Cognition 127 (2013) 439–453

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Zipfian frequency distributions facilitate word segmentation in context

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

[a] Department of Linguistics, Stanford University, United States
[b] Department of Psychology, University of California, Berkeley, United States
[c] Department of Psychology, Stanford University, United States

ARTICLE INFO

Article history:
Received 19 October 2011
Revised 20 January 2013
Accepted 4 February 2013
Available online 2 April 2013

Keywords:
Word segmentation
Statistical learning
Zipfian frequency distributions
Computational modeling

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. 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 recognizing individual words is predicted best 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. We find that computational models that implement "chunking" are more effective than "transition finding" models at reproducing this pattern of performance.

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 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 (Aslin, Saffran, & Newport, 1998; Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Hauser, Newport, & Aslin, 2001; Thiessen & Saffran, 2003). In a now-classic segmentation paradigm, Saffran, Newport et al. (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 non-words. In a similar experiment, participants demonstrated the ability to distinguish words and "part-words" – length-matched strings that also occurred in the exposure corpus, albeit with lower frequency and lower statistical consistency (Aslin et al., 1998). These studies on "statistical learning," combined with similar demonstrations with infants, suggest that learners can use the statistical structure of sound sequences to find coherent chunks in unsegmented input.

While the results of statistical learning experiments are impressive, it is still unknown how these findings relate to natural language learning (Johnson & Tyler, 2010; Yang, 2004). 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, Evans, Alibali, & Saffran, 2007), and experiments with natural languages suggest that the processes observed in artificial language experiments generalize to highly controlled natural language samples (Pelucchi, Hay, & Saffran, 2009). In addition, adults can perform sta-

* Corresponding author.
E-mail address: kurumada@stanford.edu (C. Kurumada).

0010-0277/$ - see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.cognition.2013.02.002

tistical segmentation when there is variation in sentence and word lengths (Frank, Goldwater, Griffiths, & Tenenbaum, 2010) and when languages scaled up over multiple days of exposure (Frank, Tenenbaum, & Gibson, 2013). 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").[1] 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.[2] While Zipfian distributions are ubiquitous across natural language, their consequences for learning are only beginning to be explored (Ellis & O'Donnell, 2011; Goldwater, Griffiths, & Johnson, 2006; Mitchell & McMurray, 2009; Yang, 2004).[3]

An early and influential proposal suggested that learners could succeed in statistical segmentation tasks by computing the transitional probability (TP) between syllables (Saffran, Newport et al., 1996). 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 (Frank et al., 2010; Saffran, Aslin et al., 1996; Saffran, Newport et al., 1996). Thus, in standard experiments, comparisons between TPs are easy to make because all transitions can be estimated accurately.

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 "obfuscatory") and their combination, even when possible, will be vanishingly rare ("obfuscatory toaster"). On the other hand, some combinations of frequent (monosyllabic) words have high transitional probability between them despite the presence of a word boundary (e.g. "of the"). In fact, given the collocational structure of natural language (Goldwater, Griffiths, & Johnson, 2009), the within-word transitional probabilities for low-frequency words can easily 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 (Brent, 1999; Yang, 2004). The sparsity of the transition matrices 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, on this first view, compute TPs) only to learn a small set of word forms, and hence they may not need to be particularly 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 (Blanchard, Heinz, & Golinkoff, 2010; Johnson & Jusczyk, 2001; Jusczyk, Hohne, & Bauman, 1999; Mattys & Jusczyk, 2001; Shukla, White, & Aslin, 2011). This viewpoint—that a 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.

Second, learners may rely on a more robust statistical learning method. In fact, non-TP computational proposals for statistical learning make different prediction for segmentation performance in Zipfian environments. Orbán, Fiser, Aslin, and Lengyel (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 minimum-description length (Brent & Cartwright, 1996), Bayesian (Brent, 1999; Goldwater et al., 2009), memory-based (Perruchet & Vinter, 1998), and connectionist (French, Addyman, & Mareschal, 2011) models. These models (and some corresponding psychological evidence) suggest that segmentation performance should be robust to—or even facilitated by—Zipfian distributions. (In Section 4, we provide a direct test of these predictions through a series of simulations with a variety of models.)

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 those words. 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.

"Bootstrapping" effects provide another route by which Zipfian distributions could facilitate segmentation. If a novel word occurs adjacent to a familiar word, it may be segmented more effectively because one boundary is already known (Perruchet & Tillmann, 2010). A Zipfian distribution

[1] In many languages, the top-most frequent words consist of phonologically concise function words (e.g., "the"). Hochmann, Endress, and Mehler (2010) provided an experimental evidence suggesting that 17-month-olds could distinguish function words from content words based on words' relative token frequencies.

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

[3] 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). We take it for granted that natural languages have this structure without attempting to explain its presence.

would facilitate this kind of bootstrapping effect because a small number of high-frequency words ("anchors") could create known contexts for low-frequency words (Valian & Coulson, 1988). Because bootstrapping effects are central to our predictions, below we provide a more detailed example of how they could arise.

Assume a language like the one used by Saffran, Newport et al. (1996), containing six word types (Fig. 1). When words are concatenated uniformly to make sentences, as in Fig. 1a, boundaries become unambiguous only after a certain number of word types are observed and the TPs are estimated. On the other hand, when they follow a Zipfian distribution, as in Fig. 1b, the most frequent words are repeated in nearly every sentence. If these high-frequency words are learned quickly and retained, they provide a clear context for the acquisition of lower-frequency words, as pictured in Fig. 1c.

In what follows, we distinguish two kinds of effects that have previously been labeled as "bootstrapping." *Contextual facilitation* is when an otherwise less recognizable word (e.g., *jkl* in Fig. 1c) is better segmented due to the adjacency of a well-established word (e.g., *abc*). *Contextual bootstrapping* is when hearing the sequence *abcjkl*, containing the known word *abc* and novel word *jkl*, facilitates the identification of *jkl* in the future. Under this definition, facilitation—help segmenting a word in context—is a component of bootstrapping. Bootstrapping further involves retaining that word for future use. Thus, contextual facilitation is the advantage given by high-frequency or otherwise known material in a particular context, while contextual bootstrapping is the same advantage in future contexts.

