Zipfian frequency distributions facilitate word segmentation in context

Chigusa Kurumada[a,*], Stephan C. Meylan[b], Michael C. Frank[b]

[a] *Department of Linguistics, Stanford University, United States*
[b] *Department of Psychology, Stanford University, United States*

Abstract

Word frequencies in natural language follow a highly skewed Zipfian distribution, but the consequences of this distribution for language acquisition are only beginning to be understood. Typically, learning experiments that are meant to simulate language acquisition use uniform word frequency distributions instead. We examine the effects of Zipfian distributions using two artificial language paradigms—a standard forced-choice task and a new orthographic segmentation task in which participants click on the boundaries between words in contexts. Our data show that learners can identify word forms robustly across widely varying frequency distributions. In addition, although performance in segmenting individual words is driven solely by their frequency, a Zipfian distribution facilitates word segmentation in context: The presence of high-frequency words creates more chances for learners to apply their knowledge in processing new sentences.

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

1 Introduction

Humans and other animals extract information from the environment and represent it so that they can later use the knowledge for effective recognition and inference (Fiser, 2009). One striking example of this phenomenon is that

[*] Corresponding author. Address: Department of Linguistics, Stanford University, 450 Serra Mall, Margaret Jacks Hall (Building 460), Stanford, CA 94305, United States.

Email address: kurumada@stanford.edu (Chigusa Kurumada) adults, children, and even members of other species can utilize distributional information to segment an unbroken speech stream into individual words after a short, ambiguous exposure (Saffran et al., 1996a,b; Aslin et al., 1998; Hauser et al., 2001). In a now-classic segmentation paradigm, Saffran et al. (1996b) 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"—length-matched strings that also occurred in the exposure corpus, albeit with lower frequency and lower statistical consistency. This work on "statistical learning," combined with similar demonstrations with infants, suggests that learners can use the statistical structure of sound sequences to find coherent chunks in unsegmented input (Chomsky, 1955; Harris, 1955; Hayes & Clark, 1970; Wolff, 1977; Pinker, 1984).

While the results of statistical learning experiments are impressive, it is still unknown how these findings relate to natural language learning (Yang, 2004; Johnson & Tyler, 2010). Recent research has begun to close this gap. The outputs of the statistical segmentation process are now known to be good targets for word-meaning mapping (Graf Estes et al., 2007), and experiments with natural languages suggest that the processes observed in artificial language experiments generalize to highly controlled natural language samples (Pelucchi et al., 2009). In addition, statistical segmentation has been shown to be scaled up to variation in sentence and word lengths (Frank et al., 2010b) as well as to larger lexicons (Frank et al., under review). Nevertheless, there are many links between statistical segmentation and natural language learning that need to be tested.

One key difference between standard segmentation paradigms and natural language is the distribution of word frequencies. The empirical distribution of lexical items in natural language follows a Zipfian distribution (Zipf, 1965), in which relatively few words are used extensively (e.g., "the") while most words occur only rarely (e.g., "toaster"). In a Zipfian distribution, the absolute frequency of a word is inversely proportional to its rank frequency. For this reason, this kind of distribution is often characterized as having "a long tail," in which a small number of word types have very high token frequencies while many more types have relatively low frequencies.[1] While Zipfian distributions

[1] Here and below, we make use of the distinction between word *types*—distinct word forms—and word *tokens*—individual instances of a type.

are ubiquitous across natural language,[2] their consequences for learning are only beginning to be explored (Yang, 2004; Goldwater et al., 2006; Mitchell & McMurray, 2009; Ellis & O'Donnell, 2011).

An early and influential proposal suggested that learners could succeed in statistical segmentation tasks by computing the transitional probability (TP) between syllables (Saffran et al., 1996b). Learners could then posit boundaries between units in the speech stream where TP was especially low. (The underlying intuition is that minima in TP are likely to occur at word boundaries because there is uncertainty in what words follow other words, while within words the order of syllables is predictable.) In experiments on segmentation, stimuli are generally created by randomly concatenating a small set of words with a uniform frequency distribution so that every word follows every other word, ensuring that transition matrices between individual syllables are well-populated (Saffran et al., 1996a,b; Frank et al., 2010b). Thus, in standard experiments, comparisons between TPs are easy to make because all transitions were well-estimated.

In a Zipfian language, however, the same TP procedure would result in highly sparse transition matrices. A majority of words are infrequent (e.g., "toaster" or "lucubrated") and their combination will be vanishingly rare, while some combinations of frequent words have high transitional probability between them (e.g., "of the" is very high) despite the presence of a word boundary. In fact, given the collocational structure of natural language (Goldwater et al., 2009), the within-word transitional probabilities for low-frequency words can potentially be lower than the between-word transitional probability for high-frequency words. When transitional probability models are instantiated computationally and applied to corpus data, they perform very poorly both in absolute terms and in comparison to other models (Yang, 2004; Brent, 1999). The sparsity of transition matrices in Zipfian languages may be to blame.

