Modeling online word segmentation performance in structured artificial languages

Stephan Meylan
smeylan@stanford.edu
Department of Psychology
Stanford University

Chigusa Kurumada
kurumada@stanford.edu
Department of Linguistics
Stanford University

Benjamin Börschinger
benjamin.borschinger@mq.edu.au
Department of Computing
Macquarie University

Mark Johnson
mark.johnson@mq.edu.au
Department of Computing
Macquarie University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Lexical dependencies abound in natural language. Words tend to follow particular words or word categories. Artificial language experiments exploring word segmentation generally lack such structure, however. In the present study, we explore whether simple inter-word dependencies influence the word segmentation performance of both adult learners and computational models. We use a continuous testing paradigm instead of an experiment-final test battery to reveal the trajectory of learning and to allow detailed comparison with the chosen models. Adult performance on languages with dependencies is equal or lower to those without dependencies. Three computational models capture this basic result, but vary in fit and qualitative similarity to human experimental results in different ways.

Keywords: Word segmentation; statistical learning; bigrams.

Introduction

Human learners can use distributional information to segment an unbroken speech stream into individual words after a short, ambiguous exposure (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996). Past artificial language learning experiments in this vein have typically generated words according to a uniform word frequency distribution and randomly concatenated word types to create the artificial languages. The process of learning to segment such synthesized languages may deviate significantly from learning to segment natural language, which contains asymmetric word frequency distributions, systematic dependency structure, and correlated variation in the lengths and frequencies of words.

In natural language, word-by-word transitions are governed by dependency relationships between word categories. For example, English prepositional phrases have the internal structure PP $\rightarrow$ P + NP, and an NP typically consists of a determiner and a noun. Since prepositions and determiners are lexical classes containing a relatively small number of short words, many instances of PPs often result in collocations used frequently in discourse (e.g., *in the house* or *on a map*).

Does such collocation structure make segmentation easier or more difficult? Both possibilities have some support in the literature. Frequent phrases are known to be problematic for segmentation mechanisms, especially for algorithms that rely on transitional probabilities (TPs). Due to high internal TPs, these phrases are often segmented as one unit, rather than separated into the multiple words that they contain (Goldwater, Griffiths, & Johnson, 2009). Thus, dependency structure has the potential to reduce segmentation performance.

In principle, however, there may also be the potential for increased performance in segmenting structured languages, especially if the learner is able to learn the dependency structure of the language along with the structure of individual words. For this reason, Goldwater et al. (2009)'s bigram model outperformed other segmentation models. Modeling dependency structure might also provide synergistic gains in learning: Johnson and Tyler (2010) found that an ideal learner that acquired words and word-object correspondences simultaneously was far more successful at both than each independently, but only when the learner assumed a rich collocation structure in the language.

Our current experiments test the relationship between dependency structure and segmentation performance for both human learners and computational models. We created languages with varied levels of category size asymmetry and test adult subjects' performance in word segmentation based on these languages. Two-alternative forced-choice tests administered after a discrete training phase, as used by past studies, do not produce an interpretable time course of learning. Thus, in Experiment 1 we use a new experimental paradigm that provides us with a time course of learning by testing subjects throughout the duration of exposure to stimuli (Kurumada, Meylan, & Frank, under review). In Experiment 2, we corroborate the results of the first experiment using a classic two-alternative forced choice paradigm. Both experiments, and a set of simulations with three computational models, show that asymmetric word-category sizes support adult segmentation learning considerably better than symmetric category sizes, but that performance on languages with dependencies is generally worse than performance on languages with randomly ordered words.

