Comparing Models of Phonotactics for Word Segmentation

Natalie M. Schrimpf
Department of Linguistics
Yale University
natalie.schrimpf@yale.edu

Gaja Jarosz
Department of Linguistics
Yale University
gaja.jarosz@yale.edu

Abstract

Developmental research indicates that infants use low-level statistical regularities, or phonotactics, to segment words from continuous speech. In this paper, we present a segmentation framework that enables the direct comparison of different phonotactic models for segmentation. We compare a model using phoneme transitional probabilities, which have been widely used in computational models, to syllable-based bigram models, which have played a prominent role in the developmental literature. We also introduce a novel estimation method, and compare it to other strategies for estimating the parameters of the phonotactic models from unsegmented data. The results show that syllable-based models outperform the phoneme models, specifically in the context of improved unsupervised parameter estimation. The syllable-based transitional probability model achieves a word token f-score of nearly 80%, the highest reported performance for a phonotactic segmentation model with no lexicon.

1 Introduction

One of the first language learning tasks infants must solve is the segmentation of fluent speech into words. Extensive experimental work has demonstrated that infants are able to use phonotactic restrictions (Jusczyk & Luce, 1994; Mattys et al., 1999; Mattys & Jusczyk, 2001) and other low-level statistical regularities (Saffran et al., 1996; Thiessen & Saffran, 2003; Pelucchi et al., 2009) to extract words from fluent speech before the age of one. This work has shown that infants utilize these low-level statistical regularities to segment speech during the second half of the first year of life before they have developed extensive vocabularies that could provide top-down lexical information to guide segmentation. Developmental research indicates that on average infants know fewer than 100 word types during this period (Dale & Fenson, 1996; Daland & Pierrehumbert, 2011).

One statistical cue that has received a great deal of support in experimental work on infant speech segmentation is transitional probability calculated over syllables. In foundational work, Saffran et al. (1996) found that infants are able to segment words from continuous speech using statistical regularities between syllables. Numerous subsequent studies have confirmed that infants can track transitional probabilities and use them to segment speech (Aslin et al., 1998; Thiessen & Saffran, 2003; Pelucchi et al., 2009).

Despite the extensive experimental literature demonstrating infants' sensitivity to transitional probability in an artificial language learning setting, the utility of these statistical cues in a natural language learning context is disputed. Yang (2004) shows that a segmentation strategy relying on transitional probabilities over syllables achieves very poor results on English child-directed speech, even when the input is perfectly syllabified. Yang implements the local minimum segmentation strategy proposed by Saffran et al. (1996) wherein word boundaries are posited at syllable transitions whenever the transitional probabilities at these positions are lower than at the neighboring transitions. He reports that this strategy discovers a mere 23% of target words and posits incorrect words nearly 60% of the time. Swingley (2005) argues that statistical cues calculated over syllables can provide sufficient information for infants to begin building an initial lexicon. However, the learning strategy explored by Swingley is highly conservative, reliably detecting only a small proportion of target words in the input. Overall, these results raise questions about whether syllable-based statistics can be reliably used to identify word boundaries in natural language data.

Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM, pages 19–28,
Baltimore, Maryland USA, June 27 2014. ©2014 Association for Computational Linguistics

While the experimental work emphasizes syllable-level transitional probability, recent computational modeling work and corpus analyses have primarily focused on the utility of phoneme-level statistics. A number of phonotactically-based segmentation models, focusing on the discovery of word boundaries based on phoneme-level statistics, have achieved more promising results (Adriaans & Kager, 2010; Daland & Pierrehumbert, 2011; see also Brent, 1999). For example, Brent (1999) showed that a local minimum strategy relying on phoneme bigrams correctly extracts about 50% of word tokens in English child-directed speech. Corpus analyses of child-directed speech have also highlighted the information content of phoneme-level statistics (Hockema, 2006; Jarosz & Johnson, 2013). Related work has shown that phonotactic information can improve the performance of state-of-the-art segmentation models whose primary objective is to discover the lexicon that underlies the regularities in the continuous speech signal. Again, this work has largely emphasized phoneme-level statistical cues (Blanchard & Heinz 2008, 2010), and those models that do rely on syllable structure (Johnson, 2008a; Johnson & Goldwater, 2009), do not directly encode sequential statistics between adjacent syllables of the sort investigated in the infant literature. Finally, some models assume computations are performed over syllables and that all word boundaries in the input are aligned with syllable boundaries, but provide no mechanism by which such language-specific syllabification principles could be learned (Yang, 2004; Swingley, 2005; Lignos & Yang, 2010).