The poor performance of TP-style models in corpus evaluations leaves open two theoretical possibilities for human learners. First, human learners may use statistical learning mechanisms (which compute TPs) only to learn a small set of word forms, and hence they may not need to be particularly

[2] Zipfian distributions are ubiquitous across many other phenomena (e.g., city populations) as well; even randomly generated texts exhibit a Zipfian word frequency distribution (Li, 1992). Here we take it for granted that natural languages have this structure without attempting to explain its presence.

effective (Swingley, 2005). This view is consistent with a large body of evidence suggesting that infants quickly learn to make use of lexical, prosodic, and phonotactic cues for segmentation (Mattys & Jusczyk, 2001; Jusczyk et al., 1999; Johnson & Jusczyk, 2001; Blanchard et al., 2010; Shukla et al., 2011). This viewpoint—that a heuristic, TP-based strategy allows learners to begin the segmentation process—seems to support the general prediction that segmentation should be more difficult (or at very least, not facilitated) by Zipfian frequency distributions because of the use of TPs.

Second, learners may rely on a more robust statistical learning method. In fact, non-TP computational proposals for statistical learning make a different prediction for segmentation performance in Zipfian environments. Orbán et al. (2008) propose a distinction between transition-finding models (like TP models) and "chunking" models, which look for a partition of the input stream into statistically coherent sequences. A number of recent models of word segmentation fall into the chunking category, including incremental (Brent & Cartwright, 1996), Bayesian (Brent, 1999; Goldwater et al., 2009), and memory-based (Perruchet & Vinter, 1998) models. These models (and some corresponding psychological evidence) suggest that segmentation performance should be robust to—or even facilitated by—Zipfian distributions.

One reason that Zipfian distributions might facilitate segmentation in a chunking model is because the frequent repetition of words in Zipfian languages could help learners remember them. Some chunking models hypothesize that learners store word representations in memory and match these memory representations up with the input to segment new utterances. In these models, stored representations will decay unless the corresponding word is heard frequently (Perruchet & Vinter, 1998). A Zipfian distribution makes it highly likely that a few of the most frequent words appear consistently across sentences, guaranteeing that at least a few words will be learned and retained with high reliability.

Another method by which Zipfian distributions might facilitate segmentation is via a bootstrapping effect. If a novel word occurs adjacent to a familiar word, it may be segmented more effectively because one boundary is already known. A Zipfian distribution would facilitate a bootstrapping effect because a small number of high-frequency words provide known context for many low-frequency words. We distinguish *contextual bootstrapping*—in which hearing the word sequence ABC, containing the known word A and novel word B, facilitates the identification of B in the future—and *contextual facilitation*—in which B is better segmented in this string due to the adjacency of A but is not necessarily segmented more accurately in the future.

Brent & Cartwright (1996) proposed a contextual bootstrapping model called INCDROP that segmented utterances by detecting familiar items and recognizing them as meaningful chunks, while storing the remaining chunks of the utterance as novel words. For example, if *look* were recognized as a familiar unit in the utterance *lookhere*, then the remaining portion, *here*, would be inferred as a potential lexical unit. This model, and a number that have followed it (Brent, 1999; Goldwater et al., 2009; Perruchet & Vinter, 1998), make use of contextual bootstrapping in more or less direct ways, but all suggest that knowledge of familiar words should help in recognition of new ones.

Several psychological studies have tested, with mixed results, whether known words facilitate the segmentation of nearby words. Dahan & Brent (1999) tested for contextual bootstrapping effects in adult word segmentation experiments and found some evidence for them, although primarily at the beginnings and ends of sentences. Bortfeld et al. (2005) found that 6-month-olds were able to find new words more easily when they were presented adjacent to words that were already familiar to them (e.g., the child's own name). Hollich et al. (2001), however, failed to find evidence that a familiar context (e.g., words like "flower") aided 24-month-olds in segmenting new words.

Isolated words are also often assumed to create a strong contextual bootstrapping effect, and a number of studies have investigated their role in segmentation. Brent & Siskind (2001) found that 9% of caregiver utterances consisted of words produced in isolation, and 27% of these cases were immediate repetition of words used in neighbouring utterances (e.g., "Want some milk? Milk?"). Building on this descriptive work, experimental evidence suggests that exposure to words in isolation establishes familiarity with these words, which serve as "anchors" in subsequent segmentation (Conway et al., 2010; van de Weijer, 2001; Cunillera et al., 2010; Lew-Williams et al., 2011). Thus, several lines of empirical work point toward a potential advantage of a Zipfian distribution, where a limited number of words readily acquire familiarity due to their disproportionate input frequencies.

