Distributional Cues to Word Boundaries: Context is Important

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson
Stanford University, UC Berkeley, and Brown University

1. Introduction

Word segmentation, or identifying word boundaries in continuous speech, is one of the first problems that infants must solve as they are acquiring language. A number of different weak cues to word boundaries are present in fluent speech, and there is evidence that infants are able to use many of these, including phonotactics (Mattys et al., 1999), allophonic variation (Jusczyk et al., 1999a), metrical (stress) patterns (Morgan et al., 1995; Jusczyk et al., 1999b), effects of coarticulation (Johnson and Jusczyk, 2001), and statistical regularities amongst sequences of syllables (Saffran et al., 1996a). The kinds of statistical regularities studied by Saffran et al. (1996a) allow for the possibility of language-independent word segmentation strategies, and seem to be used by infants earlier than other kinds of cues (Thiessen and Saffran, 2003). These facts have led to the proposal that strategies exploiting the statistical patterns found in sound sequences are a crucial first step in bootstrapping word segmentation (Thiessen and Saffran, 2003), and have provoked a great deal of research into statistical word segmentation using both human subjects and computational models.

Most previous work on statistical word segmentation is based on the observation that transitions from one syllable or phoneme to the next tend to be less predictable at word boundaries than within words (Harris, 1955; Saffran et al., 1996a). This observation has led to proposals that infants use statistics such as transitional probabilities or mutual information in order to segment words from speech. A number of models have been developed in an attempt to explain how these kinds of statistics can be used procedurally to identify words or word boundaries. Here, we take a different approach: we seek to identify the assumptions the learner must make about the nature of language in order to correctly segment natural language input.

Observations about predictability at word boundaries are consistent with two different kinds of assumptions about what constitutes a *word*: either a word is a unit that is statistically independent of other units, or it is a unit that helps to predict other units (but to a lesser degree than the beginning of a word predicts its end). In most artificial language experiments on word segmentation, the first assumption is adopted implicitly by creating stimuli through random concatenation of nonce words. In this paper, we use simulations to examine learning from natural, rather than artificial, language input. We ask what kinds of words are identified by a learner who assumes that words are statistically independent, or

[*] This work was partially supported by the following grants: NIH 1R01-MH60922, NIH RO1-DC000314, NSF IGERT-DGE-9870676, NSF 0631518, NSF 0631667, and the DARPA CALO project.

(alternatively) by a learner who assumes that words are partially predictive of later words. We investigate this question by developing two different Bayesian models of word segmentation incorporating each of these two different assumptions. We present the results of simulations using each of these models to segment a corpus of phonemically transcribed child-directed speech. Our simulations indicate that a learner who assumes that words are statistically independent units will tend to undersegment the corpus, whereas assuming that words predict other words leads to a more accurate segmentation. These results suggest that even in the initial stages of acquisition, language learners may need to account for more subtle statistical effects than have typically been discussed in the literature.

2. The Bayesian approach

Our approach differs from that of many other researchers, who investigate the kinds of statistical information that humans are sensitive to (Saffran et al., 1996b; Saffran et al., 1996a; Aslin et al., 1998; Johnson and Jusczyk, 2001; Thiessen and Saffran, 2003) or the kinds of architectures and algorithms that might emulate human learning (Christiansen et al., 1998; Elman, 1990; Swingley, 2005). We focus here on trying to identify some of the assumptions an ideal learner must make about the nature of language in order to successfully solve the word segmentation problem, in the spirit of Marr's (1982) computational level of analysis. In this case, the ideal learner uses Bayesian inference to combine expectations about the structure of language with the information provided by linguistic data. A previous Bayesian model of word segmentation is presented in Brent (1999); we discuss this model in more detail in Section 3.1. Venkataraman (2001) and Batchelder (2002) also propose models based on Bayesian ideas, but their goals are different (focusing on algorithmic design rather than the assumptions of the learner), and the algorithms they use introduce significant learning biases independent of their models.

