributions of this work. First, this model shows that the use of phonotactic cues improves the accuracy of existing segmentation models. These findings support the hypothesis that phonotactic cues are useful for segmentation (Mattys *et al.*, 1999; Mattys & Jusczyk, 2001). Second, the model shows that it is possible to learn language-specific phonotactic constraints while simultaneously segmenting words. These two processes feed each other with the model initially learning phonotactic constraints from entire unsegmented utterances. This helps the model segment later utterances, which consequently helps the model refine the constraints it extracts from the developing lexicon. Third, this model shows that the language-universal concept of a syllable greatly facilitates the above results, but is of little value when used on its own. Finally, we propose a general phonotactic-learning model to be embedded within a word segmenter in order to facilitate the study of the relative importance of a variety of phonotactic cues. Such a model potentially allows the systematic investigation of the contributions particular phonotactic cues and their combinations make to the segmentation process at different epochs, providing a framework within which collaborative efforts between modelers and experimentalists can obtain a deeper understanding of how infants come to segment speech.

Hereafter, we use the phrase 'word segmentation' to mean some process which adds word boundary symbols to a text that does not already contain them. A word segmentation model is a computational implementation of this process. This begs the question of what constitutes a word, which we discuss in the first section below. This paper does not directly address the problem of segmenting auditory linguistic stimuli, but any word

segmentation model could easily be plugged into a system that recognizes phonemes from speech (e.g. Mohri, 2005).

PHOCUS is an unsupervised and incremental algorithm. That is, it does not rely on pre-existing knowledge of a particular language, and it segments the corpus one utterance at a time. This is in contrast to supervised word segmentation algorithms (e.g. Teahan, McNab, Wen & Witten, 2000). Essentially, supervised learners receive correct segmentations as feedback. In practice, this amounts to supplying a lexicon beforehand since these models are typically used for segmenting text in documents written in languages that do not put spaces between their words – like Chinese. The model presented here also differs from batch segmentation algorithms (e.g. Fleck, 2008; Goldwater, 2007; Johnson & Goldwater, 2009), which process the entire corpus at least once before outputting a segmentation of the corpus. Unsupervised incremental algorithms are of special interest in modeling infant segmentation given that: (1) infants do not have an a priori lexicon; and (2) memory limitations suggest that it is unlikely that infants process large batches of linguistic information at once.

Unsupervised incremental algorithms are especially challenging to develop, as there is very little information for learners to use to make decisions at the beginning. This contrasts with supervised systems which have an a priori lexicon, and batch systems which may examine the whole corpus for trends before segmenting. Furthermore, because the algorithm is unsupervised there is no external feedback which lets it know when a particular segmentation is incorrect. Consequently if poor decisions are made, it may be impossible for unsupervised incremental algorithms to recognize the error and reverse the errors in the future. Even worse, early errors can trigger many more (a relevant example is given later).

This article first outlines a framework of word segmentation based on what is known about how children segment utterances. We also describe the Emergent Coalition Model (ECM) (Golinkoff & Hirsh-Pasek, 2006; Hollich *et al.*, 2000) of word learning, which serves as a theoretical impetus for the view of multiple, competing cues for segmentation presented here. The second section introduces the segmentation models of Brent (1999) and Venkataraman (2001), which are very similar in character. Venkataraman's model forms the basis for PHOCUS, and therefore becomes a baseline against which to compare it. The theoretical motivation for using phonotactics for segmentation is covered in the third section. The next two sections present the phonotactically enhanced segmentation model, and compare its performance to different segmenters on multiple corpora. We close with a discussion of future work and the conclusions we can draw from the current results.

A FRAMEWORK FOR WORD SEGMENTATION

Cues for segmentation

Although first investigated in the 1950s (Harris, 1954), word segmentation is a research topic that has seen a surge in popularity in the past fifteen years. Researchers have uncovered a number of cues that infants appear to use to segment speech (Saffran, Werker & Werner, 2006). We consider two classes of cues for which there is much evidence: use of familiar words and phonotactic cues.

Familiar words. Infants, like adults, can use familiar words to help them discover new words in the speech stream. However, it is not clear that infants are associating any semantic information with these word forms. This view is consistent with Jusczyk's (1993) WRAPSA hypothesis that infants first obtain phonological forms which are then filled with meaning. Bortfeld *et al*. (2005) showed that six-month-olds could use familiar words (their own name and some version of *mother*) to identify new words in utterances. Using the Head-Turn Preference Paradigm (Nelson, Jusczyk, Mandel, Myers, Turk & Gerken, 1995), Bortfeld *et al*. (2005) presented infants with a novel word that followed their name (e.g. *I like Sally's wug*). At test, infants listened longer to the word that followed their own name than to a word that followed someone else's name with the same number of syllables and the same stress pattern. Not only can children use familiar content words, Shi & Lepage (2008) showed that French-reared eight-month-olds could use frequent function morphemes, such as *des* and *mes*, to segment speech. This research supports the hypothesis that once infants recognize some words, they can use them to add new ones to their lexicon.

However, the question of how the first words are extracted is still unanswered. According to Brent & Siskind (2001), infants learn these first words from one-word utterances. If infants are predisposed to consider utterances as words, then they will add entire multisyllabic utterances to their lexicons at first. Although this results in many initial mistakes, as long as some utterances infants hear consist of one word, this strategy could be enough to bootstrap the lexicon. According to Brent & Siskind (2001), as much as 10 percent of infant-directed speech is made up of one-word utterances. It is also plausible that infants use discourse cues like rephrasing (as when parents put the same word into different places in the utterance), and that familiar phrases (such as *Look at the __ !*) would serve as indicators of a word boundary just as well as individual words.

Phonotactic cues. The phonotactics of a language are language-specific conditions that determine whether a word is well-formed or not (Chomsky & Halle, 1965; Halle, 1978). For example, although English words may contain the velar nasal [ŋ] (e.g. *sing* [sɪŋ], *Lincoln* [lɪŋkn̩]), no words in English begin with this sound. Furthermore, adult native speakers of

English would not name objects or actions with logically possible words which begin with [ŋ], such as *ngep* [ŋɛp]. They also judge nonce words beginning with [ŋ] ([ŋɛp]) as 'worse' than words which are the same in all other respects ([nɛp]) (Sapir, 1925). Many other languages allow [ŋ] word-initially, so this phonotactic constraint is specific to English.

We consider phonotactics to be synonymous with word well-formedness, and not exclusively to mean phoneme combinations. Thus, any rules that govern how subunits of words may combine to form well-formed words count as phonotactics. Consequently, even stress patterns can be considered a kind of phonotactic constraint that operates across syllables.

