Beyond Transitional Probabilities: Human Learners Impose a Parsimony Bias in Statistical Word Segmentation

Michael C. Frank
mcfrank@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Harry Tily
hjt@stanford.edu
Department of Linguistics
Stanford University

Inbal Arnon
inbal.arnon@manchester.ac.uk
Department of Linguistics
University of Manchester

Sharon Goldwater
sgwater@inf.ac.uk
School of Informatics
University of Edinburgh

Abstract

Human infants and adults are able to segment coherent sequences from unsegmented strings of auditory stimuli after only a short exposure, an ability thought to be linked to early language acquisition. Although some research has hypothesized that learners succeed in these tasks by computing transitional probabilities between syllables, current experimental results do not differentiate between a range of models of different computations that learners could perform. We created a set of stimuli that was consistent with two different lexicons—one consisting of two-syllable words and one of three-syllable words—but where transition probabilities would not lead learners to segment sentences consistently according to either lexicon. Participants' responses formed a distribution over possible segmentations that included consistent segmentations into both two- and three-syllable words, suggesting that learners do not use pure transitional probabilities to segment but instead impose a bias towards parsimony on the lexicons they learn.

Keywords: Word segmentation; statistical learning; computational modeling.

Introduction

Human adults, infants, and even members of other species have the ability to identify statistically coherent sequences in unsegmented streams of stimuli after only a very short exposure (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Hauser, Newport, & Aslin, 2001). This segmentation ability is extremely robust, operates across a wide range of modalities (Conway & Christiansen, 2005), and has been hypothesized to play an important role in early language acquisition (Kuhl, 2004). Nevertheless, relatively little is known about the computations underlying statistical segmentation.

In one influential study, Saffran, Newport, and Aslin (1996) exposed participants to a simple artificial language which consisted of six trisyllabic words concatenated together to form a continuous speech steam. After only a few minutes of exposure, participants were able to distinguish words in this language from strings that did not occur with the same frequency. They speculated that participants could succeed by computing syllable-to-syllable transitional probabilities (TPs) and segmenting the speech stream at local minima in TP.

There are many possible computations by which learners could extract coherent units from the statistical structure of the speech stream, however. Lexicon-based learners like PARSER (Perruchet & Vinter, 1998) and Bayesian lexical models (Brent, 1999; Goldwater, Griffiths, & Johnson, 2009) have also been proposed as possible models of segmentation. Though these models differ on several dimensions, all assume that learners attempt to learn a consistent lexicon—a set of word forms that is combined to form the training sequence—and they do this by preferring small lexicons composed of frequent, short words.

Two previous studies have examined whether this kind of model could provide a good fit to human learning performance. The first contrasted recognition of sub-parts of the words from a speech stream and found that PARSER, like human learners, failed to discriminate sub-parts of words after training (Giroux & Rey, 2009). The second study found that a parsimony-biased chunk-finding model better accounted for human performance across a range of experiments in the visual domain than a purely associative model (Orbán, Fiser, Aslin, & Lengyel, 2008). Thus, both of these studies suggest that human learners do not simply represent association probabilities in statistical learning.

Our current study asks what kinds of learning biases operate in statistical learning. Our study makes use of a novel language whose transition statistics support not just one but a range of possible coherent segmentations: training data could be interpreted as a sequence of sentences of six words from a lexicon of two-syllable words or a sequence of sentences of four words from a lexicon of three-syllable words (where all words appeared with approximately the same frequency). TPs for a single sentence in this language are shown in Figure 1. A learner using pure TPs to segment the language would not recover either lexicon but would instead either segment the language into sets of six-syllable words or else segment inconsistently into a mix of two- and three-syllable words. Thus, our language was designed to test whether human learners would learn more parsimonious lexicons than those implied by pure transition statistics.

Experiment 1 validates two methodological innovations: a web-based interface for data collection and a dependent measure which directly evaluates participants' word segmentation judgments. Experiment 2 uses these methods to test

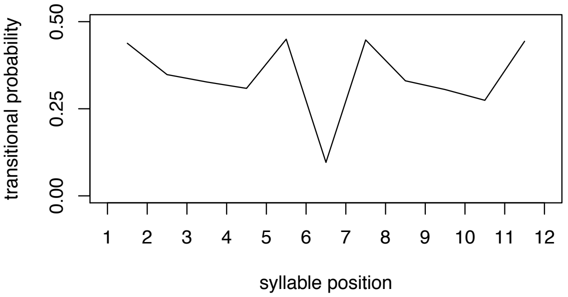

Figure 1: Average transitional probabilities between syllables in an ambiguous language from Experiment 2.

participants' segmentation judgments in the ambiguous language discussed above. We compare the distribution of participants' segmentations to the performance of two computational models—a standard TP model and a Bayesian model that looks for a parsimonious lexicon—and conclude that participants' judgments reflect the operation of a parsimony bias.

