Cross-situational learning in a Zipfian environment

Andrew T. Hendrickson
Department of Cognitive Science and Artificial Intelligence
Tilburg University

Amy Perfors
School of Psychological Sciences
University of Melbourne

Abstract

Both adults and children have shown impressive cross-situational word learning in which they leverage the statistics of word usage across many different scenes in order to isolate specific word meanings (e.g., Yu & Smith, 2007). However, relatively little is known about how this learning scales to real language. Some theoretical analyses suggest that when words follow a Zipfian distribution, as they do in natural language, it should be more difficult to learn a lexicon because of the many low-frequency words that are only observed a few times (Blythe, Smith, & Smith, 2010; Vogt, 2012). Although this effect can be mitigated somewhat by assuming mutual exclusivity (Reisenauer, Smith, & Blythe, 2013), no mathematical analyses suggest that learning in a Zipfian environment should be *easier*. In this work, we show the opposite of the predicted effect using cross-situational learning experiments with adults: when the distribution of words and meanings is Zipfian, learning is not impaired and is usually improved. Over a series of experiments, we provide evidence that this is because Zipfian distributions help people to disambiguate the meanings of the other words in the situation.

Introduction

As soon as they start learning language, children begin acquiring words at an astonishing rate. At a conservative estimate, English speakers know tens of thousands of words by the time they are adults (Bloom, 2000). This is not just a difficult feat of memory, it is also a feat of inference: the referent of each word must be individually isolated from the many potential referents supplied by the world (Quine, 1960). How do people accomplish this?

One possible mechanism is cross-situational word learning (Pinker, 1984; Gleitman, 1990; Siskind, 1996). According to this idea, people are capable of tracking the statistics of word-and-object usage across multiple different scenes. If a word statistically co-occurs more often with a certain referent, this serves as evidence that that referent is the correct one for that word. There is considerable evidence from lab-based experiments that infants and children as well as adults can use such information to learn word meanings across multiple ambiguous presentations (Yu & Smith,

2007; L. Smith & Yu, 2008; Frank, Goodman, & Tenenbaum, 2009; Yu & Smith, 2012; Vlach & Johnson, 2013; Kachergis, Yu, & Shiffrin, 2016, 2009; Yurovsky & Frank, 2015; K. Smith, Smith, & Blythe, 2011; Aussems & Vogt, 2018; Kachergis & Yu, 2018; Kachergis, 2018). However, lab-based experiments typically present words in a uniform distribution in which each word and referent is equally frequent. Although a few cross-situational word learning studies have investigated smaller differences in frequency (e.g., Vouloumanos, 2008; Kachergis et al., 2016; Kachergis & Yu, 2013), none that we are aware of have empirically explored how learning occurs when the distribution of word frequencies are highly skewed (i.e., with a few highly-frequent words and many very low-frequency ones). These distributions, called Zipfian distributions, are ubiquitous across human languages (Zipf, 1949; Piantadosi, 2014), and as our analyses in Appendix A show, occur for nouns and all words in child-directed speech.

Why is distributional shape relevant? One problem is that Zipfian distributions may make it more difficult to acquire words on the basis of cross-situational statistics. Several mathematical analyses suggest that under a range of reasonable assumptions about the nature of the learner, the size of the lexicon, and the degree of referential uncertainty in typical situations, it should be more difficult to learn in Zipfian environments than uniform ones. For instance, Blythe et al. (2010) demonstrate that the key parameter affecting learning time is the degree of referential uncertainty the learner must grapple with – the number of possible referents that each word could mean (which depends on the total number of meanings in the language as well as the average number of possible referents in the specific contexts in which it occurs). As Figure 1 shows, in a Zipfian distribution with many low-frequency words, the uncertainty about the meaning of these words remains high because they occur so rarely that it takes an incredibly long time to encounter enough contexts to sufficiently disambiguate their meaning. Indeed, even the highly frequent words may be difficult to disambiguate from each other if they occur together in nearly every context.