Studies show that infants learn phonotactic patterns of different types by roughly eight months of age (Friederici & Wessels, 1993; Jusczyk, Friederici, Wessels, Svenkerud & Jusczyk, 1993; Jusczyk, Houston & Newsome, 1999; Thiessen & Saffran, 2003). This has led researchers to propose that infants use their knowledge of word well-formedness to help them segment text. At the lowest level, infants' sensitivity to permissible allophonic variations helps them find word-like units. Jusczyk, Hohne & Baumann (1999) showed that seven-and-a-half-month-olds use their knowledge of allophonic variation to segment utterances (*nitrate* [naɪ.tʰɹeɪt] vs. *night rate* [naɪt ɹeɪt]). At the next level, knowledge of which phonemes occur together in their language assists infants in making appropriate segmentations. The idea is simple – if infants know that words do not begin with [ŋ], for example, then when faced with an utterance like *Sing it!* [sɪŋɪt], they will not be tempted to segment the utterance into words [sɪ] and [ŋɪt]. Similarly, Mattys & Jusczyk (2001) showed that nine-month-olds can segment speech by using the difference in probabilities between within-word and across-word consonant clusters. For example, the novel phrase *fang tine* [faŋ taɪn] is segmented as it is because [ŋt] does not occur within English words. On the syllabic level, infants come to identify predictable stress patterns. Jusczyk, Houston & Newsome (1999) showed that seven-and-a-half-month-olds take advantage of the trochaic stress pattern found in most words in English to segment utterances.

Although word well-formedness is logically distinct from transitional probability, a working hypothesis in probabilistic models of phonotactic learning equates them (e.g. Coleman & Pierrehumbert, 1997; Hayes & Wilson, 2008). This is because both transitional probabilities and phonotactics can be expressed in terms of conditional probability. Continuing the example above, the probability that [t] follows [ŋ] is vanishingly small. Accordingly, word-internal [ŋt] sequences are considered ill-formed (so infants posit word boundaries between them).

Finally, infants compute the transitional probabilities between syllables or phonemes to find words in speech. Saffran, Aslin & Newport (1996) showed in experiments that eight-month-olds segment utterances based

on lower transitional probabilities that existed both at the syllabic and phonemic levels of the training data. Under the working hypothesis mentioned above, transitional probabilities between syllables can be considered a kind of phonotactic constraint: well-formed multisyllabic words are those whose syllables have high transitional probabilities.

The only limits to the number and kind of phonotactic cues potentially useful for word segmentation are the number and kind of phonotactic patterns found in natural languages. In addition to the co-occurrence restrictions and stress patterns mentioned above, it is plausible that infants also make use of consonantal and vowel harmony patterns to segment speech, though to our knowledge this has not been investigated experimentally.

The challenge. The evidence suggests infants do make use of phonotactic cues to segment utterances. However, this leads to a chicken-and-egg conundrum. Since phonotactic constraints govern the well-formedness of words – as opposed to utterances – how do children learn these language-specific phonotactic constraints without a lexicon? Many phonotactic-learning models take word-sized units as input (e.g. Coleman & Pierrehumbert, 1997; Hayes & Wilson, 2008; Heinz, 2007), which is not necessarily representative of how infants approach the problem. Word segmentation models, however, take as input utterances without word boundary markers, necessitating that phonotactic constraint discovery and word discovery happen simultaneously.

One of the main contributions of this paper is that we show that, with the right model, the phonotactics of a language can be learned simultaneously as children segment words. Essentially, our model jump-starts the lexicon using isolated words as discussed above. This tiny lexicon allows the learner to infer some rudimentary language-specific phonotactic constraints, which in turn helps in segmenting additional words. Knowledge of familiar words, combined with increasingly refined phonotactic constraints, support and reinforce each other in speech segmentation. Brent & Cartwright (1996) took a step toward using phonotactic cues for word segmentation with a semi-supervised model, which learned acceptable consonant clusters at the beginning and ends of unsegmented utterances, and then used those clusters as phonotactic constraints for segmentation. Similarly, Fleck's (2008) WordEnds model segments by learning what clusters of phonemes of variable length are most predictable word-initially and word-finally. PHOCUS differs from both of these models in that it is neither learning only word-initial and word-final constraints, nor making an initial pass over the entire corpus to learn these constraints before outputting a segmentation.

How multiple cues get along. The aforementioned cues have been studied in isolation in controlled experimental contexts to determine whether they were factors in word segmentation for a particular age group (but see

Thiessen & Saffran, 2003; Toro, Nespor, Mehler & Bonatti, 2008). As a result, researchers could know neither whether infants paid more attention to one cue or another, nor at what age different cues became accessible to infants. Later research has begun to address these questions. Some cues come in before others – such as frequently occurring words (Bortfeld *et al.*, 2005) before stress patterns (Jusczyk, Houston & Newsome, 1999). This is likely because frequency matters: infants often hear a core of highly common words (their own name, *mommy*, etc.) in all positions within utterances (initial, medial and final), but hearing a variety of words would be especially useful for inferring their language's dominant stress pattern.

Another byproduct of studying cues in isolation is that thus far there has been little work on whether these cues complement each other. At the same time, cues may form a 'coalition' and come together to determine plausible segmentations; they may even compete, or interfere, with one another as new cues come on-line. For example, Thiessen & Saffran (2007) have argued that statistical cues (such as the probability that one syllable reliably follows another) precede stress cues in their use. Furthermore, to our knowledge no approach has yet attempted to uncover how the use of early segmentation cues influences the emergence of subsequent cues.

Emergent Coalition Model of word learning as it applies to segmentation

The approach adopted here is inspired by the Emergent Coalition Model (ECM) of word learning (Golinkoff & Hirsh-Pasek, 2006; Hollich *et al.*, 2000). Although word segmentation may be considered a different (though not unrelated) process to word learning, there are many similarities.

The ECM is a hybrid model of word learning that has three fundamental tenets. First, children are surrounded by multiple cues to word learning: perceptual, social and linguistic. Each type of cue is not always accessible, reliable or harnessed by the infant for word learning. Second, word-learning cues change their relative importance over time. Although a range of cues in the coalition is always available, not all cues are equally utilized in the service of word learning. Children beginning to learn words rely on a perceptual subset of the available cues in the coalition, and only later do they recruit social cues like a speaker's eye gaze and handling of an object to learn words (Hollich *et al.*, 2000). Third, the principles of word learning are emergent, changing over time. Infants may start with an immature principle of reference, such that a word will be mapped to the most salient object and not necessarily to the one the speaker is naming. Later, children sensitive to speaker intent map a word onto an object from the speaker's point of view by using the speaker's social cues.

Although the cues differ in the domain of segmentation, the same general tenets can be maintained. That is, there are multiple cues to segmentation

(familiar words and a variety of phonotactic cues), though not every cue is always accessible, reliable or harnessed. Children appear to rely on different cues across developmental time (Thiessen & Saffran, 2003). Furthermore, in the same way that the cues for word learning change over time, the later-appearing cues for segmentation may emerge from the application of the early-appearing cues. Thus, the process of segmentation itself undergoes change with development as more cues are discovered by the infant.

Each of these tenets makes empirical predictions about the developmental course of segmentation. The virtue of modeling segmentation is that it helps us understand why certain cues fall out or emerge from the use of earlier cues, potentially explaining earlier experimental results, as well as suggesting further experiments to test predictions the model makes.

What is a word?

When developing a computational procedure to segment utterances into words, one immediately faces a thorny question: What exactly constitutes a 'word'? This question has proved difficult for linguists. Matthews (1991), in a seminal book on morphology, waited until page 208 to say, 'there have been many definitions of the word, and if any had been successful I would have given it a long time ago, instead of dodging the issue until now'.

