Modeling Infant Word Segmentation

Constantine Lignos
Department of Computer and Information Science
University of Pennsylvania
lignos@cis.upenn.edu

Abstract

While many computational models have been created to explore how children might learn to segment words, the focus has largely been on achieving higher levels of performance and exploring cues suggested by artificial learning experiments. We propose a broader focus that includes designing models that display properties of infants' performance as they begin to segment words. We develop an efficient bootstrapping online learner with this focus in mind, and evaluate it on child-directed speech. In addition to attaining a high level of performance, this model predicts the error patterns seen in infants learning to segment words.

1 Introduction

The last fifteen years have seen an increased interest in the problem of how infants learn to segment a continuous stream of speech into words. Much of this work has been inspired by experiments with infants focusing on what capabilities infants have and which cues they attend to. While experimental work provides insight into the types of cues infants may be using, computational modeling of the task provides a unique opportunity to test proposed cues on representative data and validate potential approaches to using them.

While there are many potential approaches to the problem, a desirable solution to the problem should demonstrate acceptable performance in a simulation of the task, rely on cues in the input that an infant learner is able to detect at the relevant age, and exhibit learning patterns similar to those of infant learners. Most work in computational modeling of language acquisition has primarily focused on achieving acceptable performance using a single cue, transitional probabilities, but little effort has been made in that work to try to connect these learning solutions to the actual learning patterns observed in children outside of performance on short artificial language learning experiments.

In this work we present a simple, easily extended algorithm for unsupervised word segmentation that, in addition to achieving a high level of performance in the task, correlates with the developmental patterns observed in infants. We discuss the connections between the design and behavior of our algorithm and the cognitive capabilities of infants at the age at which they appear to begin segmenting words. We also discuss how our technique can easily be extended to accept additional cues to word segmentation beyond those implemented in our learner.

2 Related Work

As this paper examines the intersection of infants' capabilities and computational modeling, we discuss work in both domains, beginning with experimental approaches to understanding how infants may perform the task of word segmentation.

2.1 Infant Word Segmentation

A potential account of how infants learn to identify words in fluent speech is that they learn words in isolation and then use those words to segment longer utterances (Peters, 1983; Pinker et al., 1984). It is not clear, however, that infant-directed speech provides enough detectable words in isolation for

Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 29–38,
Portland, Oregon, USA, 23–24 June 2011. ©2011 Association for Computational Linguistics

such a strategy (Aslin et al., 1996). Whatever isolated words children do hear, they appear to attend to them; whether a word is heard in isolation is a better predictor of whether a child has learned a word than the word's frequency (Brent and Siskind, 2001).

A more plausible alternative account to assume children attend to patterns in the input, using them to identify likely word units. Much experimental work has followed from the finding that in artificial learning tasks, infants and adults appear to prefer word-like units that match statistical patterns in the input (Saffran et al., 1996b; Saffran et al., 1996a). Saffran et al. and the authors of following studies (Aslin et al., 1998; Saffran, 2001, among many others) suggest that participants used transitional probabilities to succeed in these experiments, but the actual strategy used is unclear and may even be an artifact of the perceptual system (Perruchet and Vinter, 1998; Hewlett and Cohen, 2009).

More recent work using real language data has not shown transitional probabilities to be as useful a cue as originally suggested. Lew-Williams et al. (2011) found that 9-month-old English-learning infants were not able to learn high-transitional probability words in fluent Italian speech unless those words were also presented in isolation. Given this finding and the extensive exisiting modeling work focusing on the used of transitional probabilities, we believe it is crucial to additionally explore segmentation strategies that rely on other cues in the input.

2.2 Modeling Word Segmentation

While experimental work has posited simple algorithms that infants might use to accomplish the task of word segmentation, when applied to real language data these techniques have yielded very poor results (Yang, 2004). This problem has created a challenge for researchers modeling language acquisition to suggest more sophisticated strategies that infants might use. These approaches have fallen into two primary categories: optimization-based and bootstrapping algorithm strategies.

Optimization-based strategies have focused on techniques that a learner might use to arrive at an optimal segmentation, either through a dynamic programming approach (Brent, 1999), online learning (Venkataraman, 2001), or nonparametric Bayesian inference (Goldwater et al., 2009; Johnson and Goldwater, 2009). These approaches fit within standard statistical approaches to natural language processing, defining statistical objectives and inference strategies, with the learners trying to optimize some combination of the quality of its lexicon and representations of the corpus.