Brent and Cartwright (1996) proposed a model implementing contextual bootstrapping based on sequential formation of rudimentary word chunks. Their INCDROP model 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 many others (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. In Fig. 1c, for example, recognition of the frequent words (*abc def*) is expected to provide boundaries for infrequent words (*ghi*,*jkl*,*mno*) that will bootstrap their recognition in subsequent presentations.

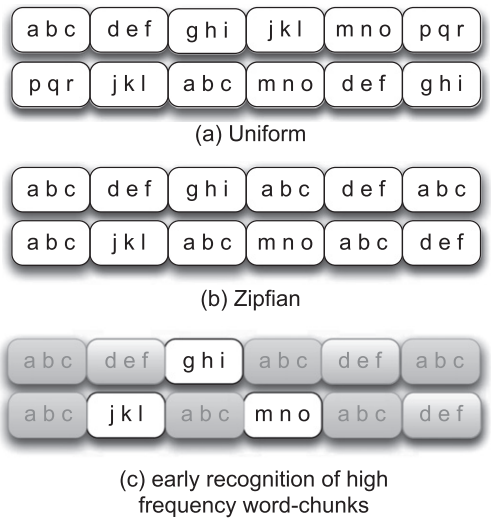

Fig. 1. Small-scale examples of a 6-word language following (a) uniform and (b) Zipfian distributions. Letters represent syllables and blocks represent words; block boundaries are pictured as a convenience but are assumed not to be available to the learner. If highly frequent words are recognized earlier, they can provide known "anchors" to facilitate segmentation of lower-frequency words, illustrated in (c).

Several psychological studies have tested whether known words facilitate the segmentation of nearby words, with mixed results. Dahan and 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, Morgan, Golinkoff, and Rathbun (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, Jusczyk, and Brent (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 (Aslin, Woodward, LaMendola, & Bever, 1996), and a number of studies have investigated their role in segmentation. Brent and 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 neighboring 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, Bauernschmidt, Huang, & Pisoni, 2010; Cunillera, Càmara, Laine, & Rodrıguez-Fornells, 2010; Lew-Williams, Pelucchi, & Saffran, 2011; van de Weijer, 2001). Thus, several lines of research 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 psychological as well as computational work leaves us with two different predictions about the effects of the Zipfian word frequency distribution in natural language on word segmentation performance. Under 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 facilitate word segmentation by using known contexts to segment novel words more effectively.

In the current study, 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. Finally we confirm the efficacy of chunking models for segmenting input with a Zipfian distribution by comparing the fit of a variety of computational models to the human data.

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, Newport et al. (1996) 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 judgements 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. (2010), 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 202 HITs from distinct individuals (a mean of 25 for each token frequency and distribution condition). Participants were paid $0.75 and the task took approximately 7–10 min.

2.1.2. Stimuli

We constructed eight 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 two, three, or four 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 100 Hz with 225 ms vowels and 25 ms consonants. Each syllable was used in one word only.[4] 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 six 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 Fig. 2.

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, as in Frank et al. (2010). The larger the number of types in the language, the smaller the number of times any given distractor appeared on average, because a larger number of types created fewer opportunities for any given set of words to occur adjacent to one another. Nevertheless, distractor frequencies were matched between Zipfian and uniform conditions: Averaged across all test items, distractor frequencies were approximately 8, 2, .5, and .2 for the 6, 12, 24, and 36 type conditions. (The effects of distractor frequency on performance for individual test trials is considered in regression analyses below.)

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

Fig. 3 illustrates accuracy of responses in the four 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., 2010), 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 eight 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 appro-

[4] To ensure the discriminability of the synthesized syllables used, we conducted an online survey in which nine participants listened to syllable pairs and judged if they were the same or different. The paired syllables were either identical or formed a minimal pair, contrasting either in their vowel or their consonant (e.g., /po/ vs. /pa/ and /pa/ vs. /ba/). The minimal pairs were distinguished correctly in 93% of trials for consonants and 99% of trials for vowels, leading to d' values of 3.92 and 5.20 respectively. While a few consonant pairs were confusable though still distinguished at levels above chance (e.g., /v/ vs. /b/, /p/ vs. /f/), the large majority of the syllables used in Experiments 1 and 2 were discriminable from each other with near perfect accuracy.

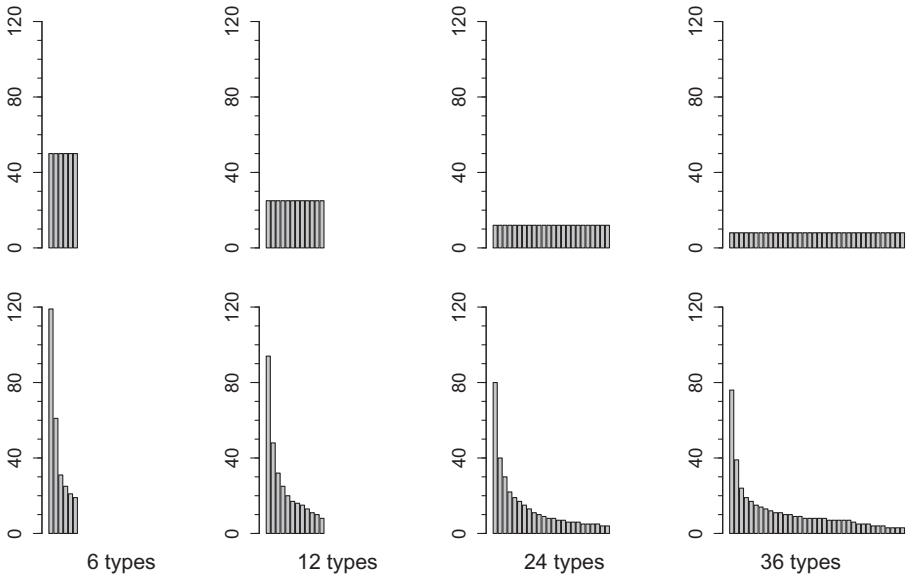

Fig. 2. Word frequencies in uniform (top) and Zipfian (bottom) conditions of Experiment 1. The horizontal axis shows distinct word types, and the vertical axis shows the frequency of each of these types.