Here we follow Dixon & Aikhenvald's (2002) illuminating discussion of words in natural language. There are phonological words, grammatical words and orthographic words. Grammatical words are defined as consisting of 'a number of grammatical elements' that cannot be separated, 'occur in a fixed order' and 'have a conventional coherence and meaning' (Dixon & Aikhenvald, 2002).[1] Conversely, a phonological word can be defined roughly as a unit of at least one syllable such that there are phonotactic constraints governing its structure, and/or some phonological rules can only apply within or between such units. One example highlighting the difference between the two types of words in English is *it's*. *It's* consists of two grammatical words (*it* and *s*), but only one phonological word (*it's*). This is because *s* is a clitic, and while it has a distinct meaning, it cannot stand on its own as a phonological word, as it does not consist of at least one syllable. Orthographic words, which are determined by a society's writing conventions, do not necessarily line up with either phonological or grammatical words, though they often line up with one or the other (Dixon & Aikhenvald, 2002).

As our computational model operates over phonetically transcribed text, and one of our goals is to examine the contribution phonotactic cues make to

[1] There are a number of possible exceptions to these criteria, but in general, the definition seems to hold.

the segmentation process, our target unit for extraction is the phonological word. If we were to use grammatical words, the model would not be learning phonotactic constraints over the correct domain and would not be developmentally appropriate, as they are not the type of words infants first acquire.

PHOCUS: A PHONOTACTIC CUE SEGMENTER

Baseline model

In this section, we describe PHOCUS. Essentially, PHOCUS is a modified version of Venkataraman's (2001) model and is similar to MBDP-Phon (Blanchard & Heinz, 2008). The code for PHOCUS, along with documentation for installation and usage is available at http://cis.udel.edu/~blanchar/research/.

What properties would a model of word segmentation have that would most resemble what an infant might do? First, the model should work incrementally, segmenting each utterance as it encounters it, rather than waiting until it has seen the entire corpus. Second, the model should not be heuristically biased such that it overlooks a possibly correct segmentation. Finally, the model should base its segmentation decisions on the lexicon it has acquired so far. Such a model allows the incorporation of word well-formedness conditions that are acquired from the current lexicon.

One additional criterion when designing a model of word segmentation is grounded in the computational domain. The model must be probabilistically sound; that is, it must describe a probability distribution over all logically possible words that sums to one. This ensures that the model functions in a more predictable fashion, making it easier to conduct analyses of the factors that effect its performance.

There is one well-known model which satisfies the aforementioned constraints: the one described by Venkataraman (2001). This model uses the idea of isolated words at its core. That is, it adds whole utterances to its lexicon when it is completely unsure of how to segment a string. It also learns the most rudimentary of logically possible phonotactic constraints: words that contain frequently observed phonemes are better than those with rare phonemes (e.g. 'words containing [n] are better than words containing [ŋ]'). While this may seem like an overly simple approach to deciding word well-formedness, Venkataraman's model, along with MBDP-1[2] (Brent, 1999), was the most accurate unsupervised word segmentation systems until Goldwater (2007). Both Brent (1999) and Venkataraman (2001) suggested that their models should be extended to

[2] MBDP stands for Model-Based Dynamic Programming. The '1' indicates Brent's desire for the development of subsequent versions.

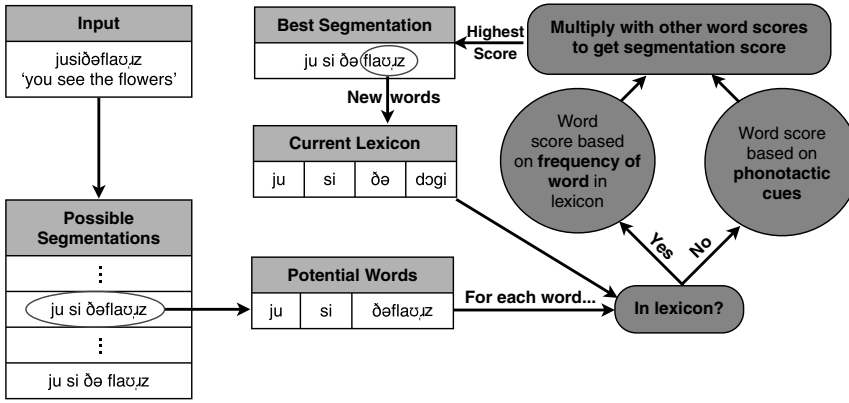

Fig. 1. PHOCUS: Venkataraman's model with *n*-grams over phonemes, $n > 1$.

incorporate more sophisticated, natural language-like phonotactic models, such as the one presented here.

Although Venkataraman's model is almost functionally identical to Brent's MBDP-1 (which Venkataraman (2001) explains is due to some of the terms in his model being approximately equivalent to the terms in Brent's (1999) model and vice versa), there is one important difference. Venkataraman's model initially assumes a uniform probability distribution over the phonemes, whereas MBDP-1's initial state assumes no well-formed probability distribution over the phonemes. Not only does this make MBDP-1 not probabilistically sound, it makes its performance much less predictable, as we discuss below (see Goldwater (2007) for additional discussion). We implemented Venkataraman's model and were able to replicate the results in the 2001 paper.

PHOCUS, illustrated in Figure 1, is very similar to Venkataraman's (2001) model. It initially assumes an empty lexicon. When given an utterance, PHOCUS chooses the most likely segmentation from all possible segmentations. The likelihood of any particular segmentation is obtained by multiplying together the probabilities of the individual words that make up the segmentation. How PHOCUS determines the likelihood of any particular word depends on whether it is familiar (i.e. exists in the current lexicon) or not. If it is familiar, its probability is equivalent to, considering all words posited so far, the percentage of words that are the familiar one. When a word is unfamiliar (i.e. not in the lexicon), PHOCUS assigns a likelihood to it based on its phonotactic well-formedness. When the most likely segmentation is determined, the frequency counts in the lexicon are updated. Consequently, any unfamiliar words in the segmentation are added to the lexicon, and are henceforth considered familiar. As a result of

this procedure, the first utterance of any corpus is added to the lexicon as a whole word, in accordance with Brent & Siskind's (2001) observations discussed earlier.

Where PHOCUS departs from Venkataraman (2001), is the phonotactic cues employed which determine a word's well-formedness. There are two phonotactic cues used to evaluate unfamiliar words. The first determines word well-formedness based on phoneme combinations. Since languages differ in the kinds of phoneme combinations that are allowed within words, this is a language-specific phonotactic that must be learned. The idea is simple: segmentations that include words with unlikely phoneme combinations are less well-formed. The second phonotactic is a universal constraint: well-formed words must have at least one SYLLABIC sound. A sound is syllabic if it is the nucleus of a syllable in a word. In English all vowels are syllabic, and there are also syllabic [l,n,ɹ] sounds (e.g. *bottle* [batl̩], *button* [bʌtn̩], *butter* [bʌtɹ̩]). As explained below, the syllabic consonants are transcribed differently from their non-syllabic counterparts in the English corpus we test PHOCUS on.[3] Unlike the phoneme combination phonotactic, this constraint is plausibly a priori, and does not need to be learned. This is because phonological words are made up of syllables, and syllables must have a nucleus.