Experiment 1

The first condition of Experiment 1 compares web-based data on a segmentation task to previously-collected lab data (Frank, Goldwater, Griffiths, & Tenenbaum, under review) on a standard 2 alternative forced choice (2AFC) test trial. The second condition evaluates a new measure of segmentation: explicit segmentation decisions. We developed a graphical paradigm in which participants heard a sentence, saw it transcribed on the screen, and were asked to click between syllables to indicate where they thought the boundaries between words were.

Methods

Participants Forty eight separate HITs (opportunities for a participant to work) were posted on Amazon's Mechanical Turk web-based crowd-sourcing platform. We received 40 HITS from distinct individuals. Participants were paid $1 for participating.

Stimuli For each condition, we constructed 16 distinct languages to be heard by different participants (to avoid item effects caused by phonological similarity of words). These languages each had a lexicon of six words (2 x two syllables, 2 x three syllables, 2 x four syllables). Words were created by randomly concatenating the syllables *ba*, *bi*, *da*, *du*, *ti*, *tu*, *ka*, *ki*, *la*, *lu*, *gi*, *gu*, *pa*, *pi*, *va*, *vu*, *zi*, and *zu*. Stimuli were synthesized using MBROLA (Dutoit, Pagel, Pierret, Bataille, & Vrecken, 1996) at a constant pitch of 100Hz with 25ms consonants and 225ms vowels. Sentences were generated by randomly concatenating words into strings of four words with no repetitions. All words had frequencies of 300 in the resulting corpus of 75 sentences.

For the 2AFC condition, part-word test stimuli (Saffran, Newport, & Aslin, 1996) were created by concatenating the first syllable of each word with the remaining syllables of another word; this created distractors which appeared in the training corpus with lower frequency than the words. For the segmentation condition, we generated 10 extra sentences according to the same uniform frequency distribution and lexicon as the training corpus.

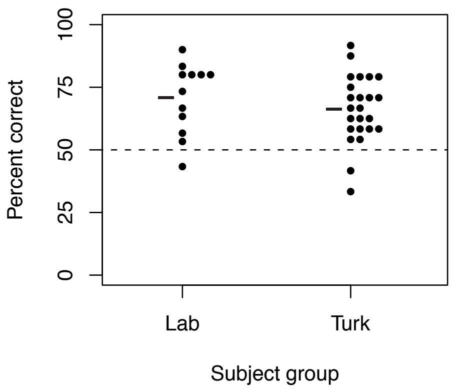

Figure 2: Average percent correct is plotted by subject for in-lab participants from Frank et al. (under review) and Mechanical Turk participants from the 2AFC condition of Experiment 1. Each point is an individual participant, bars show the mean, and the dashed line represents chance.

Procedures After selecting our HIT, our Adobe Flash interface tested that participants' sound was on and that they were able to understand our instructions by asking them to listen to a simple English word and enter it correctly. Participants were then instructed that they would listen to a set of sentences from a made-up language and then be tested on what they had learned. In order to hear each sentence during training, participants clicked a button marked "next."

In the test phase of the 2AFC condition, participants heard 24 pairs consisting of a word and a length-matched part-word and clicked a button for each to indicate which one sounded more like the language they just heard. In the segmentation condition, participants were asked to click on the breaks between words in a graphic display of a sentence. They performed one practice trial on an English sentence presented in this way ("In di an go ril las ne ver eat ba na nas") and prevented from continuing until they segmented it correctly. They then segmented 10 test sentences. Sentences were presented with each syllable separate. Each sentence was played once at the beginning of a trial, and below the sentence was a button that offered the option of hearing the sentence again.