In contrast, bootstrapping approaches (Gambell and Yang, 2004; Lignos and Yang, 2010) to word segmentation have focused on simple heuristics for populating a lexicon and strategies for using the contents of the lexicon to segment utterances. These approaches have focused on a procedure for segmentation rather than defining an optimal segmentation explicitly, and do not define a formal objective that is to be optimized.

While bootstrapping approaches have generally made stronger attempts to align with infants abilities to process the speech signal (Gambell and Yang, 2004) than other approaches, little effort has been made to connect the details of an implemented segmentation strategy with children's learning patterns since the earliest computational models of the task (Olivier, 1968). It is important to draw a contrast here between attempts to match patterns of human development with regard to word segmentation with attempts to model performance in artificial language learning experiments whose goal is to probe word segmentation abilities in humans (Frank et al., 2010). In this paper we are focused on matching the progression of development and performance in naturalistic experiments to characteristics of a segmentation strategy, an approach similar to that employed in English past tense learning (Rumelhart and McClelland, 1986; Pinker, 2000; Yang, 2002).

We will now discuss the patterns of development for children learning to segment English words, which form the motivation for the design of our segmenter.

3 Infant Performance in Word Segmentation

While the developmental patterns of English-learning infants have been broadly studied, it has been difficult to identify errors that must be caused by failures to correctly segment words and not other cognitive limitations, issues of morphological productivity, or syntactic competency issues.

Brown (1973) offers one of the most comprehensive examinations of the types of errors that young infants make regarding word segmentation. He notes that Adam's common errors included treating *it's-a, that-a, get-a, put-a, want-to*, and *at-that* as single words, as judged by various misproductions that involved these items. A possible analysis of these errors is that in addition to the high level of frequency with which those syllables co-occur, elements such as *a* and *to* do not carry any identifiable amount of stress in natural speech.

In addition to the undersegmentations that Brown identifies, Peters (1983) identifies the pattern of oversegmenting function words begin other words, including this famous dialog between a parent and child, where in the child's response *have* is pronounced in the same way as the second syllable of *behave*: Parent: Behave! Child: I *am* have!

The response by the child indicates that they have analyzed *behave* as *be have*. There are two major factors that could contribute to such an analysis: the high frequency of *be* leading to it being treated as a separate word (Saffran et al., 1996b), and the lack of stress on *be* but stress on *have* which forms a word contrary to the dominant pattern of stress in English (Cutler and Butterfield, 1992).

Infants appear to use the ends of utterances to aid segmentation, and as early at 7.5 months old they are able to recognize novel words in fluent speech if the novel words are presented at the ends of an utterance and not utterance medially (Seidl and Johnson, 2006). Thus the reliable boundaries presented by the edge of an utterance should be treated as informative for a learner.

Most crucially, the syllable seems to be the unit children use to form words. Experiments that have been performed to gauge adult and infant competency in word segmentation have been designed with the assumption that the only possible segmentation points are at syllable boundaries. That infants should be able to operate on syllables is unsurprising; infants as young as 4-days-old are able to discriminate words based on syllable length (Bijeljac-Babic et al., 1993) and phonotactic cues to syllable boundaries seem to be rapidly acquired by infants (Onishi et al., 2002). The use of the syllable in experimental work on word segmentation stands in contrast to many computational models that have operated at the phoneme level (Brent, 1999; Venkataraman, 2001; Goldwater et al., 2009). An exception to the focus on phoneme-based segmentation is the joint learning model proposed by Johnson (2008) that learns syllabification and other levels of representation jointly with word segmentation, but that model poses problems as a developmentally relevant approach in that it predicts unattested joint syllabification/segmentation errors by infants and problems as a linguistically relevant approach due to its non-phonotactic approach to learning syllabification.

From this survey, we see some relevant phenomena that a good model of infant word segmentation should replicate. (1) The learner should operate on syllables. (2) At some stage of learning, undersegmentation function word collocations (e.g., *that-a* should occur. (3) At some stage of learning, oversegmentation of function words that may begin other words (e.g., *be-have*) should occur. (4) The learner should attend to the ends of utterances as use them to help identify novel words.