Vogt (2012) points out that the problem is even worse if one assumes that the distribution of meanings (not just words) is Zipfian as well. As long as referential uncertainty is sufficiently high there is not enough information in a Zipfian distribution to identify within a lifetime the unique referents for words that occur rarely; even if it is not, learning in a Zipfian environment should be more difficult than a uniform one. Additional work by Blythe, Smith, and Smith (2016) investigates cross-situational word learning when the space of possible meanings is infinitely large. It suggests that although learning is guaranteed to occur eventually (albeit not necessarily within a lifetime), it will again be worse if the underlying frequency distribution of words is skewed rather than uniform.

It is important to note that even the most optimistic of these analyses conclude that, at best, cross-situational learning in a Zipfian environment is *no worse* than learning in a uniform environment. There is no mathematical analysis that we are aware of that suggests that cross-situational learning should be easier given a Zipfian distribution. This is a bit of a puzzle, given the profoundly Zipfian nature of language; yet the math is hard to argue with. As with any mathematical analysis, of course, the results rest on the assumptions made. In particular, these models presume that learners have perfect memory. This assumption makes sense given that their goal was to determine what *could possibly* be learned from different kinds of distributions. However, if the goal is to further explore the link to human learning, it is important to understand how learning might change when it involves agents with more limited capacities.

Why might memory capacity matter? One possibility involves an interesting potential interaction with the mutual exclusivity bias, in which learners presume that new words refer to new objects (Markman & Wachtel, 1988). Manipulating the assumption of mutual exclusivity has been

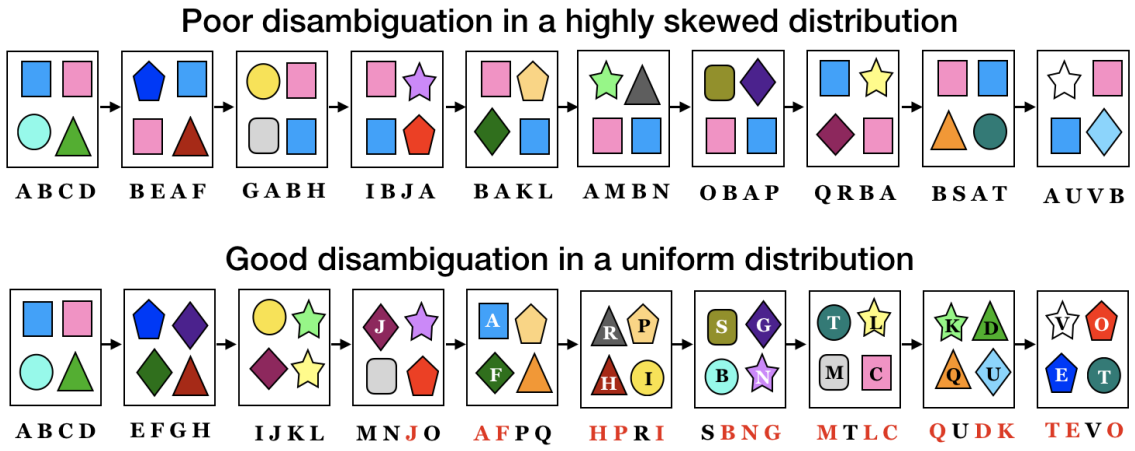

Figure 1. An illustration of why models predict superior learning when words are distributed uniformly. Each panel shows learning over ten screens/contexts, each with four objects (denoted by the shapes in the screens) and words (denoted by the letters at the bottom). **Top panel:** Objects and words are distributed in a maximally skewed way: two of them (the blue and pink squares) are highly frequent, occurring in every context, while every other word is very rare, occurring only once. In this distribution it is impossible for even an ideal learner to acquire any of the word-object mappings: the highly frequent blue and pink squares are clearly either A or B, but since they never occur apart from each other a learner cannot determine which is which. The low-frequency words do not occur frequently enough to be disambiguated either. **Bottom panel:** The same words and objects, distributed uniformly over screens. It is now possible for a learner with a perfect memory and the mutual exclusivity bias to acquire all word-object pairs. For instance, by screen four, the plum-coloured diamond is object J, because that is the only word common between both occurrences of the object. Similar reasoning enables the learner to acquire mappings for A and F on the fifth screen and H, P, and I on the sixth. Word R is learned (even though it is new) by mutual exclusivity.

