Zipfian word frequencies support statistical word segmentation

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

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

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

Abstract

Word frequencies in natural language follow a Zipfian distribution. Artificial language experiments that are meant to simulate language acquisition generally use uniform word frequency distributions, however. In the present study we examine whether a Zipfian frequency distribution influences adult learners' word segmentation performance. Using two experimental paradigms (a forced choice task and an orthographic segmentation task), we show that human statistical learning abilities are robust enough to identify words from exposures with widely varying frequency distributions. Additionally, we report a facilitatory effect of Zipfian distributions on word segmentation performance in the orthographic segmentation task, both in segmenting trained material and in generalization to novel material. Zipfian distributions increase the chances for learners to apply their knowledge in processing a new speech stream.

Keywords: Word segmentation; statistical learning; Zipfian frequency distributions.

Introduction

Humans and other animals extract information from the environment and represent it so that they can later use these representations for effective recognition and inference. One striking example of this phenomenon is that adults, children, and even members of other species can utilize statistical information to segment an unbroken speech stream into individual words after a short, ambiguous exposure (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Aslin, Saffran, & Newport, 1998; Hauser, Newport, & Aslin, 2001). In a now-classic segmentation paradigm, Saffran, Newport, and Aslin (1996) played adults a continuous stream of synthesized speech composed of uniformly-concatenated trisyllabic words. After exposure to this stream, participants were able to distinguish the original words from "part-words"—chunks that had occurred with lower frequency and lower statistical consistency. This work, combined with demonstrations with infants, suggested that statistical segmentation could be a viable method for early language learners to learn the word forms of their native language.

While the results of these experiments are impressive, the ways in which these findings can be applied to understand natural language learning are still unclear. Recent research has begun to close this gap. The outputs of this statistical segmentation process are now known to be good targets for word-meaning mapping (Graf Estes, Evans, Alibali, & Saffran, 2007), and experiments with natural language sample suggest that the processes observed in artificial language experiments generalize to highly-controlled natural language samples (Pelucchi, Hay, & Saffran, 2009). In addition, statistical segmentation has been shown robust to variation in sentence and word lengths (Frank, Goldwater, Griffiths, & Tenenbaum, 2010). Nevertheless, there are many links between statistical segmentation and natural language learning that still have not been made.

One key difference between standard segmentation paradigms and natural language is the distribution of frequencies. The empirical distribution of lexical items in natural language follows a Zipfian distribution (Zipf, 1965), in which relatively few words are used extensively ("the") while most words occur only rarely ("toaster"). In particular, the absolute frequency of a word tends to be approximately inversely proportional to its rank frequency. While Zipfian distributions are ubiquitous across natural language, their consequences for learning are only beginning to be explored (Goldwater, Griffiths, & Johnson, 2006). The current paper investigates the consequences of Zipfian frequency distributions for statistical word segmentation.

Saffran, Newport, and Aslin (1996) hypothesized that the mechanism underlying statistical word segmentation was the computation of syllable-syllable transitions. In a uniform distribution, nearly every word follows every other word, so these transition matrices are quite well-populated, but in a Zipfian language, they are very sparse. Some combinations of frequent words will have high transition probability between them (especially if they co-occur together frequently). If syllables are used in multiple words, the within-word transition probabilities for low-frequency words could be considerably lower than the between-word transition probability for high frequency words. This factor may have led to the low performance of transition-based models in computational comparisons (Yang, 2004; Brent, 1999). Thus, the first question we ask in the current study is whether human statistical learning abilities can succeed in segmenting Zipfian-distributed input.

Examining the problem from another side, however, a Zipfian distribution might actually provide *more* information for segmentation. Bannard and Lieven (2009) suggest that repetitive use of restricted types of words and word combinations in input are a strong predictor of the order of acquisition. In addition, six-month-olds can already exploit highly familiar words to segment and recognize adjoining unfamiliar words from fluent speech (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005). Thus, our second question is whether (and under what conditions) Zipfian input could facilitate word segmentation.