4 An Algorithm for Segmentation

The algorithm we propose is similar in style to previous online bootstrapping segmenters (Gambell and Yang, 2004; Lignos and Yang, 2010) but varies in a few crucial aspects. First, it inserts word boundaries in a left-to-right fashion as it processes each utterance (i.e., in temporal order), unlike previous models which have worked from the outside in. Second, it can handle cases where the segmentation is ambiguous given the current lexicon and score multiple possible segmentations. Finally, the use of word-level stress information is an optional part of the model, and not an essential part of the segmentation process. This allows us to examine the additional power that stress provides on top of a subtractive segmentation system and allows the model to generalize to languages where word-level stress is not present in the same fashion as English (e.g., French). We first discuss the individual operations the algorithm uses to segment an utterance, and then discuss how they are combined in the segmenter.

4.1 The Lexicon

The learner we propose will primarily use items in its lexicon to help identify new possible words. The structure of the lexicon is as follows:

Lexicon. *The lexicon contains the phonological material of each word that the learner has previously hypothesized. The lexicon stores a score along with each word, which the segmenter may increment or decrement.*

The score assigned to each entry in the lexicon represents the relative confidence that it is a true word of the language. Each increment simply adds to the score of an individual word and each decrement subtracts from it.

4.2 Subtractive Segmentation

Subtractive segmentation is the process of using known words to segment the speech signal, which infants appear to be able to do as young as at six months of age (Bortfeld et al., 2005).

Subtractive Segmentation. *When possible, remove a known word in the lexicon from the front of the utterance being segmented.*

One way to apply subtractive segmentation is a greedy score-based heuristic for subtractive segmentation (Lignos and Yang, 2010), such that whenever multiple words in the lexicon could be subtracted from an utterance, the entry with the highest score will deterministically be used. This greedy approach results in a "rich get richer" effect of the sort seen in Dirichlet processes (Goldwater et al., 2009). We will first discuss this approach and then later extend this greedy search to a beam search.

Figure 1 gives the implementation of subtractive segmentation in our algorithm. This algorithm results in the following properties:

Initially, utterances are treated as words in isolation. When the lexicon is empty, no word boundaries will be inserted and the full contents of each utterance will be added to the lexicon as a word.

High-frequency words are preferred. When presented with a choice of multiple words to subtract, the highest scored word will be subtracted, which will prefer higher frequency words over lower frequency words in segmentation.

Syllables between words are not necessarily considered words. Syllables that occur between subtractions are not added as words in the lexicon. For example, if *play* and *please* are in the lexicon but *checkers* is not, the utterance *play checkers please* will be correctly segmented, but *checkers* will not be added to the lexicon. Much like infants appear to do, the learner does not place as much weight on less reliable boundaries hypothesized in the middle of an utterance (Seidl and Johnson, 2006).

4.3 Incorporating Stress Information

A particularly useful constraint for defining a word, introduced to the problem of word segmentation by Yang (2004) but previously discussed by Halle and Vergnaud (1987), is as follows:

Unique Stress Constraint (USC): A word can bear at most one primary stress.

Yang (2004) evaluated the effectiveness of the USC in conjunction with a simple approach to using transitional probabilities, showing significant performance improvements. The availability of such stress cues is not, however, an uncontroversial assumption; there are no language-universal cues to stress and even within a single language automatic detection of word-level stress is still unreliable (Van Kuijk and Boves, 1999), making automatic capture of such data for simulation purposes difficult.

Before taking advantage of word-level stress information, the infant learner would need to identify the acoustic correlates to word-level stress in her language, and we will not address the specific mechanisms that an infant learner may use to accomplish the task of identifying word-level stress in this paper. Based on strong experimental evidence that infants discriminate between weakly and strongly stressed syllables and use it to group syllables into word-like units (Jusczyk et al., 1999), we assume that an infant may attend to this cue and we evaluate our model with and without it.