To apply Bayesian learning to the domain of language, we assume that the learner is exposed to an input corpus of natural language. The process of learning consists of determining some internalized representation (e.g., a grammar or lexicon) that provides a good explanation of how the observed data was generated, and also allows the learner to generate novel linguistic forms. In a statistical setting, we can state this idea formally using Bayes' rule:

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)} \propto P(d|h)P(h)$$

where d is the observed data and h is the hypothesized grammar. $P(d|h)$ (known as the *likelihood*) is the probability of the observed data given a particular hypothesis, and tells us how well that hypothesis explains the data. $P(h)$ (the *prior* probability of h) tells us how good a linguistic hypothesis h is, regardless of any data. The prior can be viewed as a learning bias: hypotheses with high prior probability may be adopted based on less evidence than hypotheses with low prior probability. Bayes' rule states that $P(h|d)$ (the *posterior* probability of h) is proportional to the product of the likelihood and the prior, with $P(d)$ (the probability of the data) acting as a normalizing constant to ensure that $P(h|d)$ sums to one over all hypotheses. The learner can compare the posterior probabilities of different hypotheses by evaluating each one according to its explanatory power (likelihood) and the learner's prior expectations.

Before defining a Bayesian model of word segmentation, we first need to decide what units of representation will be used. In the models described here, the input is represented in terms of phonemes, and the output consists of words (which are sequences of phonemes). Neither of these representations is uncontroversial; on the output side, for example, connectionist approaches typically do not learn or represent words explicitly; instead, they output boundary prediction probabilities, from which words may be reconstructed. We feel that explicit identification and representation of words is important, since learners must eventually assign meanings to words and recombine them in novel ways. As for the input side, Swingley (2005) argues in favor of a syllable-based input representation, while the connectionist model of Christiansen et al. (1998) uses a distributed representation based on phonetic features. The phoneme-based input representation we have chosen makes our model insensitive to feature-based similarity between sounds, and also abstracts away from many details of phonetic and acoustic variation. Nevertheless, it is useful because it allows us to use the same input corpus as several previous researchers (Brent, 1999; Venkataraman, 2001; Batchelder, 2002), and compare our results directly to theirs. In the future, we plan to work towards using input data with more phonetic detail.

With this choice of input and output representations, we can formulate the problem of word segmentation in Bayesian terms as follows: given an input corpus d of unsegmented utterances (i.e., strings of phonemes), each hypothesis h consists of a possible segmentation of the corpus into words. The learner's task is to identify the posterior distribution of segmentations given the observed data (or perhaps to choose a single high-probability segmentation). Notice that in this particular task, $P(d|h)$ is always 1, because for any particular segmentation, the observed data can be generated deterministically by simply concatenating all the words in the segmentation. Therefore, the posterior probability of a segmentation is directly proportional to its prior probability. In other words, the learner will prefer exactly those segmentations that best match the learner's concept of linguistic naturalness.

Using this kind of Bayesian framework, we can examine the kinds of assumptions that lead to successful learning by developing different models that define "naturalness" in different ways. Here, we consider two types of learners that make different assumptions about how words behave in natural language. One type of learner assumes that the probability of observing a particular word is statistically independent of its context (or, equivalently, that all orderings of a given set of words are equally probable). While this assumption clearly does not hold true for natural language, it yields an intuitively simple conceptual approach to word segmentation which can be roughly stated as "look for independent units of speech and identify these as words". The assumption of statistical independence between words is known in computational linguistics as a *unigram* assumption, because the probability of a corpus can be computed by multiplying together the probabilities of its unigrams, or individual words. We will refer to learners making this assumption as unigram learners.

The second type of learner we consider here treats words not as independent units, but as predictive units. This type of learner assumes that the probability of a word *does* depend on its context: words provide information that can be used to help predict future words. There are, of course, many ways in which context could be used to help predict words; the learner we have developed is based on the simplifying assumption that a word's probability is affected by only one preceding word of context. That is, each word can be used to help predict the following word, but has no statistical effect upon later words. This assumption is known as a *bigram* assumption, because frequencies of bigrams, or pairs of words, must be used when computing the probability of a corpus. We describe our bigram learner in more detail in Section 4, but first we turn to the simpler case of unigram word segmentation.

3. Unigram word segmentation

3.1. Model description

To motivate our unigram model of word segmentation, we briefly review a previous Bayesian model of word segmentation described in Brent (1999). Brent's model assumes that the goal of the learner is to identify the segmentation of the input corpus with the highest posterior probability. As in our own model, this is equivalent to finding the segmentation with the highest prior probability. Under Brent's model, the prior probability of a segmentation is defined in terms of four properties of that segmentation: the number of distinct lexical types in the segmentation, the phonemic form of each type, the frequency of each type, and the probability of the particular ordering of word tokens found in that segmentation. Crucially, this model assumes a uniform distribution over token orderings, so that the probability of any ordering of a particular set of tokens is the same as the probability of any other ordering. Since word order is irrelevant, this is a unigram learner.