Overall, the existing evidence clearly shows that there are phonotactic cues to word boundaries in spontaneous, child-directed speech. However, there are remaining questions regarding the exact nature of these cues, their reliability, and how they relate to the statistical cues explored in the infant word segmentation literature. In this paper, we investigate the computational mechanisms underlying infants' early speech segmentation abilities relying on low-level statistical regularities, or phonotactics. We present a computational framework that permits the direct comparison of segmentation predictions for alternative models of phonotactics. In particular, we compare a standard phonotactic model relying on phoneme-level bigrams to two syllable-based phonotactic models relying on transitional probabilities. Unlike previous models relying on syllabified data (Yang, 2004; Swingley, 2005; Lignos & Yang, 2010), we do not assume that word boundaries align with syllable boundaries in the input. Rather, we present a simple syllabification method that can be used to model phonotactic probability for arbitrary strings using statistics estimated from unsyllabified, unsegmented utterances. We also compare the local minimum segmentation strategy (Saffran et al., 1996; Yang, 2004) to alternatives designed to deal with the challenges of unsupervised estimation of transitional probabilities from unsegmented input.

Our focus on the early phonotactic segmentation stage differentiates our approach from many computational models emphasizing the discovery of the lexicon and higher-level language structure (Brent, 1999; Venkataraman, 2001; Swingley, 2005; Johnson, 2008a; Goldwater et al., 2009; Johnson & Goldwater, 2009; Blanchard & Heinz 2008, 2010; Lignos & Yang, 2010). It complements that of recent work investigating the use of phoneme-level statistical regularities for segmentation (Adriaans & Kager, 2010; Daland & Pierrehumbert, 2011). Our work differs from these latter approaches, however, in comparing several phonotactic models, including ones relying on the syllable-based transitional probability statistics investigated in infant research. Our work also contributes to existing segmentation work that assumes a syllabified input (Yang, 2004; Swingley, 2005; Lignos & Yang, 2010) by showing how many aspects of syllable structure can be inferred.

Our results reveal an interaction between estimation strategy and the choice of phonotactic model. The local minimum segmentation strategy works poorly in general for all models considered, but the lowest performance is achieved by the syllable-based models. However, when the same cues are used in the context of a simple, generative probability model with improved unsupervised parameter estimation, the syllable-based models substantially outperform the phoneme-based models. Indeed, the syllable-based transitional probability phonotactic model achieves a word token segmentation f-score of nearly 80%, which is the highest reported performance among purely phonotactically-based segmentation models (Adriaans & Kager, 2010; Daland & Pierrehumbert, 2011). Indeed, this performance compares favorably with state-of-the-art segmentation models that involve learning of higher level regularities, such as the lexicon and collocations (Brent, 1999; Venkataraman, 2001; Johnson, 2008a; Goldwater et al., 2009; Johnson & Goldwater, 2009), and demonstrates that good segmentation performance can be achieved by exploiting simple syllable-level phonotactic cues.

2 Segmentation Model

The proposed segmentation model defines the probability of an utterance in terms of an abstract phonotactic probability component that assigns word well-formedness probabilities to phoneme strings. The segmentation algorithm uses those probabilities to determine the maximum likelihood segmentation as defined by a simple generative model. Since the phonotactics and segmentation components are separate, they can be independently modified. This framework makes it possible to compare models of phonotactics while using the same segmentation strategy.

2.1 Probability Model

The segmentation probability model relies on the phonotactic component to assign probabilities to potential words. The probability of a segmentation w is defined in terms of a simple unigram model by multiplying the probabilities of the words $w_{1\ldots n}$ posited in that segmentation.

1) $P(w) = P(w_{1\ldots n}) = \prod_1^n P(w_i)$

$P(w_i)$ is the probability assigned by the phonotactic models, which will be defined in the next section. The various phonotactic models change how exactly $P(w_i)$ is defined, but the segmentation probability always depends directly on the word probabilities given by a particular phonotactic model. For example, for the utterance [lʊkætmi] 'lookatme', the segmentation model compares different segmentations, such as [lʊk#æ#tmi] and [lʊk#æt#mi] based on the phonotactic well-formedness of the posited words.

