Infant Word Segmentation: An Incremental, Integrated Model

Constantine Lignos
University of Pennsylvania

1. Introduction

For adults, the segmentation of continuous speech into words is generally effortless. For an infant with limited knowledge of the language being learned, the task is far from trivial. To succeed, she must identify the words of her language from a stream of input that rarely presents them in isolation and provides little feedback that she has segmented correctly.

Although the problem of infant word segmentation received only minor attention in early studies of language development (Brown, 1973; Clark, 1974; MacWhinney, 1978), it has become an active area of research in the last fifteen years primarily because of its ability to serve as a testing ground for the type of statistical information that infants can use. 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. A model of word segmentation must integrate knowledge of infants' experimental performance and developmental trajectory with a computational modeling framework to build as complete a picture of word segmentation as possible.

While the benefit of this integration should be obvious, the connection between computational models and experimental and longitudinal studies of word segmentation is thus far tenuous. The most well-known modern computational models (Goldwater et al., 2009; Johnson & Goldwater, 2009) heavily focus on the role of transitional probabilities and Bayesian inference but have not connected with infant development in a meaningful way. In those models and many others, the focus has been on Marr's *computational* level of analysis, investigating *what* a learner may be trying to optimize when trying to learn to segment rather than *how* it may reach its goal (Marr, 1983).

In contrast, the approach discussed in this paper focuses on Marr's *algorithmic* level, developing a specific, cognitively plausible strategy based on infants' performance in word segmentation experiments and comparing its predictions to the patterns observed in infant development. We use experimental evidence to suggest the operations the segmenter should perform and propose a simple learning algorithm that predicts the type of developmental changes observed in child language development. We demonstrate the effectiveness of this model at the task of word segmentation and at replicating patterns of infant development.

2. Related work

2.1. Experimental work in infant word segmentation

One simple account of how infants could learn to identify words in fluent speech is that they learn words in isolation and then use those words to segment longer utterances. It is unlikely, however, that infant-directed speech provides enough detectable words in isolation for such a strategy to succeed (Aslin et al., 1996). Words in isolation are valuable input for learners; words that appear in isolation are likely to be known by infants (Brent & Siskind, 2001), and there are heuristics for establishing whether an utterance contains only one word (Gambell & Yang, 2004). But for infants to succeed, a

* This work was presented jointly with Charles Yang at WCCFL 30. The author would like to thank Frans Adriaans, Elika Bergelson, Gene Buckley, Mitch Marcus, Daniel Swingley, Charles Yang, and audiences at CUNY Graduate Center and Yale University 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.

 Proceedings of the 30th West Coast Conference on Formal Linguistics, ed. Nathan Arnett and Ryan Bennett, 237-247. Somerville, MA: Cascadilla Proceedings Project.

more sophisticated strategy than merely identifying words in isolation is needed. Infants must attend to patterns in the input and use them to identify likely word units.

Experiments using infant and adult subjects have suggested a set of cues likely to be used in word segmentation. Much experimental work has followed from the finding that in artificial learning tasks, infants and adults appear to prefer auditory sequences that match statistical patterns in the input (Saffran et al., 1996b,a). Saffran and the authors of following studies (among *many* others, Aslin et al., 1998; Saffran, 2001) suggest that participants used transitional probabilities to learn the preference displayed in these experiments. However, the actual strategy used by participants in these is unclear and may not match the strategies suggested by the experimenters. These preferences may simply be an artifact of the perceptual system without any actual computation of transitional probabilities (Perruchet & Vinter, 1998), or they could be the result of transitional probabilities being computed without any extraction of word-like units at all (Endress & Mehler, 2009).

More recent work using more naturalistic stimuli has not shown transitional probabilities to be as robust 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. Shukla et al. (2011) found that infants were only able to segment nonce auditory words from naturalistic input when the nonce word was aligned with a prosodic boundary, failing when transitional probability information did not align with prosodic boundaries. Given these findings and the extensive existing modeling work focusing on the used of transitional probabilities, we believe it is crucial to explore segmentation strategies that rely on other information in the input.

Infants' segmentation capabilities are not limited to their ability to detect statistical cues; they use words they recognize to segment the input and use prosodic information to aid segmentation. At six months of age, infants can use familiar words such as *mommy* and their own names to segment unfamiliar words from fluent speech (Bortfeld et al., 2005). Infants appear to use the ends of utterances to aid segmentation, and as early as 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 & Johnson, 2006). The reliable boundaries presented by the edge of an utterance should thus be treated as informative for a learner.