Phoneme combinations

According to the phoneme combination cue, the likelihood of an unfamiliar word is determined by the likelihood of phoneme combinations within it. This differs from Venkataraman (2001), which only uses the likelihood of individual phonemes. The probabilities of phoneme combinations can be modeled with a traditional N-GRAM model over phonemes. An *n*-gram model is one that estimates the probability of a sequence by calculating how frequently different subsequences of phonemes (of length *n*) occur in the corpus (Jurafsky & Martin, 2008). For example, suppose PHOCUS encounters the string *He's right* [hizrait] and then considers the one-word segmentation [#hizrait#] (# is the word boundary symbol). With *n* set to two, the *n*-gram model estimates the probability of the word [#hizrait#] by multiplying the conditional probabilities of the phoneme pairs that it consists of ([#h], [hi], [iz], [zr], [ra], [ai], [it] and [t#]). The likelihood of a phoneme *n*-gram is determined by dividing its frequency by the frequency of its $(n-1)$ long prefix. PHOCUS initially assumes a uniform probability distribution over the phoneme *n*-grams. In the example above, if none of [hiz], [rait] or [hizrait] is in the lexicon, the idea is that PHOCUS may

[3] The syllabic consonants are plausibly distinguished acoustically from their non-syllabic counterparts (Toft, 2002; Xie & Niyogi, 2006).

prefer the segmentation [#hiz#rait#] over [#hizrait#] because by this point the bigram [zr] has such a low likelihood (i.e. by this point the algorithm has learned [zr] is such an unlikely combination), that it drastically reduces the overall score of [#hizrait#], but not [#hiz#rait#].

PHOCUS updates the frequency counts of the phoneme *n*-grams immediately after it updates the lexicon. The frequency counts of the phoneme *n*-grams are calculated from the lexicon, not the corpus (i.e. we measure *n*-gram frequencies from word types, not word tokens; see Venkataraman, 2001, for discussion).

We refer to the model which keeps track of phoneme *n*-grams as PHOCUS-*n* (PHOCUS-1 is identical to Venkataraman's (2001) model). In other words, the model can be made to find words and at the same time keep track of single phonemes (PHOCUS-1), phoneme pairs (PHOCUS-2) or phoneme triples (PHOCUS-3). Below we report results for PHOCUS-1, PHOCUS-2 and PHOCUS-3. We do not consider *n*-grams *n* greater than 3 since such models often run into the problem of overfitting (Jurafsky & Martin, 2008). That is, when the length of the phoneme *n*-grams is too long, the model will not see enough examples of *n*-grams of that length (as there are exponentially more possible *n*-grams as the value of *n* increases), and will not learn general enough phonotactic constraints.

Requiring syllabic sounds

The other phonotactic cue PHOCUS uses is a constraint that requires hypothetical words to have syllabic sounds. If a hypothetical word does not have a syllabic sound, it receives a likelihood of zero. Because the probability of any segmentation is the product of the probabilities of each word in it, any segmentation of an utterance which contains a word with no syllabic element receives a probability of zero. For example, the segmentation of *he's right* as [#hi#zr#ait#] would receive a likelihood of zero because the hypothetical word [zr] has no syllabic element.

We refer to the model with only this syllabic constraint, and no attention to phoneme combinations, as PHOCUS-s. In the next section we report results with PHOCUS-s, as well as its use in combination with PHOCUS-2 and PHOCUS-3.

MODEL EVALUATION

The corpora

Computational models are evaluated by studying their performance on different corpora. Generally, a model is deemed more successful if it effectively segments utterances in a variety of languages. Here we used two

child-directed corpora, one in English and one in Sesotho (Bantu), to test the model's generality.

Bernstein-Ratner (1987) corpus. The Bernstein-Ratner (1987, hereafter BR) corpus from the CHILDES database (MacWhinney & Snow, 1985) consists of 9,790 utterances containing 33,399 words of English infant-directed speech. The BR corpus is the same one that Brent (1999), Venkataraman (2001), Goldwater (2007), Fleck (2008) and Johnson & Goldwater (2009) used to evaluate their models, and it has become the de facto standard for segmentation testing ever since it was phonemicized by Brent & Cartwright (1996).

The transcription system described in Brent & Cartwright (1996) makes some unorthodox choices. In particular, complex sounds traditionally transcribed with multiple symbols are transcribed with only one. These include diphthongs and vowels followed by /ɹ/. Another decision was to use different symbols for stressed and unstressed syllabic /ɹ/ – that is, there are different symbols for the /ɹ/ in *butter* and the /ɹ/ in *bird* – though stress is not marked elsewhere in the corpus. Following Blanchard & Heinz (2008), we use a modified version of the corpus where the bi-phone symbols were split into two[4] and the syllabic /ɹ/ symbols were collapsed into one. Blanchard & Heinz (2008) showed that current segmentation models do worse on the modified BR corpus, because the models have to learn that the diphthongs always co-occur without incorrectly grouping them together into their own words.

Sesotho corpus. Johnson (2008) trimmed the Demuth (1992) corpus from the CHILDES database (MacWhinney & Snow, 1985) of speech between mother–child dyads to include only the child-directed speech. He did not convert the orthography to phonemes, because the writing system for Sesotho is nearly phonemic to begin with.[5] The final corpus contains 8,503 utterances consisting of 21,037 word tokens.

Evaluation procedure

As a general guide to a model's performance, we used a standard metric in computational linguistics: a combination of precision and recall, known as the F_0 score. Precision (also known simply as accuracy in the cognitive science community) is the percentage of items identified that are correct. Recall (also known as completeness) is the percentage of correct items identified. To illustrate the difference between these two measures, a segmentation system could achieve a boundary precision of 100% by simply

[4] Only diphthongs whose first phoneme can occur in isolation in English were split, so the vowels in *bay* and *boat* were not split.

[5] In addition to vowels, nasals sounds and the lateral liquid [l] can be syllabic in Sesotho. However, these sounds are not marked as such in the transcription, and so we treated all [l] and [n] sounds as non-syllabic.

inserting one correct boundary into the entire corpus, because 100% of the boundaries it inserted would be correct (although lacking all others). On the other hand, a segmentation model could achieve a boundary recall of 100% by inserting word boundaries between every phoneme in the corpus, because it would insert all of the correct boundaries (in spite of many extras). It is clear that neither precision nor recall is sufficient, and so the harmonic mean is used, called F_0.[6] We follow earlier researchers in reporting precision, recall and F_0 scores for word identification (as opposed to boundary), since words are the ultimate goal of the segmentation process (Brent, 1999; Goldwater, 2007).[7]

Despite representing an appropriate balance between precision and recall, F_0 can still be misleading for several reasons. First, precision and recall are measured with respect to orthographic words, though PHOCUS is trying to segment phonological words. We would like to see phonological word corpora developed in the future, but this time-consuming process is beyond the scope of this work.