2.2 Segmentation Algorithm

The segmentation algorithm computes and outputs the segmentation with the highest likelihood: $argmax_w P(w)$. The optimal segmentation is found using dynamic programming, as in several previous proposals (Brent, 1999; Venkataraman, 2001). Given an input utterance, the model considers placing word boundaries at different positions within the utterance without regard to phonotactics or syllable structure. The phonotactic probability of each posited word is calculated independently as it is considered and used to update the probability of segmentations utilizing that word. In this way, the segmentation component remains entirely divorced from the details of the phonotactic models. Crucially, this means the full space of possible segmentations is considered by the segmentation model regardless of the phonotactic model, with no a priori restrictions imposed by phonotactic or syllable constraints as to where boundaries are permitted.

3 Phonotactic Models

We implement and compare several models of phonotactics that are utilized by the segmentation component described above. While all models rely on transitional probabilities, or bigrams, as defined in (2), the unit of analysis varies between the models. One model uses phonemes and phoneme transitions, and two models incorporate syllable information: we use x to denote a generic unit. The model determines the probability of a word, $w = x_{0\ldots n+1}$ where x_0 and x_{n+1} are the word boundary symbol #, by multiplying the probabilities of all bigrams in the word.

2) $P(w) = \prod_0^n P(x_{i+1}|x_i)$

The transitional probability for the sequence $x_i x_{i+1}$ can be calculated using relative frequency estimates based on counts C in the corpus.

3) $\hat{P}(x_i|x_{i-1}) = \frac{C(x_{i-1}x_i)}{C(x_{i-1})}$

Section 4 describes strategies that we consider for estimating these parameters in an unsupervised way from unsegmented data where the only word boundaries are those that coincide with utterance boundaries.

3.1 Phoneme Model

The first phonotactic model is a standard phoneme bigram model that determines the probability of a word by multiplying the phoneme bigrams in the word (Jurafsky & Martin, 2008). For example, to calculate the phonotactic probability of the sequence [bot] as a word, this model multiplies together P(b|#)P(o|b)P(t|o)P(#|t).

3.2 Syllable-Based Models

The other two phonotactic models use syllables rather than phonemes. One model relies on transitional probabilities over syllables, and the other uses onsets and rhymes as the unit of analysis.

3.2.1 Unsupervised Syllabification

The syllabification method relies on the language universal principle of onset maximization together with an inventory of syllable onsets derived from the beginnings of utterances. When syllabifying an intervocalic sequence of consonants, this method finds the longest legal onset aligned with the right edge and places any remaining consonants in the coda of the previous syllable. Thus, a sequence like [ætmi] would be syllabified as [æt.mi] in English since [m] but not [tm] occurs utterance-initially. The only language-particular information required for this approach is knowledge of which phonemes are vowels (syllabic) and which are consonants, a limited type of information also assumed by other syllable inference models for segmentation (Johnson, 2008a; Johnson & Goldwater, 2009).

As the segmentation component posits potential words, they are passed to the phonotactic component for syllabification and phonotactic probability calculation. This differs crucially from previous work assuming a fixed syllabification of the input corpus in which word boundaries always align with syllable boundaries (Yang, 2004; Swingley, 2005; Lignos & Yang, 2010). In a setting in which syllabification must be inferred from unsegmented utterances, the learner must be capable of assigning syllabification more flexibly since word boundaries do not always align with the syllable boundaries that would be posited for the utterance as a whole. For example, the universal onset maximization principle always parses singleton consonants VCV as the onsets V.CV rather than codas VC.V. Therefore, without prior knowledge of word boundaries, the utterance [lʊkætmi] ('look at me') would be syllabified as [lʊ.kæt.mi], and if the segmentation algorithm never considered words that misaligned with these syllable boundaries, it would never extract any vowel-initial words like 'at'. Thus, a crucial feature of the current model is that syllabification takes place on a word-by-word basis as potential words are posited. The resulting syllabification for the potential word is used by the syllable-based models to assign phonotactic probability as discussed below.