The syllable appears to be the unit that young infants use when perceiving speech. Infants are able to operate on syllables essentially from birth; infants as young as 4-days-old are able to discriminate words based on syllable length (Bijeljac-Babic et al., 1993). At young ages, infants appear to perceive speech as holistic syllables rather than phoneme sequences (Bertoncini & Mehler, 1981; Jusczyk & Derrah, 1987). Additionally, infants appear to use the syllable as the relevant unit for word segmentation. Their segmentation errors align with syllable boundaries (Peters, 1983), and segmentations errors not aligned with syllable boundaries have not been reported. 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.

2.2. Developmental patterns in infant word segmentation

While the developmental patterns of English-learning infants have been broadly studied, it has been difficult to isolate errors that must be caused by failures to correctly segment words and not other cognitive limitations, issues of morphological productivity, or syntactic competency issues. As a result only a few word segmentation-specific errors have been recorded in infant development studies.

Brown (1973) offers one of the most comprehensive examinations of the types of errors that 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. These errors are likely caused by a combination of distributional and prosodic factors; the words in these collocations co-occur with high frequency and most functional elements do not carry any identifiable amount of stress in natural speech.

In addition to the undersegmentations that Brown identifies, Peters (1983) identifies a pattern of children oversegmenting function words. In one dialog between a parent and child when an adult tells a child that she "must behave" her response is "I *am* [heɪv]!". The response by the child indicates that she has 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 and the lack of stress on *be*, and stress on *have* which forms a word contrary to the dominant pattern of stress in English (Cutler & Butterfield, 1992).

2.3. Computational modeling

While computational models of infant word segmentation have focused on a number of different strategies for word segmentation, efforts to connect models and experimental accounts of infants' word segmentation capabilities have been limited. Experimenters have proposed simple statistical algorithms that infants might use to accomplish the task of word segmentation on the simple stimuli used in their experiments (e.g, Saffran et al., 1996a), but when applied to real language data these techniques have yielded very poor results (Yang, 2004). This failure challenges researchers to propose more sophisticated strategies that infants might use to take advantage of statistical information in the input.

The first effective and tractable computational model of word segmentation following the experiments of Saffran et al. (1996a) focus on segmentation strategies that could infer the best lexicon that could have produced the input using minimum description length (Brent, 1999). Later models (Borschinger & Johnson, 2011; Goldwater et al., 2009; Johnson & Goldwater, 2009; Pearl et al., 2011) build on the Bayesian properties of Brent's model, taking advantage of newer methods of approximating optimal solutions that made more sophisticated models computationally tractable. These approaches are implemented from current 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. As models of cognitive processes, all but Brent 1999 are defined at the *computational* level of analysis in that they explore what the learner is trying to optimize and what type of information is useful in learning. No claims are made for these models that the actual mechanisms and algorithms employed are the same as what occurs in the human mind or could feasibly be computed during acquisition.

Crucially, computational models of word segmentation have diverged from experimental evidence regarding the input to the infant learner. Despite the use of the syllable in experimental work on word segmentation and the evidence that young infants use the syllable as the primary unit of speech perception, most computational models intended to model young infants have operated at the phoneme level (Brent, 1999; Borschinger & Johnson, 2011; Goldwater et al., 2009; Pearl et al., 2011) with rare exceptions (Phillips & Pearl, 2012). While the learning strategies of phoneme-based models can in theory be adapted to be syllable-based, the input mismatch between infants' speech perception and phoneme-based models raises the question of whether studies of these models can provide a useful cognitive model of word segmentation even at Marr's computational level.

Other phoneme-based computational models, which acknowledge that they must be active at some later point in development where phonemic categories are better defined, have taken advantage of the phonemic representation to explore phonotactically-driven approaches, relying on diphone probabilities and an experimenter-provided estimate of the frequency of word boundaries (Daland & Pierrehumbert, 2011) or learning phonological generalizations over phoneme sequences (Adriaans & Kager, 2010).