We adopt the USC for segmentation in the following fashion:

Unique Stress Segmentation (USS). *Insert word boundaries such that no word contains two strong stresses. Do so in a lazy fashion, inserting boundaries as a last resort just before adding another syllable to the current would cause it to contain two strong stresses.*

```
$u \leftarrow$ the syllables of the utterance, initially with no word boundaries
$i \leftarrow 0$
while $i < len(u)$ do
    if $u$ starts with one or more words in the lexicon then
        Choose the highest scoring word $w$ and remove it from the front of $u$ by inserting a word boundary before and after it.
        Increment the score of $w$
        Advance $i$ to the last word boundary inserted
    else
        Advance $i$ by one syllable
    end if
end while
Add the syllables between the last boundary inserted (or the beginning of the utterance if no boundaries were inserted) and the end of the utterance as a word in the lexicon with a score of 1
```

Figure 1: Subtractive segmentation procedure

```
$u \leftarrow$ the syllables of the utterance, initally with no word boundaries
$i \leftarrow 0$
$seenStress \leftarrow$ False
while $i < len(u) - 1$ do
    if $u[i]$ is stressed then
        $seenStress \leftarrow$ True
    end if
    if $seenStress$ and $u[i+1]$ is stressed then
        Insert a word boundary between $u[i]$ and $u[i+1]$
        $w \leftarrow$ the syllables between the previous boundary inserted (or the beginning of the utterance if no boundaries were inserted) and the boundary just inserted
        Increment $w$'s score in the lexicon, adding it to the lexicon if needed
        $seenStress \leftarrow$ False
    end if
    $i \leftarrow i + 1$
end while
$w \leftarrow$ the syllables between the last boundary inserted (or the beginning of the utterance if no boundaries were inserted) and the end of the utterance
Increment $w$'s score in the lexicon, adding it to the lexicon if needed
```

Figure 2: A Unique Stress Segmentation Algorithm

This strategy is expressed in an algorithmic form in Figure 2. The learner uses USS as a last resort to prevent creating a segmentation with an impossible amount of stress in a single word. For example consider an unsegmented English utterance with the stressed syllables underlined: *Givemetheball*. Applying USS would create the following segmentation: *Givemethe ball*.

A USS-based algorithm would note the stress on the first syllable, then keep scanning until another stress is located on the fourth syllable, inserting a break between the two. *Givemethe* and *ball* would be added to the lexicon. While this is not a perfect segmentation, it can be used to aid subtractive segmentation by seeding the lexicon, even if not all entries added to the lexicon are not correct.

4.4 Combining Subtraction and Stress Information

Given our bootstrapping methodology, it is highly desirable to be able to integrate USS along with subtractive segmentation. An algorithm that combines both is shown in Figure 3.

4.5 Extending to Beam Search

The greedy segmentation proposed is limited in its ability to find a good segmentation by its reliance on local decisions. A frequent undersegmentation error of the greedy segmenter is of this type: *partof an apple*. Because *partof* has a higher score than *part* at the point in learning where this utterance is encountered, the greedy segmenter will always choose *partof*.

An alternative approach is to let the segmenter

```
$u \leftarrow$ the syllables of the utterance, initally with no word boundaries
$i \leftarrow 0$
**while** $i < len(u)$ **do**
    **if** USS requires a word boundary **then**
        Insert a word boundary and advance $i$, updating the lexicon as needed
    **else if** Subtractive Segmentation can be performed **then**
        Subtract the highest scoring word and advance $i$, updating the lexicon as needed
    **else**
        Advance $i$ by one syllable
    **end if**
**end while**
$w \leftarrow$ the syllables between the last boundary inserted (or the beginning of the utterance if no boundaries were inserted) and the end of the utterance
Increment $w$'s score in the lexicon, adding it to the lexicon if needed
```

Figure 3: An algorithm combining USS and Subtractive Segmentation

explore multiple hypotheses at once, using a simple beam search. New hypotheses are added to support multiple possible subtractive segmentations. For example, using the utterance above, at the beginning of segmentation either *part* or *partof* could be subtracted from the utterance, and both possible segmentations can be evaluated. The learner scores these hypotheses in a fashion similar to the greedy segmentation, but using a function based on the score of all words used in the utterance. The geometric mean has been used in compound splitting (Koehn and Knight, 2003), a task in many ways similar to word segmentation, so we adopt it as the criterion for selecting the best hypothesis. For a hypothesized segmentation H comprised of words $w_i \dots w_n$, a hypothesis is chosen as follows:

$$\arg\max_{H}\Big(\prod_{w_i \in H} score(w_i)\Big)^{\frac{1}{n}}$$

For any w not found in the lexicon we must assign a score; we assign it a score of one as that would be its value assuming it had just been added to the lexicon, an approach similar to Laplace smoothing.