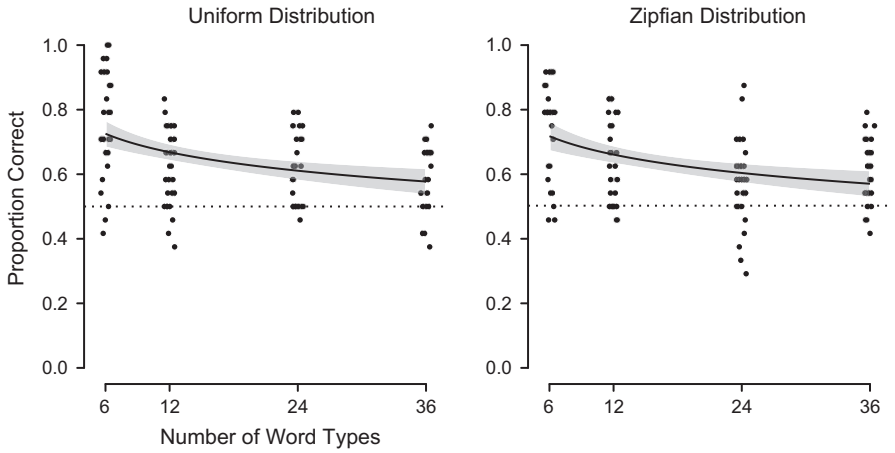

Fig. 3. Average proportion of correct responses by number of word types in the uniform and Zipfian distribution conditions. Dots represent individual participants and are jittered to avoid overplotting. Solid lines give best fit for performance as a function of log number of types, with the gray boundary representing the standard error. Dashed line represents chance (50%).

priate predictors. Our first model included effects of distribution and number of types (as well as a random effect of participant; since all manipulations were between-subjects, this was our only random effect). We found no effect of distribution ($p = .65$) but a highly significant effect of number of types ($\beta = -.020$, $p < .0001$). 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 ($\chi^2 = 9.21$, $p < .0001$). 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 log frequency of the target and distractors for each trial (calculated from the input corpus for each language; again, the logarithms were better predictors). 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 ($\chi^2(1) = .23$, $p = .63$). 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: $\beta = -.50$, $p < .01$). 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 relationship between performance and log token frequency is plotted in Fig. 4. In this final model, there was still no effect of

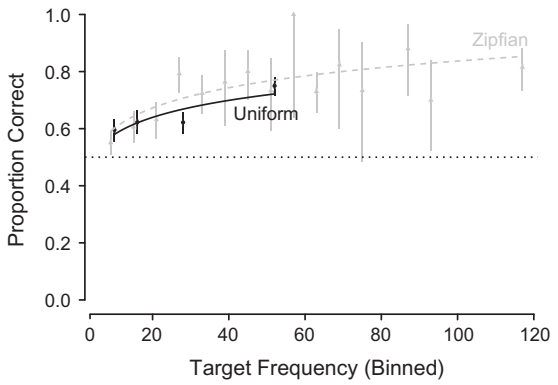

Fig. 4. Probability of a correct 2AFC answer plotted by binned token frequency. Black circles show the uniform condition, and gray triangles show the Zipfian condition. Vertical intervals show 95% confidence intervals as computed using a standard Bayesian method (with an uninformative Beta prior). Dotted line shows chance, while the dashed and solid lines give best fit lines for performance as a function of log token frequency.

distribution conditions (Zipfian $\beta = .09$, $p = .29$), though the Zipfian condition showed a slight numerical trend towards higher performance.

To summarize, participants represented target words equally well after being exposed to languages with very different frequency distributions and contingency statistics. We found robust effects of the log unigram frequency of targets and distractors, independent of distribution condition. The lack of *disadvantage* in a Zipfian condition suggests that the mechanism underlying adults' word segmentation must involve more than mere estimation of forward TPs; this claim is tested in depth in the simulation section below.

In this initial study, we did not find any bootstrapping effects for the Zipfian languages (as predicted by the chunking models): Once target and distractor frequency were accounted for, there were no further effects of condition on participants' performance. One possible reason for this lack of an effect is that the current 2AFC task tests word knowledge in isolation, and might not gauge the contextual support available in a sentential context. In addition, the design of the experiment prevented us from analyzing the contextual history of individual words (because by the end of training, all words had quite similar contextual histories). In the following section, we report experimental results based on a new paradigm, which allows us to explore the potential effects of contextual support more precisely.

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 may serve as anchors facilitating discovery of words that share boundaries with them, producing either contextual facilitation (better segmentation in known contexts) or contextual bootstrapping (better segmentation of words that have previously appeared in known contexts). Experiment 2 provides a further test of the hypothesis that Zipfian distributions could promote these kinds of effects, at least when performance is measured on items presented in context (Bortfeld et al., 2005; Cunillera et al., 2010; Dahan & Brent, 1999; Lew-Williams et al., 2011).

To conduct this test, we used an orthographic segmentation paradigm developed by Frank, Arnon, Tily, and Goldwater (2010) and Frank et al. (2013). A two-alternative forced choice compares 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 reconstruct the 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 eight conditions. Each sentence contained three to five words; we varied the number of words in sentences so that the number of word boundaries in any given sentence was not predictable.