Results and Discussion

In the 2AFC condition (N=24), we found that participants were above chance in their mean accuracy, taken as a group ($t(23) = 5.92$, $p < .0001$). Results are plotted together with data from an identical condition of Frank et al. (under review) (Experiment 2, 300 words exposure), collected from a group of participants in the lab (Figure 2). Mean performance was slightly lower for the Internet-based Turk par-

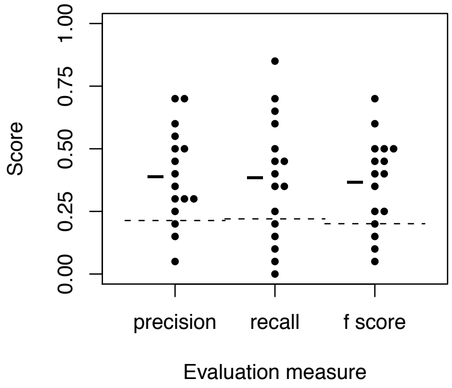

Figure 3: Token precision, recall, and F-score are plotted for individual participants in the segmentation response condition of Experiment 1. Points represent individual participants for each measure. Bars show means and dashed lines show permutation baselines.

ticipants (M=66% compared with M=71%) but not significantly so (Welch two-sample t-test for unequal sample sizes, $t(21.21) = -.92$, $p = .37$). Participants completing the learning task on their own computer via the Internet were able to perform at levels comparable to participants in an isolated room in a psychology laboratory.

In the segmentation condition (N=16), we could not analyze participants' percent correct judgments as in the 2AFC condition. Instead, we evaluated two aspects of performance. First, we asked about the correctness of the boundaries participants placed: whether these decisions corresponded to the correct segmentation (*boundary* performance). Second, we asked about whether each word in the sentence was segmented correctly at its boundaries (*token* performance).

We computed hits (correctly placed boundaries or correctly segmented tokens), misses (missed boundaries or tokens that were not segmented appropriately), and false-alarms (extra boundaries or incorrect tokens that were segmented). Precision captures the proportion of boundaries that were placed correctly and is computed as hits / (hits + false-alarms), while recall captures the total proportion of correct boundaries that were identified and is computed as hits / (hits + misses). We combined these into an F-score, a commonly used metric that is the harmonic mean of precision and recall (Goldwater et al., 2009).

Figure 3 shows token precision, recall, and F-score for participants in the segmentation condition. We calculated an empirical baseline for each measure via permutation: we repeatedly shuffled each participant's boundary decisions within each sentence at random and computed the same measures over it, then took the mean for each. We then used these empirical baselines to test whether participants were above chance in this condition and found that they were for both measures (boundary performance: one sample t-test for precision, $t(15) = 5.23$, $p = .0001$; recall, $t(15) = 6.79$, $p < .0001$; F-score, $t(15) = 8.75$, $p < .0001$, token performance: $t(15) = 3.63$, $p = .002$; recall, $t(15) = 2.71$, $p < .01$; F-score, $t(15) = 3.41$, $p < .004$), though boundary performance was better than token performance. Participants were able to understand the segmentation task and link the regularities they extracted from the exposure corpus to the response format.

Experiment 2

We made use of the two methodological innovations from Experiment 1—Internet data collection and explicit segmentation judgments—to ask about participants' responses to a language where TP did not reveal the possible lexicons of two- or three-syllable words. Instead, pure TPs predicted that participants would often segment the language into words of six-syllables and would rarely segment into words of two or three syllables. Our next experiment tests these predictions.

Methods

Participants Two-hundred and three separate experimental HITs were posted on Amazon Mechanical Turk. We received 119 HITs from distinct individuals who made segmentation decisions on every trial. Participants were paid $0.50 for participating. An addition 145 HITs in the test-only control condition were posted at $0.25 each; we received 102 HITs from distinct individuals who made segmentation decisions.

Stimuli Languages were generated using two parallel vocabularies, one of eight two-syllable words and one of six three-syllable words. These vocabularies were designed to allow overlapping segmentations where the presence of a certain word from one vocabulary did not always indicate the presence of the same set of words from the other. For example, if the three-syllable vocabulary contained ABC, the two-syllable vocabulary would contain at least either AB and two words beginning C, or BC and two words ending A. Sentences of 12 syllables were generated by choosing syllables one at a time from the set that made the sentence to the current point compatible with both vocabularies. At each point, syllables were chosen from a distribution over this set, weighted inversely to the frequency with which they had been chosen to follow the previous syllable in all sentences so far. The resulting sentences displayed probabilistic word-to-word dependencies, much as one would expect in natural language due to the syntactic relationships between words, but in no languages were there pairs of words from either vocabulary which always appeared together. We generated 30 distinct languages and synthesized them as in Experiment 1. Each language contained 25 sentences for training and 10 test sentences, sampled from the same distribution. Sentence presentation order was random.