Returning to the previous example, while the score of *partof* is greater than that of *part*, the score of *of* is much higher than either, so if both *partof an apple* and *part of an apple* are considered, the high score of *of* causes the latter to be chosen. When beam search is employed, only words used in the winning hypothesis are rewarded, similar to the greedy case where there are no other hypotheses.

In addition to preferring segmentations that use words of higher score, it is useful to reduce the score of words that led to the consideration of a losing hypothesis. In the previous example we may want to penalize *partof* so that we are less likely to choose a future segmentation that includes it. Setting the beam size to be two, forcing each hypothesis to develop greedily after an ambiguous subtraction causes two hypotheses to form, we are guaranteed a unique word to penalize. In the previous example *partof* causes the split between the two hypotheses in the beam, and thus the learner penalizes it to discourage using it in the future.

Algorithm	**Word Boundaries**		
	Precision	Recall	F-Score
No Stress Information			
Syllable Baseline	81.68	**100.0**	89.91
Subtractive Seg.	91.66	89.13	90.37
Subtractive Seg. + Beam 2	**92.74**	88.69	**90.67**
Word-level Stress			
USS Only	91.53	18.82	31.21
USS + Subtractive Seg.	93.76	**92.02**	92.88
USS + Subtractive Seg. + Beam 2	**94.20**	91.87	**93.02**

Table 1: Learner and baseline performance

5 Results

5.1 Evaluation

To evaluate the performance of our model, we measured performance on child-directed speech, using the same corpus used in a number of previous studies that used syllabified input (Yang, 2004; Gambell and Yang, 2004; Lignos and Yang, 2010). The evaluation set was comprised of adult utterances from the Brown (1973) data of the CHILDES database (MacWhinney, 2000).[1] Phonemic transcriptions of words from the Carnegie Mellon Pronouncing Dictionary (CMUdict) Version 0.7 (Weide, 1998), using the first pronunciation for each word and marking syllables with level 1 stress as strong syllables. The corpus was syllabified using onset maximization. Any utterance in which a word could not be transcribed using CMUDICT was excluded, leaving 55,840 utterances. We applied a probabilistic recall function to the lexicon to simulate the fact that a child learner will not perfectly recall all hypothesized words either due to memory limitations, variability in the input, or any other possible source of failure. We used the same function and constant as used by Lignos and Yang (2010).

To adjust the word-level stress information to better reflect natural speech, the stress information obtained from CMUdict was post-processed in the context of each utterance using the technique of Lignos and Yang (2010). For any n adjacent primary-stress syllables, only the nth syllable retains primary stress; all others are made into weak syllables. This reflects the fact that stress clash is avoided in English and that infants may not reliably detect acoustic correlates of stress in the input.

In addition to variations of our algorithm, we evaluated a baseline segmenter which marks every syllable boundary as a word boundary, treating each syllable as a word. We tested five variants of our algorithm, adding combinations of USS, subtractive segmentation, and adding beam search with a beam size of two[2] to subtractive segmentation.

Precision and recall metrics were calculated over all word boundaries over all utterances in the corpus. The segmenter's task is effectively to classify each syllable boundary as a word boundary or not. As single-syllable utterances are unambiguously a single word with no possible boundaries, they are excluded from evaluation but still given as input.

Evaluation was performed by giving each algorithm a single pass over the data set, with the performance on every utterance included in the total score. This is the most challenging metric for an online segmenter, as early mistakes made when the learner has been exposed to no data still count against it.

5.2 Performance

The performance of several variations of our algorithm is given in Table 1. The most surprising result is the high performance provided by the syllable baseline. This good performance is both an artifact of English and the metrics used to evaluate the segmenters. In English, there are larger number of monosyllabic words than in other languages, resulting in high precision in addition to the guaranteed 100% recall because it predicts every possible word boundary. The standard metric of evaluating precision and recall over word boundaries rather than words identified in each utterance also contributes to this performance; when this baseline is evaluated with a word-level precision and recall it does not perform as well (Lignos and Yang, 2010).

Subtractive Segmentation provides an improvement in utterance evaluation over the Syllable Baseline, and adding beam search to it slightly improves F-score, sacrificing precision for recall. This is to be expected from the penalization step in beam search; as the penalization penalizes some good words in addition to undesirable ones, the purification of the utterance segmentation and the lexicon comes at the cost of recall from over-penalization.

