Recession Segmentation: Simpler Online Word Segmentation Using Limited Resources[*]

Constantine Lignos, Charles Yang
Dept. of Computer and Information Science, Dept. of Linguistics
University of Pennsylvania
lignos@cis.upenn.edu, charles.yang@ling.upenn.edu

Abstract

In this paper we present a cognitively plausible approach to word segmentation that segments in an online fashion using only local information and a lexicon of previously segmented words. Unlike popular statistical optimization techniques, the learner uses structural information of the input syllables rather than distributional cues to segment words. We develop a memory model for the learner that like a child learner does not recall previously hypothesized words perfectly. The learner attains an F-score of 86.69% in ideal conditions and 85.05% when word recall is unreliable and stress in the input is reduced. These results demonstrate the power that a simple learner can have when paired with appropriate structural constraints on its hypotheses.

1 Introduction

The problem of word segmentation presents an important challenge in language acquisition. The child learner must segment a continuous stream of sounds into words without knowing what the individual words are until the stream has been segmented. Computational models present an opportunity to test the potentially innate constraints, structures, and algorithms that a child may be using to guide her acquisition. In this work we develop a segmentation model from the constraints suggested by Yang (2004) and evaluate it in idealized conditions and conditions that better approximate the environment of a child learner. We seek to determine how these limitations in the learner's input and memory affect the learner's performance and to demonstrate that the presented learner is robust even under non-ideal conditions.

2 Related Work

Most recent work in word segmentation of child-directed speech has operated within statistical optimization frameworks, particularly Bayesian approaches (Goldwater et al., 2009; Johnson and Goldwater, 2009). These models have established the state-of-the-art for the task of selecting appropriate word boundaries from a stream of unstructured phonemes. But while these models deliver excellent performance, it is not clear how they inform the *process* of acquisition.

Trying to find cognitive insight from these types of models is difficult because of the inherent mismatch in the quality and types of hypotheses they maintain during learning. Children are incremental learners (Brown, 1973), and learners relying on statistical optimization are generally not. A child's competence grows gradually as she hears and produces more and more utterances, going through predictable changes to her working grammar (Marcus et al., 1992) that statistical optimization techniques typically do not go through and do not intend to replicate.

Statistical models provide excellent information about the features, distributional cues, and priors that can be used in learning, but provide little information about *how* a child learner can use this information and how her knowledge of language develops as the learning process evolves. Previous simulations in word segmentation using the same type of distributional information as many statistical optimization-based learners but without an optimization model suggest that statistics alone are not sufficient for learning to succeed in a computationally efficient online manner; further constraints on the search space are needed (Yang, 2004).

Previous computational models have demanded tremendous memory and computational capacity from human learners. For example, the algorithm

[*] Portions of this work were adapted from an earlier manuscript, *Word Segmentation: Quick But Not Dirty*.

Proceedings of the Fourteenth Conference on Computational Natural Language Learning, pages 88–97,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

of Brent & Cartwright (1996) produces a set of possible lexicons that describe the learning corpus, each of which is evaluated as the learner iterates until no further improvement is possible. It is unlikely that an algorithm of this type is something a human learner is capable of using given the requirement to remember at the very least a long history of recent utterances encountered and constantly reanalyze them to find a optimal segmentation. Work in this tradition makes no claims, however, that these methods are actually the ones used by human learners.

On the other hand, previous computational models often underestimate the human learner's knowledge of linguistic representations. Most of these models are "synthetic" in the sense of Brent (1999): the raw material for segmentation is a stream of segments, which are then successively grouped into larger units and eventually, conjectured words. This assumption may make the child learner's job unnecessarily hard; since syllables are hierarchical structures consisting of segments, treating the linguistic data as unstructured segment sequences makes the problem harder than it actually is. For a given utterance, there are fewer syllables than segments, and hence fewer segmentation possibilities.

Modeling the corpus using hierarchical grammars that can model the input at varying levels (word collocations, words, syllables, onsets, etc.) provide the learner the most flexibility, allowing the learner to build structure from the individual phonemes and apply distributions at each level of abstraction (Johnson and Goldwater, 2009). While this results in state-of-the-art performance for segmentation performed at the phoneme level, this approach requires significant computational resources as each additional level of representation increases the complexity of learning. In addition, it is not clear that some of the intermediate levels in such an approach, such as word level collocations which are not syntactic constituents, would have any linguistic or psychological reality to a human learner.

