Word Segmentation as General Chunking

Daniel Hewlett and **Paul Cohen**
Department of Computer Science
University of Arizona
Tucson, AZ 85721
{dhewlett,cohen}@cs.arizona.edu

Abstract

During language acquisition, children learn to segment speech into phonemes, syllables, morphemes, and words. We examine word segmentation specifically, and explore the possibility that children might have general-purpose chunking mechanisms to perform word segmentation. The Voting Experts (VE) and Bootstrapped Voting Experts (BVE) algorithms serve as computational models of this chunking ability. VE finds chunks by searching for a particular information-theoretic signature: low internal entropy and high boundary entropy. BVE adds to VE the ability to incorporate information about word boundaries previously found by the algorithm into future segmentations. We evaluate the general chunking model on phonemically-encoded corpora of child-directed speech, and show that it is consistent with empirical results in the developmental literature. We argue that it offers a parsimonious alternative to special-purpose linguistic models.

1 Introduction

The ability to extract words from fluent speech appears as early as the seventh month in human development (Jusczyk et al., 1999). Models of this ability have emerged from such diverse fields as linguistics, psychology and computer science. Many of these models make unrealistic assumptions about child language learning, or rely on supervision, or are specific to speech or language. Here we present an alternative: a general unsupervised model of chunking that performs very well on word segmentation tasks. We will examine the Voting Experts, Bootstrapped Voting Experts, and Phoneme to Morpheme algorithms in Section 2. Each searches for a general, information-theoretic signature of chunks. Each can operate in either a fully unsupervised setting, where the input is a single continuous sequence of phonemes, or a semi-supervised setting, where the input is a sequence of sentences. In Section 4, we evaluate these general chunking methods on phonetically-encoded corpora of child-directed speech, and compare them to a representative set of computational models of early word segmentation. Section 4.4 presents evidence that words optimize the information-theoretic signature of chunks. Section 5 discusses segmentation methods in light of what is known about the segmentation abilities of children.

2 General Chunking

The Voting Experts algorithm (Cohen and Adams, 2001) defines the *chunk* operationally as a sequence with the property that elements within the sequence predict one another but do not predict elements outside the sequence. In information-theoretic terms, chunks have low entropy internally and high entropy at their boundaries. Voting Experts (VE) is a local, greedy algorithm that works by sliding a relatively small window along a relatively long input sequence, calculating the internal and boundary entropies of sequences within the window.

The name *Voting Experts* refers to the two "experts" that vote on possible boundary locations: One expert votes to place boundaries after sequences that have low internal entropy (also called *surprisal*), given by $H_I(seq) = -\log P(seq)$. The other places votes after sequences that have

Proceedings of the Fifteenth Conference on Computational Natural Language Learning, pages 39–47,
Portland, Oregon, USA, 23–24 June 2011.

high *branching entropy*, given by $H_B(seq) = -\sum_{c \in S} P(c|seq) \log P(c|seq)$, where S is the set of successors to seq. In a modified version of VE, a third expert "looks backward" and computes the branching entropy at locations before, rather than after, seq.

The statistics required to calculate H_I and H_B are stored efficiently using an n-gram trie, which is typically constructed in a single pass over the corpus. The trie depth is 1 greater than the size of the sliding window. Importantly, all statistics in the trie are normalized so as to be expressed in standard deviation units. This allows statistics from sequences of different lengths to be compared.

The sliding window is then passed over the corpus, and each expert votes once per window for the boundary location that best matches that expert's criteria. After voting is complete, the algorithm yields an array of vote counts, each element of which is the number of times some expert voted to segment at that location. The result of voting on the string `thisisacat` could be represented in the following way, where the number between each letter is the number of votes that location received, as in `t0h0i1s3i1s4a4c1a0t`.

With the final vote totals in place, the boundaries are placed at locations where the number of votes exceeds a chosen threshold. For further details of the Voting Experts algorithm see Cohen et al. (2007), and also Miller and Stoytchev (2008).