To summarize, previous literature leaves us with two different predictions about the effects of the Zipfian word frequency distribution in natural language on word segmentation performance. Under heuristic transition-finding models, Zipfian distributions provide sparser input, making the segmentation problem more difficult. Under chunk-finding models, Zipfian distributions provide frequent chunks that may even allow learners to engage in contextual bootstrapping: using known contexts to segment novel words more effectively. We present data from two experiments investigating adult learners' performance in artificial language word segmentation tasks that compare Zipfian and uniform frequency distributions. Our data show that learners can identify words in languages with widely varying frequency distributions, consistent with models of segmentation that posit a frequency-based chunking procedure. In addition, our data suggest that Zipfian languages provide a specific advantage for word recognition in context: in such languages, new words tend to occur next to high-frequency words that are already known.

2 Experiment 1

We first asked whether learners could learn the forms of words from unsegmented input with a Zipfian word-frequency distribution. To test this question, we made use of the paradigm originated by Saffran et al. (1996b) to measure statistical word segmentation in adult learners. In this paradigm, learners listen passively to a sample of unsegmented, monotone synthesized speech and then are asked to make two-alternative forced-choice judgments about which of two strings sounds more like the language they just heard. We used the version of this paradigm adapted by Frank et al. (2010b), which includes several features of natural language, such as silences between sentences and words of varying lengths.

2.1 Methods

2.1.1 Participants

We posted 259 separate HITs (Human Intelligence Tasks: experimental tasks for participants to work on) on Amazon's Mechanical Turk service. We received 246 HITs from distinct individuals (a mean of 30 for each token frequency and distribution condition). Participants were paid $0.75 and the task took approximately 7–10 minutes.

2.1.2 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.

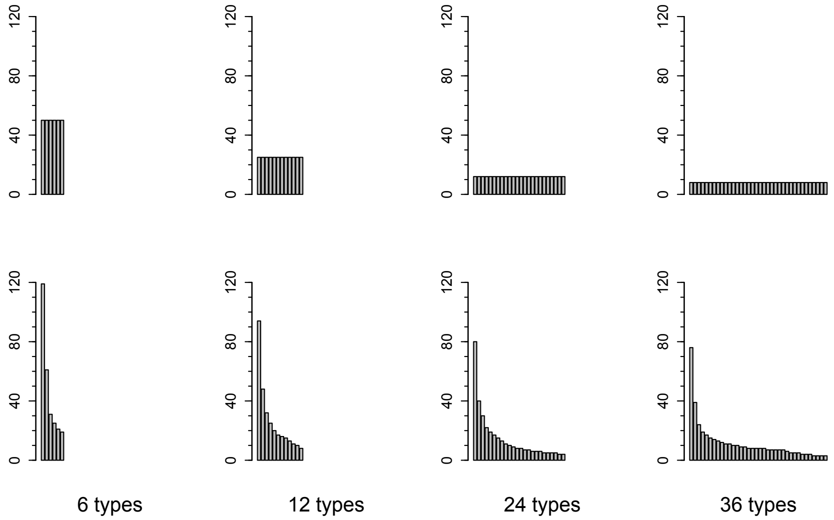

Figure 1: Word frequencies in uniform (top) and Zipfian (bottom) conditions of Experiment 1. The horizontal axis shows distinct word types, and the vertical axis shows the frequency of each of these types.

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 et al., 1996) at a constant pitch of 100 Hz with 225 ms vowels and 25 ms consonants. Each syllable was used in one word only. 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, a set of length-matched "part-words" were created for each word by concatenating the first syllable of the word with the last syllables of another word. These part-words were used as distractors; they appeared in the training input but with lower frequency than the target words.