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 line (signifying a word boundary) that could be toggled on or off. The participants were given one practice trial on an English sentence presented in the same format and prevented from continuing until they segment it

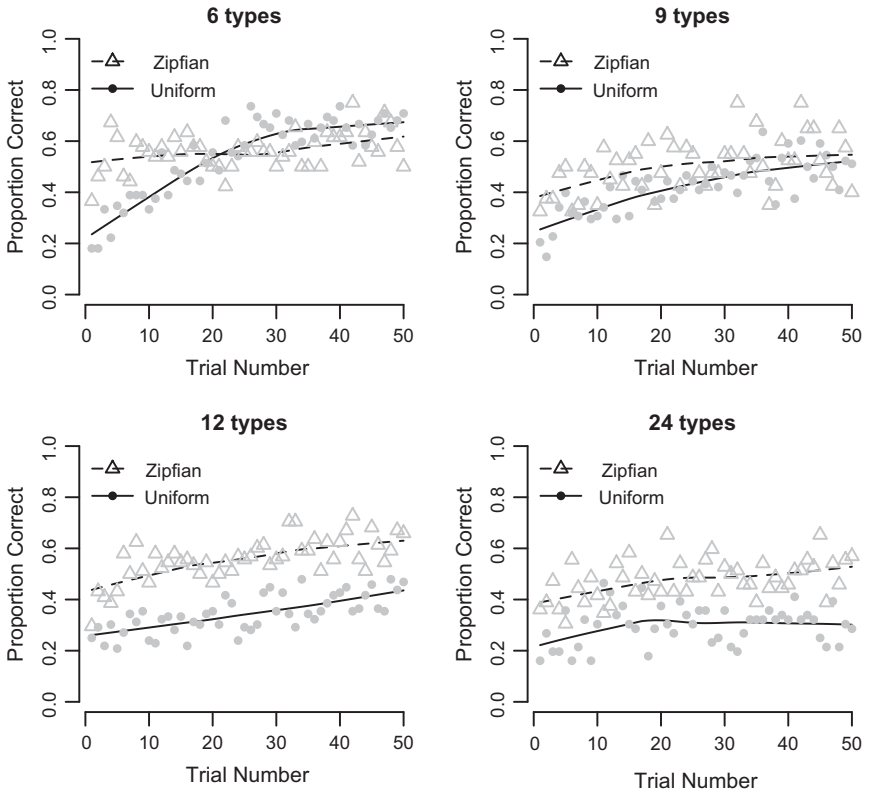

Fig. 5. Proportion of correctly segmented word tokens per sentence plotted for each condition of Experiment 2. Points represent mean F-score across individual participants for each trial; closed dots for participants from the uniform conditions and open triangles from the Zipfian conditions. Lines show a non-linear fit by a local smoother (loess, span = .75).

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 min.

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 Fig. 5.[5]

Participants who were exposed to Zipfian distributions generally achieved higher performance, especially in languages with more word types. In languages with fewer word types, participants in the uniform condition started out learning more slowly but caught up to those in the Zipfian condition; in the languages with more word types, participants in the uniform condition never caught up.

To capture this pattern of performance, we created a mixed logistic model to predict word-by-word segmentation accuracy (Table 1). We included random by-participant intercepts and by-participant slopes for log token frequency of words, assuming that participants differ in how much input they need to segment words correctly. As in Experiment 1, we found a strong main effect of log input frequency of the target word ($\beta = 0.46$, $p < 10^{-10}$).[6] The length of the target word ($\beta = -1.35$, $p < 10^{-15}$) and the

[5] 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 Fig. 5. We 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).

[6] We initially included two more factors to control additional support from sentence boundaries (Monaghan & Christiansen, 2010). These were: (1) a binary variable of seeing a target word at an edge of the current sentence or not (current boundary status) and (2) a continuous variable of the frequency of seeing a target word (type) at sentence boundaries in the past sentences (past boundary frequency). The model suggested that both of these factors were significant predictors of correct segmentation of a target word (current boundary status ($\beta = 0.72$, $p < 10^{-16}$) and past boundary frequency ($\beta = 0.39$, $p < 10^{-12}$)). However, the past boundary frequency was strongly colinear with general type frequency of the word: Words with high general frequency occur at sentence boundaries more often than other words. Therefore, we did not include this predictor in the other models we created. In the models we report below, we excluded all words that appeared at sentence boundaries in order to control the effect of current boundary status and better estimate effects of adjacency.

Table 1
Mixed logit model parameters for Experiment 2, showing contextual facilitation predictors (see text for more details).

	Name	Variance	Std. dev.	Correlation
Random effects				
Participant ID	(intercept)	0.54	0.73	
	Log token freq (target)	0.45	0.67	−0.191
	Coefficient	Std. err.	*z*-Value	*p*-Value
Fixed effects				
Intercept	1.48	0.50	2.83	<0.005**
Distribution (Zipf)	0.55	0.37	1.50	0.13
Word types (6, 9, 12, 24)	0.01	0.02	0.59	0.56
Distribution × word types	<0.01	0.02	−0.39	0.69
Log token frequency (target)	0.19	0.10	1.91	0.05
Log token frequency (previous)	0.15	0.04	4.04	5.31×10^{-5}***
Log token frequency (following)	−0.01	0.04	−0.25	0.80
Word length (syllables)	−1.34	0.10	−13.75	$<2 \times 10^{-16}$***
Sentence length (syllables)	−0.95	0.20	−4.63	3.49×10^{-7}***
Log frequency seen at boundary	0.33	0.08	4.11	3.84×10^{-5}***

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.)

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. 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. The cumulative frequency of the previous word was a significant predictor ($\beta = 0.15$, $p < 10^{-4}$): the more frequently the left neighbor 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 ($p = .8$) 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 neighbors 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 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.

Can performance in our orthographic segmentation task be compared to performance in a purely auditory task? The left–right asymmetry we observed in the contextual facilitation effects suggests that participants primarily placed boundaries in an incremental manner, moving from left to right. Followup analyses of the time-course of participants' segmentation decisions confirmed this: Participants rarely backtracked to undo decisions they had already made. This pattern indicates that behavior in the orthographic task had some similarities to auditory information processing: Both follow a sequential strategy. Nevertheless, the addition of visual information might have made segmentation easier by alleviating memory load (Frank & Gibson, 2011) or adding redundant information in a second modality.