Here, we propose a new unigram Bayesian model of word segmentation. Our model has some deep mathematical similarities to Brent's model, but has two major advantages over his model. First, in Brent's framework, it is not clear how to replace the learner's unigram assumption with the assumption that context is important. Our own framework makes this relatively easy, so that we are able to develop both unigram and bigram models, and compare the results. A second problem with Brent's model is that there is no known algorithm that can efficiently identify the best segmentation of the input. For all but the tiniest corpora, choosing the best segmentation by exhaustively evaluating the probability of every possible segmentation would be infeasible. Instead, Brent describes an *approximate* algorithm that is intended to identify a relatively high-probability segmentation, but has no guarantees of optimality. It turns out (as we show in Section 3.2) that the segmentations found by this algorithm are actually far from optimal under Brent's model. In contrast, there are well-known techniques for finding near-optimal solutions under models like ours, and we provide evidence that the algorithm we use does identify these solutions.

In our model, as in Brent's, the learner assumes that the observed (unsegmented) corpus was created according to a probabilistic generative process. The specifics of this process, and thus the probabilities assigned by the model to different segmentations, are somewhat different from Brent's. Our model assumes that the corpus was generated by generating a sequence of words $w_1 \ldots w_N$ in order and then removing the boundaries between the words.[1] The ith word in the sequence, w_i, is generated as follows:

(1) Decide if w_i is a novel lexical item.

(2) a. If so, generate a phonemic form (phonemes $x_1 \ldots x_M$) for w_i.
b. If not, choose an existing lexical form l for w_i.

Since this is a probabilistic process, we must assign probabilities to each possible choice. We do so as follows:

(1) $P(w_i \text{ is novel}) = \frac{\alpha}{n+\alpha}$, $P(w_i \text{ is not novel}) = \frac{n}{n+\alpha}$

(2) a. $P(w_i = x_1 \ldots x_M \,|\, w_i \text{ is novel}) = \prod_{j=1}^{M} P(x_j)$
b. $P(w_i = l \,|\, w_i \text{ is not novel}) = \frac{n_l}{n}$

where α is a parameter of the model, n is the number of previously generated words ($= i - 1$), and n_l is the number of times lexical item l has occurred in those n words. This model is known in Bayesian statistics as a Dirichlet process (Ferguson, 1973).

We now provide some intuition for the assumptions that are built into this model. First, notice that in Step 1, when n is small, the probability of generating a novel lexical item is fairly large. As more word tokens are generated and n increases, the relative probability of generating a novel item decreases, but never disappears entirely. This part of the model means that segmentations with too many different lexical items will have low probability, providing pressure for the learner to identify a segmentation consisting of relatively few lexical items. In Step 2a, we define the probability of a novel lexical item as the product of the probabilities of each of its phonemes. This ensures that very long lexical items will be strongly dispreferred. Finally, in Step 2b, we say that the probability of generating an instance of the lexical item l is proportional to the number of times l has already occurred. In effect, the learner assumes that a few lexical items will tend to occur very frequently, while most will occur only once or twice. In particular, our model assigns high probability to segmentations where the frequencies of lexical items follow a power-law (Zipfian) distribution, the kind of distribution that is found in natural language (Griffiths, 2006).

3.2. Simulations

All of the simulations described in this paper were performed on the same corpus used by Brent (1999), which was derived from the Bernstein-Ratner corpus (Bernstein-Ratner, 1987) in CHILDES (MacWhinney and Snow, 1985). The