To address these two questions, we compared segmentation performance in uniform and Zipfian contexts across three different large-scale web-based segmentation experiments. Since Frank, Arnon, Tily, and Goldwater (2010) provided evidence that crowd-sourcing platforms reliably repli-

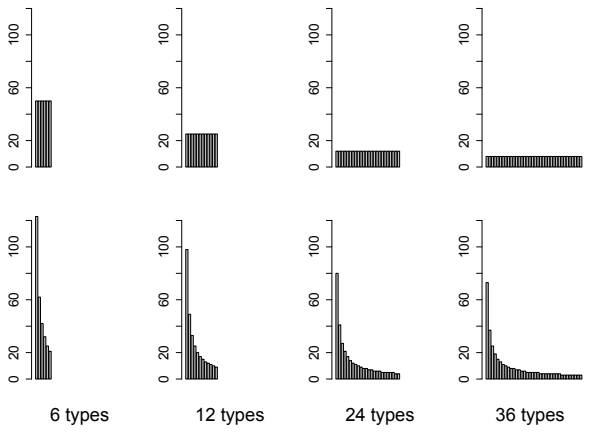

Figure 1: Word frequencies in uniform (top) and Zipfian conditions of Experiment 1.

cate lab-based experiments, we use this method to gather data across a wide range of experimental conditions. Experiment 1 tests participants in a standard 2-alternative forced-choice (2AFC) paradigm and manipulates the number of tokens in the languages used. Experiments 2 and 3 use an orthographic segmentation task and ask whether training on and testing on Zipfian-distributed materials lead to an advantage in segmenting previously heard and novel words.

Our results show that Zipfian distributions neither help nor harm segmentation performance in a traditional 2AFC paradigm (Experiment 1). In the orthographic paradigm, however, the Zipfian word frequency benefitted learners by providing them with more chances to segment familiar words (Experiment 2), which in turn helped to individuate neighboring words in the speech stream (Experiment 3). These data suggest that Zipfian frequency distributions have a scaffolding effect on segmentation that manifests at the stage where learners use acquired knowledge to segment new sentences.

Experiment 1

In Experiment 1, we use 2AFC test trials where participants are asked to distinguish a word from a part-word to test whether adult learners can learn words from input following uniform and Zipfian distributions. One additional novel feature of this experiment is that we vary the number of word *types* (distinct word forms) in the experiment from 6 all the way to 36, far higher than previously tested (Frank, Goldwater, et al., 2010). Thus, a subsidiary question is whether participants are able to identify words at above-chance levels in these more challenging environments.

Methods

Participants We posted 259 separate HIT (Human Intelligence Taks: experimental tasks for participants to work on) on Amazon's Mechanical Turk. We received 246 HITs from distinct individuals (a mean of 30 for each token frequency and distribution condition).

Stimuli We constructed 8 language conditions by controlling patterns of frequency distribution (uniform vs. Zipfian) and the numbers of word types contained in lexicon (6, 12, 24, 36 types). Within each language condition, we created 16 language variants with different phonetic material. This diversity was necessary to ensure that results did not include spurious phonological effects.

Words were created by randomly concatenating 2, 3, or 4 syllables (word lengths were evenly distributed across each language). Stimuli were synthesized using MBROLA (Dutoit, Pagel, Pierret, Bataille, & Van Der Vrecken, 1996) at a constant pitch of 100Hz with 225ms vowels and 25ms cosnonants. Each syllable was used only once. Sentences were generated by randomly concatenating words into strings of four words. The total number of word tokens was 300 and the number of sentences was 75 in all the languages. The token frequencies of words in each language were either distributed uniformly according to the total type frequency (e.g. 50 tokens each for a language with 6 word types) or given a Zipfian distribution such that frequency was inversely proportional to rank ($f \propto 1/r$). Frequency distributions for each language are shown in Figure 1.

For the test phase, "part-words" were created by concatenating the first syllable of each word with the remaining syllable of another word. These part-words were used as distractors which appeared in the training input with lower frequency than the target words.

Procedure Before the training phase began, participants were instructed to listen to a simple English word and type it in to ensure the sound is properly played and perceived. Participants then moved to the training phase, where they were instructed to listen to and learn a made-up language which they would later be tested on. To ensure compliance with the listening task for the duration of the training phase subjects needed to click a button marked next to proceed though the training sentences. In the test phase of the 2AFC condition, participants heard 24 pairs of words consisting of a target word and a length matched "part-word." After listening to each word once, they clicked a button to indicate which one sounded more familiar (or "word-like") given the language they had learned.

Results and Discussion

Figure 2 illustrates accuracy of responses in the 4 types of languages in the each of the uniform and Zipfian distribution conditions. There was not a strong numerical effect of distribution condition. Replicating previous results (Frank, Goldwater, et al., 2010), performance decreased as the number of types increased, but participants performed slightly above chance even in the most difficult 36 type condition.