To investigate whether changes in modality affected performance, we reran one condition of Experiment 2 (nine word types) without audio input. We found an overall similar pattern of results with a comparable level of performance to the results reported above. These data are consistent with the hypothesis that the phenomena we observed are similar across modalities (though they suggest that some visual statistical learning might have occurred (Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002)). Even in this "no audio" condition, however, it is impossible to know how much of the participants' word representations were formed based on visual input alone, since participants were likely reading sentences silently. Thus, despite its visual component, we believe that our current task elicits an approximation of adult learners' online segmentation behavior.

Table 2
Properties of the four models used in our simulations. Parameters column indicates the parameters that were varied rather than the total possible parameters of the model.

Model	Key reference	Chunking	Class	Params.
Forward TP	Saffran, Newport et al. (1996)	×	Simple statistical	1
PARSER	Perruchet and Vinter (1998)	✓	Memory-based	2
TRACX	French et al. (2011)	✓	Connectionist	4
Particle filter	Börschinger and Johnson (2011)	✓	Bayesian	1

To summarize, in Experiment 2 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. We were not able to estimate contextual facilitation and bootstrapping effects jointly, but our analyses suggest that facilitation effects were considerably stronger than bootstrapping effects (probably because bootstrapping requires facilitation as well as retention of the facilitated word forms). This result may explain the lack of bootstrapping effects in Experiment 1: there was no opportunity during test for facilitation effects, and weaker bootstrapping effects may not have been visible in the somewhat less sensitive two-alternative forced-choice paradigm.

4. Model simulations

In this section, we test the qualitative predictions made in the Introduction: that chunking models predict an advantage for Zipfian distributions, while transitional probability (TP) models predict a disadvantage. A transition-finding model implies a lexicon as a consequence of segmenting at low-probability transitions. A skewed word frequency distribution would result in sparse probability matrices, which are expected to give rise to a problem for strictly TP-based approaches. A lexical model, on the other hand, maintains a collection of words and word-like chunks in the form of memory representations, cue weights or probabilities. Frequent word types in a Zipfian distribution are expected to be learned and retained more easily, providing leverage in recognition of otherwise unfamiliar words in the context.

To test these predictions on a real dataset, we compare the fit of four different computational models to human data from Experiment 2. We chose as our models a forward transition-finding model (Saffran, Newport et al., 1996) and three lexical models: a memory-based model (PARSER; Perruchet & Vinter, 1998); a recognition-based connectionist model (TRACX; French et al., 2011); and a new online implementation of a probabilistic segmentation model (Goldwater et al., 2009) using an online "particle filter" inference algorithm (Börschinger & Johnson, 2011). For brevity we refer to this last model as the Particle Filter model. Properties of these models are summarized in Table 2, and details of the models are given in Appendix A.

Due to the fundamental differences in the assumptions and details of these models, the reported metrics of model fit are not meant as a formal model comparison. Rather, we present the best parameter setting for each model, providing a basic estimate of the fit to human data. Since our goal here is to show that a range of chunking models show a Zipfian advantage, rather than to decide between models, we do not provide an exposition of the differences between formalisms, though see Frank et al. (2010) for more details on some of the models.

One important feature of the models compared here is that all are "online" models: that is, they pass through the data sequentially, without storing the sentences in memory (as would be the case in a "batch" model).[7] This feature was necessary because our task is fundamentally online: performance on particular trials depended only on what has been learned in the previous sentences. To evaluate segmentation performance on individual trials, we made minor changes to several of the models. Details of these modifications as well as model parameters, their significance, and the parameter ranges tested are summarized in Appendix A.

4.1. Simulations

4.1.1. Materials

The stimuli from Experiment 2 were translated into a standardized format in which each character corresponded to a single syllable (Frank et al., 2010). To ensure convergence of estimated model performance, each model was run over 128 input files of 50 sentences, 16 in each of two distribution conditions (uniform and Zipfian) crossed with four word type conditions (6, 9, 12, and 24 types).

4.1.2. Evaluation

Scores were aggregated over the 16 input files in each condition × type combination, such that every model run yielded 400 data points: a token F-score for each of 50 sentences in each of eight condition-type combinations. For every parameter setting in each model, we calculated root mean square error (RMSE) and Pearson's product-moment correlation coefficient (Pearson's r) between these 400 datapoints and average human performance. These two metrics provide different information regarding model fit: RMSE is a metric of the absolute fit based on differences between values predicted by a model and the values actually observed, while Pearson's r characterizes the degree of correlation between the learning curves but does not punish differences in the absolute performance. These two measures together allow us to assess how well a particular model could generate the human token F-scores observed in Experiment 2.

[7] TRACX required multiple passes through the input corpus (multiple training epochs) to produce segmentation decisions, though on the final run segmentation decisions were produced continuously. For more details on simulations, see Appendices A and B.

