Modeling Human Performance in Statistical Word Segmentation

Michael C. Frank[1] (mcfrank@mit.edu), Sharon Goldwater[2] (sgwater@stanford.edu), Vikash Mansinghka[1] (vkm@mit.edu), Tom Griffiths[3] (tom_griffiths@berkeley.edu), and Joshua Tenenbaum[1] (jbt@mit.edu)

[1]Department of Brain and Cognitive Sciences, MIT; [2]Department of Linguistics, Stanford University; [3]Department of Psychology, University of California, Berkeley

Abstract

What mechanisms support the ability of human infants, adults, and other primates to identify words from fluent speech using distributional regularities? In order to better characterize this ability, we collected data from adults in an artificial language segmentation task similar to Saffran, Newport, and Aslin (1996) in which the length of sentences was systematically varied between groups of participants. We then compared the fit of a variety of computational models—including simple statistical models of transitional probability and mutual information, a clustering model based on mutual information by Swingley (2005), PARSER (Perruchet & Vintner, 1998), and a Bayesian model. We found that while all models were able to successfully complete the task, fit to the human data varied considerably, with the Bayesian model achieving the highest correlation with our results.

Keywords: Statistical learning; word segmentation; language acquisition; Bayesian modeling.

Introduction

How do young infants learn words from fluent speech? Research on this topic has identified a number of information sources which aid in word segmentation, including phonotactic, prosodic, and allophonic cues, as well as lexical knowledge (Jusczyk, 1999). However, these information sources vary across languages. Some languages (like English) have a predominant stress pattern which allows relatively robust segmentation even in the absence of lexical knowledge, while others do not.

One fact remains constant across languages: the use of a small subset of the possible sound sequences in combination to create many different meanings. This property—essentially, the existence of words which are combined together to form sentences—implies certain statistical properties of the speech stream, including an increase in predictability from the beginning of a word to its end. Recent work by Saffran and colleagues suggests that human learners can make use of these statistical properties to distinguish words from non-words in a novel artificial language (Saffran, Aslin, & Newport, 1996). Many other studies have replicated and extended these results to a variety of other domains (e.g., Fiser & Aslin, 2002) and other populations, including non-human primates (Hauser, Newport, & Aslin, 2001).

However, despite the existence of a large number of statistical models of word segmentation (reviewed in Brent, 1999), there have been relatively few attempts to integrate these models with human experimental results. One reason for this lack is the extreme simplicity of many of the human artificial language learning results: every available model of word segmentation should succeed in recovering the complete lexicon of the original Saffran, Newport, and Aslin (1996) experiment, for example. However, despite the simplicity of artificial language experiments, human adults largely do not achieve perfect performance in these tasks. In the work presented here, we take advantage of this fact by manipulating the difficulty of an artificial segmentation experiment and then evaluating a number of computational models on their fit to human performance in this paradigm.

The plan of the paper is as follows. We first describe an artificial language learning experiment with adult participants in which we parametrically varied the length of sentences in our language in an attempt to vary the difficulty of the segmentation task. In the next section, we describe the criteria for evaluation of the models and the details of the implementation of each model (in cases in which the details of a model are already described in another publication we only note where our implementation differs from that description). Because all systems tested were extremely effective at finding the correct segmentation, we compared the models on two measures: 1) the best linear fit of their performance to the experimental data, and 2) the relative contribution of target and distractor scores to the performance of each model. We conclude by discussing the relative merits of the different models in fitting human data.

While our selection of models is by no means a complete survey of the field, we have attempted to test models which have been influential or distinctive in the psychological literature. We start with models based on the suggestion by Saffran, Newport, and Aslin (1996) that boundaries between words can be effectively found through the use of simple bigram statistics. We evaluate three models of this type: local minima in transitional probability (TP); minima in TP with smoothed counts; local minima in pointwise mutual information. We then evaluate three other models which focused on finding a lexicon to fit the input corpus: a clustering model by Swingley (2005) which also uses pointwise mutual information; PARSER (Perruchet & Vinter, 1998), a memory-decay model of segmentation; and a Bayesian model in the style of Brent (1999) by Goldwater, Griffiths, and Johnson (2006).

Experimental Data

When learning a foreign language, longer sentences often seem more difficult to understand than shorter sentences. Certainly, in the limit, individually presented words are easy to learn and remember, while those presented in long sentences with no boundaries are more difficult, perhaps because of problems in segmentation. In order to test the hypothesis that segmentation performance decreases as sentence length increases, we exposed adults to sentences constructed from a simple artificial lexicon. We assigned participants to one of eight sentence-length conditions so that we could estimate the change in their performance as sentence length increased.