Second, the kinds of errors the segmenters make can be more informative than F_0. Generally, a segmenters' errors can be classified three ways. Consider the utterance *you see the doggy* [#ju#si#ðə#dɔgi#]. OVER-SEGMENTATION ERRORS are those when the segmenter segments a true word into multiple words (e.g. the segmenter segments *doggy* [dɔgi] as [#dɔ#gi#]). UNDER-SEGMENTATION ERRORS are those when the segmenter segments a sequence of true words as a single word (e.g. the segmenter guesses [#ðədɔgi#] is a single word). MIXED ERRORS are those when the segmenter segments a word which is both under-segmented and over-segmented (e.g. [#ədɔg#]).

For PHOCUS, under-segmentation errors are preferred over over-segmentation errors. This is because once the segmenter adds a word to its lexicon, nothing can ever subtract it. Consequently, it becomes more likely that this word will be segmented out of future utterances, potentially creating more and more errors. For example, if the segmenter errs by adding [dɔ] to its lexicon (instead of *doggy* [dɔgi]), it is very likely that it will segment [dɔlɻ] as [#dɔ#lɻ#], which causes [lɻ] to be added to the lexicon. On the other hand, if the segmenter adds [ðədɔgi] to its lexicon, it can overcome this error in principle by later adding [ðə] and [dɔgi] to the lexicon.[8]

---

[6] $F_0 = \dfrac{2 \times precision \times recall}{precision + recall}$.

[7] See http://cis.udel.edu/~blanchar/research/ for complete results including boundary and lexical precision, recall and F_0 scores.

[8] We calculate that PHOCUS prefers [#ðə#dɔgi#] to [#ðədɔgi#] only when the product of the lexical frequency of [ðə] and [dɔgi] divided by the square of the size of the lexicon is greater than the lexical frequency of [ðədɔgi]. Generally, if w_1, w_2 … w_n and w_1w_2 … w_n are words in the lexicon, $l(w)$ is the lexical frequency of w, and L is the sum of the

We measured each model's performance on all utterances except those in the first tenth of the corpus. We did this because we are primarily interested in the performance of the segmenter once it stabilizes. Unsupervised, incremental models spend the first several hundred utterances learning before their performance levels off and can make many errors during this learning time. If not excluded from the evaluation, these early learning errors are counted against incremental models. Considering the model's performance after its learning curve is behind it allows one to make a fair comparison between batch and incremental models, since unsupervised batch models do all their learning prior to any segmentation. Thus, for the BR corpora, we excluded the first 1000 utterances, and for Sesotho, we excluded the first 800 utterances in the results reported below.

Results

We report several comparisons of the different versions of PHOCUS to each other as well as to other models on the two corpora described above. Our main results compare PHOCUS to other incremental segmenters: Venkataraman's (2001) model (i.e. PHOCUS-1), MBDP-1 (Brent, 1999) and MBDP-Phon (Blanchard & Heinz, 2008). We also include comparison to the batch models of Goldwater (2007) and Johnson & Goldwater (2009, hereafter Johnson), since the code to run them was available from the authors and we were interested if the computationally simpler PHOCUS could achieve comparable performance. We also refer readers to the website http://cis.udel.edu/~blanchar/research/, which contains comprehensive outputs of the computational experiments, summaries of the results and more detailed error analyses.

Our main results support the conclusion of Blanchard & Heinz (2008) that phoneme combinations help incremental unsupervised models. Figure 2 shows that PHOCUS-2 achieves a higher F_0 score than Venkataraman's model (PHOCUS-1) on both the modified BR and Sesotho corpora. The fact that PHOCUS-2 also outperforms PHOCUS-3 is due to reasons discussed below.

Although PHOCUS-2 shows an improvement over PHOCUS-1, the improvement is not as great as we might expect on the modified BR corpus. To get a sense of what the maximum possible benefit of phoneme n-grams is in principle, in one experiment we trained the phonotactic component of the grammar on the BR lexicon, and then ran PHOCUS (initialized with an empty lexicon but with the mature phonotactic grammar) on the unsegmented BR corpus. With phoneme bigrams this semi-supervised

frequencies of all the words in the lexicon, PHOCUS segments utterance $[w_1 w_2 \ldots w_n]$ as $[\#w_1\#w_2 \ldots \#w_n\#]$ only if $\frac{\Pi_{1 \leqslant i \leqslant n^l(w_i)}}{L^{n-1}} > l(w_1 w_2 \ldots w_n)$.

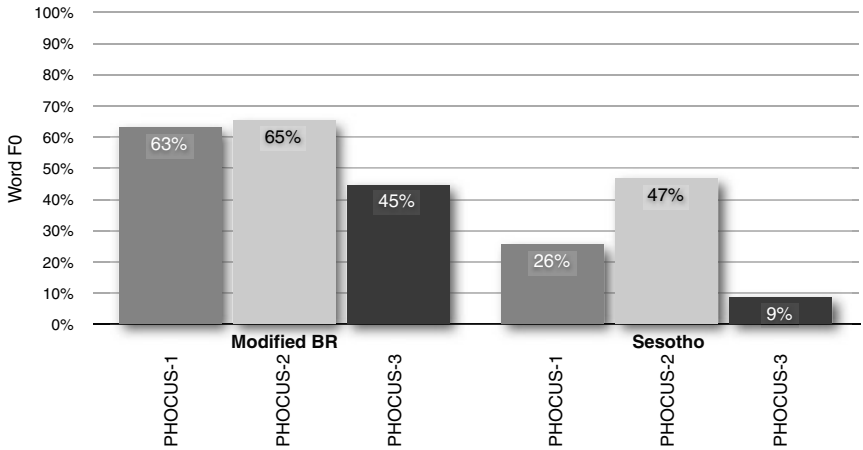

Fig. 2. F_0 of PHOCUS-1, -2 and -3 on modified Bernstein-Ratner and Sesotho corpora.

model achieved a F_0 of 73·8% and with trigrams of 80·4%. We conclude that: (1) phoneme combinations are potentially very useful for word segmentation; and (2) there must be a reason that PHOCUS does not realize this potential.

We hypothesize that the reason is that early errors condemn future segmentation choices and these snowball into increasingly many mistakes. For example, after observing the one-word utterance *block* [blak], PHOCUS-2 adds it to its lexicon. Later, since *block* [blak] is a familiar word when the learner encounters an utterance with *blocks* [blaks], it segments it as [#blak#s#]. Now *s* is considered a familiar word, and is consequently picked off everywhere. The first error creates others elsewhere and the errors compound. With PHOCUS-3, the problem is worse because an examination of its output reveals that it segments much earlier than PHOCUS-2 and the resulting snowball is much larger (hence its lower F_0). These kinds of errors are not unique to PHOCUS-2 and PHOCUS-3; they occur with all incremental unsupervised models. PHOCUS-3 does worse than PHOCUS-2 because it is prone to segment earlier and therefore more likely to make unrecoverable errors, resulting in more mistakes later on.