A number of psychologically-motivated models of word segmentation rely on the use of syllabic transitional probabilities (TPs), basing the use of TPs on experimental work in artificial language learning (Saffran et al., 1996a; Saffran et al., 1996b) and in corpus studies (Swingley, 2005). The identification of the syllable as the basic unit of segmentation is supported research in experimental psychology using infants as young as 4-days-old (Bijeljac-Babic et al., 1993), but when syllable transitional probabilities are evaluated in online learning procedures that only use local information (Yang, 2004), the results are surprisingly poor, even under the assumption that the learner has already syllabified the input perfectly. Precision is 41.6%, and recall is 23.3%, which we will show is worse than a simple baseline of assuming every syllable is a word. The below-baseline performance is unsurprising given that in order for this type of model to posit a word boundary, a transitional probability between syllables must be lower than its neighbors. This condition cannot be met if the input is a sequence of monosyllabic words for which a boundary must be postulated for every syllable; it is impossible to treat every boundary as a local minimum.

While the pseudo-words used in infant studies measuring the ability to use transitional probability information are uniformly three-syllables long, much of child-directed English consists of sequences of monosyllabic words. Corpus statistics reveal that on average a monosyllabic word is followed by another monosyllabic word 85% of time (Yang, 2004), and thus learners that use only local transitional probabilities without any global optimization are unlikely to succeed. This problem does not affect online approaches that use global information, such as computing the maximum likelihood of the corpus incrementally (Venkataraman, 2001). Since these approaches do not require each boundary be a local minimum, they are able to correctly handle a sequence of monosyllable words.

We believe that the computational modeling of psychological processes, with special attention to concrete mechanisms and quantitative evaluations, can play an important role in identifying the constraints and structures relevant to children's acquisition of language. Rather than using a prior which guides the learner to a desired distribution, we examine learning with respect to a model in which the hypothesis space is constrained by structural requirements.

In this paper we take a different approach than statistical optimization approaches by exploring how well a learner can perform while processing a corpus in an online fashion with only local information and a lexicon of previously segmented words. We present a simple, efficient approach to word segmentation that uses structural information rather than distributional cues in the input to segment words. We seek to demonstrate that even in the face of impoverished input and limited resources, a simple learner can succeed when it operates with the appropriate constraints.

3 Constraining the Learning Space

Modern machine learning research (Gold, 1967; Valiant, 1984; Vapnik, 2000) suggests that constraints on the learning space and the learning algorithm are essential for realistically efficient learning. If a domain-neutral learning model fails on a specific task where children succeed, it is likely that children are equipped with knowledge and constraints specific to the task at hand. It is important to identify such constraints to see to what extent they complement, or even replace, domain neutral learning mechanisms.

A particularly useful constraint for word segmentation, 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.

A simple example of how adult learners might use the USC is upon hearing novel names or words. Taking *Star Wars* characters as an example, it is clear that *chewb**a**cca* is one word but *d**a**rthv**a**der* cannot be as the latter bears two primary stresses.

The USC could give the learner many isolated words for free. If the learner hears an utterance that contains exactly one primary stress, it is likely it is a single word. Moreover, the segmentation for a multiple word utterance can be equally straightforward under USC. Consider a sequence $W_1 S_1 S_2 S_3 W_2$, where W stands for a weak syllable and S stands for a strong syllable. A learner equipped with USC will immediately know that the sequence consists of three words: specifically, $W_1 S_1$, S_2, and $S_2 W_2$.

The USC can also constrain the use of other learning techniques. For example, the syllable consequence $S_1 W_1 W_2 W_3 S_2$ cannot be segmented by USC alone, but it may still provide cues that facilitate the application of other segmentation strategies. For instance, the learner knows that the sequence consists of at least two words, as indicated by two strong syllables. Moreover, it also knows that in the window between S_1 and S_2 there must be one or more word boundaries.

Yang (2004) evaluates the effectiveness of the USC in conjunction with a simple approach to using transitional probabilities. The performance of the approach presented there improves dramatically if the learner is equipped with the assumption that each word can have only one primary stress. If the learner knows this, then it may limit the search for local minima to only the window between two syllables that both bear primary stress, e.g., between the two **a**'s in the sequence *l**a**nguage**a**cquisition.* This assumption is plausible given that 7.5-month-old infants are sensitive to strong/weak prosodic distinctions (Jusczyk, 1999). Yang's stress-delimited algorithm achieves the precision of 73.5% and recall of 71.2%, a significant improvement over using TPs alone, but still below the baseline presented in our results.