Experiment 1

To test the effects of dependency structure on word segmentation, we created two classes of artificial lexicons, one consisting of 12 word types and another of 8, and concatenated them to make languages with different dependency structures. Figure 1 is a diagram of the grammars that we used to produce the 12 word languages. Each sentence had four words. Three of the language types (which we refer to as "1515," "2424," and "3333") were generated via a simple finite state grammar, while the fourth (unstructured) language type was generated

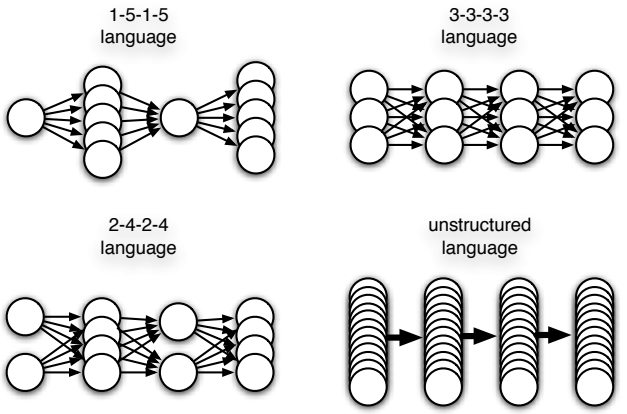

Figure 1: Schematic representations of the 4 types of languages used in Experiment 1. Circles represent individual words, arrows represent stochastic transitions.

by uniformly concatenating words with no notion of sentence position.

Each of the structured languages had a category structure such that sentences followed the form *ABCD*, where each category had a unique and non-overlapping set of words in it. The 1515 language, for example, had one word in the first category, five in the second, and so forth. We refer to such a language throughout as having "asymmetric" category sizes. In contrast, a 3333 language has "symmetric" category sizes. The 8 word languages were generated similarly, but only included three conditions ("1313," "2222," and unstructured).

We used a sentence-by-sentence orthographic segmentation paradigm (Kurumada et al., under review) to test adult learners' segmentation performance and gather detailed information about learning trajectories in these languages. Learners listened to a set of sentences; after each they were asked to click between syllables in a transcription to indicate where they thought word boundaries fell.

Methods

Participants For the 12 word languages, we posted 128 separate assignments on Amazon's Mechanical Turk (AMT) and received 124 assignments from distinct individuals. For the 8 word languages, we posted 96 assignments and received 95. Subjects were paid $1.00 for completing the task and were told that they would be paid an additional $1.00 if they performed in the top quartile of participants. The mean task duration was 14 minutes and 17 seconds.

Stimuli We created two classes of lexicons differing in the total number of word types (8 and 12). Words were made by concatenating 2 – 4 syllables from a randomly selected inventory of consonant-vowel syllables. Stimuli were synthesized with the MBROLA synthesizer (Dutoit, Pagel, Pierret, Bataille, & Van Der Vrecken, 1996) at a constant pitch of 100Hz with 225ms vowels and 25ms consonants. Each syllable appeared in only one word type in each lexicon.

For the 12-type condition, 4 languages were created by controlling how many of 12 types (distinct word forms) appeared in each of four sentences positions (Figure 1): 1515, in which the words in the first and third sentence position were drawn from categories of a single type and the second and the fourth from categories with five types each, 2424, in which the words in the first and third sentence position were drawn from categories consisting of two types and the second and fourth from categories with four types, 3333 in which the word in each sentence position was selected from a category with three types, and unstructured, in which four words were selected uniformly at random from a single category of 12 types (without replacement, to avoid in-sentence repetition).

To reflect the relationship between frequency and word length seen in natural language (Zipf, 1965), the category assignment of lexical items insured word length and frequency were inversely correlated (the shortest words appeared in the categories with the fewest types). Within each condition, 16 language variants were made using different phonemic inventories to control for unwanted phonological effects. The total number of word tokens per subject was 240 and the number of sentences was 60 in all languages. The total token frequencies of words were 60-12-60-12 in the 1515 condition, 30-15-30-15 in the 2424 condition, 20-20-20-20 in the 3333 condition; in the randomized condition each word appeared in each sentence position 5 times. Note that there was no discrete testing phase separate from the training phrase; rather, subjects were tested in the continuous paradigm on their knowledge of the language as they learned it.