shown to impact learning in multiple cross-situational word learning studies (Suanda & Namy, 2012; Yurovsky, Yu, & Smith, 2013; Halberda, 2003; Yurovsky & Yu, 2008; Kachergis, Yu, & Shiffrin, 2012). Furthermore, from a theoretical perspective, mutual exclusivity supports cross-situational learning: as Reisenauer et al. (2013) demonstrate, as long as referential uncertainty is sufficiently small, a mutual exclusivity bias removes "the undesirable increase in learning times that arises from non-uniform confounder distributions." The key, as before, is the degree of referential uncertainty given the capacity of a specific learner. If a learner has perfect memory, referential uncertainty depends only on the distribution of words and referents. However, for learners with *imperfect* memories, referential uncertainty is also affected by any memory biases that might exist: words that have been remembered act to decrease the referential uncertainty of the remaining items. Since people tend to remember more frequent items, then having some highly frequent items to act as "anchors" may impose enough of a learning advantage to overcome the other limitations imposed by Zipfian distributions. We test this possibility here.

The situation gets more complicated when one realizes that there are multiple kinds of robust Zipfian effects in language. The cross-situational literature has focused on what one might call Zipfian *frequency* distributions in which the most common words are very frequent and there are many extremely low-frequency items; this focus makes sense given the role that distribution shape plays in disambiguation. But there is also a well-known effect, sometimes called Zipf's law of ab-

breviation, in which the most frequent words are also shorter (Sigurd, Eeg-Olofsson, & van Wiejer, 2004; Strauss, Grzybek, & Altmann, 2006; Bentz & Ferrer-i-Cancho, 2015). This phenomenon is typically explained in terms of principles of minimum effort (Zipf, 1949), optimal coding (Ferrer-i-Cancho, Bentz, & Seguin, 2015), or efficient information transfer (Piantadosi, Tily, & Gibson, 2011b). Though the law of abbreviation does improve word segmentation performance (Kurumada, Meylan, & Frank, 2013), it has not typically been considered as an issue when it comes to cross-situational learning. However, if memory plays a role in making Zipfian distributions easier to learn from, then word length may be relevant as well. If shorter words are easier to remember than long ones, then having short *and* highly-frequent words might confer the most benefit of all: those items would be learned fastest of all, increasing referential certainty more quickly.

In this paper we experimentally investigate the issue of how people learn words cross-situationally when the distribution of those words and meanings is Zipfian, as it is in real life. In Experiment 1, we investigate how the shape of the distribution affects cross-situational learning when vocabularies are small and learned to completion. In this case, people are able to acquire the full set of words equally quickly regardless of the nature of the frequency distribution. In Experiment 2, we expand on these findings in a second experiment that increases memory demands by increasing the number of words in the vocabulary and also varies the length of words to test for the impact of Zipf's law of abbreviation. We find that with even more words in the vocabulary people learn *better* in a Zipfian environment, and that this occurs regardless of whether the word lengths follow Zipf's law of abbreviation or not. This experiment suggests that Zipfian distributions are actually helpful to learning. In order to explore why this is the case, in Experiment 3 we present the items individually rather than cross-situationally, thus removing referential ambiguity during training. This manipulation eliminates the advantage for Zipfian distributions, supporting the hypothesis that Zipfian distributions improve learning by aiding disambiguation during training.

Taken together, our results suggest that for people, word learning is generally *easier* in a Zipfian environment than a uniform one. How do these findings reconcile with previous work suggesting that for a learner with no memory limitations, Zipfian environments should usually be easier? We suggest that, taken together, research in this area is converging on the idea that Zipfian environments are helpful for humans because of our limited memory capacity: the highly-frequent words lower the referential uncertainty associated with the unknown words. We conclude by discussing the broader implications of these results.