[1] In our descriptions here of both the unigram and bigram models, we omit certain details that are required to account for the presence of utterance boundaries in the input corpus. These details can be found in Goldwater et al. (2006).

```
(a)                       (b)                        (c)

yu want tu si D6 bUk      yuwant tu si D6bUk         yu want tu si D6 bUk
lUk D*z 6 b7 wIT hIz h&t  lUk D*z 6b7 wIT hIz h&t    lUk D*z 6 b7 wIT hIz h&t
&nd 6 dOgi                &nd 6dOgi                  &nd 6 dOgi
yu want tu lUk &t DIs     yu wanttu lUk&tDIs         yu want tu lUk&t DIs
lUk &t DIs                lUk&tDIs                   lUk&t DIs
h&v 6 drINk               h&v6 drINk                 h&v 6 drINk
oke nQ                    oke nQ                     oke nQ
WAts DIs                  WAtsDIs                    WAts DIs
WAts D&t                  WAtsD&t                    WAts D&t
WAt Iz It                 WAtIzIt                    WAtIz It
lUk k&n yu tek It Qt      lUk k&nyu tek ItQt         lUk k&nyu tek It Qt
tek It Qt                 tek ItQt                   tek It Qt
yu want It In             yuwant It In               yuwan t It In
pUt D&t an                pUt D&t an                 pUt D&t an
D&t                       D&t                        D&t
```

Figure 1: Segmentation of the first 15 utterances in the corpus, according to (a) the correct segmentation, (b) our unigram model, and (c) our bigram model. See the Appendix for a key to the ASCII phoneme encoding.

original corpus contains orthographic transcriptions of utterances directed at 13- to 23-month-olds; Brent removed disfluencies and non-words and used a phonemic dictionary to convert the remaining words into a phonemic representation. The resulting corpus consists of 9790 utterances, with a total of 33399 word tokens belonging to 1321 types. The average number of words per utterance is 3.41, and the average number of phonemes per word is 2.87. In the input to the model, utterance boundaries (corresponding to pauses) are provided, but utterance-internal word boundaries are removed. The utterance-internal word boundaries are used only to evaluate the performance of the system.

In order to evaluate the performance of our unigram model, we need to introduce a procedure that can identify high-probability segmentations of the input corpus. We used a stochastic search procedure known as *Gibbs sampling*, which works by iteratively performing small random perturbations to the current segmentation (inserting or removing one boundary at a time). This algorithm produces samples from the posterior distribution of segmentations defined by the model. A good approximation to the optimal segmentation can be found by collecting a large number of samples and choosing the one with the highest probability; in practice, we found that different samples produced qualitatively and quantitatively similar results. Our evaluation is therefore based on a single sample taken after 20,000 iterations of the sampler. In the results discussed here, the parameter α was set to 20; other values of α yielded qualitatively similar results. For more details of the sampling algorithm and results for other values of α, see Goldwater et al. (2006).

Some example utterances showing the segmentation found by our unigram model are given in Figure 1(b). As these utterances illustrate, the units identified as words by our unigram model often consist of sequences of two or more actual words concatenated together. The system seems to be quite accurate when it proposes a boundary, it simply doesn't propose enough. To quantify these results, we computed the system's accuracy in terms of *precision* and *recall* (also known

Table 1: Accuracy of the two unigram models.

	P	R	BP	BR	LP	LR
Brent	**67.0**	**69.4**	80.3	**84.3**	53.6	51.3
GGJ	61.9	47.6	**92.4**	62.2	**57.0**	**57.5**

Note: Measures are precision and recall on word tokens (P, R), boundaries (BP, BR), and lexicon entries (LP, LR). In all tables, bold indicates the best scoring model.

as *accuracy* and *completeness*):

$$\begin{aligned}\text{Precision} &= 100 * \frac{\text{number of correct items found}}{\text{number of items found}} \\ \text{Recall} &= 100 * \frac{\text{number of correct items found}}{\text{number of true items}}\end{aligned}$$

For example, the recall on word tokens is the percentage of tokens in the true segmentation that were correctly identified in the model's segmentation (where a token is counted as correct only if both boundaries are correct). We calculated precision and recall on ambiguous boundaries (i.e., all possible boundary locations except at utterance boundaries), word tokens, and word types (i.e., lexicon entries). The results are shown in Table 1, with scores from Brent's model provided as a comparison.[2] The scores confirm our qualitative observations: boundary precision is very high for our model, but boundary recall is very low. As a result, overall token precision and recall are both lower than in Brent's model. Lexicon precision and recall are actually better than Brent's, but our low token accuracy is an indication that errors are often made on the most frequent words.