Our analysis used mixed effects logistic regression (Gelman & Hill, 2006) fit to the entire dataset. This model attempted to predict performance on individual trials; we used model comparison to find the appropriate predictors. Our first model included effects of distribution and number of types;

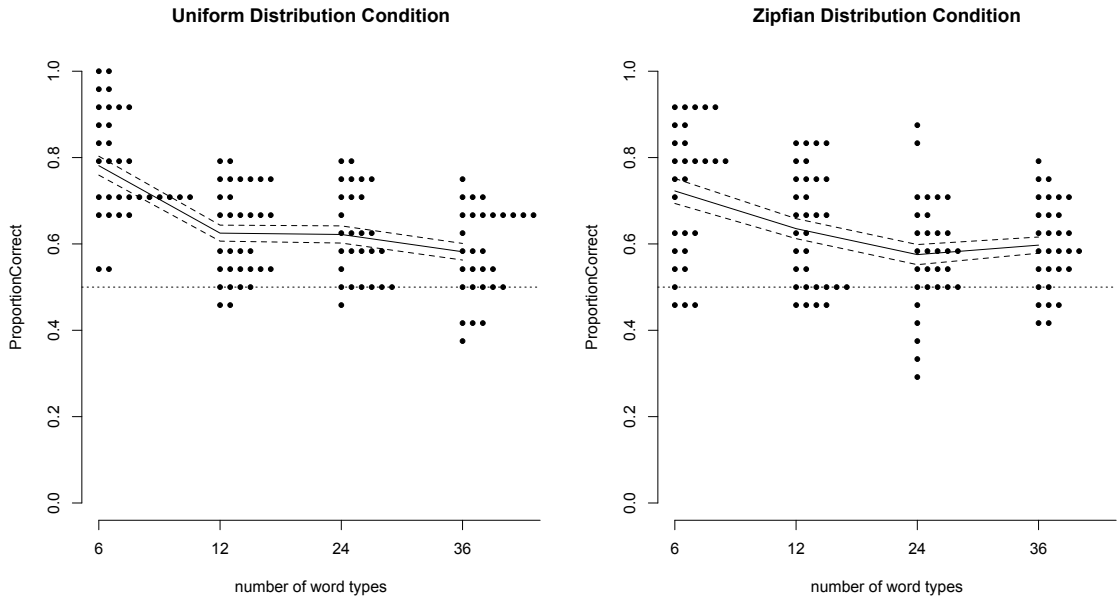

Figure 2: Average proportion correct trials by number of word types in the uniform and Zipfian distribution conditions. Dots represent individual participants and are stacked to avoid overplotting. Solid line represents means, dashed lines represent standard errors, and the dotted line represents chance (50%).

we found no effect of distribution ($\beta = -.226$, $p = .12$) but a highly significant effect of number of types ($\beta = -.021$, $p < .0001$). Further exploration revealed that better model fit was given by the logarithm of number of types as a predictor rather than raw number of types ($\chi^2 = 9.49$, $p < .0001$). Thus, the log number of types was the only significant predictor of performance in this model.

In our second set of models, we introduced as additional trial-level predictors the frequency of the target and distractors for each trial (calculated from the input corpus for each language). In this model, we found that once these factors were added, there was no gain in model fit from log number of types ($\chi^2(1) = .11$, $p = .74$). Instead, the only significant effects were a positive coefficient on log tokens (the more times a word is heard, the better performance gets: $\beta = .35$, $p < .0001$), a negative coefficient on log distractor tokens (the more times a distractor is heard in the corpus, the worse performance gets: $\beta = -.51$, $p = .003$) and a positive interaction of the two (bad distractors are worse if the target is low frequency: $\beta = .14$, $p = .003$). The general relation here is plotted in Figure 3, showing mean proportion of accuracy according to the input frequency of the target words. In this final model, there was still no effect of distribution conditions (i.e., uniform vs. Zipf) ($\beta = .05$, $p = .49$).

To summarize: participants represented target words equally well after being exposed to languages with radically different frequency distributions and contingency statistics. The only factors that mattered in 2AFC test trials were the log frequency of targets and distractors, independent of what context they were heard in. In a Zipfian condition, some words have significantly higher and lower frequency than those in a uniform condition, which could create a skewed distribution of transition probabilities between lexical items. However, our results indicate that 2AFC accuracy for a word is predicted based predominantly on the (uni-gram) word frequency in input, not on the distribution of the context.