Experiment 1: Learning a small vocabulary in Zipfian and Uniform environments

Our first study provides an initial exploration of how distributional shape affects cross-situational word learning by comparing performance with a small vocabulary of only 12 words.[1] People were randomly assigned to one of two conditions that varied according to the distributional shape of the items. In the UNIFORM condition, all items occurred equally often. In the ZIPFIAN condition, the items followed an approximate power law in which the most frequent items were very common and many occurred very rarely.

[1]The planned sample size, and exclusion criteria for this experiment were preregistered (`https://aspredicted.org/blind.php/?x=fi8zd4`) along with the Bayesian analyses of test data. No frequentist analyses were preregistered.

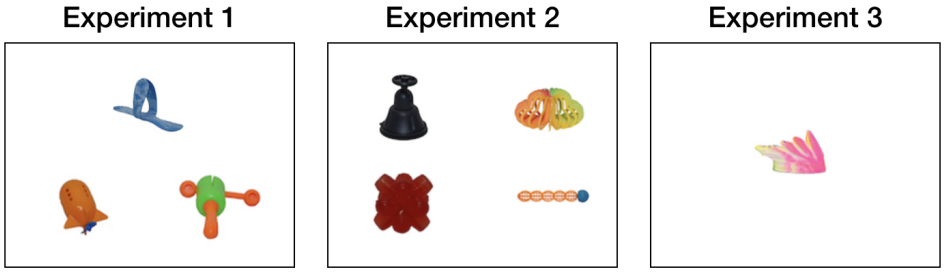

Figure 2. Sample training screens from each of the experiments in this paper. In Experiment 1 participants saw three images per screen, in Experiment 2 there were four, and in Experiment 3 there was only one. In all experiments the images were drawn from the same set of photos of novel but realistic objects and were taken from the NOUNS image corpus (Horst & Hout, 2015).

Method

Participants. Complete data was collected from 101 participants recruited from Amazon's Mechanical Turk and paid US$3 for the ten minute task. 36.7% were female, with ages ranging from 18 to 60 years old (mean: 34.2). 97% were from the US. Of these, 22 were excluded for failing performance checks: 9 for never improving above chance performance at test (7 in the UNIFORM condition) and 13 (4 in the UNIFORM condition) for failing to follow the task instructions (as described below), leaving 79 in the final analysis. Participants were randomly assigned to the UNIFORM condition (31 people) and the ZIPFIAN condition (48 people).

Materials. Visual stimuli consisted of 12 color images of novel objects, three examples of which are shown in Figure 2. The images were photos of novel but realistic objects, taken from the NOUNS image corpus (Horst & Hout, 2015). Each image was 150x150 pixels displayed against a white background. Audio stimuli in all conditions consisted of 12 novel one-syllable pseudo-words. We designed the words to sound English-like and be maximally distinct from each other. They were generated by the AT&T Natural Voices Text-to-Speech tool (Crystal voice). Appendix B contains the full list of images and words used in the study.

Procedure. In order to ensure that everyone could hear the audio stimuli and understand English, before the main experiment began participants were required to listen to an audio recording of two random English words ("boat" and "swim") and report them in a text box. People were then given instructions in which they were asked to learn the name of each of the objects they saw. They were told that each object had only one name, but they would never be shown one object at a time. They were also asked not to write down the objects since we were interested in how people use their memory to learn. After the instructions, people had to correctly answer comprehension questions about the task before they were permitted to begin. The experiment consisted of nine blocks, each containing a training and test block.

Training block. A training block consisted of 12 screens, each with three objects displayed and three words played, resulting in 36 total stimulus presentations per training block. The objects were arranged in a pyramid with one centered on top and two below it as in Figure 2. Everything was randomized across participants, including the mapping of labels to objects, the order of presentation of objects and labels within a block, and the location of objects on the screen. When each screen began, one of the words played 500ms after the objects appeared. Participants were asked to click on the object they thought each word referred to. After they did so, the next word was played after 500ms. No feedback was given.