Methods

Participants We tested 96 MIT students and members of the surrounding community, but excluded 5 participants from the final sample based on performance greater than two standard deviations below the population mean.

Materials Each participant in the experiment heard a unique and randomly generated sample from an artificial language. The lexicon of this language was generated by concatenating 18 syllables (*ba, bi, da, du, ti, tu, ka, ki, la, lu, gi, gu, pa, pi, va, vu, zi, zu*) into six words, two with two syllables, two with three syllables, and two with four syllables. Sentences in the language were created by randomly concatenating words together without adjacent repetition of words. Each participant heard a randomly generated language sample consisting of 1200 words.

Participants were randomly placed in one of eight sentence length conditions (1, 2, 3, 4, 6, 8, 12, or 24 words per sentence). All speech in the experiment was synthesized using the MBROLA speech synthesizer (Dutoit, Pagel, Pierret, Bataille, & van der Vrecken, 1996) with the us3 diphone database, in order to produce an American male speaking voice. All consonants and vowels were 25 and 225ms in duration, respectively. The fundamental frequency of the synthesized speech was ~100 Hz. No breaks were introduced into the sentences: the synthesizer created equal co-articulation between every phone. There was a 500ms break between each sentence in the training set. Test materials consisted of a word from the lexicon paired with a part-word distractor (a set of syllables of the same length which also appeared—with lower frequency—in the corpus).

Procedure. Participants were given instructions that they were going to listen to a nonsense language for 15 minutes, after which they would be tested on how well they learned the words of the language. All participants listened on headphones in a quiet room. After they had heard the training set, they were instructed to make forced choice decisions between pairs of words from the test set by indicating which one of the two sounded more like a word in the language they just heard. No feedback was given during testing.

Results

Performance by condition is shown in Figure 1. We observed a significant main effect of sentence length on performance (F[7,88]=5.57, p < .001), resulting from the gradual decrease in performance as sentences grew longer.

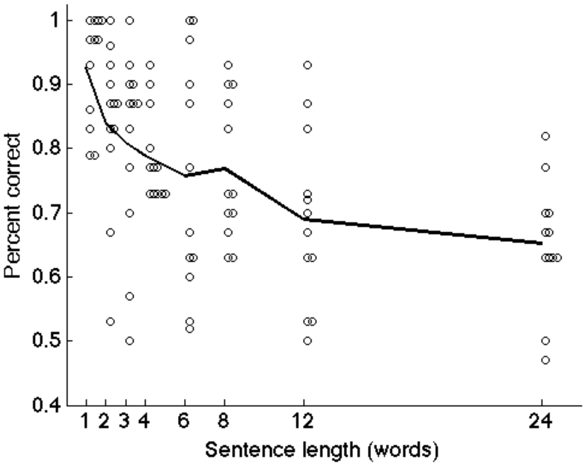

Figure 1. Segmentation performance as a function of sentence length. Dots show mean performance for individuals.

Computational Modeling

In order to evaluate models of word segmentation, we compared their performance on our experimental materials to that of our adult participants.

Materials We compiled a corpus of ten randomly generated training sets in each of the eight sentence length conditions. Each training set was of the same length as those presented to our experimental participants (1200 words) and was accompanied by 30 pairs of test items, the same number of trials as our participants received. Test items were, as in the experimental section, words in the generating lexicon of the training set or part-word distractors.

Evaluation Our metric of evaluation was simple: each model was required to generate a score of some kind for each of the two forced-choice test items. We turned these scores into probabilities by applying the Luce choice rule (Luce, 1963):

$$p_{choice}(\mathrm{x}) = \frac{s(\mathrm{x})}{s(\mathrm{x}) + s(\mathrm{y})}$$

where $s(x)$ and $s(y)$ denote the scores of the two words in the forced choice. (Note that in the case where the scores are probabilities under the language, this is exactly what we should do to condition on the fact that one outcome of the two forced choice options must be in the language). Having produced a choice probability for each test trial, we then averaged these probabilities across test trials and training corpora to produce a set of average choice probabilities for each sentence-length condition.