Experiment 2

Experiment 2 tests our hypothesis about possible facilitative effect of a Zipfian word distribution on segmentation of a speech stream via a different method. Because a 2AFC asks only about a comparison between a particular target-distractor pair, we hypothesized that effects of distribution might be more obvious in a paradigm where words were presented in context during testing. To test this hypothesis, we use an orthographic segmentation task developed by Frank, Goldwater, et al. (2010). In this task, participants were trained on either a Zipfian or a uniform distributions and later asked to give explicit judgments as to where in a sentence they would place word boundaries. We predicted that, in this paradigm, participants would be able to segment speech more accurately when exposed to a Zipfian distribution during training and test.

Methods

Participants 149 separate HITs were posted on Mechanical Turk. We received 127 complete HITs from distinct individuals. Participants were paid $0.50 for participation. To ensure participants' attention to the task, we applied an incentive payment system where participants were told they would receive an additional $1.00 if they scored in the top quartile.

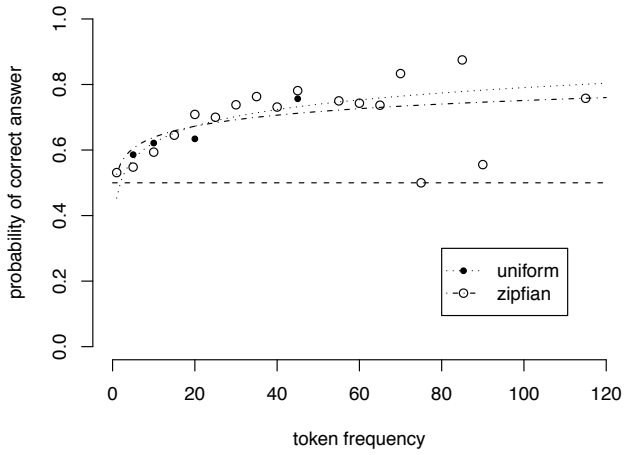

Figure 3: Probability of a correct 2AFC answer plotted by binned token frequency. Filled circles indicate uniform condition, while open circles indicate Zipfian condition. Dashed line shows chance, while the dotted and alternating lines give best fit lines for performance as a function of log token frequency.

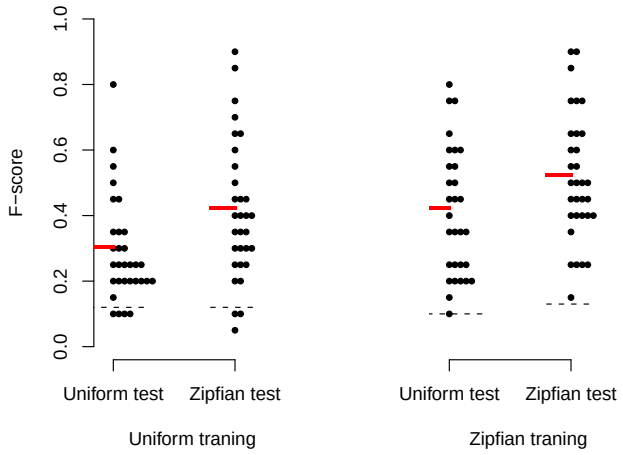

Figure 4: Token F-scores (a measure of segmentation performance for individual words) plotted for each condition of Experiment 2. Points represent individual participants, bars show means, and dashed lines show a permutation baseline.

Stimuli The process of generating stimuli was nearly identical to the 6 word condition in Experiment 1. Six words were generated following either a uniform or Zipfian distribution. Six hundred word tokens were presented in 150 sentences in the training phase. For the test phase, 10 additional sentences were created according to one of the two frequency distributions; the same lexicon was used to generate the training corpus.

To examine the effects of frequency distribution at the different stages of segmentation, we applied a 2x2 factorial design. Subjects were divided into four groups according to the frequency distributions at the training phase (uniform vs. Zipfian training) and at the test phase (uniform vs. Zipfian test). In each case, sentences were generated by sampling words from either a uniform frequency distribution or one that was generated via the same classic Zipfian formulation given above ($f \propto 1/r$).

Procedure The training section of this experiment was identical to that of Experiment 1 (though twice as long; approximately 15 minutes). In the test phase, participants were asked to click on the breaks between syllables to indicate word boundaries. They were given one practice trial on an English sentence presented in the same format and prevented from continuing until they segment it correctly. At test, sentences were presented visually, with each syllable separated by a toggleable button. All the syllables were spelled with two letters representing a consonant and a vowel respectively (e.g., *ka*, *pi*, *ta*, *bu*). Each sentence was also played back at the beginning of the trial.