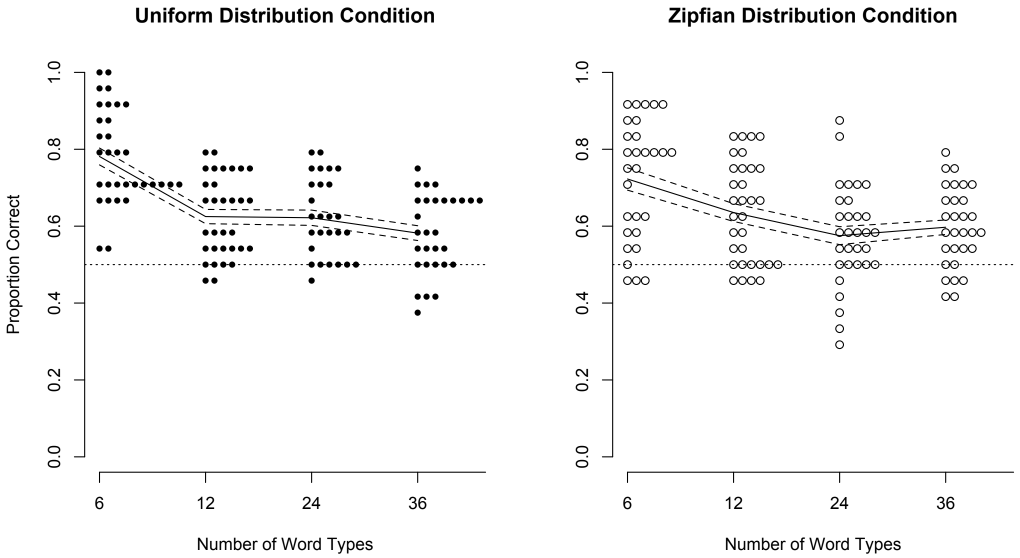

Figure 2: Average proportion of correct responses by number of word types in the uniform and Zipfian distribution conditions. Open and closed dots represent individual participants and are stacked to avoid overplotting. Solid, dashed, and dotted lines represent means, standard errors, and chance (50%),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

2.1.3 Procedure

Before the training phase began, participants were instructed to listen to a simple English word and type it in to ensure that sound was being played properly on the participants' system. Participants then moved to the training phase, where they were instructed to listen to 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" after each sentence to proceed through the training phase. 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") in the language they had learned.

2.2 Results and Discussion

Figure 2 illustrates accuracy of responses in the 4 types of languages in each of the uniform and Zipfian distribution conditions. There was not a strong numerical effect of the distribution condition. Replicating previous results (Frank et al., 2010b), performance decreased as the number of types increased, but participants performed slightly above chance even in the most difficult 36-type condition; this is a surprising and intriguing result given that each word in the uniform condition was heard on average only 8 times.

We conducted a mixed-effects logistic regression analysis (Breslow & Clayton, 1993; Gelman & Hill, 2006; Jaeger, 2008), fit to the entire dataset to avoid issues of multiple independent comparisons. This model attempted to predict the odds of correct answers on individual trials; we then used comparison between models to find the appropriate predictors. Our first model included effects of distribution and number of types; we found no effect of distribution (Zipfian distribution $p > .1$) 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 the overall log number of types in the language ($\chi^2(1) = .11$, $p > .7$). Instead, there were two main effects: a positive coefficient on log token frequencies (the more times a word is heard, the better performance gets: $\beta = .35$, $p < .0001$), and a negative coefficient on log distractor tokens (the more times a distractor is heard in the corpus, the worse performance gets: $\beta = -.51$, $p < .01$). We also found a positive interaction of the two (bad distractors are worse if the target is low frequency: $\beta = .14$, $p < .01$). The general relation is plotted in Figure 3, showing mean proportion of accuracy according to the log input frequency of the target words. In this final model, there was still no effect of distribution conditions (i.e., uniform vs. Zipfian) ($\beta = .05$, $p > .4$).

To summarize, participants represented target words equally well after being exposed to languages with very different frequency distributions and contingency statistics. The only factors that affected performance were the log frequency of targets and distractors, independent of distribution condition.

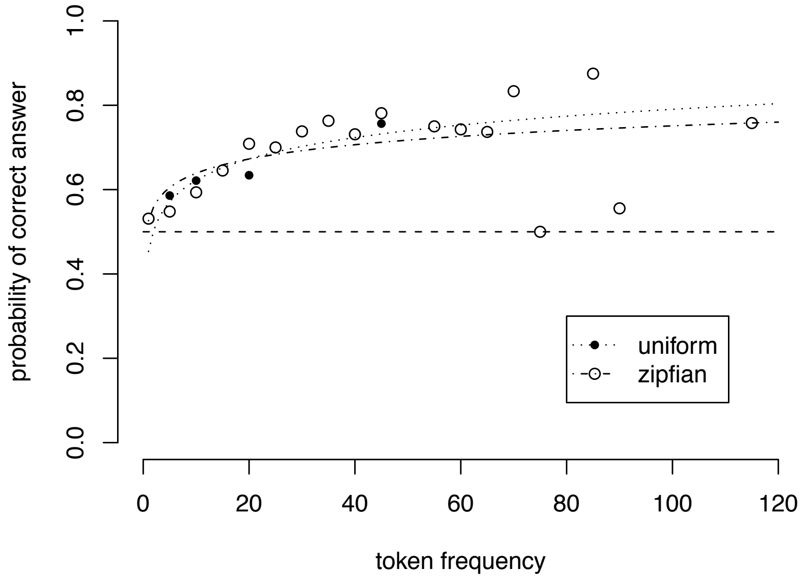

Figure 3: Probability of a correct 2AFC answer plotted by binned token frequency. Closed dots indicate uniform condition, and open dots indicate Zipfian condition. Dashed line shows chance, while the dotted and alternating lines give best fit lines for performance as a function of log token frequency.