2.1 Generality of the Chunk Signature

The information-theoretic properties of chunks upon which VE depends are present in every non-random sequence, of which sequences of speech sounds are only one example. Cohen et al. (2007) explored word segmentation in a variety of languages, as well as segmenting sequences of robot actions. Hewlett and Cohen (2010) demonstrated high performance for a version of VE that segmented sequences of utterances between a human teacher and an AI student. Miller and Stoytchev (2008) applied VE in a kind of bootstrapping procedure to perform a vision task similar to OCR: first to chunk columns of pixels into letters, then to chunk sequences of these discovered letters into words. Of particular relevance to the present discussion are the results of Miller et al. (2009), who showed that VE was able to segment a continuous audio speech stream into phonemes. The input in that experiment was generated to mimic the input presented to infants by Saffran et al. (1996), and was discretized for VE with a Self-Organizing Map (Kohonen, 1988).

2.2 Similar Chunk Signatures

Harris (1955) noticed that if one proceeds incrementally through a sequence of letters and asks speakers of the language to list all the letters that could appear next in the sequence (today called the *successor count*), the points where the number *increases* often correspond to morpheme boundaries. Tanaka-Ishii and Jin (2006) correctly recognized that this idea was an early version of branching entropy, one of the experts in VE, and they developed an algorithm called Phoneme to Morpheme (PtM) around it. PtM calculates branching entropy in both directions, but it does not use internal entropy, as VE does. It detects change-points in the absolute branching entropy rather than local maxima in the standardized entropy. PtM achieved scores similar to those of VE on word segmentation in phonetically-encoded English and Chinese.

Within the morphology domain, Johnson and Martin's HubMorph algorithm (2003) constructs a trie from a set of words, and then converts it into a DFA by the process of minimization. HubMorph searches for *stretched hubs* in this DFA, which are sequences of states in the DFA that have a low branching factor internally, and high branching factor at the edges (shown in Figure 1). This is a nearly identical chunk signature to that of VE, only with successor/predecessor count approximating branching entropy. The generality of this idea was not lost on Johnson and Martin, either: Speaking with respect to the morphology problem, Johnson and Martin close by saying "We believe that hub-automata will be the basis of a general solution for Indo-European languages as well as for Inuktitut." [1]

2.3 Chunking and Bootstrapping

Bootstrapped Voting Experts (BVE) is an extension to VE that incorporates knowledge gained from prior segmentation attempts when segmenting new input, a process known as *bootstrapping*. This

[1] Inuktitut is a polysynthetic Inuit language known for its highly complex morphology.

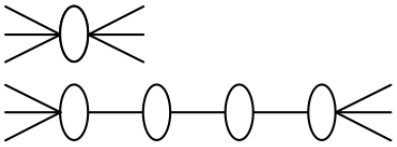

Figure 1: The DFA signature of a *hub* (top) and *stretched hub* in the HubMorph algorithm. Figure from Johnson and Martin (2003).

knowledge does not consist in the memorization of whole words (chunks), but rather in statistics describing the beginnings and endings of chunks. In the word segmentation domain, these statistics effectively correspond to phonotactic constraints that are inferred from hypothesized segmentations. Inferred boundaries are stored in a data structure called a *knowledge trie* (shown in Figure 2), which is essentially a generalized prefix or suffix trie.

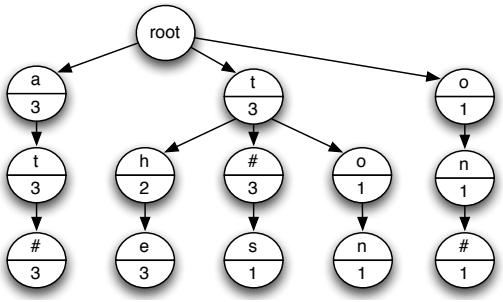

Figure 2: A portion of the knowledge trie built from `#the#cat#sat#on#the#mat#`. Numbers within each node are frequency counts.

BVE was tested on a phonemically-encoded corpus of child-directed speech and achieved a higher level of performance than any other unsupervised algorithm (Hewlett and Cohen, 2009). We reproduce these results in Section 4.

3 Computational Models of Word Segmentation

While many algorithms exist for solving the word segmentation problem, few have been proposed specifically as computational models of word segmentation in language acquisition. One of the most widely cited is MBDP-1 (Model-Based Dynamic Programming) by Brent (1999). Brent describes three features that an algorithm should have to qualify as an algorithm that "children could use for segmentation and word discovery during language acquisition." Algorithms should learn in a completely unsupervised fashion, should segment incrementally (i.e., segment each utterance before considering the next one), and should not have any built-in knowledge about specific natural languages (Brent, 1999).