3.2.2 Syllable Model

The first syllable-based model is one in which bigram transitional probabilities are calculated over syllables. These transitional probabilities are precisely those discussed earlier as having played a prominent role in the infant segmentation literature. The phonotactic probability of a posited word is calculated by multiplying the transitional probabilities of all syllable bigrams in the word, including an assumed initial and final #. For example, if the segmentation component posits a potential word such as [lʊkætmi] 'lookatme', this sequence is first syllabified using the procedure described earlier as [lʊ.kæt.mi]. Then the phonotactic probability of this potential word is calculated by multiplying together the syllable-based bigram probabilities: P(lʊ|#)P(kæt|lʊ)P(mi|kæt)P(#|mi). As before, relative frequency estimates calculated from unsegmented input data (automatically syllabified using the unsupervised syllabification method described earlier) provide a starting point for parameter estimation. Estimation strategies are discussed in depth in Section 4.

3.2.3 Onset Rhyme Model

In addition to the phoneme level and syllable level bigram models, we consider an intermediate model that makes use of the main subconstituents of syllables: onsets and rhymes. Recall that the syllabification procedure relies on identifying maximal onsets, whereas rhymes are composed of the remaining material in the syllable. So these constituents are already available during the syllabification procedure, and this phonotactic model operates over these smaller constituents, rather than over entire syllables. The syllable-based model operates over indivisible syllable units, while this models treats syllables as combinations of smaller subconstituents.

Once a sequence is syllabified (separating onsets and rhymes), this model uses bigrams over these units to determine word probabilities. Consider again the potential word [lʊkætmi] 'lookatme'. This sequence is first syllabified into onsets and rhymes as [l.ʊ.k.æt.m.i]. Then its phonotactic probability is calculated by multiplying together the bigram probabilities: P(l|#)P(ʊ|l)P(k|ʊ)P(æt|k)P(m|æt)P(i|m)P(#|i). As before, relative frequency estimates are calculated from an (automatically syllabified) unsegmented version of the input corpus.

4 Estimation

Inferring the parameters of these models in an unsupervised way from unsegmented utterances presents a number of challenges. First, a generative model relying on these parameters must be able to accommodate elements and sequences of elements that have not previously been encountered. This includes unseen phonemes, onsets, rhymes, syllables, and unseen sequences of these units. A second difficulty for the generative model arises specifically in the context of segmentation due to the number of boundaries encountered in the input data. In an unsegmented corpus there are no boundaries within an utterance. The only evidence for word boundaries comes from boundaries at the beginnings and ends of utterances. The effect is that the total number of boundaries is lower than the number that must be inferred by the learner, and the overall probability of boundaries is underrepresented in the input data. We considered several estimation methods to overcome these effects.

4.1 Local Minimum Strategy

In previous research (Saffran et al., 1996) it has been suggested that word boundaries are placed at troughs in transitional probability so that a boundary is inserted between two elements when the transitional probability of those elements is lower than the probability of the neighboring transitions. This strategy captures the fact that word boundaries are more likely to occur between elements that have a low probability of occurring together. Since this strategy does not incorporate transitional probabilities into a generative segmentation model, it provides a simple way around the estimation challenges discussed above. We include it for comparison to previous results relying on syllable-based transitional probabilities (Yang, 2004).

4.2 Adjusted Boundary Count Strategy

We also introduce a novel, simple method for adjusting the estimates of transitional probabilities based on input data that underrepresents word boundaries. This method directly adjusts the parameter estimates in order to increase the overall likelihood of word boundaries. The main insight behind this estimation strategy is that observed bigram counts (of co-occurring phonemes, syllables, or onsets and rhymes) in the input data are overestimated since a proportion of them are in reality separated by word boundaries in the desired segmentation. For a given proportion $p_{\#}$ (a parameter of this estimation method), the bigram counts of co-occurring elements (phonemes, syllables, or onsets/rhymes) are systematically decreased by a factor of (1- $p_{\#}$) and for each context *c*, are reallocated to the transitional probability of P(# | *c*). The formula below illustrates how this adjustment works for arbitrary contexts *c* and proportion $p_{\#}$. The probability of each possible element e_i that can follow *c* is decreased by a factor of $p_{\#}$ as shown in (4). The total probability taken away from all continuations of *c* is used to increase the probability of P(# | *c*) as shown in (5).

4) $P(e_i|c) = \frac{C(ce_i)}{C(c)}(1 - p_{\#})$

5) $P(\#|c) = \frac{C(c\#)}{C(c)} + p_{\#}(1 - \frac{C(c\#)}{C(c)})$

Consider an example for the context *x*, with three bigrams observed in the input: c(*xy*) = 10, c(*xz*) = 6, and c(*x#*) = 4. The relative frequency estimates for these transitional probabilities are 0.5, 0.3, and 0.2 respectively. The adjusted count method takes away $p_{\#}$ of the *xy* and *xz* counts and reallocates them to *x#*. For $p_{\#}$= 0.5, for example, the new estimates would be 0.25, 0.15, and 0.6. The adjustment works analogously for every context for each of the units of analysis.