A variation on phoneme-based segmentation is the joint learning model proposed by Johnson & Goldwater (2009) which assumes the input consists of sequences of phonemes but learns syllabification and other levels of representation jointly with word segmentation. Although it provides a useful application of powerful joint learning techniques, their model poses serious problems if considered as a model of infant development. It predicts unattested joint syllabification/segmentation errors by infants, and its memorization-based approach to learning syllabification does not align with any linguistic theory's concept of permissible syllables of a language. It also predicts that infants' phonological learning proceeds in the opposite order to what has been observed; the model requires fine phonemic categories *before* learning to recognize syllables. Finally, the framework itself can only connect weakly to any linguistic hierarchical representation as the model allows only one sort of hierarchical word structure–syllables or morphemes, but not both–and requires arbitrary linguistically and psychologically unattested levels of collocations within the utterance that are disconnected from syntactic constituents.

Bootstrapping approaches to word segmentation (Gambell & Yang, 2004; Lignos & Yang, 2010) have focused on simple heuristics for populating a lexicon and strategies for using the contents of the

lexicon to segment utterances, defining themselves at the *algorithmic* level of analysis as cognitively plausible algorithms for learning. 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 & 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).[1] In this paper we focus 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 & McClelland, 1986; Pinker, 1999; Yang, 2002). The use of a model at the algorithmic level of analysis allows for much tighter integration with experimental work in word segmentation; rather than suggesting what the learner could be optimizing, we build our learner from evidence of the algorithmic processes infants are using and demonstrate that its behavior matches that of infants.

2.4. Requirements for a model of word segmentation

From the above survey child language development studies, we can identify the most important phenomena that a good model of infant word segmentation should replicate. First, a learner modeling early development should operate on syllabified input. Second, at early stages of learning, undersegmentation of function word collocations (e.g., *that-a*) should occur. Third, at later stages of learning, oversegmentation of function words that can begin other words (e.g., *be have*) should occur. Finally, the learner should attend to the ends of utterances and use them to help identify novel words.

3. An algorithm for segmentation

In this section we describe the mechanisms of an algorithm for word segmentation first presented in Lignos 2011, discussing the individual operations the algorithm uses to segment an utterance, their connections to experimental work, and how they are combined in the segmenter.

This algorithm is similar in style to previous online bootstrapping segmenters (Gambell & Yang, 2004; Lignos & 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 that work from the outside in or begin with random segmentations. Second, it is able to resolve cases where the segmentation is ambiguous given the current lexicon by considering multiple possible segmentations of an utterance. 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). Unlike incremental models that allow memory of recent utterances to affect the segmentation of the current utterance (Pearl et al., 2011), this algorithm is incremental and online in the strictest sense; it sees each utterance in the data set once and cannot carry forward anything other than its lexicon as it processes more data.

Full implementation details, including pseudocode for all variants, are given in Lignos 2011. An open-source reference implementation is available from the author.

3.1. Lexicon

The lexicon is the central data structure for the segmenter; the learner's goal is to populate the lexicon with the highest quality words and use those words to segment utterances. The structure of the lexicon is as follows:

Lexicon. *The lexicon contains the phonological content (syllables) 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.*

[1] It is important to differentiate between our goal of matching patterns of long-term infant development and attempts to model short-term performance in artificial language learning experiments such as Frank et al. 2010.

The score assigned to each entry in the lexicon represents the relative confidence that it is a true word of the language. Each increment adds to the score of an individual word and each decrement (penalization) subtracts from it.[2]

3.2. Subtractive segmentation

Subtractive segmentation is the process of using words in the lexicon to segment the speech signal. Infants appear to be able to perform subtractive segmentation from at least six months of age for very salient words (Bortfeld et al., 2005). Subtractive segmentation is applied iteratively working from left to right within the utterance:

Subtractive Segmentation. *When possible, remove a known word in the lexicon from the remaining utterance.*

If a word can be subtracted from the beginning of the utterance, it is removed from the utterance and subtractive segmentation is applied to the remaining utterance. Otherwise, the segmenter tries to apply subtractive segmentation at the next syllable, moving rightward until a word already in the lexicon is found or the end of the utterance is reached.

The simplest way to resolve ambiguity in subtractive segmentation is a greedy score-based heuristic for subtractive segmentation (Lignos & 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. As the score of each word used is incremented this greedy approach results in a "rich get richer" effect of the sort seen in Dirichlet processes (Goldwater et al., 2009).