Next we examine the effect of adding the language-universal phonotactic that all words must consist of at least one syllabic sound. Every version of PHOCUS improves dramatically after this addition, with PHOCUS-3s obtaining the highest F_0 of 80·8% on the BR corpus (Figure 3). The reason for the improvement is that this language-universal constraint eliminates the over-segmentation errors described above. For example, the [s] in [blaks] cannot be peeled off as its own word because it is not a syllabic

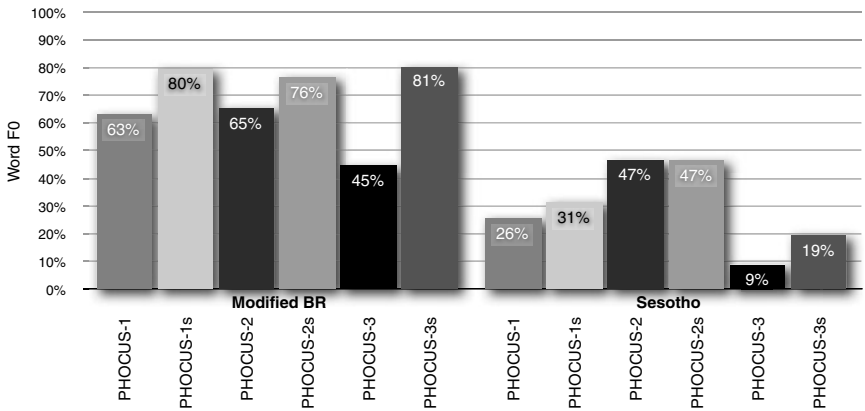

Fig. 3. F_0 of all PHOCUS models on modified Bernstein-Ratner and Sesotho corpora.

element. This reduces the overall number of exclusive over-segmentation errors. In other words, this constraint makes the learner more conservative in introducing words into the lexicon, which makes it less prone to make irreversible, costly mistakes at the beginning.

Although the 'require syllabic' constraint greatly improves the performance of the PHOCUS models that use phoneme combinations, it is almost entirely ineffective by itself. When we ran a version of PHOCUS that included the syllabic constraint, but which assigned unfamiliar words a small constant probability[9] instead of one based on phoneme *n*-grams, we found that the best F_0 obtained for a variety of different constant values was 19·30%. In fact, when we ran a version of PHOCUS with neither the 'require syllabic' constraint nor the phoneme combinations, the output was identical. This is because, when assigning a constant probability to unfamiliar words, longer words receive the same probability as shorter words, so there is no incentive to segment an utterance in such a way that it contains unfamiliar words and, consequently, single-word segmentations become very likely. The only type of utterance that this model segments is one made up of multiple utterances that the model has already added to its lexicon. Therefore, PHOCUS without any phoneme combinations exclusively makes under-segmentation errors. As the 'require syllabic' constraint only helps prevent over-segmentation errors, we conclude that it is the combination of the language-universal syllabic constraint and the simultaneous learning of language-specific phoneme combinations that results in the high level of performance of PHOCUS-3s.

[9] The probability was chosen to be small to ensure that familiar words would still be more likely than unfamiliar ones.

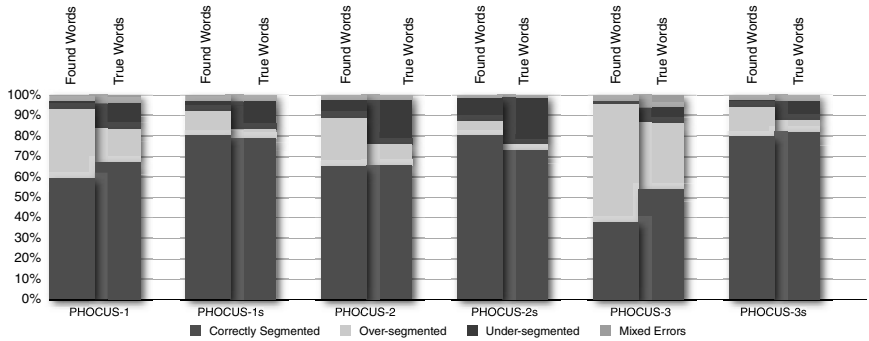

Fig. 4. Found word and true word errors for PHOCUS models on modified Bernstein-Ratner corpus. 'Correct' for found words is precision, and 'correct' for true words is recall.

Interestingly PHOCUS-1s beats PHOCUS-2s in terms of F_0, and an error analysis reveals why. As can be seen in Figure 4, 34·0% of the words PHOCUS-1 finds are over-segmentation errors, and ~3% are under-segmentation errors. For PHOCUS-2, only 25·4% of found words are over-segmentation errors and 8% of found words are under-segmentation errors. Consequently, there are substantially fewer consonant-only over-segmentations for the 'require syllabic' constraint to prevent when added to PHOCUS-2, and thus the F_0 improvement is less pronounced than with PHOCUS-1.

Next we compare the top-performing version of PHOCUS (PHOCUS-3s) on the modified BR corpus to the top-performing versions of the unsupervised batch algorithms developed by Goldwater and Johnson (Figure 5). PHOCUS-3s outperforms Goldwater (81% vs. 71% F_0). While PHOCUS-3s does not come out ahead of Johnson's best adaptor grammar, it does begin to close the gap between incremental and batch systems (81% vs. 86% F_0).

However, the above comparisons ought to be interpreted with some caution. First, the models are not implementing the same set of cues. For example, both Goldwater and Johnson's models build in sensitivity to frequent word collocations to reduce the number of exclusive under-segmentation errors which PHOCUS currently does not. Second, neither model is sensitive to phoneme combinations, though Johnson's adaptor grammars include the concept of a syllable and require every word to have one. We expect that Goldwater's and Johnson's results will also improve if they include more phonotactic cues like PHOCUS. Also, while we are explicit that the aim of our model is to segment phonological words, other researchers have not been explicit about exactly what kinds of words their models aim to segment. In the case of phonological words, some

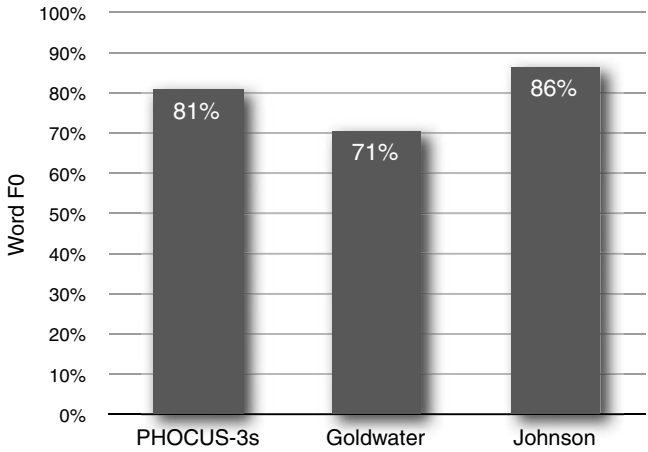

Fig. 5. F_0 of PHOCUS-3s and competing models on Bernstein-Ratner corpus.

of the errors PHOCUS makes may not really be errors – for example the determiner *a* [ə] is frequently under-segmented (e.g. *a boy* is segmented as [#əbɔ#], but it is plausible that this is actually one phonological word, with the determiner sticking to the noun in the same way that *s* sticks to *it* in *it's*. Until a corpus is uncontroversially segmented into phonological words, such issues will go unaddressed.