However, the word segmentation paradigm Brent describes as "completely unsupervised" is actually *semi-supervised*, because the boundaries at the beginning and end of each utterance are known to be true boundaries. A fully unsupervised paradigm would include no boundary information at all, meaning that the input is, or is treated as, a continuous sequences of phonemes. The MBDP-1 algorithm was not designed for operation in this continuous condition, as it relies on having at least some true boundary information to generalize.

MBDP-1 achieves a robust form of bootstrapping through the use of Bayesian maximum-likelihood estimation of the parameters of a language model. More recent algorithms in the same tradition, including the refined MBDP-1 of Venkataraman (2001), the WordEnds algorithm of Fleck (2008), and the Hierarchical Dirichlet Process (HDP) algorithm of Goldwater (2007), share this limitation. However, infants are able to discover words in a single stream of continuous speech, as shown by the seminal series of studies by Saffran et al. (1996; 1998; 2003). In these studies, Saffran et al. show that both adults and 8-month-old infants quickly learn to extract words of a simple artificial language from a continuous speech stream containing no pauses.

The general chunking algorithms VE, BVE, and PtM work in either condition. The unsupervised, continuous condition is the norm (Cohen et al., 2007; Hewlett and Cohen, 2009; Tanaka-Ishii and Jin, 2006) but these algorithms are easily adapted to the semi-supervised, incremental condition. Recall that these methods make one pass over the entire corpus to gather statistics, and then make a second pass to segment the corpus, thus violating Brent's requirement of incremental segmentation. To adhere to the incremental requirement, the algorithms simply must segment each sentence as it is seen, and then update their trie(s) with statistics from that sentence. While VE and PtM have no natural way to store true boundary information, and so cannot ben-

efit from the supervision inherent in the incremental paradigm, BVE has the knowledge trie which serves exactly this purpose. In the incremental paradigm, BVE simply adds each segmented sentence to the knowledge trie, which will inform the segmentation of future sentences. This way it learns from its own decisions as well as the ground truth boundaries surrounding each utterance, much like MBDP-1 does. BVE and VE were first tested in the incremental paradigm by Hewlett and Cohen (2009), though only on sentences from a literary corpus, George Orwell's *1984*.

4 Evaluation of Computational Models

In this section, we evaluate the general chunking algorithms VE, BVE, and PtM in both the continuous, unsupervised paradigm of Saffran et al. (1996) and the incremental, semi-supervised paradigm assumed by bootstrapping algorithms like MBDP-1. We briefly describe the artificial input used by Saffran et al., and then turn to the broader problem of word segmentation in natural languages by evaluating against corpora drawn from the CHILDES database (MacWhinney and Snow, 1985).

We evaluate segmentation quality at two levels: boundaries and words. At the boundary level, we compute the Boundary Precision (BP), which is simply the percentage of induced boundaries that were correct, and Boundary Recall (BR), which is the percentage of true boundaries that were recovered by the algorithm. These measures are commonly combined into a single metric, the Boundary F-score (BF), which is the harmonic mean of BP and BR: $\text{BF} = (2 \times \text{BP} \times \text{BR})/(\text{BP} + \text{BR})$. Generally, higher BF scores correlate with finding correct chunks more frequently, but for completeness we also compute the Word Precision (WP), which is the percentage of induced words that were correct, and the Word Recall (WR), which is the percentage of true words that were recovered exactly by the algorithm. These measures can naturally be combined into a single F-score, the Word F-score (WF): $\text{WF} = (2 \times \text{WP} \times \text{WR})/(\text{WP} + \text{WR})$.

4.1 Artificial Language Results

To simulate the input children heard during Saffran et al.'s 1996 experiment, we generated a corpus of 400 words, each chosen from the four artificial words from that experiment (`dapiku`, `tilado`, `burobi`, and `pagotu`). As in the original study, the only condition imposed on the random sequence was that no word would appear twice in succession. VE, BVE, and PtM all achieve a boundary F-score of 1.0 whether the input is syllabified or considered simply as a stream of phonemes, suggesting that a child equipped with a chunking ability similar to VE could succeed even without syllabification.