4.3 Smoothing

We also utilized rudimentary smoothing techniques to allow the generative model to deal with unknown sequences. We chose a simple method that allocated non-zero probability to unseen sequences while minimally disrupting the estimates computed using the adjusted boundary count strategy, since our primary concern was in exploring the effects of this novel re-estimation strategy. For all models, add-lambda smoothing (Jurafsky & Martin, 2008) with a value of 0.001 was used. For the syllable-based models this total value was allocated to all unseen bigrams in order to avoid over-allocation of probability to the numerous combinations of unseen syllabic units.

4.4 Iterative Re-estimation

After estimating the transitional probabilities from the unsegmented corpus, the above strategies can be used to compute the optimal segmentation of the input corpus in a single pass. In addition to the above strategies, we also investigated a greedy, iterative re-estimation strategy that makes multiple passes through the corpus. This estimation method takes the output of the above methods and uses it to re-estimate (smoothed and adjusted) parameters for the phonotactic models. It then recomputes the optimal segmentation of the corpus based on the new parameters and repeats until convergence. This method is motivated by previous segmentation work highlighting the effectiveness of greedy re-estimation techniques (Brent, 1999; Venkataraman, 2001; Goldwater et al., 2009; Johnson & Goldwater, 2009). As noted in previous work, such greedy re-estimation has the potential to infer additional word boundaries based on commitments made to word boundaries on earlier passes.

5 Experiments

5.1 Corpus

The experiments for all the models were run on the Brent (1999) version of the Bernstein-Ratner (1987) corpus of English child-directed speech consisting of phonetically transcribed utterances. This corpus has been widely used for evaluating segmentation models. Other models evaluated on this corpus include those of Brent (1999), Venkataraman (2001), Blanchard and Heinz (2008), and Johnson and Goldwater (2009).

5.2 Evaluation

Precision, recall, and f-scores of both word tokens and boundaries were used to evaluate performance. For the models with iterative re-estimation, the reported performance scores are taken from the iteration after convergence. This typically happened after 5-10 iterations.

5.3 Results and Discussion

Table 1 summarizes the word boundary and word token f-scores for all models, while Table 2 presents the precision and recall scores for the best-performing adjusted count models and the local minimum models.

Focusing first on the local minimum estimateion strategy, there are several noteworthy effects. First, our results with local minima for the syllable-level transitional probabilities achieves very similar word token precision and recall to that reported by Yang (2004), who examined a different corpus of child-directed English. The word token precision and recall of our model is 40.2% and 23.7%, respectively, while Yang reported 41.6% and 23.3%, respectively, for his experiments. This corroborates Yang's finding that the local minima estimation strategy for syllable-level transitional probabilities works very poorly, this time showing that this level of performance can be achieved with simultaneous inference of syllabification. As Table 2 shows, the poor performance can be attributed to poor recall, which the low boundary recall and high precision illustrate most clearly. As Yang discusses, the fatal flaw for this approach is that it categorically fails to segment monosyllabic words, which account for an overwhelming majority of words in child-directed speech. This is because local minima must, by definition, be separated by at least one transition with a higher bigram probability, which is not treated as a boundary. Indeed, the proportion of monosyllables is so high that a baseline strategy that simply posits word boundaries at all syllable boundaries achieves a word token f-score of 58.0% using the minimally-supervised syllabification procedure described here[1]. The high performance of the monosyllabic baseline highlights the ineffectiveness of the local minimum strategy but also indicates that syllable structure provides a significant amount of information about word boundaries in English, even if this syllable structure is automatically inferred from unsegmented input using minimal prior knowledge.