Stimuli for the 8-type condition were similar to those in the 12-type condition except that we created only 3 languages: 1313, with the first and third word position drawn from categories with a single type and the third and fourth position drawn from categories with three types in each, 2222, in which each word position is selected from a category with two types, and randomized, where each sentence was composed from four words randomly selected from the total 8 word types. 32 phonetic variants of each language were generated to control for phonological effects. The per-subject exposure was 240 tokens over 60 sentences, and the total token token frequencies were 60-20-60-20 in the 1313 condition, and 30-30-30-30 in the 2222 condition; in the unstructured condition each word appeared about 7-8 times in each position.

Procedure Before the experimental trials began, participants were instructed to listen to and transcribe a short, common English word to confirm that their computer's audio system was working. Participants were then instructed that they would be presented with 60 consecutive sentences in a novel artificial language. They were informed that from the beginning they would need to click the boundaries between syllables presented on screen to indicate what they thought the boundaries between words were in each trial. While they would not know the words at first, they were told that they would be able to discern patterns and recognize at least some

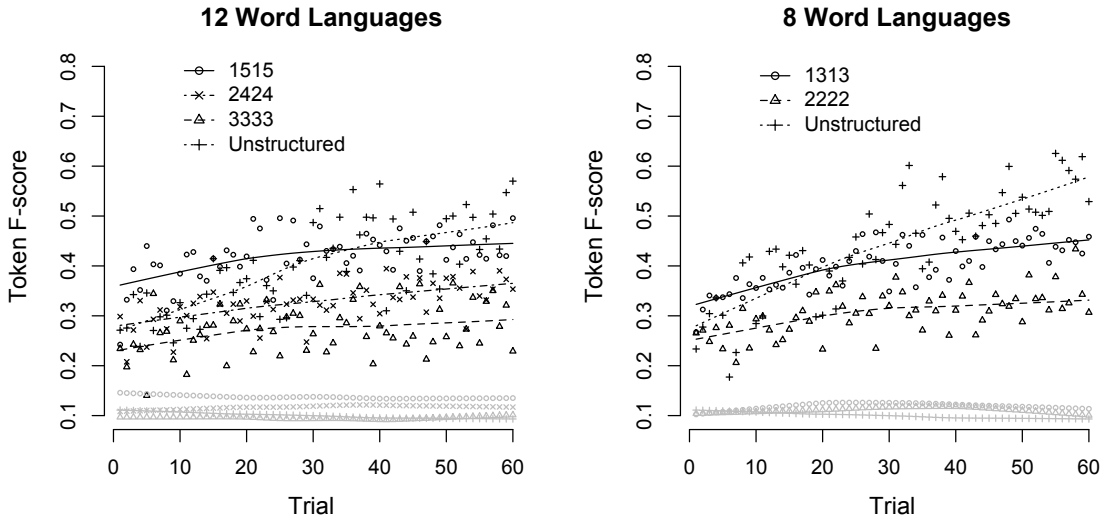

Figure 2: Token F-scores (a measure of segmentation performance for individual words) plotted for Experiment 1. Symbols represent mean performance for each condition (see legend) on a single trial, across participants. Black lines show a smoothed estimate of performance (using lowess). Gray lines at bottom show permutation baselines.

of the words as the experiment progressed. Each trial contained both an audio presentation and an orthographic transcription with selectable boundaries between each syllable. After clicking segments in a sentence subjects clicked a button marked "next" to proceed to the next trial.

Results and Discussion

To assess subjects' segmentation performance, we calculated the precision, recall, and F-score (the harmonic mean of precision and recall, as in Goldwater et al., 2009). This metric was computed by subject and trial, both for boundary decisions and tokens.[1] Figure 2 shows token F-score aggregated across subjects for each language and condition.

All conditions showed some learning. Unstructured languages were learned best, and learning was faster in the 8 word than the 12 word languages. Trajectories of the structured conditions maintained a relatively constant ordering: the more asymmetrical the categories in the language were, the better participants learned.