4.2 CHILDES: Phonemes

To evaluate these algorithms on data that is closer to the language children hear, we used corpora of child-directed speech taken from the CHILDES database (MacWhinney and Snow, 1985). Two corpora have been examined repeatedly in prior studies: the Bernstein Ratner corpus (Bernstein Ratner, 1987), abbreviated BR87, used by Brent (1999), Venkataraman (2001), Fleck (2008), and Goldwater et al. (2009), and the Brown corpus (Brown, 1973), used by Gambell and Yang (2006).

Before segmentation, all corpora were encoded into a phonemic representation, to better simulate the segmentation problem facing children. The BR87 corpus has a traditional phonemic encoding created by Brent (1999), which facilitates comparison with other published results. Otherwise, the corpora are translated into a phonemic representation using the CMU Pronouncing Dictionary, with unknown words discarded.

The BR87 corpus consists of speech from nine different mothers to their children, who had an average age of 18 months (Brent, 1999). BR87 consists of 9790 utterances, with a total of 36441 words, yielding an average of 3.72 words per utterance. We evaluate word segmentation models against BR87 in two different paradigms, the incremental paradigm discussed above and an unconstrained paradigm. Many of the results in the literature do not constrain the number of times algorithms can process the corpus, meaning that algorithms generally process the entire corpus once to gather statistics, and then at least one more time to actually segment it. Results of VE and other algorithms in this unconstrained setting are presented below in Table 1. In this test, the general chunking algorithms were given one continuous corpus with no boundaries, while the results for

bootstrapping algorithms were reported in a semi-supervised condition.

Algorithm	BP	BR	BF	WP	WR	WF
PtM	0.861	**0.897**	0.879	0.676	0.704	0.690
VE	0.875	0.803	0.838	0.614	0.563	0.587
BVE	**0.949**	0.879	**0.913**	**0.793**	**0.734**	**0.762**
MBDP-1	0.803	0.843	0.823	0.670	0.694	0.682
HDP	0.903	0.808	0.852	0.752	0.696	0.723
WordEnds	0.946	0.737	0.829	NR	NR	0.707

Table 1: Results for the BR87 corpus with unconstrained processing of the corpus. Algorithms in italics are semi-supervised.

In the incremental setting, the corpus is treated as a series of utterances and the algorithm must segment each one before moving on to the next. This is designed to better simulate the learning process, as a child would normally listen to a series of utterances produced by adults, analyzing each one in turn. To perform this test, we used the incremental versions of PtM, VE, and BVE described in Section 3, and compared them with MBDP-1 on the BR87 corpus. Each point in Figure 3 shows the boundary F-score of each algorithm on the last 500 utterances. Note that VE and PtM do not benefit from the information about boundaries at the beginnings and endings of utterances, yet they achieve levels of performance not very inferior to MBDP-1 and BVE, which do leverage true boundary information.

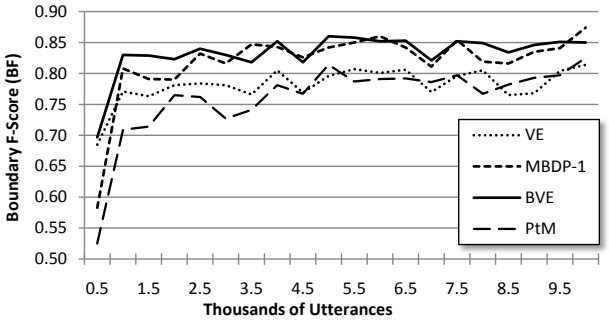

Figure 3: Results for three chunking algorithms and MBDP-1 on BR87 in the incremental paradigm.

We also produced a phonemic encoding of the BR87 and Bloom73 (Bloom, 1973) corpora from CHILDES with the CMU pronouncing dictionary, which encodes stress information (primary, secondary, or unstressed) on phonemes that serve as syllable nuclei. Stress information is known to be a useful factor in word segmentation, and infants appear to be sensitive to stress patterns by as early as 8 months of age (Jusczyk et al., 1999). Results with these corpora are shown below in Figures 4 and 5. For each of the general chunking algorithms, a window size of 4 was used, meaning decisions were made in a highly local manner. Even so, BVE outperforms MBDP-1 in this arguably more realistic setting, while VE and PtM rival it or even surpass it. Note that the quite different results shown in Figure 3 and Figure 4 are for the same corpus, under two different phonemic encodings, illustrating the importance of accurately representing the input children receive.