The models' performances on Sesotho highlight the importance of testing acquisition models on data from a variety of languages because the results can be so different than from what is obtained with English corpora. For example, MBDP-Phon outperforms all other models on the Sesotho corpus, as shown in Figure 6. As MBDP-Phon does not start with a uniform distribution over the phoneme *n*-grams, it is not probabilistically sound, which makes determining why it performs better on Sesotho difficult. Also, PHOCUS-2s and PHOCUS-2 are about the same on Sesotho (but requiring a syllabic sound makes a difference for PHOCUS-1s vs. PHOCUS-1, and for PHOCUS-3s vs. PHOCUS-3). This is again due to PHOCUS-2 making fewer over-segmentation errors than either PHOCUS-1 or PHOCUS-3 (Figure 7). It may appear there is not as reliable of an improvement when adding the language-universal cue when PHOCUS is evaluated on Sesotho. This is because the over-segmentation errors that the models make in Sesotho are almost entirely with single vowel sounds. As such, the 'require syllabic' constraint does not prevent these early over-segmentations from snowballing into a massive problem. For example, the most common isolated word in the corpus is [e] 'yes, what', and of course many words contain this sound as well, which results in much

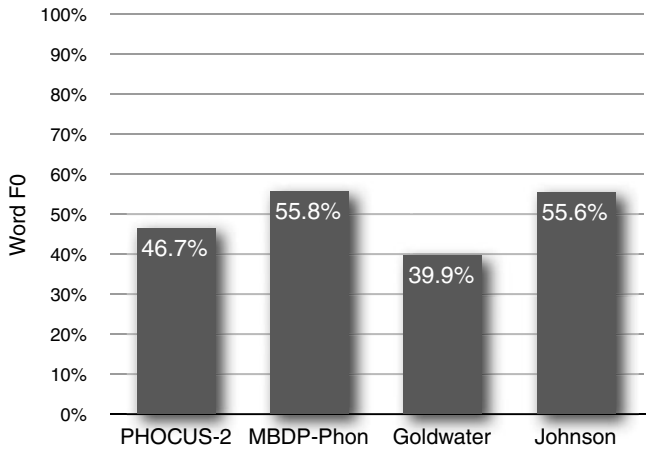

Fig. 6. F_0 of PHOCUS-2 and competing models on Sesotho corpus.

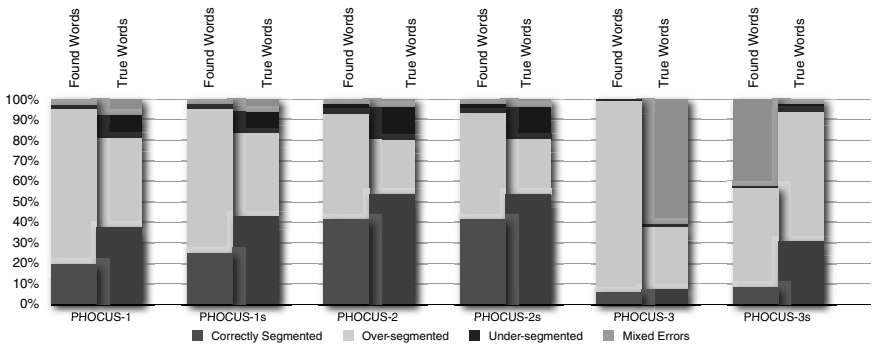

Fig. 7. Found word and true word errors for PHOCUS models on Sesotho corpus. 'Correct' for found words is precision, and 'correct' for true words is recall.

over-segmentation (e.g. [ee] 'no, this' is segmented as [#e#e#] 224 times). Despite these complications, the same general result as seen with the BR corpus can be reported about Sesotho: adding richer phonotactic cues improves the performance of incremental segmenters.

In conclusion, these results show that the performance of an incremental, unsupervised segmentation model greatly improves when it is equipped with both a language-specific phonotactic learning component (here phoneme *n*-grams) and a language universal phonotactic constraint (words have at least one syllabic element). We have also shown how these two components work together – language-specific knowledge of phoneme

n-grams helps to correctly identify likely positions of word boundaries, and the language-universal constraint helps prevent earlier errors that derail the learning process. Furthermore, PHOCUS-2s and PHOCUS-3s compare favorably to state-of-the-art unsupervised batch algorithms.

SIGNIFICANCE OF FINDINGS AND FUTURE WORK

The findings reported above are significant to researchers from many disciplines who are interested in word segmentation. Our findings suggest that knowledge of even simple phonotactic constraints is useful for segmentation. Specifically, having a phonotactic component that keeps track of likely phoneme combinations within words helps the model's performance in at least two ways. First, the model can learn which phonemes are more likely to start and end words, because they will be parts of bigrams or trigrams that contain the word boundary symbol. Second, the model can make decisions about the well-formedness of novel words by evaluating the probabilities of the phoneme combinations within words. This is analogous to the infants in the Mattys & Jusczyk (2001) experiment who segment [faŋ taɪn] *fang tine* properly by realizing that [ŋt] is not a valid phoneme combination within English words.

Our results also suggest that a plausible language universal phonotactic – well-formed words have at least one syllabic sound – helps the cue above by reducing the number of errors made in the learning curve that later prove to be costly. The language-universal cue seems to especially help in languages like English, which do not contain many one-word utterances where the word is a single vowel sound, unlike Sesotho. This result is consistent with the claim within the Emergent Coalition Model that there are multiple cues, which can reinforce (and compete with) each other. Generally, the analysis of the performance of PHOCUS adds support to the results from developmental studies which suggest that infants use multiple sources of phonotactic information to aid word segmentation.

Although it is standard in computational linguistics to evaluate the worth of models on their performance on multiple corpora, it is also important to look at the fundamental properties of the models in relation to the task at hand. In the case of modeling infant language segmentation, it is important for the model to utilize the types of cues that infants do, and combine them in a way that does not conflict with data on how infants process language. Under these two criteria, an incremental segmenter that makes use of both phonotactic and familiar word cues is a desirable model of how infants segment speech. The fact that PHOCUS – an incremental segmenter which uses phonotactic cues – comes close to (and in some cases surpasses) the performance of batch segmenters lends support to the idea that PHOCUS is on the right track.

One aspect of the model we would like to change in the future is the relationship between the implementation of the familiar word cue and the phonotactic cues. Currently, the model only relies on the phonotactic cues when a hypothetical word is unfamiliar. Ultimately, we would like to implement a model that, like the Emergent Coalition Model, considers all available cues simultaneously, perhaps some more than others. In the computational linguistics literature, a common approach for many tasks which involve multiple sources of information is to weight them and then to determine the 'best' weights for each of the various sources. However, such models are exclusively either supervised (and so weights can be updated appropriately and confidently) or batch (so trends in the data can be established and then optimal weights can be assigned). In general, it is unknown how to assign and update weights for unsupervised and incremental models. The problem of how to assign and update weights incrementally when there is very little information and no feedback is an open problem in computer science and computational linguistics.