Procedures Procedures were identical to the segmentation condition of Experiment 1. Participants in the test-only control condition received no training sentences.

Results and Discussion

Participants produced a wide range of segmentations, from those which segmented every three syllables to those which

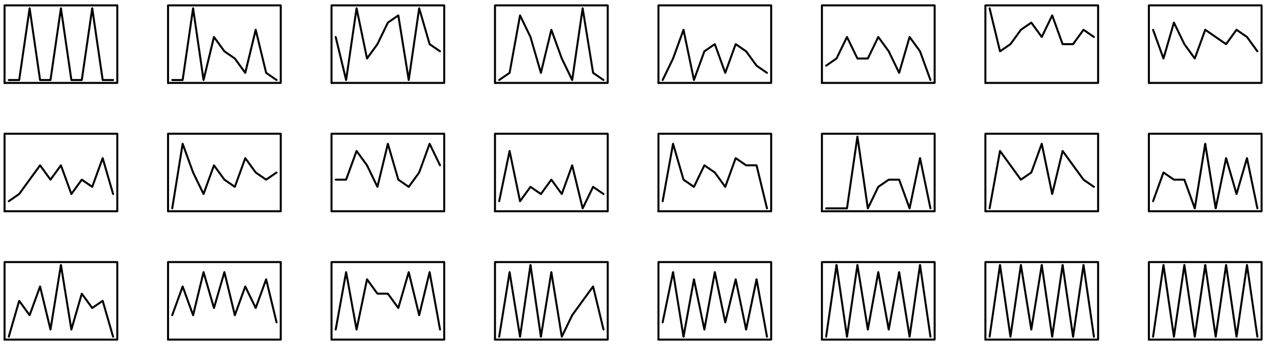

Figure 4: Twenty four participants in Experiment 2, uniformly sampled along the dimension of 2-segmentation F-score. Plots show average probability of placing a boundary at each location in a sentence. Top left shows three-segmenters (three peaks separating four three-syllable plateaus), while bottom right shows two-segmenters (five peaks separating six two-syllable plateaus).

segmented every two syllables. Sample responses are shown in Figure 4. While there was an overall trend towards 2-consistent segmentations, a wide variety of segmentations were observed. Contrary to the predictions of the TP account, there were almost no segmentations into words of six syllables and there were a considerable number of segmentations into words of two and three syllables.

We evaluated participants' performance on the same measures used in Experiment 1: precision, recall, and F-score for both boundaries and tokens. Rather than using a single correct segmentation, we calculated these measures for both the 2-syllable lexicon and the 3-syllable lexicon (Figure 5), showing the distribution of responses on the continuum between a perfect 2-segmentation and a perfect 3-segmentation.

One possible alternative explanation of our finding could be that learners have a bias towards segmenting consistently (e.g., because of the trochaic, bisyllabic structure of English) even without taking into account the structure of the languages they heard. However, results from the first trial of the test-only condition had a very different distribution than those who underwent training (Figure 5). Without training, performance was similar to a randomized baseline in which participants' judgments for each sentence were shuffled randomly. Although there was some learning during test for participants in the test-only condition, there was very little change in the distribution of responses during test for those participants who underwent training.

Our results are inconsistent with the hypothesis that participants segmented on the basis of TPs. Instead, the distribution of participants' responses shows a bias towards segmentations that were consistent with a more parsimonious lexicon than that produced by segmenting at low transition probabilities.

Models

To formalize the intuitions motivating Experiment 2, we evaluated a TP model and a lexicon-finding model on the experimental stimuli. We then evaluated the segmentations produced by these models on the same criteria that we used for the human participants.

Transitional probability model

For each language, we calculated TP for each pair of syllables that appeared in the training portion of the corpus. We computed TP as $P(s_2|s_1) = C(s_1,s_2)/\sum_{s' \in S} C(s_1,s')$ where $C(s_1,s_2)$ refers to the count of instances of the string $s_1 s_2$.

Earlier proposals for TP models called for segmenting at local minima in TP (Saffran, Newport, & Aslin, 1996). However, this method produces only a single possible segmentation for a given sentence and provides no plausible explanation for how participants could have given such different responses for such similar languages. Thus, we chose to convert the TPs for test sentences into decision boundaries via a simple threshold operation: we inserted a boundary in a test sentence every time TP was below a threshold value in that sentence. Rather than picking a single threshold value, we assumed that participants might have a range of threshold values and that this range might explain the variation between participants we observed. Therefore we created a separate segmentation for each language for each threshold value from zero to one at an interval of .1.