Furthermore, our results with the phoneme bigram local minimum strategy (47.1% word token f-score) corroborate Brent's (1999) finding that this method achieves a roughly 50% word token f-score (Brent did not provide exact numbers). The improvement in performance is not surprising given the above discussion about the prevalence of monosyllabic words: local minima defined over the smaller phoneme units do not automatically rule out the possibility of segmenting short words. We also demonstrate that the onset-rhyme model achieves performance similar to that of the syllable bigram model using the local minima strategy. Finally, the results with iterative re-estimation show that further refinement of the posited word boundaries can lead to some improvement, but none of the local minimum models surpass 53% word token f-score, and the syllable-based models perform substantially worse. Overall, these partial results are consistent with the trend suggested by previous work that the syllable-level bigrams examined in the infant studies provide little information about word boundaries in natural language data when the local minimum strategy is used.

However, a different picture emerges when the performance of the adjusted count strategy is considered. The fact that the local minimum strategy is ineffectual is already clear from the comparison with the monosyllabic baseline; however, the results for the adjusted counts estimation strategy reveal that it is possible to ex-

[1] In contrast, Lignos & Yang (2010) report a word token f-score of 78.9% for this baseline for already syllabified input. The difference between these baselines highlights how much more difficult the segmentation task is when the syllabification must be inferred from unsegmented input.

	$p_{\#}$= 0		$p_{\#}$= 0.35		$p_{\#}$= 0.5		$p_{\#}$= 0.6		$p_{\#}$= 0.75		$p_{\#}$= 0.99		LM	
	WF	BF	WF	BF	WF	BF	WF	BF	WF	BF	WF	BF	WF	BF
P	13.0	10.2	34.7	51.9	40.3	60.6	**49.9**	**69.2**	45.9	68.8	13.9	50.1	47.1	64.5
OR	15.4	17.9	28.7	43.3	37.1	55.8	42.2	62.0	**58.4**	**76.0**	52.3	71.4	27.9	44.1
S	10.7	3.1	12.7	8.6	14.2	12.4	15.9	16.3	20.7	26.1	**74.1**	**84.1**	29.8	51.0
P-IR	13.0	10.2	34.7	51.9	40.3	60.6	**50.7**	**69.6**	46.9	69.6	9.9	47.0	52.9	70.5
OR-IR	19.8	29.1	36.8	54.7	47.7	67.7	53.4	72.8	**63.8**	**79.8**	37.1	62.1	42.3	62.3
S-IR	10.9	3.8	13.3	10.5	15.2	15.0	16.8	18.7	23.1	31.4	**79.8**	**88.0**	27.2	43.9

Table 1: Word token (WF) and boundary (BF) f-scores for all models. The columns in the first section of the table represent different settings of the $p_{\#}$ parameter, with highest performance for each adjusted count model shown in bold. $p_{\#}$ values were selected to show a representative range of performance. P = phoneme model; OR = onset-rhyme model; S = syllable model; IR = iterative re-estimation; LM = local minimum strategy. The best performing local minimum model is shaded.

	Adjusted Count Estimation				Local Minimum Estimation			
	WP	WR	BP	BR	WP	WR	BP	BR
P-IR	50.3	51.1	68.8	70.4	53.4	52.4	71.5	69.5
OR-IR	63.8	63.8	79.9	79.8	44.2	40.5	66.5	58.6
S-IR	85.2	75.0	97.0	80.6	40.4	20.5	94.0	28.6

Table 2: Word precision (WP), word recall (WR), boundary precision (BP), and boundary recall (BR) scores for selected models. For the adjusted count estimation models, the results for the best performing parameter value are shown (P-IR: 0.6; OR-IR: 0.75; S-IR: 0.99).

extract substantially more information about word boundaries from syllable-based models when these cues are used in the context of a generative model and better methods are used for unsupervised estimation of these parameters. In fact, using the adjusted counts estimation method with the optimal parameter settings, the reverse trend is observed, wherein the phoneme-level bigrams perform worse than the syllable-based models, and syllable-level bigrams perform best of all, reaching word token f-scores of nearly 80%. Crucially, both the onset-rhyme and the syllable bigram models achieve levels of performance that surpass the monosyllabic baseline. In the case of the syllable bigram, the improvement in word token f-score is more than 20% when iterative re-estimation is used and more than 15% when segmentation is performed in only a single pass through the corpus.

The phoneme-based models perform about as well whether adjusted counts or local minimum estimation is used. However, compensation for the underrepresentation of word boundaries in the input is crucial to the syllable-based models. These models surpass the local minimum estimation models only when the $p_{\#}$ parameter compensates sufficiently for the input bias against word boundaries. As shown in Table 1, without any compensation ($p_{\#}$= 0), all models perform terribly. This is because utterance boundaries provide very little evidence of word boundaries, and the models estimated directly from such input massively undersegment. It is only at higher settings of the parameter that performance improves. As expected, the optimal parameter value increases with the granularity of the unit over which bigrams are computed. This makes sense since boundaries are more likely to fall between larger units than between smaller units.