We analyzed token-by-token segmentation performance separately for the 8 word and the 12 word languages using mixed-effects logistic models (Gelman & Hill, 2006; Jaeger, 2008). The dependent variable in these models was whether a particular token had been segmented correctly. In the 12 word languages, we found a strong positive main effect of the log input frequency of that token ($\beta = .19$, $p < 10^{-6}$) and a negative effect of word length ($\beta = -.56$, $p < 10^{-6}$), as well as significant effects of the 3333 test condition ($\beta = -.69$, $p < .05$) and the 2424 test condition ($\beta = -.62$, $p < .05$) with respect to the unstructured condition. The 1515 condition also had a negative effect effect on token segmentation ($\beta = -.29$), though this was not reliably different than the unstructured condition ($p > .3$). As in the 12 word languages, in the 8 word languages there was a positive main effect of the log input frequency of that token ($\beta = .26$, $p < 10^{-6}$) and negative effect of word length ($\beta = -.68$, $p < 10^{-6}$). There were also significant effects of the 1313 condition ($\beta = -.71$, $p < .05$) and the 2222 condition ($\beta = -.95$, $p < .01$).

To summarize: in both the 12 word and the 8 word languages we observed the best performance in the unstructured condition and the worst performance in the condition where types are symmetrically distributed across categories/sentence positions.

Experiment 2

Experiment 2 was conducted to ensure that the effects of dependency structure could be captured in a classic two-alternative forced choice task, where subjects were asked to distinguish between a word from the language and a distractor.

Methods

Participants For the 12 word languages, 144 assignments were posted on Amazon Mechanical Turk, of which we received 133 from distinct individuals. For the 8-type condi-

[1] In our example sentence ("indiangorrillaseatbananas"), we compute these measures for a participant who gave the segmentation "indian|gorillas|eatbana|nas." Computing word boundaries, the participant would have 2 hits, 1 miss, and 1 false alarm, leading to precision of .66 (hits / (hits + false alarms)), and recall of .66 (hits / (hits + misses)), for an F-score of .66. On the other hand, for word tokens, the participant would have 2 hits ("indian" and "gorillas"), 2 misses ("eat" and "bananas") and 2 false alarms ("eatbana" and "nas"), for precision of .5, recall of .5, and F-score of .5.

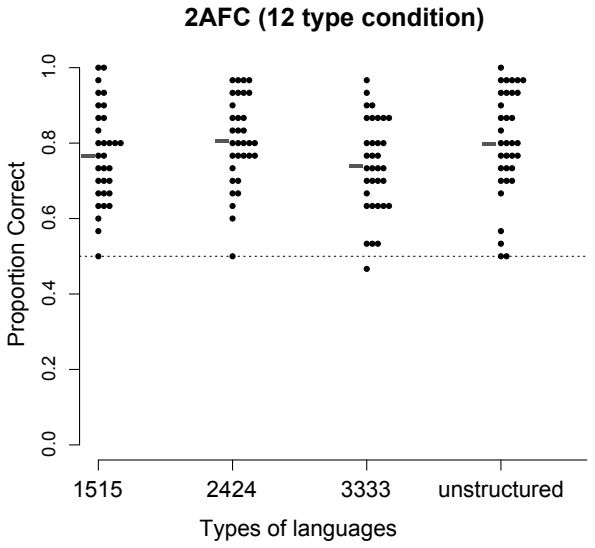

Figure 3: Proportion of correct two-alternative forced choice test trials for the languages in the 12-type condition in Experiment 2. Each point represents and individual participant, while the red bar shows the condition mean.

tion, 96 assignments were posted, of which we received 89. The same payment method as in Experiment 1 was used.

Stimuli The process of generating stimuli was nearly identical to the procedures presented in Experiment 1. For the training phase, 150 input sentences (totaling 600 tokens) were generated. A discrete test phase consisted of 30 pairs of a target word from the language and a length-matched distractor word, composed of syllables randomly selected from that language's syllabic inventory. Such test trials were intended to test if subjects had reliably learned which words belonged to the language.

Procedure Participants were instructed that they would listen to 150 sentences in a novel artificial language, then take a short test on what they'd learned about the language. Training sentences were presented aurally; when the audio file finished playing subjects needed to press "next" to advance to the next trial. In the test phase they were asked to choose which of the two options "sounded like it came from the language." For the test trials, the two options were presented aurally and subjects could replay the audio if desired.