While USS alone is clearly not a sufficiently rich segmentation technique, it is important to note that it is a high precision indicator of word boundaries, suggesting that stress information can be useful to the learner even when used in this simple way. More importantly, USS contributes unique information to subtractive segmentation, as the utterance F-score of subtractive segmentation improves from 90.37 to 92.88.

While the performance numbers show that the segmenter performs competently at the task, the more significant question at hand is whether the errors committed by the learner match developmental patterns of infants. As the design of the segmenter predicts, the main error types of the Subtractive Seg-

[1] A separate set of previous studies have used a corpus selected by Brent (1999) for evaluation. Due to length limitations and the fact that the results presented here cannot be meaningfully compared to those studies, we only present results on the Brown (1973) data here.

[2] As larger beam sizes did not lead to any benefits, partly because they do not straightforwardly allow for penalization, we do not report results for larger beam sizes.

mentation + USS algorithm fall into two classes:

Function word collocations. For example, the third highest-scored non-word in the lexicon is *that'sa*, congruent with observations of function word collocations seen in children (Brown, 1973).

Oversegmentation of function words. The greedy approach used for segmenting the words of highest score results in function words being aggressively segmented off the front of words, for example *a nother*. The highest scored non-word in the lexicon is *nother* as a result.

Adding beam search reduces the number of function word collocations in the segmenter's output; the learner's most commonly penalized lexicon entry is *isthat*. However, beam search also penalizes a lot of words, such as *another*. Thus the strategy used in beam search predicts an early use of function word collocations, followed by later oversegmentation.

6 Discussion

In the discussion of related work, we identified two major paradigms in modeling word segmentation: optimization and bootstrapping approaches. The algorithm presented here combines elements of both. Its behavior over time and across utterances is that of a bootstrapping learner, but when processing each utterance it selects a segmentation based on a simple, cognitively plausible beam search.

By using a beam search of the kind suggested, it is easy to see how a variety of other cues could be integrated into the learning process. We have given a simple function for selecting the best hypothesis that only relies on lexicon scores, but more sophisticated functions could take multiple cues into account. For example it has been observed that 7-month-olds attend more to distributional cues while 9-month-olds attend more to stress cues (Thiessen and Saffran, 2003). A learner in which the weight placed on stress cues increases as the learner receives more data would match this pattern. Other research has suggested a more complex hierarchy of cues (Mattys et al., 2005), but how the weighting of the various cues can be adjusted with more input remains an open question.

A crucial frontier in word segmentation is the expansion of evaluation to include other languages. As with many other tasks, creating solutions that perform well in a broad variety of languages is important but has not yet been pursued. Future work should attempt to match developmental patterns in other languages, which will require adding morphological complexity to the system; the techniques developed for English are unlikely to succeed unchanged in other languages.

Comparing with other algorithms' published results is difficult because of varying choices of data sets and metrics. For example, other syllable-based algorithms have evaluated their performance using word-level, as opposed to boundary-level, precision and recall (Gambell and Yang, 2004; Lignos and Yang, 2010). We have adopted the more popular boundary-based metric here, but there is no way to directly compare with work that does not use syllabified input. The variety of possible evaluation metrics obviates the need for a longer-form exploration of how existing approaches perform when evaluated against varying metrics. Additionally, a more standard set of evaluation data in many languages would greatly improve the ability to compare different approaches to this task.

7 Conclusion

The work presented here represents a step toward bringing together developmental knowledge regarding word segmentation and computational modeling. Rather than focusing on cues in artificial learning experiments which may or may not generalize to the natural development of word segmentation in children, we have shown how a simple algorithm for segmentation mimics many of the patterns seen in infants' developing competence. We believe this work opens the door to a promising line of research that will make a stronger effort to see simulations of language acquisition as not just an unsupervised learning task but rather a modeling task that must take into account a broad variety of phenomena.

8 Acknowledgments

I would like to thank Charles Yang and Mitch Marcus for many enlightening discussions regarding this work. The author was supported by an NSF IGERT grant to the University of Pennsylvania Institute for Research in Cognitive Science.

References

R.N. Aslin, J.Z. Woodward, N.P. LaMendola, and T.G. Bever. 1996. Models of word segmentation in fluent maternal speech to infants. *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, pages 117–134.