Boundary-finding approaches

One approach to segmentation employs simple bigram statistics to measure the relationship between units such as syllables or phonemes. This approach is originally due to Harris (1951) but has been the focus of much recent interest. We chose syllables as the primary level of analysis for our models; all syllables had the same structure (consonant-vowel), so there was no difficulty in segmenting words into syllables. We examined three models of this type, beginning with the suggestion of Saffran, Newport, and Aslin (1996) to use local minima in transitional probability as word boundaries. In order to ascertain that our problems in fitting human data did not stem from the 0 and 1.0 TPs found in our corpus, we further tested a transitional probability model with smoothed counts. We also tested a model using pointwise mutual information (MI), a bidirectional measure of association.

Transitional probability We calculated transitional probability (TP) by creating bigram syllable counts over the training sentences in our corpus with a symbol appended to the beginning and end of each sentence to indicate a boundary. The transitional probability of a syllable *b* given *a* was defined as:

$$p(b \mid a) = \frac{p(a,b)}{\sum_{y \in V} p(a,y)}$$

where $p(\mathrm{a},b)$ was the probability of the bigram *ab* appearing and V was the complete set of bigrams observed in the corpus.

The score of a word under this model was defined as the minimum transitional probability within that word (as in Saffran, Newport, and Aslin, 1996). However, given that in our stimuli, transitional probabilities between syllables in the words in the language were equal to one, the same probabilities for targets and distractors would have been computed if the dependent measure were the product of the probabilities within a word rather than the minimum.

Smoothed transitional probability We additionally calculated transitional probabilities using a simple add-lambda smoothing scheme in order to eliminate zero counts for unobserved bigrams. We did this by calculating the probability of a bigram $p(a,b)$ as:

$$p(a,b) = \frac{count(a,b) + \lambda}{\lambda \cdot |V| + \sum_{x \in V} count(a,x)}$$

In other words, we incremented each count by a small constant, lambda, and then divided by lambda times the number of words in the vocabulary. We tested using a range of values for lambda but found equivalent results for all values, thus we report values with a standard value, $\lambda = 1$.

Point-wise mutual information Mutual information is sometimes suggested as an alternative to transitional probability for computing the association strength between syllable pairs (Swingley, 2005; Brent, 1999). Pointwise mutual information is defined as:

$$MI(a,b) = \log_2 \frac{p(a,b)}{\sum_{y \in V} p(a,y) \sum_{x \in V} p(x,b)}$$

We create scores for words using the same method as we used with TP above: taking local minima in MI across words. In a less uniform corpus, this measure would differ significantly from the result of summing mutual information across words. However, given that part-words spanned exactly one word boundary and all syllables appeared in only one word, summing mutual information produced the same result as taking local minima.

Lexicon-finding approaches

We evaluated three other models. These models were distinguished by the assumption that a lexicon—a list of words—is the fundamental representation to be optimized with respect to the input corpus. The first was a recent model by Swingley (2005), which clusters syllables based on their frequency and the mutual information of their syllables. The second was a Bayesian model of segmentation proposed by Goldwater, Griffiths, and Johnson (2006). The third was PARSER (Perruchet & Vinter, 1998), a model based on simple memory principles of decay and interference. We describe each briefly because the relevant details are available in the respective publications.

Swingley (2005) This model is a heuristic clustering model which calculates n-gram statistics and pointwise mutual information over a corpus, then takes as words those strings which exceed a certain threshold value both in their frequency and in the mutual information of their constituent bisyllables. In order to run the model on the language of our experiment, we added support for four syllable words. We then defined the score of a string under the model (given some input corpus) as the maximum threshold value at which that string appeared in the lexicon found by the model. In other words, the highest-scoring strings were those that had the highest percentile rank both in mutual information and in frequency. It should be noted that, unlike the next two models, Swingley's model relies on purely local and word-based statistics (frequency and MI); thus, unlike either PARSER or GGJ2006, a word's score is unrelated to the size of the lexicon.

Goldwater, Griffiths, & Johnson (2006) This model uses Bayesian inference to optimize a lexicon with respect to an observed corpus. Its lexicon is generated according to a Dirichlet process, a probability distribution which gives higher probability to small lexicons containing short words. We use the implementation of the unigram model described in GGJ2006 since there were no bigram syllable dependencies in our materials.

The score for a word under this model was the posterior probability of the word, estimated using a Gibbs sampler as in the original paper. Because the posterior probability of the correct solution was normally so high (indicating a high degree of confidence in the solution the model found), we ran the Gibbs sampler using a range of temperatures to encourage the model to consider alternate solutions. (Temperature is a parameter which controls the degree to which the Gibbs sampler prefers more probable lexicons, with higher temperature indicating greater willingness to consider lower-probability lexicons). The model had one further parameter: the parameter of the Dirichlet process, α, which was kept constant at the value used in the original paper.