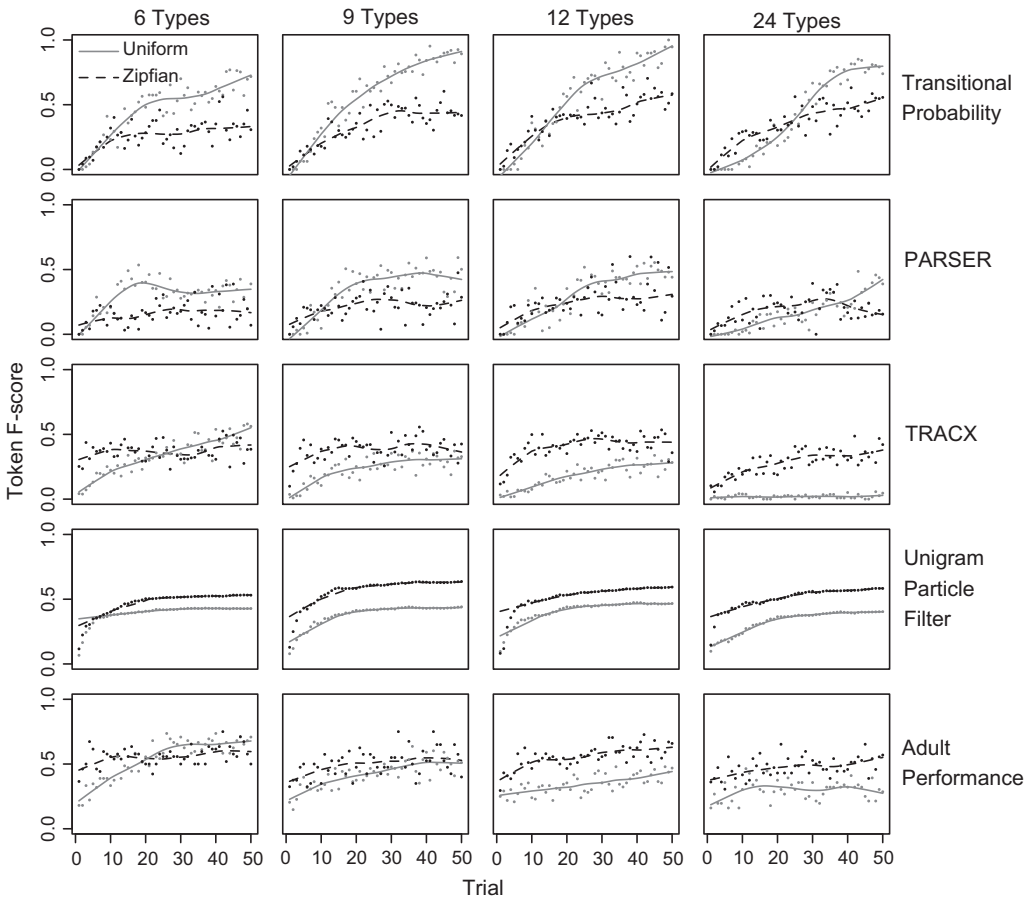

Fig. 6. Best-fitting model simulations (RMSE). Dots represent mean F-score for each trial; lines show a non-linear fit by a local smoother (loess, span = .4).

4.2. Comparison results

Fig. 6 shows the performance of each model as well as the human performance observed in Experiment 2. We show learning curves with the best RMSE for ease of interpretation; those curves with the highest correlation values (presented in Appendix B) often had radically different scales than the human data.

As hypothesized, the transition-finding model demonstrated better performance on input with a uniform word frequency distribution in all the four conditions (6, 9, 12, and 24 word types).[8] On the other hand, two of the chunking models—TRACX and the Particle Filter—performed better in the Zipfian conditions, and successfully captured the overall characteristics of the human data. Table 3 shows the results of comparison (the lowest RMSE and the highest Pearson's *r*) between the models and the human data. TRACX and the Particle Filter fit human performance better than the other two models.

Table 3
Comparison between models and human data for Experiment 2.

Model	RMSE	Pearson's *r*	Zipfian advantage?
Transitional probability	0.24	0.50	
PARSER	0.28	0.66	?
TRACX	0.21	**0.79**	✓
Particle filter	**0.12**	0.65	✓

Further exploration revealed that PARSER's performance was modulated by a free parameter: its forgetting rate (the rate at which lexical chunks decayed from its memory store). It performed better based on a Zipfian distribution (patterning with the other chunking models) when items in memory decayed quickly. On the other hand, when chunks remained in memory longer, PARSER showed higher performance in the uniform condition. A Zipfian frequency distribution provided more leverage when the model required more exposure to maintain a word chunk, or when there were more words to be learned. Together with recent findings in implementing resource limitation in modeling human segmentation performance (Frank et al., 2010), our results from PARSER suggest that

[8] We further explored a Bayesian variation of the TP model that uses smoothing of transition counts to approximate memory limitations (Frank et al., 2010). When the amount of smoothing was relatively small, results showed an advantage in uniform conditions; when more smoothing was applied, performance was higher in Zipfian conditions. This pattern was congruent with the results of the PARSER model: increasing memory limitations led to a Zipfian advantage. Nevertheless, those parameter settings which produced a Zipfian advantage also led to very poor overall fit to the data (high RMSE and high variability in performance across conditions).

we can better understand the significance of word frequency distributions by taking into account some notions of memory or resource limitation.

To summarize: Under the evaluation scheme we used, the chunking models segmented a Zipfian language better than a uniform language, successfully simulating the patterns observed in the human data. The pattern was reversed for the transition-finding model—it segmented a uniform language better than a Zipfian language. This result suggests that sensitivity to statistically coherent chunks plays an important role in segmenting a language with a skewed frequency distribution.

5. General discussion

We presented two artificial language word segmentation experiments as well as simulations with four models, 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 lexical frequency was accounted for, we observed no remaining effect of distribution condition, suggesting that the sparsity of Zipfian languages posed no problem for learners. In the simulations, we found that the best fitting models were largely driven by consistent exposure to frequent chunks (Frank et al., 2010; Frank et al., 2013; Perruchet & Vinter, 1998), supporting a "chunking" view of statistical learning.

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. 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 (although bootstrapping effects were smaller).

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 both 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 relatively lower frequency common 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, Goodman, & Tenenbaum, 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.

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, Meylan, and Frank (2011).

Appendix A. Model and simulation details

A.1. Forward transitional probability

We implemented an online version of a simple forward transition probability (TP) model (Aslin et al., 1998; Saffran, Aslin et al., 1996; Saffran, Newport et al., 1996). We calculated conditional probability as:

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

where a and b are unigram syllable counts, $c(ab)$ is a count of the bigram ab, and V is the set of all bigrams. Sentence boundaries were not treated as a pseudo-syllable; unigram and bigram counts were calculated only within sentences.

In previous research, transitional probabilities have generally been computed over all the data in a given corpus. In contrast, our model updated unigram counts, bigram counts, and transitional probabilities at the end of each sentence. This was done to simulate a continuous time course of learning for comparison with the human learners. The TP model had only one free parameter, the threshold under which word boundaries were imposed. We systematically tested threshold increments of .025 in the range of 0–1, placing boundaries at all transitions under the threshold.