Results and Discussion

To evaluate participants' segmentation performance, we relied on precision and recall, and their harmonic mean, F-score. These same measures are used in computational studies of segmentation and in previous work (Goldwater, Griffiths, & Johnson, 2009; Brent, 1999). We computed precision and recall for both boundaries and for word tokens.[1] Token F-scores in each condition are plotted in Figure 4. We removed 3 participants with F-scores of exactly 0 or 1. As in Frank, Arnon, et al. (2010), we calculated an empirical baseline for each measure via permutation. We repeatedly shuffled boundary placement responses for each sentence and computed the same measures (precision, recall, and F-score). The mean values of baseline token F-scores in each condition are indicated as dashed lines in Figure 4.

Because participant mean F-scores were normally distributed but trial-level F-scores were not, and because we had no trial-level predictors in this experiment, we used a simple linear model to predict participants' mean token F-scores. We found reliable main effects of the training condition ($\beta = 0.12$, $p = .02$) and the test condition ($\beta = 0.11$, $p = .02$) and no interaction ($\beta = -0.017$, $p = 0.80$). The boundary scores exhibited the same patterns and the same

[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 prevision of .5, recall of .5, and F-score of .5.

Table 1: Mean token F-scores and boundary F-scores for overall segmentation performance in Experiment 2.

Input	Test	Token F	Boundary F
Uniform	Uniform	0.29	0.52
	Zipf	0.41	0.61
Zipf	Uniform	0.41	0.63
	Zipf	0.51	0.68

pattern of statistical significance (see Table 1 for means): main effects of training ($\beta = .10$, $p = .02$) and test ($\beta = .08$, $p = .04$), and no interaction ($\beta = -.03$, $p = .64$).

The critical finding in this experiment is that even when participants were tested with uniform materials, if they had been trained on Zipfian-distributed items, they still performed better. Participants used their knowledge of the high-frequency items they learned during training to help segment lower-frequency items, bringing total performance above performance in the uniform-uniform condition. The additional effect of Zipfian testing materials then follows logically. A Zipfian distribution at test makes over-learned, high-frequency items even more prevalent, and increases the chance that they are adjacent to unlearned words.

Due to the small number of word types in Experiment 2, even the low frequency items were still heard 40 times. Thus, in Experiment 3 we test the hypothesis that Zipfian contexts support better segmentation of truly novel material.

Experiment 3

Experiment 3 replicated Experiment 2, but for each test sentence, we added a single novel item. If identification of familiar words improves segmentation accuracy of adjoining words, we would expect better identification of novel words when participants were both trained and tested using Zipfian materials.

Methods

Participants 158 separate HITs were posted on Amazon's Mechanical Turk. We received 121 complete HITs from distinct individuals. Participants were paid $0.50 for their participation and we again added bonus payments for participants in the top quartile.

Stimuli Sentences for training phase were created identically to Experiment 2. At test, we generated 10 new words of varying length (2, 3, and 4 syllables) based on syllables that appeared in the training corpus. To ensure each syllable was used only once in the lexicon despite the enlarged lexicon (6 training items + 10 novel test items), an additional vowel was added to the phonemic inventory. We added one new word in a sentence-internal position in each test sentence. With the additional word, there were 5 test sentences of length 4 and 5 test sentences of length 5.

Procedures Procedures were identical to Experiment 2.

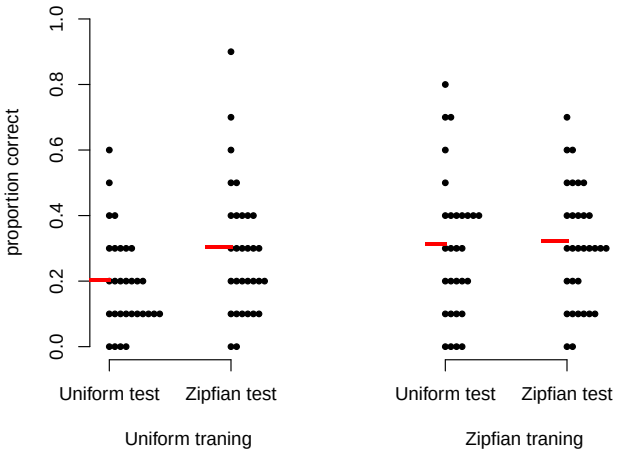

Figure 5: Proportion correct for the identification of new words in Experiment 3; plotting conventions are as in Figure 4.