The improvement of the transitional probability-based approach when provided with a simple linguistic constraint suggests that structural constraints can be powerful in narrowing the hypothesis space so that even sparse, local information can prove useful and simple segmentation strategies can become more effective.

It should be noted that the classification of every syllable as "weak" or "strong" is a significant simplification. Stress is better organized into hierarchical patterns constructed on top of syllables that vary in relative prominence based on the domain of each level of the hierarchy, and generally languages avoid adjacent strong syllables (Liberman and Prince, 1977). We later discuss a manipulation of the corpus used by Yang (2004) to address this concern.

Additionally, there are significant challenges in reconstructing stress from an acoustic signal (Van Kuijk and Boves, 1999). For a child learner to use the algorithm presented here, she would need to have mechanisms for detecting stress in the speech signal and categorizing the gradient stress in utterances into a discrete level for each syllable. These mechanisms are not addressed in this work; our focus is on an algorithm that can succeed given discrete stress information for each syllable. Given the evidence that infants can distinguish weak and strong syllables and use that information to detect word boundaries (Jusczyk et al., 1999), we believe that it is reasonable to assume that identifying syllabic stress is a task an infant learner can perform at the developmental stage of word segmentation.

4 A Simple Algorithm for Word Segmentation

We now present a simple algebraic approach to word segmentation based on the constraints suggested by Yang (2004). The learner we present is algebraic in that it has a lexicon which stores previously segmented words and identifies the input as a combination of words already in the lexicon and novel words. No transitional probabilities or any distributional data are calculated from the input. The learner operates in an online fashion, segmenting each utterance in a primarily left-to-right fashion and updating its lexicon as it segments.

The USC is used in two ways by the learner. First, if the current syllable has primary stress and the next syllable also has primary stress, a word boundary is placed between the current and next syllable. Second, whenever the algorithm is faced with the choice of accepting a novel word into the lexicon and outputting it as a word, the learner "abstains" from doing so if the word violates USC, that is if it contains more than one primary stress. Since not all words are stressed, if a word contains no primary stresses it is considered an acceptable word; only a word with *more that one* primary stress is prohibited. If a sequence of syllables has more than one primary stress and cannot be segmented further, the learner does not include that sequence in its segmentation of the utterance and does not add it to the lexicon as it cannot be a valid word.

The algorithm is as follows, with each step explained in further detail in the following paragraphs.

For each utterance in the corpus, do the following:

1. As each syllable is encountered, use *Initial Subtraction* and *USC Segmentation* to segment words from the beginning of the utterance if possible.

2. If unsegmented syllables still remain, apply *Final Subtraction*, segmenting words iteratively from the end of the utterance if possible.

3. If unsegmented syllables still remain, if those syllables constitute a valid word under the USC, segment them as a single word and add them to the lexicon. Otherwise, abstain, and do not include these syllables in the segmentation of the sentence and do not add them to the lexicon.

Initial Subtraction. If the syllables of the utterance from the last segmentation (or the start of the utterance) up to this point matches a word in the lexicon but adding one more syllable would result in it not being a known word, segment off the recognized word and increase its frequency. This iteratively segments the longest prefix word from the utterance.

USC Segmentation. If the current and next syllables have primary stress, place a word boundary after the current syllable, treating all syllables from the last segmentation point up to and and including the current syllable as a potential word. If these syllables form a valid word under the USC, segment them as a word and add them to the lexicon. Otherwise, abstain, not including these syllables in the segmentation of the sentence and not adding them to the lexicon.

Final Subtraction. After initial subtraction and USC Segmentation have been maximally applied to the utterance, the learner is often left with a sequence of syllables that is not prefixed by any known word and does not have any adjacent primary stresses. In this situation the learner works from right to left on the remaining utterance, iteratively removing words from the end of the utterance if possible. Similar to the approach used in Initial Subtraction, the longest word that is a suffix word of the remaining syllables is segmented off, and this is repeated until the entire utterance is segmented or syllables remain that are not suffixed by any known word.

The ability to abstain is a significant difference between this learner and most recent work on this task. Because the learner has a structural description for a word, the USC, it is able to reject any hypothesized words that do not meet the description. This improves the learner's precision and recall because it reduces the number of incorrect predictions the learner makes. The USC also allows the learner keep impossible words out of its lexicon.