Results and Discussion

In the 8 word langauges, performance was at ceiling, with all condition means above 90%. As such there was no meaningful variance across languages and we do not discuss this condition further.

In the 12 word languages, the pattern of mean performance across conditions replicated our findings in Experiment 1 (Figure 3). Although there was massive variation across participants, performance was higher in the unstructured and asymmetric conditions; performance was lowest in the symmetric condition.

We analyzed trial-by-trial 2AFC performance for the 12 word languages using a mixed-effects logistic model. There was a weak negative effect of trial number ($\beta = -.02$, $p < 10^{-4}$; recall all learning happened in this experiment after the trial phase) and word length, though unlike in the first experiment longer word length facilitated segmentation in the 2AFC ($\beta = .40$, $p < 10^{-13}$). The 3333 condition had a significant negative effect on performance ($\beta = -.38$, $p < .05$). Thus, 2AFC results support the negative effect of symmetrical category sizes seen in Experiment 1.

Online Segmentation Models

Our goal in modeling human performance in Experiment 1 was to understand whether the pattern of performance shown by humans reflects particular limitations on human learning. We thus selected a range of models that have been suggested by previous literature to fit human performance (Frank, Goldwater, Griffiths, & Tenenbaum, 2010): an incremental version of a transitional probability (TP) learner; PARSER, a heuristic, memory-based model (Perruchet & Vinter, 1998); and a new online implementation of a probabilistic segmentation model (Börschinger & Johnson, 2011).

Models and Parameters

TP-based model The transitional probability model is a boundary-finding approach that segments on the basis of statistically less likely transitions from syllable to syllable under the premise that within-word syllable TPs are higher than those at word boundaries (Saffran, Aslin, & Newport, 1996). In the present study, we calculate syllable bigram counts at the end of each sentence and calculate TP as

$$p(a|b) = \frac{c(a,b)}{\sum_{y \in V} c(a,y)} \quad (1)$$

where a and b are syllables, $c(a,b)$ is the count of the bigram ab, and V is all bigrams observed up to that point in the corpus. Sentence boundaries, which contain potentially useful information, were limited by a special symbol and treated as syllables. We systematically varied the threshold (0 – 1 in .1 increments) at which a TP was low enough to constitute a word boundary, and placed boundaries after all syllables where TP was below the threshold.

PARSER PARSER (Perruchet & Vinter, 1998) is a lexicon-finding model that makes use of a persistent collection of segments whose weights increase when encountered in the inputand decay with exposure to new data lacking such segments. Found items similar to ones that are already in the collection also prompt a decrease in the weights of similar segments. The PARSER model has a high number of adjustable parameters, including initial weight, max segment length for consideration, the decay rate of material in the