We would also like to see segmentation models which make use of a variety of phonotactic cues. As discussed earlier, phonotactic constraints can encompass more than just ordering restrictions over phonemes, and infants seem to use many types of constraints for segmentation. To this end, we argued that a more general concept of word well-formedness is appropriate: a well-formed word is made up of frequently occurring subsequences of units. These units can be syllables, phonemes/phones or even bundles of phonological features. Additionally, the subsequences could be of any length, including one, or even non-contiguous (e.g. in order to describe vowel or consonantal harmony in languages like Finnish or Navajo; see Heinz, 2007). Once implemented, this generalized notion of word well-formedness allows a model to keep track of different cues shown to be useful by previous researchers: transitional probabilities between syllables (Saffran *et al.*, 1996), phonotactic constraints (Mattys & Jusczyk, 2001), allophonic variation (Jusczyk, Hohne & Baumann, 1999) and stress (Jusczyk, Houston & Newsome, 1999).

CONCLUSION

Three important findings have emerged from the development of this model for infant speech segmentation. First, evidence from the computer simulations conducted here suggests that both language-specific and language-universal phonotactic constraints are useful for word segmentation, and that language-specific constraints can be learned at the same time that the model segments speech. Second, incremental models with a phonotactic component come close to achieving (and in some cases surpass) the same level of accuracy as state-of-the-art batch models. Given that

infants are more likely to process their input incrementally, the computational complexity of batch models may not be necessary for the task of segmentation. This suggests that the present model has some psychological reality. Finally, the research program outlined here investigates the utility of different types of phonotactic cues to word segmentation, shows how to quantitatively evaluate how such cues interact with one another and highlights an area of common interest shared by language acquisition researchers and computational linguists.

REFERENCES

Bernstein-Ratner, N. (1987). The phonology of parent–child speech. In K. Nelson & A. van Kleeck (eds), *Children's language, Volume 6*, 159–74. Hillsdale, NJ: Erlbaum.

Blanchard, D. & Heinz, J. (2008). Improving word segmentation by simultaneously learning phonotactics. In *12th Conference on Computational Natural Language Learning*, 65–72. Morristown, NJ: Association for Computational Linguistics.

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

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

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

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

Cole, R. & Jakimik, J. (1980). A model of speech perception. In R. Cole (ed.), *Perception and production of fluent speech*, 136–63. Hillsdale, NJ: Lawrence Erlbaum Associates.

Coleman, J. & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In *Proceedings of the Third Meeting of the Association for Computational Linguistics SIGPHON*, 49–56. Somerset, NJ: Association for Computational Linguistics.

Cutler, A. & Carter, D. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language* **2**, 133–42.

Demuth, K. (1992). Acquisition of Sesotho. In D. Slobin (ed.), *The cross-linguistic study of language acquisition, Volume 3*, 557–638. Hillsdale, NJ: Lawrence Erlbaum Associates.

Dixon, R. M. W. & Aikhenvald, A. Y. (2002). Word: A typological framework. In R. M. W. Dixon & A. Y. Aikhenvald (eds), *Word: A cross-linguistic typology*, 1–41. Cambridge: Cambridge University Press.

Fleck, M. M. (2008). Lexicalized phonotactic word segmentation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, 130–38. Morristown, NJ: Association for Computational Linguistics.

Friederici, A. & Wessels, J. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception and Psychophysics* **54**, 287–95.

Goldwater, S. (2007). Nonparametric Bayesian models of lexical acquisition. Unpublished doctoral dissertation, Brown University, Department of Cognitive and Linguistic Sciences.

Golinkoff, R. & Hirsh-Pasek, K. (2006). Baby wordsmith: From associationist to social sophisticate. *Current Directions in Psychological Science* **15**, 30–33.

Halle, M. (1978). Knowledge unlearned and untaught: What speakers know about the sounds of their language. In M. Halle, J. Bresnan & G. A. Miller (eds), *Linguistic theory and psychological reality*, 294–303. Cambridge, MA: MIT Press.

Harris, Z. (1954). Distributional structure. *Word* **10**, 146–62.
Hayes, B. & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* **67**, 379–440.
Heinz, J. (2007). Inductive learning of phonotactic patterns. Unpublished doctoral dissertation, University of California, Los Angeles, Department of Linguistics.
Hollich, G., Hirsh-Pasek, K., Golinkoff, R., Brand, R. J., Brown, E., Chung, H. L., *et al.* (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development* **65**, i–vi, 1–123.
Johnson, M. (2008). Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of the Association for Computational Linguistics, SIGMORPHON*, 20–27. Morristown, NJ: Association for Computational Linguistics.
Johnson, M. & Goldwater, S. (2009). Improving nonparameteric Bayesian inference: Experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 317–25. Morristown, NJ: Association for Computational Linguistics.
Jurafsky, D. & Martin, J. (2008). *Speech and language processing*, 2nd edn. Upper Saddle River, NJ: Prentice-Hall.
Jusczyk, P. (1993). From general to language specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics* **21**, 3–28.
Jusczyk, P., Friederici, A., Wessels, J., Svenkerud, V. Y. & Jusczyk, A. M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language* **32**, 402–420.
Jusczyk, P., Hohne, E. & Baumann, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics* **61**, 1465–76.
Jusczyk, P., Houston, D. & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology* **39**, 159–207.
MacWhinney, B. & Snow, C. (1985). The child language data exchange system. *Journal of Child Language* **12**, 271–95.
Matthews, P. (1991). *Morphology*, 2nd edn. Cambridge: Cambridge University Press.
Mattys, S. & Jusczyk, P. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition* **78**, 91–121.
Mattys, S., Jusczyk, P., Luce, P. & Morgan, J. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology* **38**, 465–94.
Mohri, M. (2005). Statistical natural language processing. In M. Lothaire (ed.), *Applied combinatorics on words*, 210–40. Cambridge: Cambridge University Press.
Nelson, D. K., Jusczyk, P., Mandel, D., Myers, J., Turk, A. & Gerken, L. (1995). The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development* **18**, 111–16.
Saffran, J., Aslin, R. & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science* **274**, 1926–28.
Saffran, J., Werker, J. & Werner, L. (2006). The infant's auditory world: Hearing, speech, and the beginnings of language. In R. Siegler & D. Kuhn (eds), *6th edition of the handbook of child development, Volume 2*, 58–108. New York: Wiley.
Sapir, E. (1925). Sound patterns in language. *Language* **1**, 37–51.
Shi, R. & Lepage, M. (2008). The effect of functional morphemes on word segmentation in preverbal infants. *Developmental Science* **11**, 407–413.
Teahan, W. J., McNab, R., Wen, Y. & Witten, I. H. (2000). A compression-based algorithm for Chinese word segmentation. *Computational Linguistics* **26**(3), 375–93.
Thiessen, E. & Saffran, J. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology* **39**, 706–716.
Thiessen, E. & Saffran, J. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development* **3**, 73–100.

Toft, Z. (2002). The phonetics and phonology of some syllabic consonants in Southern British English. In Z. Toft (ed.), *Papers on phonetics and phonology: The articulation, acoustics and perception of consonants, Volume 28*, 111–144.

Toro, J. M., Nespor, M., Mehler, J. & Bonatti, L. L. (2008). Finding words and rules in a speech stream: Functional differences between vowels and consonants. *Psychological Science* **19**, 137–144.

Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics* **27**, 352–72.

Xie, Z. & Niyogi, P. (2006). Robust acoustic-based syllable detection. In INTERSPEECH-2006, paper 1327-Wed1BuP.6. Accessed at www.isca-speech.org/archive/interspeech_2006/