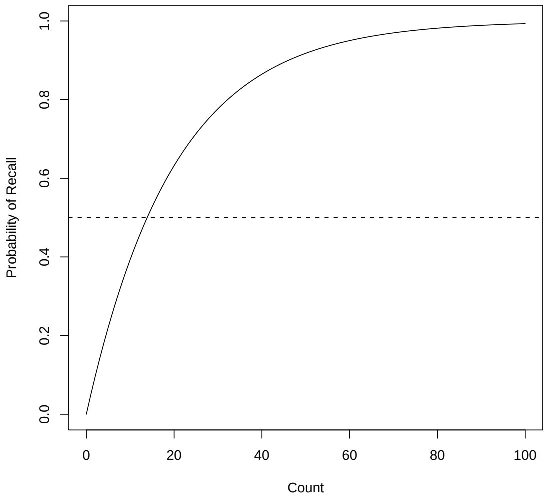

Figure 1: The selected probabilistic memory function for $\alpha = 0.05$. The dashed line at 0.05 represents the threshold above which a word is more likely than not to be recalled, occurring at a count of approximately 14.

5 A Probabilistic Lexicon

To simulate the imperfect memory of a child learner, we use a simple exponential function to generate the probability with which a word is retrieved from the lexicon:

$$p_r(word) = 1.0 - e^{-\alpha c(word)}$$

$p_r(word)$ is the probability of a $word$ being retrieved, α is a constant, and $c(word)$ is the number of times the word has been identified in segmentations thus far. This type of memory function is a simplified representation of models of humans' memory recall capabilities (Anderson et al., 1998; Gillund and Shiffrin, 1984). This memory function for the value of $\alpha = 0.05$, the value used in our experiments, is given in Figure 1. We later show that the choice of α has little impact on the learner's segmentation performance, and thus the more or less arbitrary selection of a value for α is of little consequence.

When the algorithm attempts to subtract words from the beginning or end of an utterance, it may miss words in the lexicon due to this probabilistic retrieval. The learner only has one opportunity to recall a word in a given utterance. For example, in the utterance *P.EH1.N S.AH0.L* (*pencil*), if the learner has *P.EH1.N* and *P.EH1.N S.AH0.L* in its lexicon but *P.EH1.N* is more frequent, it may fail to recall *P.EH1.N S.AH0.L* when examining the second syllable but succeed in recognizing *P.EH1.N* in the first. Thus it will break off *P.EH1.N* instead of *P.EH1.N S.AH0.L.* This means the learner may fail to reliably break off the longest words, instead breaking off the longest word that is successfully recalled.

While probabilistic memory means that the learner will fail to recognize words it has seen before, potentially decreasing recall, it also provides the learner the benefit of probabilistically failing to repeat previous mistakes if they occur rarely.

Probabilistic word recall results in a "rich get richer" phenomenon as the learner segments; words that are used more often in segmentations are more likely to be reused in later segmentations. While recent work from Bayesian approaches has used a Dirichlet Process to generate these distributions (Goldwater et al., 2006), in this learner the reuse of frequent items in learning is a result of the memory model rather than an explicit process of reusing old outcomes or generating new ones. This growth is an inherent property of the cognitive model of memory used here rather than an externally imposed computational technique.

6 Evaluation

Our computational model is designed to process child-directed speech. The corpus we use to evaluate it is the same corpus used by Yang (2004). Adult utterances were extracted from the Brown (1973) data in the CHILDES corpus (MacWhinney, 2000), consisting of three children's data: Adam, Eve, and Sarah. We obtained the phonetic transcriptions of words from the Carnegie Mellon Pronouncing Dictionary (CMUdict) Version 0.6 (Weide, 1998), using the first pronunciation of each word. In CMUdict, lexical stress information is preserved by numbers: 0 for unstressed, 1 for primary stress, 2 for secondary stress. For instance, *cat* is represented as *K.AE1.T, catalog* is *K.AE1.T.AH0.L.AO0.G,* and *catapult* is *K.AE1.T.AH0.P.AH2.L.T.* We treat primary stress as "strong" and secondary or unstressed syllables as "weak."