Test block. After each training block participants were immediately shown the instructions for the test block. During testing people were shown all 12 images on the screen. The images were arranged in a random order and positioned in a grid with three rows and four columns. They then heard 12 words, also in a random order; for each word people were instructed to click on the object they thought the word referred to. After a pause of 500ms the next word would play. No feedback was provided. After each test block participants proceeded to the next training block by pressing the "Next" button when they were ready.

Exclusion criteria. Although participants were asked not to write anything down, since they were not physically present in the lab we could not visually confirm this. In order to exclude those that did, we removed participants according to a preregistered criterion. Reasoning that near-perfect performance was unlikely in the first few blocks without writing anything down, we excluded all participants who were correct on at least 11 of the 12 test items in either the first or second test block. These 13 participants (9 from the ZIPFIAN condition) were removed from all analyses.

Conditions. In all conditions people had to learn 12 word-object mappings, distributed across 12 screens in each training block. Each participant was shown a random allocation of labels to objects and a different random presentation order within each block. The conditions differed only in the distribution of word and referent frequencies. In the UNIFORM condition, all 12 words had the same frequency during a training block (three times each). In the ZIPFIAN condition, the distribution was highly skewed, varying in frequency from one to 11 presentations per word with a best fitting power law of $x^{-0.97}$. To eliminate item effects, each participant in this condition received a different random mapping of labels and objects to frequency, so that the same label or object was not always higher or lower frequency. Appendix B shows the distribution of word frequencies in the Zipfian condition.

Results

Our main question in this study was whether people in the ZIPFIAN condition learn all of the words in the vocabulary more slowly than in the UNIFORM condition. To evaluate this, in each test block we calculated the proportion of participants in each condition who achieved greater than 90% performance (i.e., who got 11 or 12 correct responses out of 12). The left panel of Figure 3 shows the proportion of people who had reached this criteria as a function of test block. Although participants did improve over time, there appears to be no difference in performance between the ZIPFIAN and UNIFORM distribution conditions: a preregistered Bayesian t-test found evidence against ($BF = 3.74$) a difference in the number of blocks to criteria between the ZIPFIAN (M = 5.08, sd = 2.87) and UNIFORM (M = 4.74, sd = 2.83) distribution conditions. [2] The analogous frequentist analysis shows the same pattern, a t-test found no significant effect of distribution condition on the number of blocks until criteria ($t(77) = 0.52, p = 0.61$) and a mixed effects logistic regression predicting the proportion of participants who exceeded the threshold with a random intercept per participant also found no effect of distribution ($\beta = 0.74, SE = 0.86, p = 0.40$) but a significant effect of block ($\beta = 1.2, SE = 0.12, p < 0.0005$). The same qualitative pattern of results are found when a random slope is added for each participant.

[2] Throughout this paper, Bayesian statistical analyses based on model comparison using Bayes Factors will be presented. Bayes Factors have clear interpretations (Kass & Raftery, 1995) and provide advantages over traditional frequentist analyses (Rouder, Speckman, Sun, Morey, & Iverson, 2009; Ly, Verhagen, & Wagenmakers, 2016). All Bayesian analyses were done using the BayesFactor package (v 0.9.12) in R.