A.2. PARSER

The PARSER model (Perruchet & Vinter, 1998) is organized around a dynamic collection of chunks in a working memory. It explicitly represents the proposed lexicon in this buffer, maintaining a discrete weight for each item rather than a set of weights between sub-lexical chunks. PARSER serially scans the input stream of syllables and either chunks adjacent syllables according to the items maintained in the working memory, or chunks them randomly if they do not yet exist. Each chunk kept in the working memory decays at a constant rate, similar items interfere with each other, and chunks whose weights fall under a certain threshold are removed from the collection. We used the word-by-word chunking decisions, as determined by the weight of items in the working memory to simulate segmentation decisions. As correct lexical items accrued weight through exposure and incorrect ones received only limited or no support, more of the tokens in the input stream are correctly recognized as trials progress. In the current simulation, the weights are updated after every word to model human segmentation decisions.

The PARSER model as described has six free parameters, of which we varied two: the amount each item's weight decreases with the presentation of new material (.005–.1 in increments of .005) and the amount each item deceases as a result of interference between items (.004, .005, or .006). The other four parameters were left unchanged: the maximum number of primitives chunks in an item (3), the threshold of perception for an item (1), the gain in weight for reactivation (.5), and the initial weight of new words (1).

A.3. TRACX

TRACX is an implicit chunk recognition model based on a connectionist account of sequence learning (French et al., 2011). The learning algorithm relies on the recognition of previously encountered subsequences (chunks) in the input using an autoassociative neural network. Network error of a sequence is inversely proportional to the number of times those chunks have been seen together previously in the input. As the model proceeds through a string, units that comprise a sequence of syllables recognized as sequential are moved to a single representation in a hidden layer, and association is then assessed between that multisyllable item in the hidden layer and the next syllable. In this way, TRACX maintains a distributed representation of a probable lexicon encoded as weighted associations between syllables and syllable sequences. The codebase was adapted to output the parse, word-by-word, as guided by the network weights at the end of the last training epoch.

The model as described has four parameters: the number of repeated exposures to the data (epochs), the threshold for what qualifies as a chunk (the error criterion), the adjustment rate of the neural network (the learning rate), and the proportion of instances with which backpropagation takes place (the reinforcement threshold). For the current experiment, epochs were varied in increments of one from four to eight, criteria between .2 and .6 in increments of .1, learning rates from .02 to .06 in increments of .01, and the reinforcement threshold between .15 and .35 in increments of .05.

A.4. Bayesian lexical model

We chose a unigram model as an example of an ideal observer model that uses Bayesian statistics to assess the probability of different segmentation hypotheses. Börschinger and Johnson (2011) re-implemented a popular Bayesian word segmentation model (Goldwater et al., 2009) to use a particle filter (Doucet, Godsill, & Andrieu, 2000) rather than Gibbs sampling to estimate the posterior distribution over proposed segmentation hypotheses. As in Goldwater et al. (2009), the model defines a generative model for words and segments and uses Bayesian inference to establish the parameters of that generative model for the optimal segmentation of a text. A Dirichlet process governs the distribution of proposed lexicons, enforcing a distribution that favors smaller lexicons of shorter words.

The use of a particle filter in the inference stage allows for single-pass incremental processing of the input that yields a time course of learning comparable to the human timecourses in Experiment 2. The particle filter sequentially approximates a target posterior distribution with a number of weighted point samples (particles), updating each particle and its weight in light of each succeeding observation. A high number of poorly performing particles prompts a resampling from existing particles, with a higher probability of sampling from the better performing particles. This model has one free parameter in the current simulations, the number of particles used in the inference, in this case $2^{0–8}$. The concentration parameter for the Dirichlet process, a hyperparameter of the model, was set to the number of types for each input (6, 9, 12, and 24).

Appendix B. Additional simulation results

As illustrated in Figs. 6 and 7, the TP model showed a preference for input with a uniform word frequency distribution, and an overall higher level of performance than that of human subjects. The parameter setting with the best correlation, a TP threshold of .3, showed an advantage for the uniform condition across languages of all sizes. While the model reached peak performance less rapidly in both conditions as the number of types increased, scores were considerably higher than human performance on the same task. The parameter setting with the best RMSE, at a threshold of .95, displayed a similar uniform advantage across language sizes, along with very high F-scores. Deviating further from the observed human data, performance increased as the number of types increased.

As shown in Table 3, the Bayesian lexical model had the lowest RMSE, the second highest correlation, and demonstrated higher performance on Zipfian-distributed input across languages of various sizes. In both the parameter setting with the best correlation (32 particles) and the best RMSE (four particles), the model showed slightly higher performance than human subjects, and was especially resilient to the increase in the number of types. Through-