3.3. Updating the lexicon

Words enter the lexicon with a score of 1 when they are produced by the segmentation of an utterance. When there are no words in the lexicon that can be subtracted from the utterance, the segmenter simply adds the entire utterance to the lexicon. When subtractive segmentation can be applied to an utterance, the lexicon is updated in two ways. New words that are created by the segmentation are added to the lexicon, and the score of each word in the lexicon that used to segment a new utterance is increased.

For example, as the lexicon starts empty, the first utterance of our evaluation *big drum* is added as a single two-syllable word in the lexicon (*bigdrum*) with a score of one.[3] If the next utterance the learner encounters is *big drum boom*, the learner will subtract the word *bigdrum* which is in its lexicon, produce the segmentation *bigdrum boom*, increment the score of *bigdrum* to two, and add *boom* to its lexicon with a score of one.

However, not all words in a segmentation are equally trustworthy. If a word touches the end of an utterance, as *boom* does in the previous example, at least one of the boundaries of the word is correct, as an utterance boundary must be a word boundary. A variation on adding all new words to the lexicon is to add a *Trust* feature to the learner that only adds words to the lexicon if they touch the end of the utterance.

3.4. Considering multiple utterance segmentations

Applying subtractive segmentation greedily when multiple words can be subtracted often results in suboptimal segmentations. Consider the utterance *Is that a broom?* from the Adam dataset of Brown (1973) in the CHILDES database. While the segmenter attempts to maintain the highest quality lexicon possible, inevitably non-words like *isthat* which are collocations of other words are added to the lexicon, so the segmenter must decide whether this utterance begins with *is* or *isthat*. Early on in learning, *isthat* often has a higher score than *is*, so a greedy approach will subtract the incorrect word.

---

[2] The scores may be thought of as raw counts or probabilities; their normalization has no impact on the segmenter's behavior as long as all words' scores are transformed in the same way.

[3] For ease of reading, we give example data in orthographic form, but the input to the learner is phonemic syllables.

A simple solution is to let the segmenter explore multiple hypotheses at once, effectively using beam search.[4] New hypotheses are added to support multiple possible subtractive segmentations. For example, using the utterance above, at the beginning of segmentation either *is* or *isthat* could be subtracted from the utterance, and both possible segmentations can be evaluated. The learner scores these hypotheses in a fashion similar to a greedy approach but uses a function based on the score of all words used in the utterance rather than deciding using one word at a time. The geometric mean has been used in compound splitting (Koehn & Knight, 2003), a task in many ways similar to word segmentation, so it used as the criterion for selecting the best hypothesis. For a hypothesized segmentation H comprised of words $w_i \ldots w_n$, a hypothesis is chosen as follows: $\arg\max_H (\prod_{w_i \in H} score(w_i))^{\frac{1}{n}}$. In other words, the segmentation with the highest geometric mean of the scores of its words is selected. For any w not already 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 *isthat* is greater than that of *is*, the score of *that* is much higher than either, so if both *isthat a broom* and *is that a broom* are considered, the high score of *that* 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, guarantees 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.

3.5. Discussion

This behavior of this algorithm has in the following properties, which as discussed in Section 2 align with what has been observed in infant development.

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. Thus, the model predicts that the first (and easiest) words to be acquired will be the words that appear in isolation. Additionally, phrases that appear often as the sole content of an utterance (e.g., *all gone*) will initially be treated as a single word.

High-frequency units are preferred in segmentations. When presented with a choice of multiple items in the lexicon to subtract, the highest scored word will be subtracted, giving preference to the most frequent units in segmentation. In early stages of learning, high frequency word collocations will be treated as single words and reused often. In later stages of learning where the scores of words are representative of their frequency in the input and penalization has been applied to reduce the scores of frequent collocations, overuse of these functional elements will lead to undersegmentation errors.

New words are recognized more easily when they are adjacent to utterance boundaries. When the *Trust* feature is used, 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 (Seidl & Johnson, 2006), the learner does not place as much weight on less reliable boundaries hypothesized in the middle of an utterance.

[4] This varies somewhat from typical applications of beam search in that the beam is not used as a less expensive approximation of an exhaustive search. It is used to let the segmenter explore multiple hypotheses within a very small space. This is why beam sizes of larger than two are of little or no benefit to this learner.