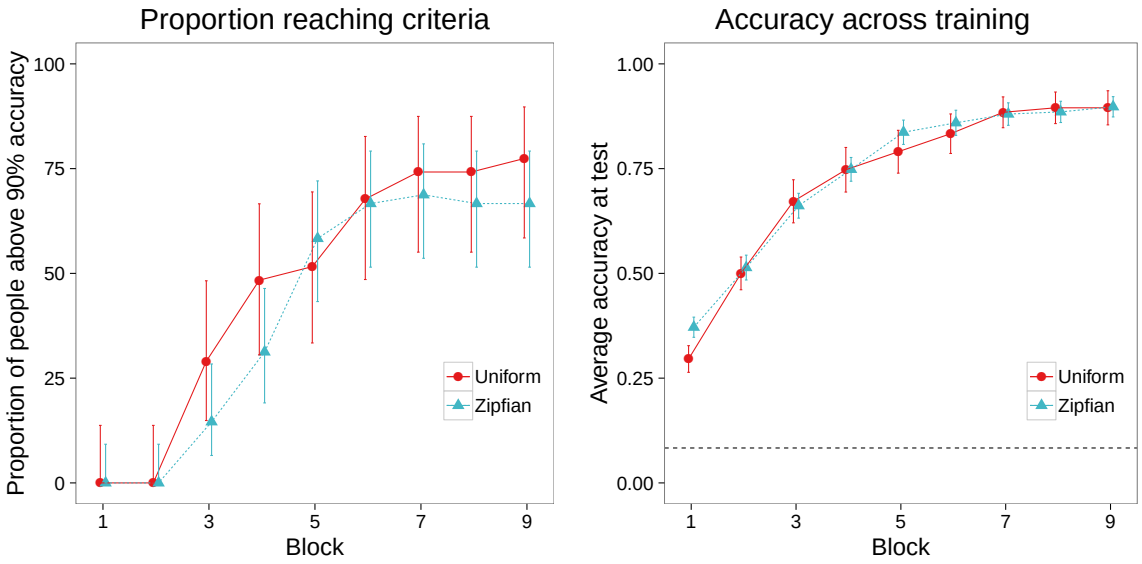

Figure 3. Performance in Experiment 1 learning a small vocabulary. **Left:** Ability to acquire the full vocabulary is measured based on the proportion of participants within each condition who exceeded the 90% accuracy criteria in the test phase after each block of training. There is no difference between the UNIFORM and ZIPFIAN conditions. Error bars reflect 95% confidence intervals. **Right:** Mean learning is measured based on average accuracy across all test items at each block. As before, there is no difference between the UNIFORM and ZIPFIAN conditions. Error bars reflect standard errors.

So far our results suggest that there is no difference in learning based on distributional shape, but that may be because we analyzed what proportion of people were able to learn the full vocabulary. It is theoretically possible that people might not differ in the time to acquire all of the words while still finding one distribution easier to achieve high *average* performance in. In order to test this, we compare average accuracy across test blocks, shown in the right panel of Figure 3. It suggests that there was no difference in mean accuracy between the UNIFORM and ZIPFIAN conditions. Consistent with this, a preregistered Bayesian linear regression predicting accuracy with block (as a continuous predictor) and a random intercept for each participant was preferred over a model with block, distribution, and a random intercept ($BF = 3.2$). The analogous frequentist mixed-effect regression with a random intercept per participant also found no significant effect of distribution ($\chi^2(1) = 0.143, p = 0.705$) but a significant effect of block ($\chi^2(1) = 602.16, p < 0.0005$).[3]

These results suggest that people learn the word meanings no more slowly from a Zipfian distribution than a Uniform distribution. This performance seems to contradict the prediction of the simulation analyses (Blythe et al., 2010; Vogt, 2012; Blythe et al., 2016) that argue Zipfian distributions of words and objects should significantly hinder learning. However, the vocabulary size of only 12 words was quite small by cross-situational word learning standards. We deliberately chose that so that we could evaluate the time needed to acquire *all* of the words, but it did mean that the task was far simpler than a typical experiment, much less the real world. As such, the lack of difference between conditions may have emerged due to ceiling effects in performance rather than

[3]The same qualitative pattern of results was found in a mixed effects regression containing both a random intercept and slope per participant.

any more interesting factors. In addition, for simplicity all of the words involved were one-syllable in length, which means that the experiment evaluated the effect of Zipfian *frequency* distributions but not the Zipfian law of abbreviation (in which the more frequent words tend to be shorter). We address these limitations in the next section.