3 Experiment 2

If learners accumulate evidence for words as they appear in the input, they should detect some words earlier than others based on token input frequencies. When presented in a sentential context, these early representations of some words may serve as anchors facilitating discovery of words that share boundaries with them. We referred to these kinds of effects as contextual facilitation and contextual bootstrapping, with facilitation referring to effects on segmentation of words in an initial known context, and with bootstrapping referring to effects of seeing a word in a known context on later segmentation performance. Experiment 2 tests the hypothesis, formed on the basis of previous work (Dahan & Brent, 1999; Bortfeld et al., 2005; Cunillera et al., 2010; Lew-Williams et al., 2011), that Zipfian contexts could promote these kinds of effects.

To conduct this test, we used an orthographic segmentation paradigm developed by Frank et al. (2010a, under review). A 2AFC asks only about a comparison between a particular target and its paired distractor; this method might hence be relatively insensitive to contextual effects. In contrast, the orthographic segmentation paradigm—where participants click on a transcript of a sentence to indicate where they think word boundaries fall—might be more sensitive to the kind of contextual effects we were looking for.

In our version of this orthographic segmentation task, participants were exposed to a language following either a Zipfian or a uniform distribution. After hearing each sentence, they were asked to give explicit judgements as to where they would place word boundaries. The experiment consists of 50 sentences (trials) and no discrete test phase—instead each sentence gave us information about participants' knowledge of the language, allowing us to construct a time course of learning for each participant and condition.

3.1 Methods

3.1.1 Participants

We posted 281 separate HITs on Mechanical Turk. We received 250 complete HITs from distinct individuals. Participants were paid $0.50 for participation. Because of the increased complexity of the task, we applied an incentive payment system to ensure participants' attention: they were told they would receive an additional $1.00 if they scored in the top quartile.

3.1.2 Stimuli

The process of generating stimuli was nearly identical to the 8 conditions in Experiment 1. Four word type conditions (with 6, 9, 12, and 24 word types, respectively) were generated and crossed with the two distribution patterns (uniform or Zipfian). These languages were used to generate 200 word tokens in 50 sentences. We chose to reduce the maximum number of word types (24 vs. 36) due to the complexity of the task and more limited overall amount of input. Participants were randomly assigned to one of the 8 conditions. Each sentence contained 3–5 words; we varied the number of words in sentences so that there was not a predictable number of word boundaries in any given sentence.

3.1.3 Procedure

After a synthesized sentence was played, participants were asked to indicate word boundaries in a corresponding transcription presented visually.

Each syllable was separated by a button that could be toggled. The participants were given one practice trial on an English sentence presented in the same format and prevented from continuing until they segment it correctly. All the syllables were spelled with one letter representing a consonant followed by one or two letters depending on the length of the vowel (e.g., *ka*, *ta*, *pee*). Participants could play back each sentence as many times as needed. Average time spent on the 50 trials was 16 minutes.

3.2 Results and Discussion

We were interested in participants' performance on individual words based on the words' frequencies and contexts. We thus created a binary dependent variable for success in segmenting each word: 1 if the word was segmented correctly (with a boundary at each edge and no boundaries at any internal syllable breaks) and 0 otherwise. Average segmentation results across trials are shown in Figure 4.[3] Participants who were exposed to Zipfian distributions generally achieved higher performance, especially in languages with more word types. Participants in the Zipfian condition outperformed those who heard languages with uniform distributions from the earliest trials on. When the lexicon contained only six types of words, participants who were exposed to a uniform distribution achieved a comparable level of performance by the time they finished, but participants hearing uniform token distributions never caught up when languages had more types.

We created a mixed logistic model to predict word-by-word performance (Table 1). As in Experiment 1, we found a strong main effect of log input frequency of the target word ($\beta = 0.46$, $p < 10^{-10}$). The length of the target word ($\beta = -1.35$, $p < 10^{-15}$) and the length of the sentence ($\beta = -1.0$, $p < 10^{-7}$) were significant predictors of correct segmentation of the target word. (The large effect of word length is likely due to the fact that longer words contain more syllables and hence more opportunities for incorrectly placed boundaries.)