Less expected is the fact that the optimal parameter values are high compared to the empirical rates of word boundaries in the true segmentation of the input corpus. For example, the true rate of utterance-internal word boundaries is around 30% at the phoneme level, yet the optimal $p_{\#}$ value for phoneme bigrams is around 60%. The reason for this is that our generative model, like that of a number of previous models discussed in the literature (Brent, 1999; Venkataraman, 2001; Goldwater et al., 2009), has an inherent undersegmentation bias. Due to the way the phonotactic models are defined, there is a cost for every additional word boundary posited in the segmentation. This is because positing a boundary corresponds to the generation of an additional symbol, #, which otherwise does not have to be generated. Since generating a # is never done with 100% probability, doing so always incurs a cost relative to a segmentation where no such # has to be generated. The high optimal settings of the $p_{\#}$ parameter reflect this inherent bias and enable the estimation procedure to compensate not only for the underrepresentation of word boundaries in the input but also for this bias in the generative model.

6 Conclusions

We compared segmentation models that rely on phoneme transitions to models that make use of syllable structure. The results indicate that syllable-based statistics are valuable for segmentation. We also showed that it is possible to utilize this structure successfully with limited prior knowledge of the target language by using a simple syllabification strategy inferred from unsegmented utterances. The performance of the syllable-based models also demonstrates that it is possible to achieve good segmentation results without the use of a lexicon. Another contribution of this work is a novel estimation procedure that addresses some challenges of unsupervised segmentation. We showed that adjusting parameter estimates inferred from unsegmented input is essential for achieving good performance.

The strong performance of the syllable level bigram phonotactic model has a number of implications. First, it demonstrates that the kind of statistical regularities that infants have been consistently shown to be sensitive to in artificial experimental stimuli do provide a substantial amount of information about word boundaries in natural language data, at least in English. This lends significant credibility to the claim that sensitivity to such statistical regularities plays a crucial role in infants' early language development (contra Yang 2004). This result also highlights the role that sensitivity to richer phonological information, beyond the level of phonemes, plays in language learning, a result that is echoed in much recent work on the modeling of phonotactic well-formedness of isolated words (Hayes & Wilson, 2008; Albright, 2009; Daland et al., 2011). A consistent finding of this work has been that access to abstract structure and robust generalization mechanisms is crucial to the modeling of human phonotactic knowledge. While our results are compatible with these conclusions, our results cannot confirm that it is syllable structure *per se* that improves segmentation since the syllable-based models have several co-occurring advantages. In addition to abstract structure, they can track longer and more complex dependencies. Nonetheless, these results motivate further investigation into the role that richer models of phonotactics may play in word segmentation and into the precise mechanisms responsible for improved segmentation using syllable structure. Particularly critical is exploration of phonotactically-based segmentation models for languages besides English, for which phonotactic cues hold significant promise (Jarosz & Johnson, 2013) given the relatively low performance of state-of-the-art lexicon-building models (Johnson 2008b).

Another important direction for future work is investigating how early, phonotactically-based segmentation interacts with subsequent learning of higher-level structure, including the lexicon. Johnson (2008a) and Johnson & Goldwater (2009) have already demonstrated that syllable structure provides valuable information in this context; however, their models relied on very different syllable regularities than those investigated here, and the consequences of these differences should be explored in future work.

Goldwater et al. (2009) showed that a number of proposed segmentation models have an undersegmentation bias that can be avoided by simultaneously modeling statistical dependencies between words. They proposed a Bayesian prior to favor a smaller lexicon and showed that otherwise unigram models introduce a severe undersegmentation bias due to the possibility of matching empirical probabilities by memorizing utterances as words. Note that the same is not true of syllable-based models since the hypothesis space does not permit memorization of utterances, and the size of the syllable inventory, unlike a lexicon, remains relatively stable under different segmentations. Thus, the syllable-based models are not subject to the same kind of undersegmentation bias. Interestingly, the syllable bigram model surpasses the performance of the word bigram model proposed by Goldwater et al. (word token f-score 72.3) given sufficient compensation for its undersegmentation bias. However, this level of performance requires adjustment of the $p_{\#}$ parameter to compensate for the cost of generating additional boundaries. Although parameters are common in computational models (for example, Goldwater et al. used a $p_{\#}$ parameter to modulate the prior distributions in their Bayesian models), they do not provide a particularly satisfying explanation for why infants are compelled to break up the speech stream into smaller units (words). Further work is needed to determine how undersegmentation biases are ultimately overcome by children.