4. Results

4.1. Evaluation

To evaluate the algorithm presented, we must confirm that the model can perform well at the task of segmentation but more importantly validate that it behaves as predicted and matches the developmental trajectory of infants. We measured performance of the presented algorithm on child-directed speech, using the same corpus used in a number of previous studies that used syllabified input (Yang, 2004; Gambell & Yang, 2004; Lignos & Yang, 2010). The evaluation set consisted of adult utterances from the Brown (1973) data of the CHILDES database (MacWhinney, 2000). Phonemic transcriptions of words were generated using a modified version of the Carnegie Mellon Pronouncing Dictionary (CMUdict) Version 0.7 created by adding common words of infant-directed speech and removing stress from pronunciations for the unstressed functional elements given by Selkirk 1984. The first pronunciation for each word was used, and syllables with level 1 stress were marked as stressed syllables. The corpus was syllabified using onset maximization, and any utterance in which a word could not be transcribed using CMUdict was excluded, yielding a total of 56,755 utterances.

In addition to variations of the presented algorithm, we evaluated several baseline segmenters. The *Utterance* baseline treats every utterance as a single word. The *Random* segmenter randomly labels each syllable boundary as a word boundary with .50 probability, while the *Oracle Random* baseline randomly labels each syllable boundary as a word boundary with the true probability of a boundary in this data set (.802). The *Syllable* segmenter labels every syllable boundary as a word boundary.

We evaluated three variants of our algorithm, each building on the previous and adding additional features as described in Section 3: greedy subtractive segmentation (*Subtractive* in Table 1), greedy subtractive segmentation with *Trust* (*+Trust*), and multiple hypothesis subtractive segmentation with *Trust* (*+Multiple Hypotheses*). The baselines and algorithm variants were evaluated by their ability to detect word boundaries and the quality of the words identified by the segmenter in each utterance.

To evaluate the reliability of detecting word boundaries, we computed the hit rate, false alarm rate, and A-prime (A') metrics.[5] We selected A-prime instead of d-prime because it requires fewer assumptions about the distribution of segmenter performance (Donaldson, 1993). These metrics are evaluated by comparing each syllable boundary[6] in the segmenter's output to the canonical segmentation[7] of the input, checking whether the segmenter's output and the canonical version agree on whether a word boundary should be placed there.

In other words, for the purpose of evaluating boundaries we cast the segmenter's task as detecting whether each syllable boundary is a word boundary or not. When the canonical segmentation gives a word boundary at a given syllable boundary, if the output of the segmenter labels a word boundary there it is a true positive (TP) and otherwise a false negative (FN). If the canonical segmentation does not give a word boundary at a given syllable boundary, if the segmenter labels a word boundary there it is a false positive (FP) and otherwise a true negative (TN). The hit rate, false alarm rate, and A-prime metrics are calculated from theses scores as given below.

Hit rate, which is also called recall, gives the rate at which the segmenter identified the word boundaries given by the canonical segmentation. False alarm rate gives the rate at which the segmenter misidentified word-medial syllable boundaries in the canonical segmentation as word boundaries.

---

[5] The advantages of using these signal detection metrics as opposed to the more commonly used information retrieval (IR) metrics (precision, recall, and F-score) for evaluating word boundary quality are discussed in Adriaans & Kager 2010. In brief, signal detection metrics provide a much clearer evaluation of performance against random or trivial baselines in cases of uneven labels.

[6] As the input is divided into utterances, utterance-initial and utterance-final syllable boundaries are unambiguously word boundaries and are not considered in evaluation. A consequence of this is that single-syllable utterances are unambiguously one word and are excluded from evaluation.

[7] In this work and in all other corpus simulations of word segmentation, we use orthographic spaces as the canonical segmentation of the input. In English, this aligns well with syntactic/morphological/semantic word boundaries and is consistent with the definition of word segmentation as a process whose goal is to produce units for the learner's grammar to operate on. It is possible to evaluate segmentation using prosodic word boundaries. However, that evaluation would be problematic in that learning that canonical segmentation would be of little use to the learner; determiners would be segmented as a single word with adjacent nouns, with no process for learning to separate them. Additionally, the undersegmentation errors reported by Brown (1973) do not align with prosodic word boundaries.