PARSER We implemented the PARSER model described in Perruchet and Vintner (1998). This model is organized around a lexicon, that is, a set of words and scores for each word. The model receives input sentences and parses them according to the current lexicon and then adds sequences to the lexicon at random from the parsed input. Each lexical item decays at a constant rate and similar items interfere with each other. The model as described has six parameters: the maximum length of an added sequence, the weight threshold for a word being used to parse new sequences, the forgetting and interference rates, the gain in weight for reactivation, and the initial weight of new words. Because of the large number of parameters in this model, it was not possible to complete an exhaustive search of the parameter space; however, we experimented with a variety of different combinations of interference and forgetting rates and maximum sequence lengths without finding any major differences in forced-choice performance. Therefore, we report results using the same parameter settings used by Perruchet and Vintner.

We made one modification to the model to allow it to run on our data: rather than iterating through the entire input corpus, our implementation of the model iterated through each sentence until reaching the end and then began anew at the beginning of the next sentence. Scores for words under this model were the average scores from the lexicon obtained by averaging the lexicons from 20 PARSER runs on each training set.

Comparison 1: Linear fit to experimental data

Because all of the models we evaluated gave high choice probabilities to the targets (indicating near-perfect segmentation), absolute performance was not useful in comparing models. Instead, we examined performance across the eight sentence-length conditions relative to the performance of adult participants. In other words, we were interested in whether the models extracted a similar amount of information from a corpus with e.g., three-word sentences relative to what they extracted from a corpus with two-word sentences. In order to compute the relationship between conditions in different models relative to the human data, we first scaled the performance of each model using a linear regression, finding the best linear adjustment of the scale of the curve from each model. We then computed simple correlation coefficients (*r* values) between the best fit of each model and the experimental data. See figure 2 for results.

We found that GGJ2006 best fit our experimental data, succeeding in particular in modeling the decrease in

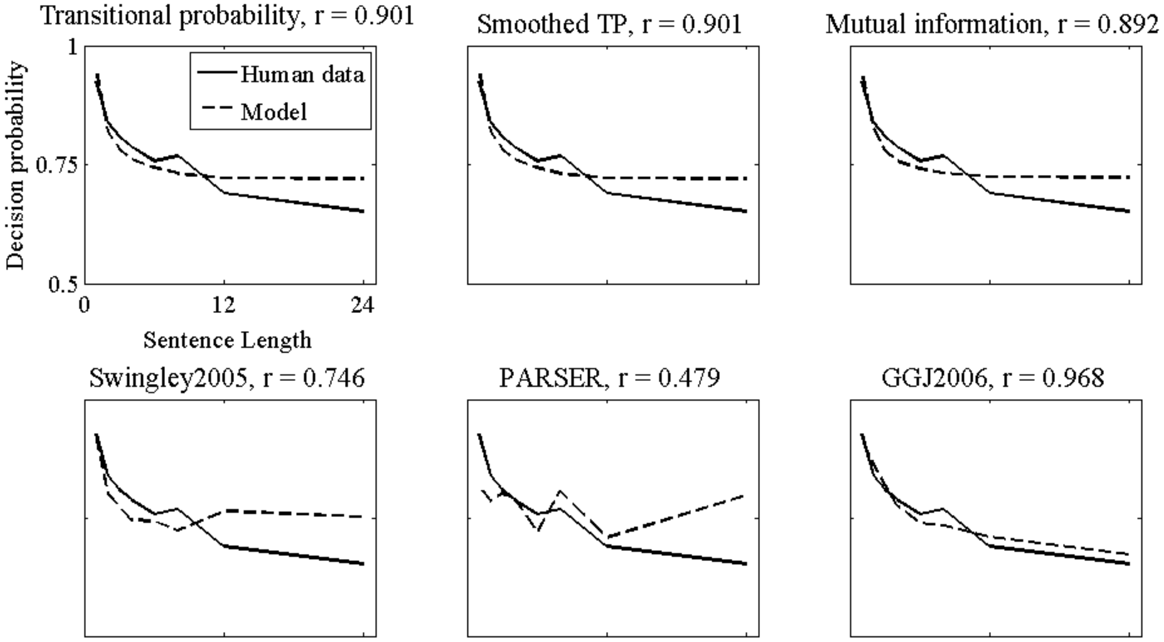

Figure 2. Best linear fit of each model's performance to human data, graphed by sentence length. The vertical axis represents decision probabilities for models and percentage correct for human data; the horizontal, sentence length.

information between sentences of length 12 and sentences of length 24. Here we plot results from this model at temperature 5, but for temperatures between 50 and 3, there was relatively little difference in *r* values (ranging predictably between .920 and .968).