3.3. Discussion

Upon reflection, we should not be surprised at the kind of segmentation found by our model. Recall that a basic assumption of this model is that words have the same probability *regardless of context*. However, this assumption is clearly violated in the corpus. For example, the empirical probability of the word *that* in our data is .024 (i.e., 2.4% of word tokens are the word *that*). Following the word *what's*, the probability of *that* rises to .46, but after the word *to*, the probability of *that* is only .0019. In other words, a single word of context can create variations in probability of more than two orders of magnitude! Since these variations are contrary to the unigram assumption of the model, the only way the system can capture strong word-to-word dependencies is by assuming that sequences of strongly non-independent words are actually single words. The system tends to make this kind of error on the most frequent words precisely because their high frequency provides a great deal of evidence against independence.

Of course, this analysis raises the question of why Brent's unigram model does not produce the same kinds of errors as our own model. The answer lies

[2] Results from Brent's system were obtained using an implementation by Anand Venkataraman available at http://www.speech.sri.com/people/anand/.

Table 2: Negative log probabilities (x 1000) under each unigram model of the true segmentation and the segmentation found by each algorithm.

Seg:	True	Brent	GGJ
Brent	208.2	217.0	**189.8**
GGJ	222.4	231.2	**200.6**

Table 3: Accuracy of the two unigram models on the permuted corpus.

	P	R	BP	BR	LP	LR
Brent	77.0	86.1	83.7	97.7	60.8	53.0
GGJ	**94.2**	**97.1**	**95.7**	**99.8**	**86.5**	**62.2**

in the algorithm used to identify a good segmentation. It turns out that Brent's algorithm finds a segmentation that is actually very far from optimal under his model. While we do not know exactly what segmentation *is* optimal, we can at least compare the probabilities of the two segmentations we have (the one found by our system and the one found by his), as calculated under Brent's model. Table 2 shows the results of these calculations, which indicate that both of the unigram models assign higher probability to the undersegmented solution than to either the solution found by Brent's algorithm, or the correctly segmented corpus.

To provide evidence that our own algorithm is able to identify a near-optimal segmentation, we created an artificial corpus consisting of all the same words and utterance lengths as the original corpus, but with the words permuted at random. Since word order has been randomized, this corpus conforms to the model's expectation that context has no effect on word probabilities. When we used this corpus as input to our algorithm, we found that segmentation performance improved markedly, as shown in Table 3. Brent's system improved on this corpus as well, but to a much lesser extent, again indicating problems with his algorithm.

So far, we have provided evidence that, for two different unigram models of word segmentation, the optimal segmentation of a natural language corpus identifies many common sequences of words as single words. It is natural to ask whether undersegmentation is the result of an optimal segmentation strategy under any model that assumes independence between words, regardless of other properties of the model. A thorough discussion of this question is beyond the scope of this paper, but we have shown elsewhere using formal analysis that undersegmentation is indeed a general property of unigram models (Griffiths et al., 2006). For all but tiny corpora, any reasonable assumptions about word shapes, lexicon size, and token frequencies have less influence on the probabilities of different segmentations than the assumption of independence between words.

4. Bigram word segmentation

In the previous section, we discussed empirical and theoretical evidence that defining words as statistically independent units leads to undersegmentation of natural language. We now ask whether modifying this assumption can lead to better segmentation. We address this question by developing a different model in which words are assumed to help predict other words. In particular, this model assumes that the probability of a word depends on a single previous word of context, so the unit of dependency is a pair of words, or bigram.

4.1. Model description

Like our unigram model, our bigram model defines the probability of a segmentation by assuming that it was generated as a sequence of words $w_1 \ldots w_N$ using a probabilistic process. Unlike the unigram model, w_i is generated using a process that takes into account the previous (already generated) word in the sequence, w_{i-1}:

(1) Decide whether the pair (w_{i-1}, w_i) will be a novel bigram type.

(2) a. If so,

i. Decide whether w_i will be a novel unigram type.

ii. a. If so, generate a phonemic form (phonemes $x_1 \ldots x_M$) for w_i.

b. If not, choose an existing lexical form l for w_i.

b. If not, choose a lexical form l for w_i from among those that have already been observed following w_{i-1}.

Notice that Step 2a, which creates the second word of a novel bigram, invokes the unigram generative process described in Section 3.1. The unigram process in Step 2a generates a set of word types which the bigram process in Steps 1–2 assembles into bigrams.