R.N. Aslin, J.R. Saffran, and E.L. Newport. 1998. Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9(4):321.

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

H. Bortfeld, J.L. Morgan, R.M. Golinkoff, and K. Rathbun. 2005. Mommy and me. *Psychological Science*, 16(4):298.

M.R. Brent and J.M. Siskind. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81(2):B33–B44.

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

R. Brown. 1973. *A First Language: The Early Stages.* Harvard Univ. Press, Cambridge, Massachusetts 02138.

A. Cutler and S. Butterfield. 1992. Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, 31(2):218–236.

M.C. Frank, S. Goldwater, T.L. Griffiths, and J.B. Tenenbaum. 2010. Modeling human performance in statistical word segmentation. *Cognition*.

T. Gambell and C. Yang. 2004. Statistics learning and universal grammar: Modeling word segmentation. In *First Workshop on Psycho-computational Models of Human Language Acquisition*, page 49.

S. Goldwater, T.L. Griffiths, and M. Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*.

M. Halle and J.R. Vergnaud. 1987. *An essay on stress*. MIT Press.

D. Hewlett and P. Cohen. 2009. Word segmentation as general chunking. In *Psychocomputational Models of Language Acquisition Workshop (PsychoCompLA)*, July 29, 2009.

M. Johnson and S. Goldwater. 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*, pages 317–325. Association for Computational Linguistics.

M. Johnson. 2008. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *46th Annual Meeting of the ACL*, pages 398–406. Citeseer.

P.W. Jusczyk, D.M. Houston, and M. Newsome. 1999. The Beginnings of Word Segmentation in English-Learning Infants. *Cognitive Psychology*, 39(3-4):159–207.

P. Koehn and K. Knight. 2003. Empirical methods for compound splitting. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, pages 187–193. Association for Computational Linguistics.

C. Lew-Williams, B. Pelucchi, and J. Saffran. 2011. Isolated words enhance statistical learning by 9-month-old infants. In *Budapest CEU Conference on Cognitive Development 2011*.

C. Lignos and C. Yang. 2010. Recession Segmentation: Simpler Online Word Segmentation Using Limited Resources. In *Proceedings of CoNLL-2010*, pages 88–97.

B. MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates.

S.L. Mattys, L. White, and J.F. Melhorn. 2005. Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology-General*, 134(4):477–500.

D.C. Olivier. 1968. *Stochastic grammars and language acquisition mechanisms: a thesis*. Ph.D. thesis, Harvard University.

K.H. Onishi, K.E. Chambers, and C. Fisher. 2002. Learning phonotactic constraints from brief auditory experience. *Cognition*, 83(1):B13–B23.

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

A.M. Peters. 1983. *The units of language acquisition*. CUP Archive.

S. Pinker, Harvard University. The President, and Fellows of Harvard College. 1984. *Language learnability and language development*. Harvard University Press Cambridge, MA.

S. Pinker. 2000. *Words and rules: The ingredients of language*. Harper Perennial.

D.E. Rumelhart and J.L. McClelland. 1986. *Parallel distributed processing: Explorations in the microstructure of cognition.* MIT Press, Cambridge, MA.

J.R. Saffran, R.N. Aslin, and E.L. Newport. 1996a. Statistical Learning by 8-month-old Infants. *Science*, 274(5294):1926.

J.R. Saffran, E.L. Newport, and R.N. Aslin. 1996b. Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language*, 35(4):606–621.

J.R. Saffran. 2001. Words in a sea of sounds: The output of infant statistical learning. *Cognition*, 81(2):149–169.

A. Seidl and E.K. Johnson. 2006. Infant word segmentation revisited: edge alignment facilitates target extraction. *Developmental Science*, 9(6):565–573.

E.D. Thiessen and J.R. Saffran. 2003. When cues collide: Use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental Psychology*, 39(4):706–716.

D. Van Kuijk and L. Boves. 1999. Acoustic characteristics of lexical stress in continuous telephone speech. *Speech Communication*, 27(2):95–111.

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

R.L. Weide. 1998. The Carnegie Mellon Pronouncing Dictionary [cmudict. 0.6].

C.D. Yang. 2002. *Knowledge and learning in natural language*. Oxford University Press, USA.

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