	Word Boundaries			Word Tokens		
Baseline	Hit Rate	FA Rate	A'	Precision	Recall	F-score
Utterance	0	0	-	0.045	0.010	0.016
Random	0.500	0.493	0.506	0.459	0.321	0.378
Oracle Random	0.803	0.803	0.500	0.607	0.608	0.607
Syllable	1.0	1.0	-	0.692	0.827	0.753
Algorithm	Hit Rate	FA Rate	A'	Precision	Recall	F-score
Subtractive	**0.992**	0.776	0.795	0.746	0.854	0.797
+Trust	0.960	0.469	0.860	0.817	0.866	0.841
+Multiple Hypotheses	0.953	**0.401**	**0.875**	**0.832**	**0.867**	**0.849**

Table 1: Learner and baseline performance

To evaluate the quality of the words in each utterance, we use the traditional information retrieval metrics of precision, recall (which is identical to hit rate), and F-score (also known as F1). These metrics, defined below, capture how well the segmenter has identified the words contained in each utterance. The equations for these metrics are given below.

$$\text{Hit rate } (H): \; H = \frac{TP}{TP+FN} \qquad \text{Precision } (P): \; P = \frac{TP}{TP+FP}$$

$$\text{False alarm rate } (F): \; F = \frac{FP}{TN+FP} \qquad \text{Recall } (R): \; R = \frac{TP}{TP+FN}$$

$$\text{A-prime } (A'): \; A' = \frac{1}{2} + \frac{(H-F)(1+H-F)}{4H(1-F)} \qquad \text{F-score } (F1): \; F1 = 2\frac{PR}{P+R}$$

For example, assume that *is that a lady* is segmented into three words as *isthat a lady*. The words *a* and *lady* are true positives (words in both the output and canonical segmentation), *isthat* is a false positive (word in the output that was not in the canonical segmentation), and *is* and *that* are false negatives (in the canonical segmentation but not in the output). This would lead to a precision of $2/3$ and a recall of $2/4$.

Evaluation was performed by giving each algorithm a single pass over the data set. The learner segments each utterance when it is seen without being able to consult previous utterances, and the performance on every utterance included in the total score. This is the most difficult metric for an online segmenter; early mistakes made when the learner has been exposed to little data are still counted as errors.

4.2. Performance

The performance of several variations of the presented algorithm and baselines is given in Table 1. Evaluation using the A' metric shows the differences between the baselines and algorithm performance most clearly. Simple baselines that do not discriminate based on the input either have undefined (effectively zero) A' scores as they create hits just as often as false alarms; the Syllable segmenter inserts all possible boundaries while the Utterance segmenter inserts none. Random segmenters provide chance discrimination, even in the Oracle Random case where the true probability of word boundaries is given to the baseline method.

While these numbers are given to quantify the quality of segmenter's output, as with most cognitive modeling work it is difficult to assess what an appropriate level of performance is. For the purposes of this study, it is sufficient to note that boundary discrimination and the quality of extracted words increase with the extensions to the simple subtractive segmentation model and that all variants of the algorithm perform better than baselines.

4.3. Behavior during learning

Analyzing the actual output of the algorithm is of much greater interest than merely quantifying the quality of its segmentation. An important characteristic of a faithful model of acquisition is that it make the same type of errors over time that infants do. To characterize the type of errors the segmenter makes

Early Errors			Late Errors		
Error	**Example**	**Frequency**	**Error**	**Example**	**Frequency**
oh	over	209	a	away	441
a	away	184	oh	over	194
thats-a	-	101	some	something	101
thank-you	-	45	any	anyone	77
some	something	39	all	always	67
all	always	31	every	everyone	60
any	anyone	31	in	inside	57
it's-a	-	30	on	onto	53
why-don't	-	28	-ty	pretty	41
don't-know	-	26	be	become	40
at-the	-	24	more	anymore	39
put-the	-	24	huh	honey	37

Table 2: Most frequent word errors in early and late stages of learning

	Error Categories			
Stage	**Func. Words**	**Func. Coll.**	**Content Coll.**	**Other**
Early	44.2%	37.0%	8.5%	10.4%
Late	70.6%	1.0%	1.2%	27.3%

Table 3: Distribution of error tokens across categories in early and late stages of learning.

at different points in time during learning, we labeled the 100 most frequent incorrect words produced by the Multiple Hypotheses variant of the algorithm over two stages of learning: the first and last 10,000 utterances processed by the segmenter. These errors represent words predicted in an utterance that are different than the canonical segmentation, for example when *is that* is segmented as *isthat* one error word is counted. Although many of these "errors" may be palatable analyses (e.g., *something* as *some thing*, *another* as *a nother*) we consider them errors as the output diverges from an adult's segmentation of the same utterance. The most frequent errors words and their frequencies are shown in Table 2.