Lexical model

We also ran the unigram Bayesian Lexical model described in Goldwater et al. (2009). This model is a probabilistic model which uses Bayesian inference to search the space of segmentations of the training corpus, evaluating each segmentation on the parsimony of the lexicon that would have created it. The structure of the model makes a segmentation more probable when it results in fewer, shorter lexical items (though also when the segmentation itself contains fewer word tokens, which leads to a trade-off).

As in the TP model, it was important to investigate the range of segmentations that were available under this model. When we ran a standard Markov-chain monte carlo algorithm using the parameter set from previous simulations, we found

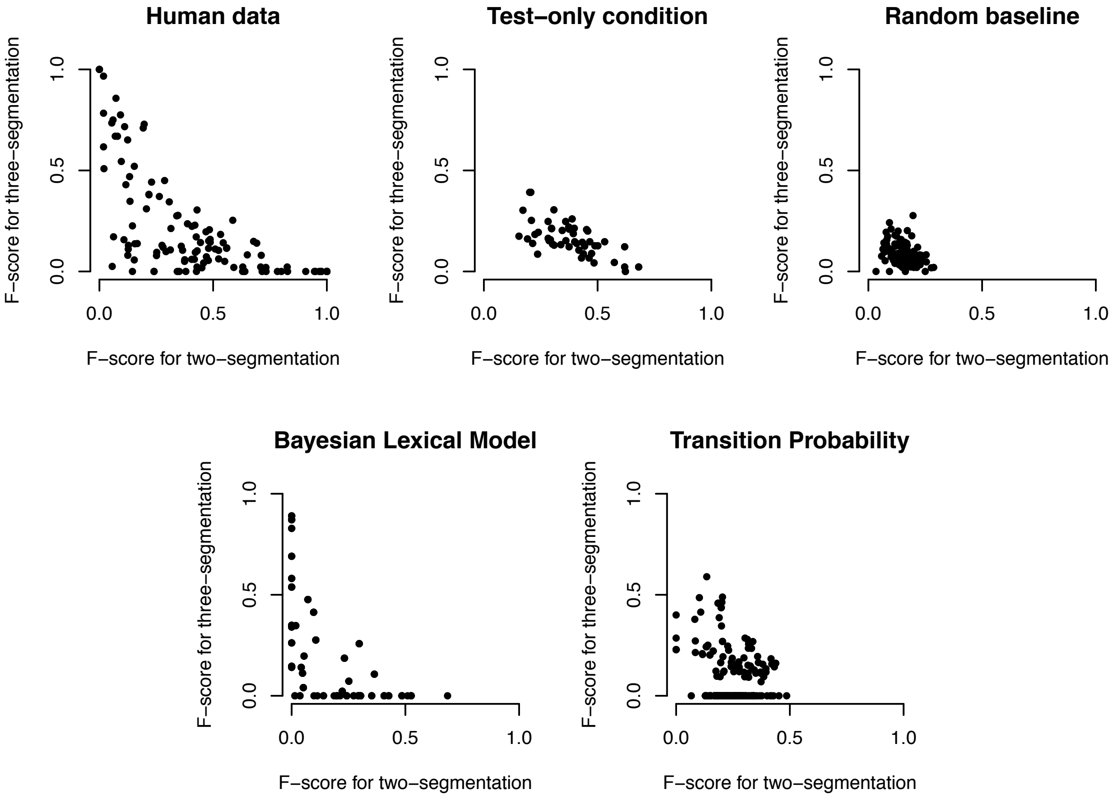

Figure 5: Participant and model token F-scores for Experiment 2. Three-syllable token F-scores are plotted by their two-syllable token F-scores. Each dot represents a single participant or a single model run.

Table 1: Kullback-Leibler divergence between the distribution of human experimental data and other data.

Model	Token F	Boundary F
Test-only condition	4.01	3.45
Random baseline	7.26	9.45
Lexical model	**2.07**	**3.16**
Transitional probability	4.62	3.72

Table 2: Log probability of consistent segmentations under the Lexical model.

Syllables per word	Log probability
6	-594.28
4	-932.92
3	-530.62
2	-697.07
1	-1127.20
unsegmented	-1907.20

that it converged to a segmentation that preferred a lexicon of three-syllable words. In order to investigate a broader range of segmentations, we manipulated the temperature of inference in the model by exponentiating posterior probabilities at a range of values. (This manipulation is a standard technique for allowing sampling algorithms to explore a hypothesis space more broadly, rather than converging to the single highest-probability answer.) With slightly higher temperatures, our sampler explored a broad range of possible segmentations. We report results for temperature = 2 although results for a temperature of 3 were comparable.