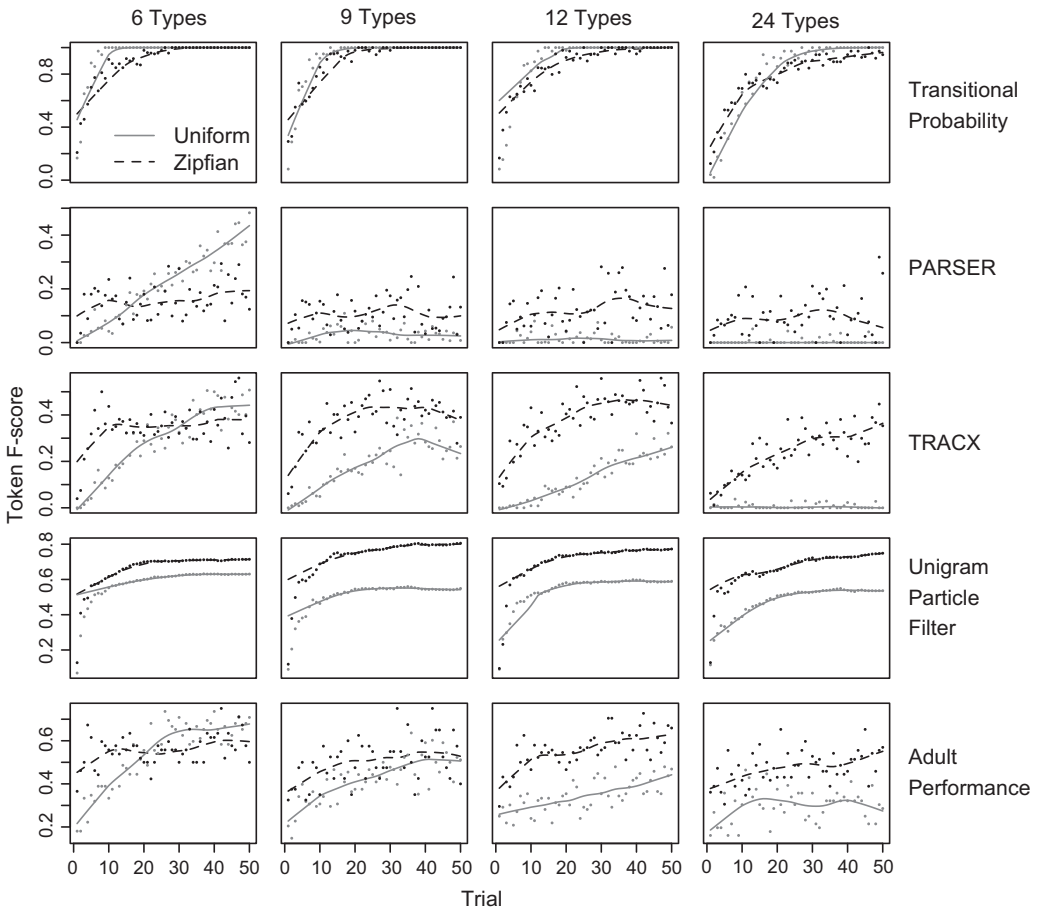

Fig. 7. Best-fitting model simulations (Pearson's *r*). Dots represent mean F-score for each trial; lines show a non-linear fit by a local smoother (loess, span = .4).

out the parameter space, the Bayesian lexical model learned better from input with a Zipfian word frequency distribution than from input with a uniform word frequency distribution.

TRACX had the highest correlation and the second lowest RMSE. The best parameter setting, at a learning rate of .2, five epochs, a criterion of .6, and a reinforcement threshold of .45, showed a preference for Zipfian data in three of four conditions. It also demonstrated a faster decay in performance in the uniform condition as the number of types in the language increased, a pattern qualitatively consistent with, but quantitatively more pronounced than, the human data. In the 24 type performance is at floor in the uniform condition while people achieve F-scores in the .25–.3 range. The parameter setting with the lowest RMSE, at a learning rate of .02, eight epochs, a criterion of .5, and a reinforcement threshold of .35, shows a similar gradual decrease in segmentation performance on the Zipfian data as the number of types increases, but a rapid drop in performance in the uniform condition.

PARSER grouped either with the forward transitional probability model or with the other lexical models depending on the forgetting rate, the rate at which the weights of items decayed in the model's working memory as new materials were observed. At the parameter settings with the highest correlation with human data, the model performed better on the input with the Zipfian word frequency distribution in the 9, 12, and 24 type languages. This highest Pearson's r came at a forget rate of .06 and an interference rate of .004. The highest correlation came at absolute scores much lower (from 0 to .3) than human performance (from .2 to about .8), however. Learning curves from the Zipfian data showed a decrease in performance in the last 10 trials was not characteristic of the human learners. Performance was at floor for both the 12 and 24 type uniform conditions, presumably because the high forget rate removed items from the working memory before they accrued any weight. High frequency items in the Zipfian condition, on the other hand, were recognized and maintained in the working memory. The lowest RMSE parameters were a forget rate of .005 and an interference rate of .004. With these parameters, PARSER performed better on the input data with a uniform word frequency distribution, but the model in the Zipfian condition showed

only a marginal increase in performance as a result of learning, while the model in the uniform condition learned to segment better than human subjects in the 9, 12 and 24 type languages.

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.

Aslin, R. N., Woodward, J., LaMendola, N., & Bever, T. (1996). Models of word segmentation in fluent maternal speech to infants. *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, 117–134.

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.

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

Bortfeld, H., Morgan, J., Golinkoff, R., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech stream segmentation. *Psychological Science, 16*, 298–304.

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

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

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.

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. R. (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*, 165–185.

Doucet, A., Godsill, S., & Andrieu, C. (2000). On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing, 10*, 197–208.

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.

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.

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

Frank, M.C., 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. C., Tenenbaum, J. B., & Gibson, E. (2013). Learning and long-term retention of large-scale artificial languages. *PLoS ONE, 8*, e52500.

Frank, M. C., & Gibson, E. (2011). Overcoming memory limitations in rule learning. *Language, Learning, and Development, 7*, 130–148.

Frank, M. C., Goldwater, S., Griffiths, T. L., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition, 117*, 107–125.

Frank, M. C., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science, 20*, 579–585.

French, R., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review, 118*, 614–636.

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* (Vol. 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*, 254–260.

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.

Hochmann, J.-R., Endress, A. D., & Mehler, J. (2010). Word frequency as a cue for identifying function words in infancy. *Cognition, 115*, 444–457.

Hollich, G., Jusczyk, P., & Brent, M. R. (2001). How infants use the words they know to learn new words. *Proceedings of the 25th annual Boston University conference on language development* (Vol. 1, pp. 353–364). Cascadilla Press.

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. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language, 44*, 548–567.

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

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

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83*, B35–B42.

Kurumada, C., Meylan, S. C., & Frank, M. C. (2011). Zipfian word frequencies support statistical word segmentation. In *Proceedings of the 33rd annual meeting of the cognitive science society*.

Lew-Williams, C., Pelucchi, B., & Saffran, J. R. (2011). Isolated words enhance statistical language learning in infancy. *Developmental Science, 14*, 1323–1329.

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *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.

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: Modelling psycholinguistic effects in speech segmentation. *Journal of Child Language, 37*, 545–564.

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., & Tillmann, B. (2010). Exploiting multiple sources of information in learning an artificial language: Human data and modeling. *Cognitive Science, 34*, 255–285.

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

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

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). 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.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology, 39*, 706–716.

Valian, V., & Coulson, S. (1988). Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language, 27*, 71–86.

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

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.