Errors were divided into the following categories: function word oversegmentation (e.g., *a way* instead of *away*), function word collocations (e.g., *it's-a*), content word collocations (e.g. *fell-down*), and other errors which were not of a consistent pattern (e.g., *tu mmy* for *tummy*). The number of tokens in each category is given in Table 3. A Chi-squared test shows that the distribution of errors in each category across the two stages are significantly different ($p < .0001$). The early time period errors include more function word collocations and content word collocations, while the later utterances show significantly more oversegmentation of function words. This change in the distribution is consistent with the predictions made by the structure of the algorithm; penalization of words that lead to losing hypotheses reduces the number of collocations while rich-get-richer scoring results in high frequency elements being overused with more exposure.

Allowing the segmenter to consider multiple hypotheses 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 many correct words, such as *another* in favor of *a nother*. This confirms that this mechanism corrects for an early use of function word collocations but later causes oversegmentation errors using functional elements. As discussed earlier, this is the developmental trajectory documented in Brown 1973 and Peters 1983.

4.4. Discussion

The results given here demonstrate that this simple algorithm performs well at the task of word segmentation and its behavior over time is similar to that of infants. However, such a simple model only gives a partial view of what the learner must learn. A natural extension of the algorithm given here is to ask what language-specific segmentation information can be learned using the lexicon that the learner has acquired.

English learning infants use stress as a part of their segmentation strategy starting around 9 months (Jusczyk et al., 1993; Thiessen & Saffran, 2003). The learner's lexicon inferred using the strategy given in this paper contains a large proportion (80%) of multisyllabic words with initial stress. By generalizing from the stress pattern in the inferred lexicon, the learner can apply a segmentation strategy that can use stress information. Similarly, the lexicon may be used to infer phonotactic generalizations which may be useful in scoring hypothesized segmentations or determining the probability of a novel word.

A crucial frontier in word segmentation is the expansion of evaluation to include other languages. As with many other tasks, creating solutions that perform robustly 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. This will require adding morphological complexity to the system; the techniques developed for English are unlikely to succeed unchanged in other languages.

5. Conclusion

In this paper we have demonstrated that a simple, online, algorithmic-level model of word segmentation can learn to segment words and predicts aspects of the developmental changes seen in infants as they acquire language. This work integrates with experimental and longitudinal infant studies at a deeper level than previous models and works without the use of any complicated or cognitively implausible learning mechanisms. Our focus on a solution at Marr's algorithmic level has allowed us to build a model that can align directly with infant behaviors and be used to predict developmental patterns. While computational level models have been useful in defining the outline of possible solutions, we believe that further work at the algorithmic level will allow us to reach a much deeper understanding of the cognitive mechanisms and linguistic representations involved in language acquisition.

References

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

Aslin, Richard N., Julide Z. Woodward, Nicholas P. LaMendola & Thomas G. Bever (1996). Models of word segmentation in fluent maternal speech to infants. *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 117–134.

Aslin, Richard N., Jenny R. Saffran & Elissa L. Newport (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science* 9:4, 321–324.

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

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

Borschinger, Benjamin & Mark Johnson (2011). A particle filter algorithm for Bayesian wordsegmentation. *Proceedings of the Australasian Language Technology Association Workshop 2011*, Canberra, Australia, 10–18.

Bortfeld, Heather, James L. Morgan, Roberta Michnick Golinkoff & Karen Rathbun (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science* 16:4, 298–304.

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

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

Brown, Roger (1973). *A First Language: The Early Stages*. Harvard University Press, Cambridge, Massachusetts, USA.

Clark, Ruth (1974). Performing without competence. *Journal of Child Language* 1:01, 1–10.

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

Daland, Robert & Janet B. Pierrehumbert (2011). Learning diphone-based segmentation. *Cognitive Science* 35:1, 119–155.

Donaldson, Wayne (1993). Accuracy of d-prime and a-prime as estimates of sensitivity. *Bulletin of the Psychonomic Society* 31:4, 271–274.

Endress, Ansgar D. & Jacques Mehler (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language* 60:3, 351–367.