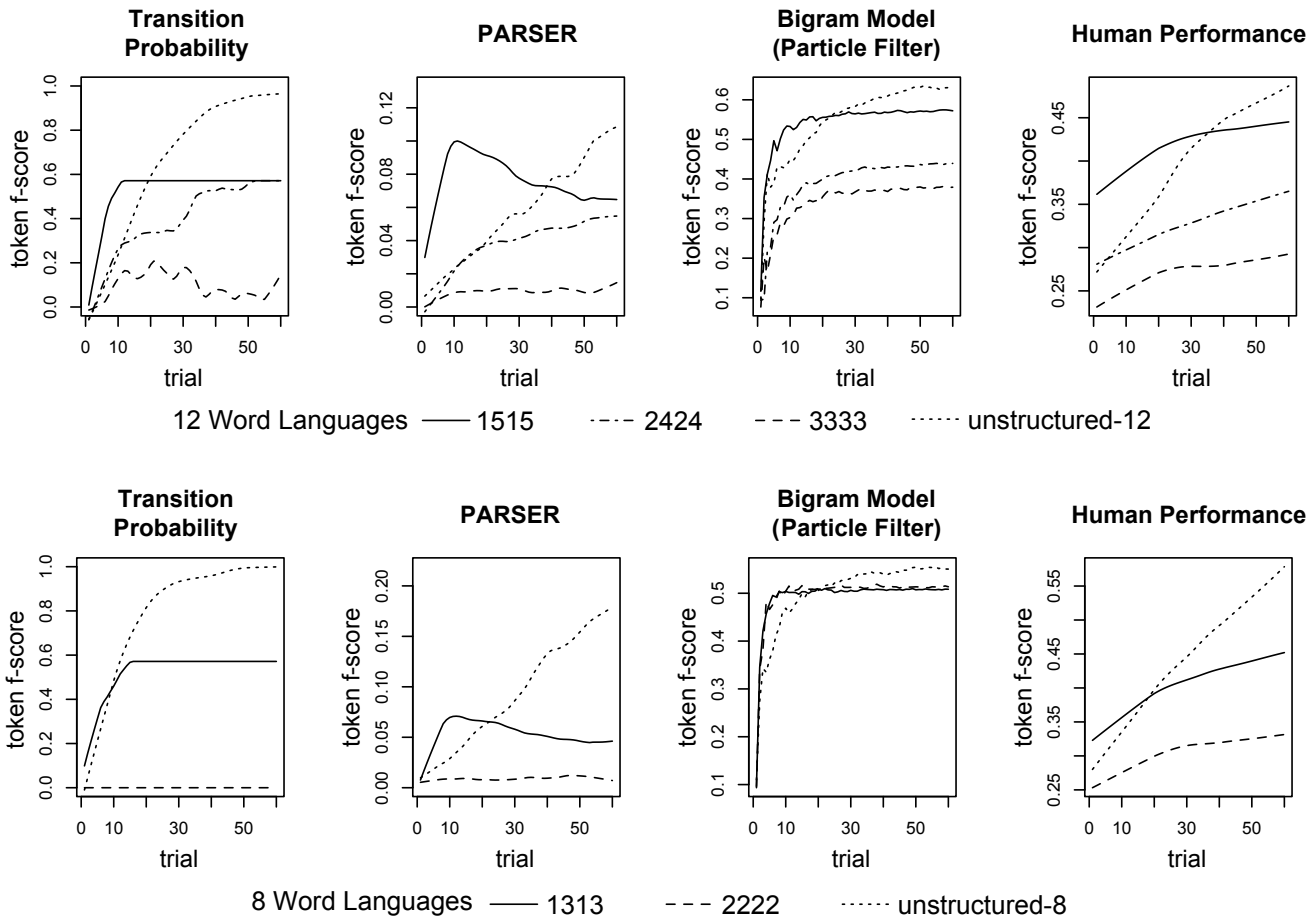

Figure 4: Mean token F-scores for the best-fitting parameter set for each of the three models on the 12-type and 8-type conditions of Experiment 1, plotted alongside human performance. Model results are scaled to human performance; as such, axes change from model to model.

stored lexicon, the interference weight, a threshold to determine whether a parse should be considered a new word, and the reactivation gain. We adapted the model to output sentence-by-sentence segmentations, taking the initial parse of a new sentence (guided by the weights of the collection in memory) as the output. For our simulations here, decay rate was most important, so we systematically manipulated this parameter. Among the other parameters, we used a max segment length for consideration of 3, an initial weight of 1, an interference rate of .005, a reactivation gain of .005, and a threshold of consideration as a new segment of 1.

Bigram Model with Particle Filter Inference The bigram model with particle filter inference is a new version of the Bayesian bigram model of Goldwater et al. (2009) that uses a particle filter, rather than a Gibbs sampler to estimate the posterior distribution over segmentations (Börschinger & Johnson, 2011). As in the original Bayesian bigram model, the lexicon for the particle filter is generated using a Dirichlet process, enforcing a probability distribution which gives higher probability to smaller lexicons with shorter words, and with a small number of high-probability collocations. Unlike the original version of the lexical model, however, the particle filter inference algorithm allows the model to be run incrementally via a single pass through a corpus, producing incremental segmentations for our evaluation along the way. In a particle filter, each particle constitutes a different set of segmentation hypotheses (see Börschinger & Johnson, 2011 for details). In our current study, we manipulated the number of different hypotheses that the model could track, with fixed concentration parameters of $\alpha_0 = 12$ and $\alpha_1 = 24$ in the 12 word languages and $\alpha_0 = 8$ and $\alpha_1 = 16$ in the 8 type languages.