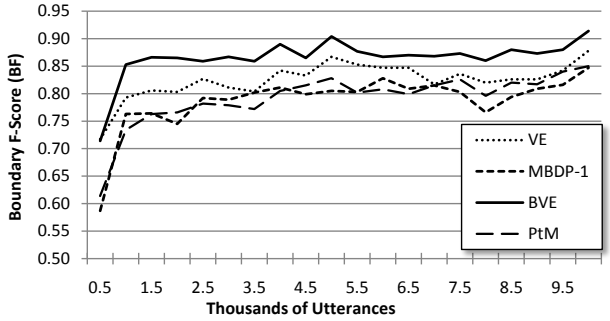

Figure 4: Results for chunking algorithms and MBDP-1 on BR87 (CMU) in the incremental paradigm.

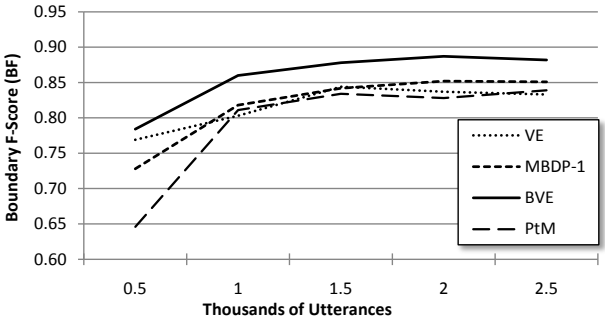

Figure 5: Results for chunking algorithms and MBDP-1 on Bloom73 (CMU) in the incremental paradigm.

4.3 CHILDES: Syllables

In many empirical studies of word segmentation in children, especially after Saffran et al. (1996), the problem is treated as though syllables were the basic units of the stream to be segmented, rather than phonemes. If we assume children can syllabify their

phonemic representation, and that word boundaries only occur at syllable boundaries, then word segmentation becomes a very different, and potentially much easier, problem. This must be the case, as the process of syllabification removes a high percentage of the potential boundary locations, and all of the locations it removes would be incorrect choices. Table 2 supports this argument. In the CHILDES corpora examined here, over 85% of the words directed to the child are monosyllabic. This means that the trivial All-Locations baseline, which segments at every possible location, achieves an F-measure of 0.913 when working with syllabic input, compared to only 0.524 for phonemic input.

Gambell and Yang (2006) present an algorithm for word segmentation that achieves a boundary F-score of 0.946 on correctly syllabified input. In order to achieve this level of performance, Gambell and Yang use a form of bootstrapping combined with a rule called the "Unique Stress Constraint," or USC, which simply requires that each word contain exactly one stressed syllable. Gambell and Yang developed this algorithm partially as a response to a hypothesis put forward by Saffran et al. (1996) to explain their own experimental results. Saffran et al. concluded that young infants can attend to the transitional probabilities between syllables, and posit word boundaries where transitional probability (TP) is low. The TP from syllable X to syllable Y is simply given by:

$$P(Y|X) = \text{frequency of } XY/\text{frequency of } X \quad (1)$$

While TP is sufficient to explain the results of Saffran et al.'s 1996 study, it performs very poorly on actual child-directed speech, regardless of whether the probabilities are calculated between phonemes (Brent, 1999) or syllables. Because of the dramatic performance gains shown by the addition of USC in testing, as well as the poor performance of TP, Gambell and Yang conclude that the USC is required for word segmentation and thus is a likely candidate for inclusion in Universal Grammar (Gambell and Yang, 2006).

However, as the results in Table 2 show, VE is capable of slightly superior performance on syllable input, without assuming any prior constraints on syllable stress distribution. Moreover, the performance of both algorithms is also only a few points above the baseline of segmenting at every possible boundary location (i.e., at every syllable). These results show the limitations of simple statistics like TP, but also show that segmenting a sequence of syllables is a simple problem for more powerful statistical algorithms like VE. The fact that a very high percentage of the words found by VE have one stressed syllable suggest that a rule like the USC could be emergent rather than innate.