For each word, the phonetic segments were grouped into syllables. This process is straightforward by the use of the principle "Maximize Onset," which maximizes the length of the onset as long as it is valid consonant cluster of English, i.e., it conforms to the phonotactic constraints of English. For example, *Einstein* is *AY1.N.S.T.AY0.N* as segments and parsed into *AY1.N S.T.AY0.N* as syllables: this is because */st/* is the longest valid onset for the second syllable containing *AY0* while */nst/* is longer but violates English phonotactics. While we performed syllabification as a preprocessing step outside of learning, a child learner would presumably learn the required phonotactics as a part of learning to segment words. 9-month old infants are believed to have learned some phonotactic constraints of their native language (Mattys and Jusczyk, 2001), and learning these constraints can be done with only minimal exposure (Onishi et al., 2002).

Finally, spaces and punctuation between words were removed, but the boundaries between utterances–as indicated by line breaks in CHILDES–are retained. Altogether, there are 226,178 words, consisting of 263,660 syllables. The learning material is a list of unsegmented syllable sequences grouped into utterances, and the learner's task is to find word boundaries that group substrings of syllables together, building a lexicon of words as it segments.

We evaluated the learner's performance to address these questions:

- How does probabilistic memory affect learner performance?
- How much does degrading stress information relied on by USC segmentation reduce performance?
- What is the interaction between the probabilistic lexicon and non-idealized stress information?

To evaluate the learner, we tested configurations that used a probabilistic lexicon and ones with perfect memory in two scenarios: Dictionary Stress, and Reduced Stress. We create the Reduced Stress condition in order to simulate that stress is often reduced in casual speech, and that language-specific stress rules may cause reductions or shifts in stress that prevent two strong syllables from occurring in sequence. The difference between the scenarios is defined as follows:

Dictionary Stress. The stress information is given to the learner as it was looked up in CMUdict. For example, the first utterance from the Adam corpus would be *B.IH1.G D.R.AH1.M* (*big drum*), an utterance with two stressed monosyllables (SS). In most languages, however, conditions where two stressed syllables are in sequence are handled by reducing the stress of one syllable. This is simulated in the reduced stress condition.

Reduced Stress. The stress information obtained from CMUdict is post-processed in the context of each utterance. For any two adjacent primary stressed syllables, the first syllable is reduced from a strong syllable to a weak one. This is applied iteratively from left to right, so for any sequence of n adjacent primary-stress syllables, only the nth syllable retains primary stress; all others are reduced. This removes the most valuable clue as to where utterances can be segmented, as USC segmentation no longer applies. This simulates the stress retraction effect found in real speech, which tries to avoid adjacent primary stresses.

Learners that use probabilistic memory were allowed to iterate over the input two times with access to the lexicon developed over previous iterations but no access to previous segmentations. This simulates a child hearing many of the same words and utterances again, and reduces the effect of the small corpus size used on the learning process. Because the probabilistic memory reduces the algorithm's ability to build a lexicon, performance in a single iteration is lower than perfect memory conditions. In all other conditions, the learner is allowed only a single pass over the corpus.

The precision and recall metrics are calculated for the segmentation that the learner outputs and the lexicon itself. For an utterance, each word in the learner's segmentation that also appears in the gold standard segmentation is counted as correct, and each word in the learner's segmentation not present in the gold standard segmentation is a false alarm. F-score is computed using equally balanced precision and recall (F_0). The correct words, false words, and number of words in the gold standard are summed over the output in each iteration to produce performance measures for that iteration.

Precision, recall, and F-score are similarly computed for the lexicon; every word in the learner's lexicon present in the gold standard is counted as correct, and every word in the learner's lexicon not present in the gold standard is a false alarm. These computations are performed over word types in the lexicon, thus all words in the lexicon are of equal weight in computing performance regardless of their frequency. In the probabilistic memory conditions, however, the memory function defines the probability of each word being recalled (and thus being considered a part of the lexicon) at evaluation time.

In addition to evaluating the learner, we also implemented three baseline approaches to compare the learner against. The *Utterance* baseline segmenter assumes every utterance is a single word. The *Monosyllabic* baseline segmenter assumes every syllable is a single word. The *USC* segmenter inserts word boundaries between all adjacent syllables with primary stress in the corpus.

6.1 Results

The performance of the learner and baseline segmenters is given in Table 1. While the Utterance segmenter provides expectedly poor performance, the Monosyllabic segmenter sets a relatively high baseline for the task. Because of the impoverished morphology of English and the short words that tend to be used in child-directed speech, assuming each syllable is a word proves to be an excellent heuristic. It is unlikely that this heuristic will perform as well in other languages. Because the USC segmenter only creates segmentation points where there are words of adjacent primary stress, it is prone to attaching unstressed monosyllabic function words to content words, causing very low lexicon precision (13.56%).