References

Adriaans, Frans and Kager, René. 2010. Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language* 62(3): 311-331.

Albright, Adam. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology*, 26(1): 9-41.

Aslin, Richard N., Saffran, Jenny R., & Newport, Elissa L. 1998. Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9, 321-324.

Bernstein-Ratner, Nan. 1987. The phonology of parent child speech. *Children's Language*, 6: 159-174.

Blanchard, Daniel and Heinz, Jeffrey. 2008. Improving word segmentation by simultaneously learning phonotactics. In *Conll '08: Proceedings of the 12th Conference on Computational Natural Language Learning.* Stroudsburg, PA: Association for Computational Linguistics.

Blanchard, Daniel, Heinz, Jeffrey and Golinkoff, Roberta. 2010. Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language,* 37(3): 487-511.

Brent, Michael R. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34(1-3): 71-105.

Daland, Robert and Pierrehumbert, Janet B. 2011. Learning Diphone-Based Segmentation. *Cognitive science,* 35(1). Wiley Online Library. 119–155.

Daland, Robert, Hayes, Bruce, White, James, Garellek, Marc, Davis, Andrea and Norrmann, Ingrid. 2011. Explaining sonority projection effects. *Phonology*, 28(2): 197-234.

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

Goldwater, Sharon, Griffiths, Thomas L. and Johnson, Mark. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1): 21-54.

Hayes, Bruce & Wilson, Colin. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3): 379-440.

Hockema, Stephen A. 2006. Finding Words in Speech: An Investigation of American English. *Language Learning and Development,* 2(2). Psychology Press. 119-146.

Jarosz, Gaja and Johnson, J. Alex. 2013. The Richness of Distributional Cues to Word Boundaries in Speech to Young Children. *Language Learning and Development*, 9(2): 175-210.

Johnson, Mark. 2008a. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Johnson, Mark. 2008b. Unsupervised Word Segmentation for Sesotho Using Adaptor Grammars. *Proceedings of the 10th Meeting of ACL SIGMORPHON.* Columbus, OH: Association of Computational Linguistics.

Johnson, Mark & Goldwater, Sharon. 2009. Improving nonparametric Bayesian inference: Experiments on unsupervised word segmentation with adaptor grammars. In *NAACL '09: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics.* Boulder, CO: Association for Computational Linguistics.

Jurafsky, Daniel & Martin, James H. 2008. Speech and language processing, 2nd edition. Upper Saddle River, NJ: Prentice-Hall.

Jusczyk, Peter W. & Luce, Paul A. 1994. Infants' Sensitivity to Phonotactic Patterns in the Native Language. *Journal of Memory and Language*, 33(5): 630-645.

Lignos, Constantine and Yang, Charles. 2010. Recession Segmentation: Simpler Online Word Segmentation Using Limited Resources. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning.* (CoNLL '10). Association for Computational Linguistics. .

Mattys, Sven L. and Jusczyk, Peter W. 2000. Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78(2): 91-121.

Mattys, Sven L., Jusczyk, Peter W., Luce, Paul A., and Morgan, James L. 1999. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive psychology*, 38(4): 465-494.

Newport, Elissa L. and Aslin, Richard N. 2004. Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive psychology*, 48(2): 127-162.

Pelucchi, Bruna, Hay, Jessica F., and Saffran, Jenny R. 2009. Learning in reverse: Eight-month-old infants track backward transitional probabilities. *Cognition*, 113(2): 244-247.

Saffran, Jenny R., Aslin, Richard N., and Newport, Elissa L. 1996. Statistical learning by 8-month-old infants. *Science*, 274(5294): 1926-1928.

Swingley, Daniel. 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50(1): 86-32.

Thiessen, Erik D. and Saffran, Jenny R. 2003. When cues collide: use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental psychology,* 39(4): 706.

Venkataraman, Anand 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics*, 27(3): 352-372.

Yang, Charles D. 2004. Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, 8(10): 451-456.