Algorithm	BP	BR	BF
TP	0.416	0.233	0.298
TP + USC	0.735	0.712	0.723
Bootstrapping + USC	0.959	0.934	0.946
Voting Experts	0.918	0.992	0.953
All Points	0.839	1.000	0.913

Table 2: Performance of various algorithms on the Brown corpus from CHILDES. Other than VE and All Points, values are taken from (Gambell and Yang, 2006).

4.4 Optimality of the VE Chunk Signature

It is one thing to find chunks in sequences, another to have a theory or model of chunks. The question addressed in this section is whether the chunk signature – low internal entropy and high boundary entropy – is merely a good detector of chunk boundaries, or whether it characterizes chunks, themselves. Is the chunk signature merely a good detector of word boundaries, or are words those objects that maximize the signal from the signature? One way to answer the question is to define a "chunkiness score" and show that words maximize the score while other objects do not.

The chunkiness score is:

$$Ch(s) = \frac{H_f(s) + H_b(s)}{2} - \log Pr(s) \quad (2)$$

It is just the average of the forward and backward boundary entropies, which our theory says should be high at true boundaries, minus the internal entropy between the boundaries, which should be low. $Ch(s)$ can be calculated for any segment of any sequence for which we can build a trie.

Our prediction is that words have higher chunkiness scores than other objects. Given a sequence, such as the letters in this sentence, we can generate other objects by segmenting the sequence in every

possible way (there are 2^{n-1} of these for a sequence of length n). Every segmentation will produce some chunks, each of which will have a chunkiness score.

For each 5-word sequence (usually between 18 and 27 characters long) in the Bloom73 corpus from CHILDES, we generated all possible chunks and ranked them by their chunkiness. The average rank of true words was the 98.7th percentile of the distribution of chunkiness. It appears that syntax is the primary reason that true chunks do not rank higher: When the word-order in the training corpus is scrambled, the rank of true words is the 99.6th percentile of the chunkiness distribution. These early results, based on a corpus of child-directed speech, strongly suggest that words are objects that maximize chunkiness. Keep in mind that the chunkiness score knows nothing of words: The probabilities and entropies on which it is based are estimated from continuous sequences that contain no boundaries. It is therefore not obvious or necessary that the objects that maximize chunkiness scores should be words. It might be that letters, or phones, or morphemes, or syllables, or something altogether novel maximize chunkiness scores. However, empirically, the chunkiest objects in the corpus are words.

5 Discussion

Whether segmentation is performed on phonemic or syllabic sequences, and whether it is unsupervised or provided information such as utterance boundaries and pauses, information-theoretic algorithms such as VE, PtM and especially BVE perform segmentation very well. The performance of VE on BR87 is on par with other state-of-the-art semi-supervised segmentation algorithms such as WordEnds (Fleck, 2008) and HDP (Goldwater et al., 2009). The performance of BVE on corpora of child-directed speech is unmatched in the unconstrained case, to the best of our knowledge.

These results suggest that BVE provides a single, general chunking ability that that accounts for word segmentation in both scenarios, and potentially a wide variety of other cognitive tasks as well. We now consider other properties of BVE that are especially relevant to natural language learning. Over time, BVE's knowledge trie comes to represent the distribution of phoneme sequences that begin and end words it has found. We now discuss how this knowledge trie models phonotactic constraints, and ultimately becomes an emergent lexicon.

5.1 Phonotactic Constraints

Every language has a set of constraints on how phonemes can combine together into syllables, called phonotactic constraints. These constraints affect the distribution of phonemes found at the beginnings and ends of words. For example, words in English never begin with /ts/, because it is not a valid syllable onset in English. Knowledge of these constraints allows a language learner to simplify the segmentation problem by eliminating many possible segmentations, as demonstrated in Section 4.3. This approach has inspired algorithms in the literature, such as WordEnds (Fleck, 2008), which builds a statistical model of phoneme distributions at the beginnings and ends of words. BVE also learns a model of phonotactics at word boundaries by keeping similar statistics in its knowledge trie, but can do so in a fully unsupervised setting by inferring its own set of high-precision word boundaries with the chunk signature.