[3] The measure we used here is known as "token recall" in the literature on evaluating segmentation models (Brent, 1999; Goldwater et al., 2009). Other work in this area has used precision and recall for tokens, as well as precision and recall measured for individual boundary judgments. We computed each of these measures, as well as the harmonic mean of precision and recall for each (F-score). The overall picture for all of the measures was almost identical to Figure 4. We thus focus on token recall, a measure that is related to comprehension (since the overall number of tokens correctly segmented will determine how many of them can be recognized and interpreted).

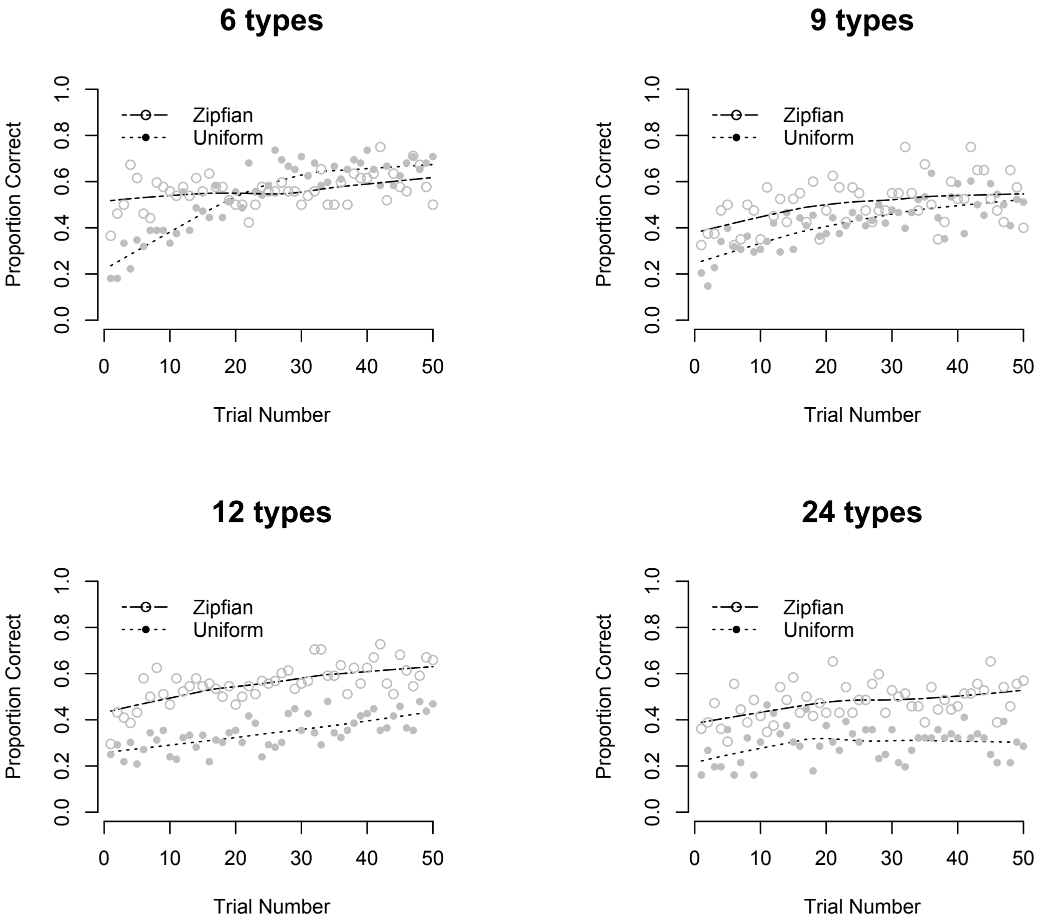

Figure 4: Proportion of correctly segmented word tokens per sentence plotted for each condition of Experiment 2. Dots represent mean F-score across individual participants for each trial; closed dots for participants from the uniform conditions and open dots from the Zipfian conditions. Lines show a non-linear fit by a local smoother (loess).

Table 1: One mixed logit model for Experiment 2, showing contextual facilitation effects but not contextual bootstrapping effects (see text for more details).

Random effects					
	Name	Variance	Std.Deviation	Correlation	
Participant ID	(intercept)	0.54	0.73		
	Log token freq (target)	0.45	0.67	-0.191	
Fixed effects					
	Coefficient	Std. Error	z-value	p-value	
Intercept	1.54	0.50	3.06	< 0.003	**
Distribution (Zipf)	0.28	0.35	0.79	0.43	
Word types (6,9,12,24)	0.01	0.02	0.69	0.49	
Distribution × Word types	< 0.01	0.02	0.05	0.96	
Log token frequency (target)	0.46	0.07	6.77	1.33×10^{-11}	***
Log token frequency (previous)	0.15	0.03	4.31	1.63×10^{-5}	***
Log token frequency (following)	< 0.01	0.03	0.04	0.97	
Word length (syllables)	-1.35	0.10	-13.75	$< 2 \times 10^{-16}$	***
Sentence length (syllables)	-1.00	0.20	-5.12	3.02×10^{-7}	***