With both perfect memory and dictionary stress information, the learner attains an F-score of 86.69%, with precision (83.78%) lower than recall (89.81%). First, we consider the effects of probabilistic memory on the learner. In the Dictionary Stress condition, using probabilistic memory decreases F_o by 1.15%, a relatively small impact given that with the setting of $\alpha = 0.05$ the learner must use a word approximately 14 times before it can retrieve it with 50% reliability and 45 times before it can retrieve it with 90% reliability. In the first iteration over the data set, 17.87% of lexicon lookups for words that have been hypothesized before fail. The impact on F_0 is caused by a drop in recall, as would be expected for a such a memory model.

To examine the effect of the α parameter for probabilistic memory on learner performance, we plot the utterance and lexicon F_0 after the learner iterates over the corpus once in the Probabilistic Memory, Dictionary Stress condition. As Figure 2 shows, the choice of α has little effect on the utterance F_0 through most of a broad range from 0.01 to 0.9. Because the setting of α determines the number of times a word must be hypothesized before it can reliably be recalled, it expectedly has a significant effect on lexicon F_0. The selection of $\alpha = 0.05$ for our experiments is thus unlikely to have had any significant bearing on the utterance segmentation performance, although for lower values of α precision is favored while for larger values recall is favored. Larger values of α imply the learner is able to recall items after fewer exposures. While a larger value of α would have yielded higher performance in lexicon performance, it also assumes much more about the learner's memory capabilities.

The Reduced Stress condition also has only a

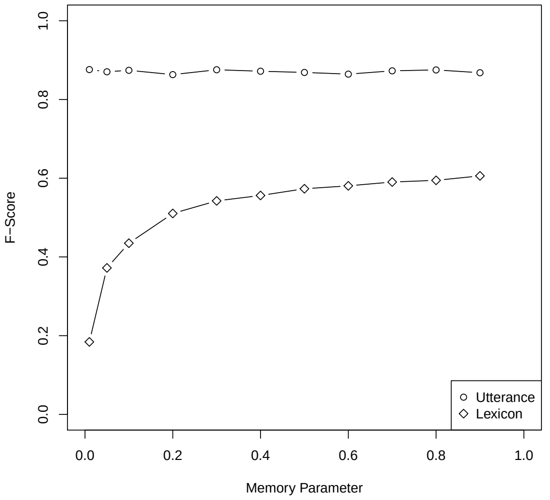

Figure 2: Learner utterance and lexicon F-scores after two iterations when α is varied in the Probabilistic Memory, Dictionary Stress condition

	Perfect Memory, Dictionary Stress	Perfect Memory, Reduced Stress
USC Seg.	114,333	0
Initial Sub.	65,800	164,989
Final Sub.	5,690	14,813
Total	185,823	179,802

Table 2: Number of segmentations performed by each operation: USC Segmentation, Initial Subtraction, and Final Subtraction.

	Utterances			**Lexicon**		
Segmenter	Precision	Recall	F_0	Precision	Recall	F_0
Utterance	18.61%	4.67%	7.47%	3.57%	30.35%	6.39%
Monosyllabic	73.29%	85.44%	78.90%	55.41%	43.88%	48.97%
USC	81.06%	61.52%	69.95%	13.56%	66.97%	22.55%
Perfect Memory, Dictionary Stress	83.78%	89.81%	86.69%	67.72%	58.60%	62.83%
Perfect Memory, Reduced Stress	82.32%	85.81%	84.03%	39.18%	50.08%	43.97%
Prob. Memory, Dictionary Stress	84.05%	87.07%	85.54%	72.34%	30.01%	42.42%
Prob. Memory, Reduced Stress	84.85%	85.24%	85.05%	41.13%	22.91%	29.43%

Table 1: Baseline and Learner Performance. Performance is reported after two iterations over the corpus for probabilistic memory learners and after a single iteration for all other learners.

small impact on utterance segmentation performance. This suggests that the USC's primary value to the learner is in constraining the contents of the lexicon and identifying words in isolation as good candidates for the lexicon. In the Reduced Stress condition where the USC is not directly responsible for any segmentations as there are no adjacent primary-stressed syllables, the learner relies much more heavily on subtractive techniques. Table 2 gives the number of segmentations performed using each segmentation operation. The total number of segmentations is very similar between the Dictionary and Reduced Stress conditions, but because USC Segmentation is not effective on Reduced Stress input, Initial and Final Subtraction are used much more heavily.