5.2 An Emergent Lexicon

VE does not represent explicitly a "lexicon" of chunks that it has discovered. VE produces chunks when applied to a sequence, but its internal data structures do not represent the chunks it has discovered explicitly. By contrast, BVE stores boundary information in the knowledge trie and refines it over time. Simply by storing the beginnings and endings of segments, the knowledge trie comes to store sequences like `#cat#`, where # represents a word boundary. The set of such bounded sequences constitutes an emergent lexicon. After segmenting a corpus of child-directed speech, the ten most frequent words of this lexicon are *you, the, that, what, is, it, this, what's, to*, and *look*. Of the 100 most frequent words, 93 are correct. The 7 errors include splitting off morphemes such as *ing*, and merging frequently co-occurring word pairs such as *do you*.

6 Implications for Cognitive Science

Recently, researchers have begun to empirically assess the degree to which segmentation algorithms accurately model human performance. In particular,

Frank et al. (2010) compared the segmentation predictions made by TP and a Bayesian Lexical model against the segmentation performance of adults, and found that the predictions of the Bayesian model were a better match for the human data. As mentioned in Section 4.3, computational evaluation has demonstrated repeatedly that TP provides a poor model of segmentation ability in natural language. Any of the entropic chunking methods investigated here can explain the artificial language results motivating TP, as well as the segmentation of natural language, which argues for their inclusion in future empirical investigations of human segmentation ability.

6.1 Innate Knowledge

The word segmentation problem provides a revealing case study of the relationship between nativism and statistical learning. The initial statistical proposals, such as TP, were too simple to explain the phenomenon. However, robust statistical methods were eventually developed that perform the linguistic task successfully. With statistical learning models in place that perform as well as (or better than) models based on innate knowledge, the argument for an impoverished stimulus becomes difficult to maintain, and thus the need for a nativist explanation is removed.

Importantly, it should be noted that the success of a statistical learning method is not an argument that nothing is innate in the domain of word segmentation, but simply that it is the learning *procedure*, rather than any specific *linguistic knowledge*, that is innate. The position that a statistical segmentation ability is innate is bolstered by speech segmentation experiments with cotton-top tamarins (Hauser et al., 2001) that have yielded similar results to Saffran's experiments with human infants, suggesting that the ability may be present in the common ancestor of humans and cotton-top tamarins.

Further evidence for a domain-general chunking ability can be found in experiments where human subjects proved capable of discovering chunks in a single continuous sequence of non-linguistic inputs. Saffran et al. (1999) found that adults and 8-month-old infants were able to segment sequences of tones at the level of performance previously established for syllable sequences (Saffran et al., 1996). Hunt and Aslin (1998) measured the reaction time of adults when responding to a single continuous sequence of light patterns, and found that subjects quickly learned to exploit predictive subsequences with quicker reactions, while delaying reaction at subsequence boundaries where prediction was uncertain. In both of these results, as well as the word segmentation experiments of Saffran et al., humans learned to segment the sequences quickly, usually within minutes, just as general chunking algorithms quickly reach high levels of performance.

7 Conclusion

We have shown that a domain-independent theory of chunking can be applied effectively to the problem of word segmentation, and can explain the ability of children to segment a continuous sequence, which other computational models examined here do not attempt to explain. The human ability to segment continuous sequences extends to non-linguistic domains as well, which further strengthens the general chunking account, as these chunking algorithms have been successfully applied to a diverse array of non-linguistic sequences. In particular, BVE combines the power of the information-theoretic chunk signature with a bootstrapping capability to achieve high levels of performance in both the continuous and incremental paradigms.

8 Future Work

Within the CHILDES corpus, our results have only been demonstrated for English, which leaves open the possibility that other languages may present a more serious segmentation problem. In English, where many words in child-directed speech are mono-morphemic, the difference between finding words and finding morphs is small. In some languages, ignoring the word/morph distinction is likely to be a more costly assumption, especially for highly agglutinative or even polysynthetic languages. One possibility that merits further exploration is that, in such languages, morphs rather than words are the units that optimize chunkiness.

Acknowledgements

This work was supported by the Office of Naval Research under contract ONR N00141010117. Any views expressed in this publication are solely those of the authors and do not necessarily reflect the views of the ONR.