We used this model to investigate a contextual facilitation effect: that high familiarity with particular items would improve segmentation accuracy for their neighboring words. To test this hypothesis, we included the cumulative log frequency—number of times heard in the input prior to the target word—of the words on the both sides of the target words as predictors.[4] The cumulative frequency of the previous word was a significant predictor ($\beta = 0.15$, $p < 10^{-4}$): the more frequently the left neighbour word had been heard so far, the more likely it was for the target word to be segmented correctly. The absence of a similar effect on the right-hand side ($p > .9$) may be due to the directionality of the segmentation process. Participants in our task might be placing boundaries moving from the left edge (the onset of a sentence) to the right edge, making the information from the preceding word more important.

We next used the model to test for a contextual bootstrapping effect: that having been seen in supportive contexts (e.g., next to high-frequency items) leads to better segmentation in future exposures. To do so, we constructed another model which included a predictor that measured the degree of support given by the previous contexts in which the target word had been seen. This predictor was composed of the average log frequency of all the words that had appeared on either side of the target word prior to the current exposure. The frequency-based predictors we used to investigate the two contextual effects—contextual facilitation and bootstrapping—are highly collinear and cannot be tested in a single model (Gelman & Hill, 2006; Jaeger, 2008). For this test, we thus removed the contextual facilitation predictors.

If being flanked by high-frequency neighbours can improve recognition, words that have neighbors with higher average frequency should be segmented more correctly than those which have a history of adjacency with low-frequency words. As with the contextual facilitation predictors, our model showed such an effect for the words on the left of the target word ($\beta = 0.18$, $p = .014$) but not for the words on the right ($\beta = -.03$, $p = .72$). Both contextual facilitation and contextual bootstrapping models dramatically increased goodness-of-fit compared to models that did not include contextual predictors (ps $< 10^{-16}$), but the contextual facilitation model had

[4] Note that this predictor is only available for words that fall in the middle positions of sentences, hence the dataset used in this and following models is a subset of the full dataset. Coefficients for effects shared across both models were comparable, however.

overall lower Akaike's Information Criterion values (AIC: 13,331 vs. 13,344 respectively, with the same number of parameters in each model), suggesting that it fit the data somewhat better.

To summarize, we found highly reliable effects of contextual facilitation and contextual bootstrapping. As in Experiment 1, however, there was no overall effect of distribution condition (uniform vs. Zipfian) beyond frequency effects at the token level.

4 General Discussion

We presented two artificial language word segmentation experiments, comparing performance in word recognition and word segmentation in languages with uniform and Zipfian frequency distributions. Both experiments showed that the major determinant of performance was the frequency with which words were heard. Once frequency was accounted for, we observed no remaining effect of distribution condition, suggesting that the sparsity of Zipfian languages posed no problem for learners. Thus, our results support a view of "statistical learning" that—although sensitive to statistical coherence—is largely a process driven by consistent exposure to frequent chunks (Perruchet & Vinter, 1998; Frank et al., 2010b, under review).

Nevertheless, when we examined word segmentation in context, we saw that performance for Zipfian languages was considerably higher. This result highlighted a simple fact about Zipfian languages: in these languages, listeners are repeatedly exposed to a small number of high-frequency words, giving them many chances to learn these words and use them in segmenting incoming sentences. When the words were uniformly distributed, learners could not reliably segment sentences until they became sufficiently familiar with the entire lexicon (Experiment 2). The highly skewed distribution of word frequencies thus supports an efficient entry into the task of word segmentation.

Furthermore, our results suggest that established familiarity with high-frequency words helps learners segment adjacent material. We distinguished two effects stemming from this observation: contextual facilitation effects—in which adjacent high-frequency words help learners segment words in the moment—and contextual bootstrapping effects—in which a history of these supportive contexts leads to longer-term learning. In our dataset, we saw reliable evidence for both types of effects, explaining the overall advantage that learners had in the Zipfian conditions.

Our results are thus compatible with previous work on contextual facilitation and bootstrapping (Bortfeld et al., 2005; Brent & Siskind, 2001; Cunillera et al., 2010; Lew-Williams et al., 2011). In fact, they may suggest a way to reconcile some conflicting developmental results. Since contextual facilitation and bootstrapping effects are small relative to direct frequency effects, these effects may have been easier to observe in the Bortfeld et al. (2005) study, which used very high-frequency names, rather than the Hollich et al. (2001) study, which used familiar but much lower frequency nouns. Nevertheless, more research with infants and children is necessary to understand whether contextual effects play a large role in children's early word segmentation performance.