The probabilities associated with the bigram generative process are

(1) $P((w_{i-1}, w_i)$ is a novel bigram $| \, w_{i-1} = l') = \frac{\beta}{n_{l'}+\beta}$

$P((w_{i-1}, w_i)$ is not a novel bigram $| \, w_{i-1} = l') = \frac{n_{l'}}{n_{l'}+\beta}$

(2) a. i. $P(w_i$ is a novel word $| \, (w_{i-1}, w_i)$ is a novel bigram$) = \frac{\gamma}{b+\gamma}$

$P(w_i$ is not a novel word $| \, (w_{i-1}, w_i)$ is a novel bigram$) = \frac{b}{b+\gamma}$

ii. a. $P(w_i = x_1 \ldots x_M \, | \, w_i$ is a novel word$) = \prod_{j=1}^{M} P(x_j)$

b. $P(w_i = l \, | \, w_i$ is not a novel word$) = \frac{b_l}{b}$

b. $P(w_i = l \, | \, (w_{i-1}, w_i)$ is not a novel bigram and $w_{i-1} = l') = \frac{n_{(l',l)}}{n_{l'}}$

where β and γ are parameters of the model, l' is the lexical form of w_{i-1}, $n_{l'}$ and $n_{(l',l)}$ are the number of occurrences in the first $i-1$ words of the unigram l' and the bigram (l', l), b is the number of bigram types in the first $i-1$ words, and b_l is the number of those types whose second word is l. This model is known as a hierarchical Dirichlet process (Teh et al., 2005).

The intuition behind this model is similar to that of the unigram model. Step 1 says that the more times l' has been generated, the less likely a new word will be generated following it; this limits the number of bigram types. Step 2a is like

Table 4: Accuracy of our bigram model as compared to the unigram models.

	P	R	BP	BR	LP	LR
Brent	67.0	69.4	80.3	**84.3**	53.6	51.3
GGJ (unigram)	61.9	47.6	**92.4**	62.2	57.0	57.5
GGJ (bigram)	**79.4**	**74.0**	**92.4**	83.5	**67.9**	**58.9**

the unigram generative process, except that the probabilities are defined in terms of bigram types instead of unigram tokens. The idea is that some words combine more promiscuously into bigrams than others: If l has been generated in many different contexts already, it is more likely to be generated in this new context. Finally, in Step 2b, the probability of generating l following l' is proportional to the number of times this pair has been generated already, which leads to a preference for power-law distributions over the second item in each bigram.

4.2. Simulations and discussion

For our simulations, we used the same input corpus as in the unigram simulations, and a similar Gibbs sampling algorithm to identify a high-probability solution. The results reported here are with $\beta = 10$ and $\gamma = 1000$. As illustrated in Figure 1(c), the segmentation found by our bigram model contains far fewer errors than the segmentation found by our unigram model, and undersegmentation is much less prevalent. Table 4 shows that our bigram model outperforms both unigram models on almost all measures, in several cases by a wide margin. This improvement can be attributed to a large increase in boundary recall relative to the unigram model, with no loss in precision. In other words, the bigram model proposes more word boundaries and is just as accurate with those proposals.

When the bigram model does make errors, they often fall into one of two categories. First, a few multi-word sequences are still treated as single words. Second, oversegmentation often occurs at morpheme boundaries. The 100 most frequent lexical items found by the model include `z`, `s`, `IN`, `i`, and `t`, which correspond to plural, progressive, diminutive/adjectival, and past tense suffixes. These kinds of errors are not surprising given the similar statistical properties of word boundaries and morpheme boundaries. It is possible that the kind of information used by this model (patterns of sound sequences, word frequencies, etc.) is sufficient to distinguish between morphemes and words, if used in the proper way. However, it is plausible that additional sources of information (e.g., semantics) may be required.

5. Conclusion

In this paper, we have investigated the problem of word segmentation using a Bayesian modeling approach. We have presented two different kinds of models, each of which can be seen as an ideal learner whose goal is to identify words in continuous speech. The difference between these models lies in their assumptions about how words behave. The unigram model assumes that all possible word orderings are equally likely, i.e., that the next word is statistically independent of the previous word. In contrast, the bigram model assumes that the identity of the previous word can be used to help predict the current word. In simulations using these models, we found that the unigram model proposed far too few boundaries, often identifying common word sequences as individual words. We have argued that this behavior results from a mismatch between the independence assumptions in the model and the strong word-to-word dependencies that are found in realistic input corpora. When these dependencies are taken into account, as in our bigram model, word segmentation improves markedly. The importance of considering word-to-word dependencies has not been revealed by previously proposed unigram models because of biases introduced by their learning algorithms, which prevent these models from finding optimal segmentations.