Results and Discussion

Results for both models are shown in Figure 5, bottom. The transitional probability model failed to capture the spread of human results: nearly all segmentations it found were comparable in F-score for 2- and 3-segmentation, and no segmentation was over an F-score of .5 on either measure. The Lexical model came closer to capturing the distribution of responses, though it was not as effective at finding 2-segmentations as the human participants, suggesting a possible role for a trochaic bias. Unlike the TP model, however, its probability landscape was truly multi-modal, finding relatively high probability segmentations with 2, 3, and 6 syllables per word (Table 2).

We measured the differences between the distributions of responses across human participants and models using Kullback-Leibler divergence—an information-theoretic measure of the difference between a true distribution and an approximation of that distribution—to quantify the number of bits between distributions (MacKay, 2003). In order to convert sets of observations into smooth distributions, we convolved them with a Gaussian kernel with a constant kernel width. This manipulation produced a smooth density which could be effectively compared using KL divergence.[1] Results are shown in Table 1. The Lexical model showed the lowest divergence from the human response distribution, while the TP model was closer to the empirical baseline in its divergence from the human distribution.

General Discussion

We presented two studies of statistical word segmentation. The first study introduced two methodological innovations, web-based data collection and explicit segmentation judgments. We used these new methods in the second study to test whether human learners faithfully learned the transitional probabilities of an ambiguous language or whether they gave a segmentation that was more consistent with one of the two possible lexicons that generated the training corpus. We found that the distribution of participants' responses was not consistent with the distribution of segmentations produced by segmenting according to a TP model. Thus, our results provide evidence that human learners do not simply encode transitional or associative statistics but instead impose some kind of bias on what they learn.

This bias could be either a bias for consistent word lengths or for a parsimonious lexicon. A model which searched for lexicons with small lexicons consisting of highly frequent, short words produced a distribution similar to that produced by the human learners. Nonetheless, the Lexical model preferred a lexicon with three-syllable words, unlike human learners who preferred to segment into two-syllable words; and the Lexical model assigned a high probability to a segmentation into two words of six syllables each, while participants rarely produced this segmentation. Frank et al. (under review) found that models with memory limitations provided a better fit to human performance, suggesting that one possible explanation for these differences is the increased difficulty for human learners of remembering longer words.

The language used in Experiment 2 has a number of limitations. First, unlike recent studies (Frank et al., under review; Giroux & Rey, 2009), the competing lexicons we used in this study were composed of words of homogenous length, leading to stimuli that could be perceived as isochronous. Second, the size of the lexicons was relatively small and the restrictions on sentences were tight, leading to a small number of possible sentences. Our ongoing work attempts to address both of these issues.

Results in the statistical learning literature have rightly been interpreted as showing that human learners are sensitive to associative and transitional statistics in their environment. But these interpretations should not be confused with the conclusion that learners compute these particular—or any—transition statistics. Instead, future research on statistical learning should attempt to characterize both human learning biases and the computations that give rise to them.

Acknowledgments

Thanks to Richard Aslin, Noah Goodman, Elissa Newport, Josh Tenenbaum, and Ed Vul for valuable discussion. MCF was supported by a Jacob Javits Graduate Fellowship and NSF DDRIG #0746251.

References

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

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 24–39.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Fourth International Conference on Spoken Language Processing.*

Frank, M., Goldwater, S., Griffiths, T., & Tenenbaum, J. (under review). Modeling human performance in statistical word segmentation.

Giroux, I., & Rey, A. (2009). Lexical and sub-lexical units in speech perception. *Cognitive Science*, *33*, 260–272.

Goldwater, S., Griffiths, T., & Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21–54.

Hauser, M., Newport, E., & Aslin, R. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*, 53–64.

Kuhl, P. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, *5*, 831–843.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms.* Cambridge, UK: Cambridge University Press.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*, 2745-2750.

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

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926.

Saffran, J., Newport, E., & Aslin, R. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, *35*, 606–621.

[1] Because both the TP model and the Lexical model produced a significant number of segmentations that failed to place any boundaries—for the TP model this was due to extreme threshold values, and for the Lexical model this was due to convergence issues in the online sampler we used—we excluded all model runs that failed to make any segmentation decisions.