Interestingly, the next most effective models were the boundary-finding models using MI and TP. These models produce curves that are noticeably too shallow, changing little in performance between sentences with length 12 and those with length 24; however, they do show the same dependency between sentence length and score as the human data do. Surprisingly, once TP, smoothed TP, and MI were fit to the human data, there was no appreciable quantitative difference between the models' fit, suggesting that these models' underlying difficulty in fitting this dataset may be a fundamental deficit of the bigram statistic approach.

Finally, both Swingley2005 and PARSER performed very poorly on this task, producing choice probabilities that did not decrease as utterance length increased. One issue in the Swingley model is that because it relies on percentile rankings of frequency rather than raw frequency, its performance can vary highly with very small changes in frequency. In addition, because the model is deterministic, this noise could not be averaged out by multiple runs through the input corpus.

PARSER was similarly variable in its performance, but for different reasons. In any given run, PARSER very rarely assigned any score to a given distractor; thus it was necessary to run the model a large number of times on each different corpus in order to estimate choice probabilities despite its intent to be a single-pass, online segmentation system. Run in this fashion, PARSER is actually quite similar to a probabilistic model such as GGJ2006 or Brent (1999) in that it is an algorithm for sampling a posterior distribution over lexicons, albeit one that incorporates a number of free parameters and ad hoc approximations.

In the following section we examine further the reasons why performance differed between models by examining the contribution of target and distractor scores to each model's choice probabilities.

Comparison 2: Target and distractor scores

Why did some models match the drop in human performance as sentence length increased, while others did not? One way in which models differed from one another was whether the score assigned to targets and distractors changed as the sentence length changed. For example, because TPs within words were always 1, target scores under the two TP models remained constant no matter what the sentence length. In order to investigate this factor more systematically, we plotted target and distractor scores for each model, normalized by the maximal target score (so that all scores varied between 0 and 1). See figure 3 for results.

Although we have no empirical data which address whether participants make errors at longer sentence lengths because targets are less attractive or because distractors are more so, it seems plausible that both are true. However, in each of the four models based on bigram statistics, nearly all of the change in performance across sentence lengths was caused by changes in the score assigned to distractor elements. In contrast, there was almost no change in the score of target items. This lack of change in target word scores seems unrealistic in a psycholinguistic model. Words

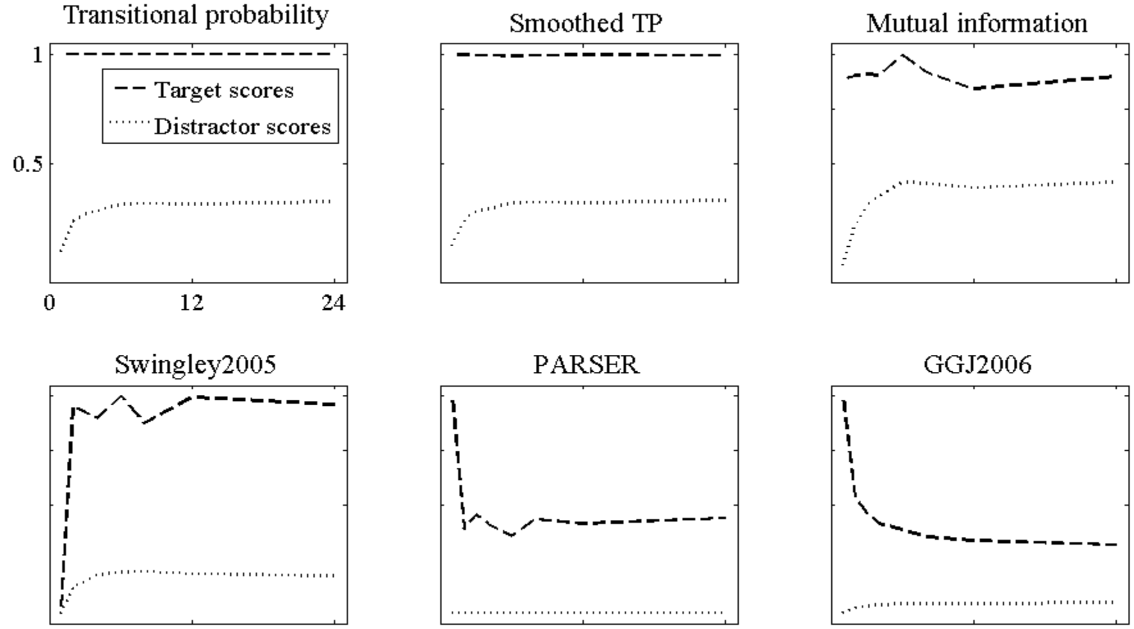

Figure 3. Target and distractor scores for each model, normalized by highest target probability. The vertical axis is in arbitrary units; the horizontal represents sentence length.