Frank, Michael C., Sharon Goldwater, Thomas L. Griffiths & Joshua B. Tenenbaum (2010). Modeling human

performance in statistical word segmentation. *Cognition* 117:2, 107–125.
Gambell, Timothy & C. Yang (2004). Statistics, learning, and universal grammar: Modeling word segmentation. *First Workshop on Psycho-computational Models of Human Language Acquisition*, 49–52.
Goldwater, Sharon, Thomas L. Griffiths & Mark Johnson (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition* 112:1, 21–54.
Johnson, Mark & Sharon Goldwater (2009). Improving nonparameteric Bayesian inference: Experiments on unsupervised word segmentation with adaptor grammars. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, 317–325.
Jusczyk, Peter W. & Carolyn Derrah (1987). Representation of speech sounds by young infants. *Developmental Psychology* 23:5, 648–654.
Jusczyk, Peter W., Anne Cutler & Nancy J. Redanz (1993). Infants' preference for the predominant stress patterns of english words. *Child Development* 64:3, 675–687.
Koehn, Philipp & Kevin Knight (2003). Empirical methods for compound splitting. *Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, vol. 1, 187–193.
Lew-Williams, Casey, Bruna Pelucchi & Jenny R. Saffran (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science* 14:6, 1323–1329.
Lignos, Constantine (2011). Modeling infant word segmentation. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Portland, Oregon, USA, 29–38.
Lignos, Constantine & Charles Yang (2010). Recession segmentation: Simpler online word segmentation using limited resources. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, Uppsala, Sweden, USA, 88–97.
MacWhinney, Brian (1978). The acquisition of morphophonology. *Monographs of the society for research in child development* 43:1/2, 1–123.
MacWhinney, Brian (2000). *The CHILDES Project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ, US.
Marr, David (1983). *Vision: A computational approach*. Freeman & Co., San Francisco, CA, USA.
Olivier, Donald C. (1968). *Stochastic grammars and language acquisition mechanisms*. Ph.D. thesis, Harvard University.
Pearl, Lisa, Sharon Goldwater & Mark Steyvers (2011). Online learning mechanisms for Bayesian models of word segmentation. *Research on Language & Computation* 8:2–3, 107–132.
Perruchet, Pierre & Annie Vinter (1998). PARSER: A model for word segmentation. *Journal of Memory and Language* 39, 246–263.
Peters, Ann M. (1983). *The Units of Language Acquisition*. Cambridge University Press.
Phillips, Lawrence & Lisa Pearl (2012). Less is more in Bayesian word segmentation: When cognitively plausible learners outperform the ideal. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, Cognitive Science Society, Sapporo, Japan.
Pinker, Steven (1999). *Words and rules: The ingredients of language*. Basic Books, New York, NY, USA.
Rumelhart, David E. & James L. McClelland (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. MIT Press, Cambridge, MA.
Saffran, Jenny R. (2001). Words in a sea of sounds: The output of infant statistical learning. *Cognition* 81:2, 149–169.
Saffran, Jenny R., Richard N. Aslin & Elissa L. Newport (1996a). Statistical learning by 8-month-old infants. *Science* 274:5294, 1926–1928.
Saffran, Jenny R., Elissa L. Newport & Richard N. Aslin (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language* 35:4, 606–621.
Seidl, Amanda & Elizabeth K. Johnson (2006). Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science* 9:6, 565–573.
Selkirk, Elisabeth O. (1984). *Phonology and Syntax: The Relation between Sound and Structure*. MIT Press, Cambridge, MA, USA.
Shukla, Mohinish, Katherine S. White & Richard N. Aslin (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences* 108:15, 6038–6043.
Thiessen, Erik D. & Jenny 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.
Yang, Charles D. (2002). *Knowledge and learning in natural language*. Oxford University Press, New York, NY, USA.
Yang, Charles D. (2004). Universal grammar, statistics or both? *Trends in Cognitive Sciences* 8:10, 451–456.

Proceedings of the 30th West Coast Conference on Formal Linguistics

edited by Nathan Arnett
and Ryan Bennett

Cascadilla Proceedings Project Somerville, MA 2012

ISBN 978-1-57473-454-6 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Lignos, Constantine. 2012. Infant Word Segmentation: An Incremental, Integrated Model. In *Proceedings of the 30th West Coast Conference on Formal Linguistics*, ed. Nathan Arnett and Ryan Bennett, 237-247. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2821.