Results and Discussion

Token and boundary F-scores for overall segmentation performance are shown in Table 2. We again fit linear models to the token and boundary F-score data. In token F-score, the effect of training condition was slightly attenuated ($\beta = .08$, $p = .07$) but the effect of test condition was still equivalent ($\beta = .13$, $p = .005$), and there was still no significant interaction term, though there was a negative coefficient value, indicating some sub-additivity ($\beta = -.07$, $p = .29$). Both training ($\beta = .10$, $p = .02$) and test ($\beta = .089$, $p = .03$) effects were still significant in boundary F-score data , and there was no significant interaction, though the coefficient was again negative ($\beta = -.047$, $p = .42$).

We next analyzed generalization data: we coded each of the ten generalization trials (one novel word per sentence) as a binary variable: 1 if the word was segmented correctly, 0 otherwise. Participant means are plotted in Figure 5. We then used a mixed logistic model to predict this variable on the basis of training and test condition, including a random effect of participant. (We chose a mixed model here in order to avoid the issue of computing a linear regression over a non-normally distributed DV). As in the overall test (and Experiment 2), we found main effects of training ($\beta = .61$, $p < .02$) and ($\beta = .57$, $p = .03$), with a negative but non-significant interaction term ($\beta = -.52$, $p = .15$).

To summarize: Experiment 3 replicates the findings from Experiment 2 and highlights a benefit of segmentation within a Zipfian language: if a learner hears a novel word, that word is much more likely to be flanked by known words.

General Discussion

The results of three experiments indicate that a Zipfian distribution of word frequency lends support to statistical word

Table 2: Mean token F-scores and boundary F-scores for overall segmentation performance in Experiment 3.

Input	Test	Token F	Boundary F
Uniform	Uniform	0.26	0.47
	Zipf	0.39	0.56
Zipf	Uniform	0.35	0.57
	Zipf	0.41	0.61

segmentation by scaffolding learners active use of acquired knowledge. In Experiment 1, we found that learning performance in a 2AFC task was neither helped nor hurt by a Zipfian frequency distribution. While participants could have used high frequency words to figure out adjacent material during training, they appear not to have. Instead, the only factor that predicted learning was a word's log frequency.

In contrast, in Experiments 2 and 3, when participants needed to engage actively in segmenting a sentence in our orthographic segmentation paradigm, we saw reliable effects of Zipfian training and testing materials. Crucially, even when participants were tested on uniform sentences, they were still able to use the knowledge acquired in the Zipfian training condition to succeed. The only way they could have done this is to use higher-frequency, over-learned items to segment out lower-frequency (Experiment 2) or novel (Experiment 3) words at test. Inferences using well-known neighboring materials for segmentation are part and parcel of what it means to segment a sentence. In general, when learners know a single word well, it can provide extra leverage in segmenting the entire sentence, including any novel material. In contrast, in the word/part-word comparison paradigms that have traditionally been used to evaluate segmentation accuracy (Saffran, Newport, & Aslin, 1996), this benefit is absent.

Our results provide evidence that the frequency structure of natural language input provides a natural scaffolding for statistical word segmentation. In the past 15 years, studies on distributional learning have established that some aspects of early word learning can be viewed as resulting from learners capacity for statistical learning. The current study suggests that the process of word segmentation in naturalistic contexts may be supported by the structure of the environment. We hope that future research continues to investigate aspects of artificial languages in order to explore the interaction of human cognition and the natural language learning environment.

Acknowledgments

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

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321-324.

Bannard, C., & Lieven, E. (2009). Repetition and reuse in child language learning. *Formulaic language*, *2*.

Bortfeld, H., Morgan, J., Golinkoff, R., & Rathbun, K. (2005). Mommy and me. *Psychological Science*, *16*(4), 298.

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 speechsynthesizers 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., Arnon, I., Tily, H., & Goldwater, S. (2010). Beyond transitional probabilities: Human learners impose a parsimony bias in statistical word segmentation. In *Proceedings of the 31st annual meeting of the cognitive science society.*

Frank, M., Goldwater, S., Griffiths, T., & Tenenbaum, J. (2010). Modeling human performance in statistical word segmentation. *Cognition*.

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

Goldwater, S., Griffiths, T., & Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems 18* (pp. 459–466). Cambridge, MA: MIT Press.

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

Graf Estes, K. M., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? *Psychological Science*, *18*(3), 254.

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.

Pelucchi, B., Hay, J., & Saffran, J. (2009). Statistical learning in a natural language by 8-month-old infants. *Child development*, *80*(3), 674–685.

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.

Yang, C. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences*, *8*(10), 451–456.

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