differ in their frequency and acceptability, and if models of segmentation are to make contact with models of the lexicon they should be able to take this fact into account.

Perhaps this difference is ultimately a difference between boundary-finding models and models which search for a globally good lexicon (such as GGJ2006 and PARSER). The performance of these latter two models is characterized by a steep dip in target probabilities as sentence length increases, corresponding to a decrease in confidence in the target words. However, unlike PARSER (inasmuch as we could determine given our computational constraints), GGJ2006 also increased the probability of distractor items as sentence length increased. This combined drop in target probability and increase in distractor probability led to the close fit between GGJ2006's performance and our participant data.

Conclusions

We collected data on word segmentation by adults in an artificial language task in which words were presented as part of unsegmented sentences. Our data showed a clear and dramatic decrease in performance as sentence length increased. In order to better characterize the mechanisms involved in segmentation, we attempted to fit a variety of computational models to our participants' data. We initially implemented three boundary-finding models based on bigram statistics: local minima in unsmoothed transitional probability, smoothed TP, and mutual information. We additionally tested three other models which focused not on finding boundaries but on developing a lexicon: a clustering model by Swingley (2005); PARSER, an online model by Perruchet & Vintner (1998); and a Bayesian model by Goldwater, Griffiths, and Johnson (2006).

We found that the Bayesian model—GGJ2006—achieved a significantly higher fit to the human empirical data than any of the other models, reflecting the change in both target and distractor probabilities under the model across different sentence length conditions. Surprisingly, the next highest-performing models were the three based on simple bigram statistics: TP, smoothed TP, and MI. These models were very nearly equivalent to one another once they had been fit to the data, and they correctly predicted the decrease in performance as sentence lengths increased, although the shape of the predicted curve did not exactly match the empirical data. Finally, neither PARSER nor Swingley2005 adequately fit the human data, although they failed for different reasons. Swingley2005's weakness was its extreme sensitivity to small differences in frequency combined with its deterministic character. PARSER's weakness, in contrast, may have been its excessive variability, which made evaluating the scores of distractor items quite difficult.

The success of the Bayesian model in fitting our empirical data suggests several conclusions, both about modeling and about segmentation. First, two of the lexicon-finding models, PARSER and GGJ2006, have the property of assigning varying scores to target words depending on the model's degree of confidence in that word; this property seems useful in models which make contact with other aspects of the word learning task and should be pursued in future modeling. Second, the close correspondence between the predictions of the Bayesian model and the human data suggests that there may be a congruence between the assumptions of the model and the assumptions of the human word learning system.

Acknowledgments

The first author was supported by a Jacob Javits Graduate Fellowship; this research was also supported by NSF grant #0631518. The authors would like to acknowledge Ted Gibson and the members of Tedlab and Cocosci for helpful comments.

References

Brent, M. R. (1999). An Efficient, Probabilistically Sound Algorithm for Segmentation and Word Discovery. *Machine Learning, 34*(1), 71-105.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & van der Vrecken, O. (1996). The MBROLA Project: Towards a Set of High-Quality Speech Synthesizers Free of Use for Non-Commercial Purposes. *Proccedings of the International Conference on Spoken Language Processing, 3*, 1393-1396.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, 99*(24), 15822-15826.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2006). Contextual Dependencies in Unsupervised Word Segmentation. *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*.

Harris, Z. S. (1951). *Methods in structural linguistics*: University of Chicago Press [Chicago.

Hauser, M. D., Newport, E. L., & Aslin, R. N. (2001). Segmentation of the speech stream in a human primate: statistical learning in cotton-top tamarins. *Cognition, 78*, B53-B64.

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

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush & E. Galanter (Eds.), *Handbook of Mathematical Psychology*. New York: Wiley.

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

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science, 274*(5294), 1926-1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language, 35*(4), 606-621.

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