Our results are not incompatible with the possibility that infants use transitional probabilities or other local statistics to identify word boundaries. However, they do imply that statistics and strategies that are sufficient for segmenting the kinds of stimuli found in most behavioral experiments will not necessarily be sufficient for completely segmenting natural language. Our findings suggest the possibility that human learners may exploit statistical information in more subtle ways than have typically been investigated, and we hope that this work will provide a source of further hypotheses that can be tested through experiments.

Appendix: phoneme encoding

Consonants

ASCII	Ex.	ASCII	Ex.
D	THe	h	Hat
G	Jump	k	Cut
L	bottLe	l	Lamp
M	rhythM	m	Man
N	siNG	n	Net
S	SHip	p	Pipe
T	THin	r	Run
W	WHen	s	Sit
Z	aZure	t	Toy
b	Boy	v	View
c	CHip	w	We
d	Dog	y	You
f	Fox	z	Zip
g	Go	~	buttON

Vowels

ASCII	Ex.
&	thAt
6	About
7	bOY
9	flY
A	bUt
E	bEt
I	bIt
O	lAW
Q	bOUt
U	pUt
a	hOt
e	bAY
i	bEE
o	bOAt
u	bOOt

Rhotic Vowels

ASCII	Ex.
#	ARe
%	fOR
(	hERE
)	lURE
*	hAIR
3	bIRd
R	buttER

References

Richard Aslin, Jenny Saffran, and Elissa Newport. 1998. Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9:321–324.

Eleanor Batchelder. 2002. Bootstrapping the lexicon: A computational model of infant speech segmentation. *Cognition*, 83:167–206.

Nan Bernstein-Ratner. 1987. The phonology of parent-child speech. In K. Nelson and A. van Kleeck, editors, *Children's Language*, volume 6. Erlbaum, Hillsdale, NJ.

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

Morten Christiansen, Joseph Allen, and Mark Seidenberg. 1998. Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13:221–268.

Jeff Elman. 1990. Finding structure in time. *Cognitive Science*, 14:179–211.

Thomas Ferguson. 1973. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230.

Sharon Goldwater, Thomas Griffiths, and Mark Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *Proceedings of COLING/ACL*, Sydney.

Thomas Griffiths, Sharon Goldwater, and Mark Johnson. 2006. Relating model-based word segmentation and nonparametric Bayesian statistics. Unpublished manuscript.

Thomas Griffiths. 2006. Power-law distributions and nonparametric Bayes. Unpublished manuscript.

Zelig Harris. 1955. From phoneme to morpheme. *Language*, 31:190–222.

Elizabeth Johnson and Peter Jusczyk. 2001. Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, 44:548–567.

Peter Jusczyk, Elizabeth Hohne, and Angela Bauman. 1999a. Infants' sensitivity to allophonic cues for word segmentation. *Perception and Psychophysics*, 61:1465–1476.

Peter Jusczyk, Derek Houston, and Mary Newsome. 1999b. The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, 39:159–207.

Brian MacWhinney and Catherine Snow. 1985. The child language data exchange system. *Journal of Child Language*, 12:271–296.

David Marr. 1982. *Vision: A Computational Approach*. Freeman & Co., San Francisco.

Sven Mattys, Peter Jusczyk, Paul Luce, and James Morgan. 1999. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, 38:465–494.

James Morgan, Katherine Bonamo, and Lisa Travis. 1995. Negative evidence on negative evidence. *Developmental Psychology*, 31:180–197.

Jenny Saffran, Richard Aslin, and Elissa Newport. 1996a. Statistical learning in 8-month-old infants. *Science*, 274:1926–1928.

Jenny Saffran, Elissa Newport, and Richard Aslin. 1996b. Word segmentation: the role of distributional cues. *Journal of Memory and Language*, 35:606–621.

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

Yee Whye Teh, Michael Jordan, Matthew Beal, and David Blei. 2005. Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems 17*. MIT Press, Cambridge, MA.

Erik Thiessen and Jenny 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.

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