7 Discussion

The design of the segmenter presented here suggests that both the quality of memory and the structural purity of the input would be critical factors in the learner's success. Our results suggest, however, that using probabilistic memory and a less idealized version of stress in natural language have little impact on the performance of the presented learner. They do cause the learner to learn much more slowly, causing the learner to need to be presented with more material and resulting in worse performance in the lexicon evaluation. But this slower learning is unlikely to be a concern for a child learner who would be exposed to much larger amounts of data than the corpora here provide.

Cognitive literature suggests that limited memory during learning may be essential to a learner in its early stages (Elman, 1993). But we do not see any notable improvement as a result of the probabilistic memory model used in our experiments, although the learner does do better in the Reduced Stress condition with Probabilistic Memory than Perfect Memory. This should not be interpreted as a negative result as we only analyze a single learner and memory model. Adding decay to the model such that among words of equal frequency those that have not been used in segmentation recently are less likely to be remembered may be sufficient to create the desired effect.

The success of this learner suggests that the type of "bootstrapping" approaches can succeed in word segmentation. The learner presented uses USC to identify utterances that are likely to be lone words, seeding the lexicon with initial information. Even if these first items in the lexicon are of relatively low purity, often combining function words and content words into one, the learner is able to expand its lexicon by using these hypothesized words to segment new input. As the learner segments more, these hypotheses become more reliable, allowing the learner to build a lexicon of good quality.

The subtraction approaches presented in this work provide a basic algorithm for to handling segmentation of incoming data in an online fashion. The subtractive heuristics used here are of course not guaranteed to result in a perfect segmentation even with a perfect lexicon; they are presented to show how a simple model of processing incoming data can be paired with structural constraints on the hypothesis space to learn word segmentation in a computationally efficient and cognitively plausible online fashion.

8 Conclusions

The learner's strong performance using minimal computational resources and unreliable memory suggest that simple learners can succeed in unsupervised tasks as long as they take advantage of domain-specific knowledge to constrain the hypothesis space. Our results show that, even in adversarial conditions, structural constraints remain powerful tools for simple learning algorithms in difficult tasks.

Future work in this area should focus on learners that can take advantage of the benefits of a probabilistic lexicon and memory models suited to them. Also, a more complex model of the type of stress variation present in natural speech would help better determine a learner that uses USC's ability to handle realistic variation in the input. Our model of stress reduction is a worst-case scenario for USC segmentation but is unlikely to be an accurate model of real speech. Future work should adopt a more naturalistic model to determine whether the robustness found in our results holds true in more realistic stress permutations.

Acknowledgements

We thank Kyle Gorman, Josef Fruehwald, and Dan Swingley for their helpful discussions regarding this work. We are grateful to and Mitch Marcus and Jana Beck for their feedback on earlier versions of this paper.

References

J.R. Anderson, D. Bothell, C. Lebiere, and M. Matessa. 1998. An integrated theory of list memory. *Journal of Memory and Language*, 38(4):341–380.

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

M.R. Brent and T.A. Cartwright. 1996. Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61(1-2):93–125.

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.

J.L. Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.

G. Gillund and R.M. Shiffrin. 1984. A retrieval model for both recognition and recall. *Psychological Review*, 91(1):1–67.

E.M. Gold. 1967. Language identification in the limit. *Information and control*, 10(5):447–474.

S. Goldwater, T. Griffiths, and M. Johnson. 2006. Interpolating between types and tokens by estimating power-law generators. *Advances in Neural Information Processing Systems*, 18:459.

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.

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.

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.W. Jusczyk. 1999. How infants begin to extract words from speech. *Trends in Cognitive Sciences*, 3(9):323–328.

M. Liberman and A. Prince. 1977. On stress and linguistic rhythm. *Linguistic Inquiry*, 8(2):249–336.

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

G.F. Marcus, S. Pinker, M. Ullman, M. Hollander, T.J. Rosen, F. Xu, and H. Clahsen. 1992. Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, 57(4).

S.L. Mattys and P.W. Jusczyk. 2001. Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78(2):91–121.

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

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.

D. Swingley. 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50(1):86–132.

LG Valiant. 1984. A theory of the learnable. *Communications of the ACM*, 27(11):1142.

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

V.N. Vapnik. 2000. *The nature of statistical learning theory*. Springer.

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