The contrast between the two paradigms we used—word recognition judgments and explicit orthographic word segmentation—highlights an important assumption of previous work on segmentation: that the goal of word learning is to attain a large vocabulary of word types. In fact, language learners are likely pursuing multiple simultaneous goals. One is to build a vocabulary of word types; the other is to interpret word tokens as they are heard (Frank et al., 2009). The higher performance we observed in the Zipfian conditions of Experiment 2 was a consequence of this distinction. While Zipfian contexts did not have any particular effects on segmentation accuracy per se, the fact that new material in these conditions tended to contain many high-frequency tokens means that segmentation was considerably more accurate. Thus, Zipfian languages support word segmentation in context, allowing learners to begin parsing and interpreting the language they hear much more quickly than they would otherwise be able to.

5 Acknowledgements

Thanks to T. Florian Jaeger, Noah Goodman, Josh Tenenbaum, and the members of the Stanford Language and Cognition Lab for valuable discussion. An earlier version of this work was reported to the Cognitive Science Society in Kurumada et al. (2011).

References

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

Blanchard, D., Heinz, J., & Golinkoff, R. (2010). Modeling the contribution of phonotactic cues to the problem of word segmentation. *Journal of Child Language*, *37*, 487–511.

Bortfeld, H., Morgan, J., Golinkoff, R., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech stream segmentation. *Psychological Science*, *16*, 298.

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93–125.

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

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, 33–44.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*, 9–25.

Chomsky, N. (1955). *The logical structure of linguistic theory* volume 53. Springer.

Conway, C., Bauernschmidt, A., Huang, S., & Pisoni, D. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition*, *114*, 356–371.

Cunillera, T., Càmara, E., Laine, M., & Rodríguez-Fornells, A. (2010). Words as anchors: Known words facilitate statistical learning. *Experimental Psychology*, *57*, 134–141.

Dahan, D., & Brent, M. (1999). On the discovery of novel wordlike units from utterances: An artificial-language study with implications for native-language acquisition. *Journal of Experimental Psychology: General*, *128*, 165.

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* (pp. 1393–1396). Philadelphia, PA volume 3.

Ellis, N. C., & O'Donnell, M. B. (2011). Robust language acquisition: An emergent consequence of language as a complex adaptive system. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*.

Fiser, J. (2009). The other kind of perceptual learning. *Learning Perception*, *1*, 69–87.

Frank, M., Arnon, I., Tily, H., & Goldwater, S. (2010a). 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. L., & Tenenbaum, J. B. (2010b). Modeling human performance in statistical word segmentation. *Cognition*, *117*, 107–25.

Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578.

Frank, M. C., Tenenbaum, J. B., & Gibson, E. (under review). Learning and long term retention of large scale artificial languages, .

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*, 254.

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, *31*, 190–222.

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.

Hayes, J. R., & Clark, H. H. (1970). Experiments in the segmentation of an artificial speech analogue. In *Cognition and the development of language* (pp. 221–234). Wiley.

Hollich, G., Jusczyk, P., & Brent, M. (2001). How infants use the words they know to learn new words. In *Proceedings of the 25th Annual Boston University Conference on Language Development* (p. 353). Cascadilla Press volume 1.

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

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

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*.

Jusczyk, P., Hohne, E., & Bauman, A. (1999). Infants' sensitivity to allophonic cues for word segmentation. *Attention, Perception, & Psychophysics*, *61*, 1465–1476.

Kurumada, C., Meylan, S. C., & Frank, M. C. (2011). Zipfian word frequencies support statistical word segmentation. *Cognitive Science Conference*, .

Lew-Williams, C., Pelucchi, B., & Saffran, J. (2011). Isolated words enhance statistical language learning in infancy. *Manuscript in press, Developmental Science*, .

Li, W. (1992). Random texts exhibit zipf's-law-like word frequency distribution. *Information Theory, IEEE Transactions on Information Theory*, *38*, 1842–1845.

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, *78*, 91–121.

Mitchell, C., & McMurray, B. (2009). On leveraged learning in lexical acquisition and its relationship to acceleration. *Cognitive Science*, *33*, 1503–1523.

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.

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

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

Pinker, S. (1984). *Language learnability and language development* volume 193. Harvard University Press.

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

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

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 6038–6043.

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

van de Weijer, J. (2001). The importance of single-word utterances for early word recognition. *Early lexicon acquisition: normal and pathological development*, *2*.

Wolff, J. G. (1977). The discovery of segments in natural language. *British Journal of Psychology*, *68*, 97–106.

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

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