Model Results

For each model, we fit a single free parameter to human data by choosing the value that maximized the proportion of the variance in the human data accounted for by that model, across all conditions and datapoints. For the TP model, that parameter was the threshold at which a transitional probability was considered a boundary; for PARSER it was the for-

getting rate; and for the particle filter bigram model, it was the number of particles. Figure 4 shows performance for the best-fitting parameter for each of the three models alongside the human data. The vertical scale of models is adjusted in order to emphasize the similarities and differences in the time-course of learning rather than the absolute level of performance achieved.

All models captured the final ordering of conditions with the exception of the particle filter in the 8 word languages, which failed to distinguish between the two conditions with dependencies, and the TP model in the 12 languages, which failed to distinguish between conditions with high and intermediate asymmetry in category sizes. In all models the asymmetric category size condition shows the most early learning, but is at some point overtaken by the unstructured condition. In both the TP model and the particle filter, learning in the asymmetric category size condition levels off around the tenth trial (of 60), with limited further gains in learning, a pattern not observed in the experimental data.

As a result of taking scores on the basis of model fit, however, there were large absolute differences in F-scores among some of the models: TP model F-scores were consistently too high (range: 0 – 1), PARSER F-scores were too low (0 – .2), and particle filter scores were approximately correct (0 – .6). Thus, although all of the models were able to capture the relative ordering of conditions, this came at the price of an absolute fit to the magnitude of the data. Importantly, the experimental method that we used allowed us to gather learning curves, and so—unlike previous work (Frank et al., 2010)—we were able to make absolute as well as relative comparisons between models and data.

General Discussion

We began by noting a difference between natural languages and the artificial languages that have previously been used to study the phenomenon of 'statistical word segmentation.' Natural languages have complex inter-word dependencies, while previous experiments have purposefully created languages that lacked such dependencies. While this initial simplification was useful, it is uncertain whether dependency structure would be positive or negative for segmentation performance. Our experiments suggest that adults learn better from a language *without* dependencies than one with them.

Our method of using learning curves provided for a higher-resolution examination of the dynamics of both human word segmentation and of models instantiating hypotheses about the mechanisms of segmentation. The measurement of learning across time allows for a richer investigation of the performance for each model, revealing the baseline from which models initialize, their rate of learning, and their final attainment after having been exposed to the full set of stimuli.

In our estimation of whether the mechanisms of 'statistical learning' can scale to the task of learning the lexicon of a natural language, we must take into account the difficulties posed by dependency structure. Nevertheless, our studies—both experimental and computational—showed that there was an advantage for languages with an asymmetrical dependency structures. Languages with variability in the number of types assigned to categories facilitated segmentation, perhaps by providing frequent tokens that served as anchors for segmentation (Kurumada et al., under review). Thus, as suggested by our previous work, it may be the case that high frequency material facilitates language learning by promoting segmentation of adjacent material.

Acknowledgments

Thanks to the members of the Language and Cognition Lab for valuable discussion.

References

Börschinger, B., & Johnson, M. (2011, December). A particle filter algorithm for bayesian wordsegmentation. *Proceedings of the Australasian Language Technology Association Workshop 2011*, 10–18.

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. In *Proceedings of the fourth international conference on spoken language* (Vol. 3, pp. 1393–1396). Philadelphia, PA.

Frank, M., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–25.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

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

Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Johnson, E., & Tyler, M. (2010). Testing the limits of statistical learning for word segmentation. *Developmental Science*, *13*(2), 339–345.

Kurumada, C., Meylan, S. C., & Frank, M. C. (under review). Zipfian frequency distributions facilitated word segmentation in context.

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

Saffran, J. R., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926.

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.

Zipf, G. (1965). *Human behavior and the principle of least effort: An introduction to human ecology*. New York, Hafner.