References

Richard N. Aslin, Jenny R. Saffran, and Elissa L. Newport. 1998. Computation of Conditional Probability Statistics by 8-Month-Old Infants. *Psychological Science*, 9(4):321–324.

Nan Bernstein Ratner, 1987. *The phonology of parent-child speech*, pages 159–174. Erlbaum, Hillsdale, NJ.

Lois Bloom. 1973. *One Word at a Time*. Mouton, Paris.

Michael R. Brent. 1999. An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery. *Machine Learning*, (34):71–105.

Roger Brown. 1973. *A first language: The early stages*. Harvard University, Cambridge, MA.

Paul Cohen and Niall Adams. 2001. An algorithm for segmenting categorical time series into meaningful episodes. In *Proceedings of the Fourth Symposium on Intelligent Data Analysis*.

Paul Cohen, Niall Adams, and Brent Heeringa. 2007. Voting Experts: An Unsupervised Algorithm for Segmenting Sequences. *Intelligent Data Analysis*, 11(6):607–625.

Margaret M. Fleck. 2008. Lexicalized phonotactic word segmentation. In *Proceedings of The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 130–138, Columbus, Ohio, USA. Association for Computational Linguistics.

Michael C Frank, Harry Tily, Inbal Arnon, and Sharon Goldwater. 2010. Beyond Transitional Probabilities : Human Learners Impose a Parsimony Bias in Statistical Word Segmentation. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.

Timothy Gambell and Charles Yang. 2006. Statistics Learning and Universal Grammar: Modeling Word Segmentation. In *Workshop on Psycho-computational Models of Human Language*.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2009. A Bayesian Framework for Word Segmentation: Exploring the Effects of Context. *Cognition*, 112(1):21–54.

Sharon Goldwater. 2007. *Nonparametric Bayesian models of lexical acquisition*. Ph.D. dissertation, Brown University.

Zellig S. Harris. 1955. From Phoneme to Morpheme. *Language*, 31(2):190–222.

Marc D. Hauser, Elissa L. Newport, and Richard N. Aslin. 2001. Segmentation of the speech stream in a non-human primate: statistical learning in cotton-top tamarins. *Cognition*, 78(3):B53–64.

Daniel Hewlett and Paul Cohen. 2009. Bootstrap Voting Experts. In *Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence*.

Daniel Hewlett and Paul Cohen. 2010. Artificial General Segmentation. In *The Third Conference on Artificial General Intelligence*.

Ruskin H. Hunt and Richard N. Aslin. 1998. Statistical learning of visuomotor sequences: Implicit acquisition of sub-patterns. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, Mahwah, NJ. Lawrence Erlbaum Associates.

Howard Johnson and Joel Martin. 2003. Unsupervised learning of morphology for English and Inuktitut. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003)*, pages 43–45.

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

Teuvo Kohonen. 1988. *Self-organized formation of topologically correct feature maps*.

Brian MacWhinney and Catherine E Snow. 1985. The child language data exchange system (CHILDES). *Journal of Child Language*.

Matthew Miller and Alexander Stoytchev. 2008. Hierarchical Voting Experts: An Unsupervised Algorithm for Hierarchical Sequence Segmentation. In *Proceedings of the 7th IEEE International Conference on Development and Learning*, pages 186–191.

Matthew Miller, Peter Wong, and Alexander Stoytchev. 2009. Unsupervised Segmentation of Audio Speech Using the Voting Experts Algorithm. *Proceedings of the 2nd Conference on Artificial General Intelligence (AGI 2009)*.

Jenny R. Saffran and Erik D. Thiessen. 2003. Pattern induction by infant language learners. *Developmental Psychology*, 39(3):484–494.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport. 1996. Statistical Learning by 8-Month-Old Infants. *Science*, 274(December):926–928.

Jenny R. Saffran, Elizabeth K Johnson, Richard N. Aslin, and Elissa L. Newport. 1999. Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1):27–52.

Kumiko Tanaka-Ishii and Zhihui Jin. 2006. From Phoneme to Morpheme: Another Verification Using a Corpus. In *Proceedings of the 21st International Conference on Computer Processing of Oriental Languages*, pages 234–244.

Anand Venkataraman. 2001. A procedure for unsupervised lexicon learning. In *Proceedings of the Eighteenth International Conference on Machine Learning*.