Experiment 2: Learning a larger vocabulary as distributional shape varies

This experiment addresses the open questions left by the previous one by increasing the number of words in the vocabulary (to 28) as well as the amount of referential uncertainty (by presenting four objects and words per screen rather than three). As both the number of words (Blythe et al., 2016) and referential uncertainty (Reisenauer et al., 2013) increase, learning the words from the Zipfian distribution is predicted to be more difficult. As is typical in cross-situational word learning experiments, but unlike Experiment 1, participants were not expected to learn all the words in the vocabulary and thus were only tested at the end of the full training.[4]

Increasing the number of words provides the opportunity to manipulate the length of the words and test the impact on learning of the second aspect of Zipfian distributions in language: the law of abbreviation. This resulted in three conditions. In the UNIFORM condition, all items occurred equally often. In the ZIPFIANFREQUENCY condition, the items followed an approximate power law in which the most frequent items were very common and many occurred very rarely. However, the length of words was randomized relative to word frequency. The ZIPFIANLENGTH condition had the same highly skewed frequency distribution and the most frequent words were also the shorter ones, while the infrequent ones were the longest. It thus most closely parallels the environmental characteristics of human speech.

Method

Participants. 240 participants were recruited from Amazon's Mechanical Turk and paid US$3 for the ten minute task. Complete data was collected for 239 participants. 56.5% were female, with ages ranging from 19 to 79 years old (mean: 36.1). 94.8% were from the US. Of these, nine were excluded for failing the checks described below, this left 230 people in the final analysis. Participants were randomly assigned to the UNIFORM condition (76 people), the ZIPFIANFREQUENCY condition (76 people), and the ZIPFIANLENGTH condition (78 people).

Materials. Visual stimuli consisted of 32 color images of novel objects from the NOUNS image corpus (Horst & Hout, 2015). 28 were used as the training and test items, and four as "check" items for determining exclusion criteria, as explained below. Each image was 150x150 pixels displayed against a white background. Audio stimuli consisted of 32 novel pseudo-words ranging in length from one to three syllables. 12 words had one syllable, ten had two syllables, and six had three syllables; the four check words also had three syllables.

Procedure. As in Experiment 1, participants were given an audio check, basic instructions, and an instruction check before beginning the experiment.

Training phase. The first phase consisted of 71 screens. On each, four objects were displayed and four words were played, all in random order, resulting in 284 word presentations during training.

[4]The planned sample size, exclusion criteria, and Bayesian statistical analyses of the test data were preregistered for this experiment and Experiment 3 (`https://aspredicted.org/blind.php/?x=4ak9mh`). The frequentist analyses and analyses of training data were not preregistered.

The objects were arranged in a two by two grid. The timing was identical to Experiment 1 and as before no feedback was provided. After every 10 screens participants took a short break.

Test phase. Immediately after the training phase, the instructions for test phase were presented. In this phase, people were shown all 32 images (the 28 items of interest plus the 4 check items) in a grid with four rows and eight columns. The position of images within the grid was randomized. They then heard 32 words, also in a random order; for each word participants were asked to click on the object they thought the word referred to. After a pause of 500ms the next word would play. No feedback was provided.

Exclusion criteria. In order to ensure that participants were not writing anything down, we incorporated four predefined "check items" into the experiment which we expected that participants would be unable to answer correctly during the final test block unless they had been writing. Each check item appeared only *once* in the middle of the entire 284-word training and consisted of four of the 32 words and images; the check items appeared no different to the participants. These items were included in the final 32-item test screen. In order to ensure that we were comparing performance on the same 28-item set, the same four words and objects were used as the check items for all people (see Appendix B). The other 28 words and images were randomly assigned for each participant. During training each of the check items occurred on a screen with three non-check items.

Our reasoning was that a person who was not writing down words would be unable to correctly identify these items at test, given that they occurred in the middle of a long training and amidst 28 other words to be learned, all in the space of ten minutes or so. Even if the participants could narrow their options via mutual exclusivity, chance performance would still be far below 50%. We therefore excluded any participant in any condition who got more than two of the four check items correct at test and did not analyze their data further. All the analyses in the paper exclude the check items, incorporating only the 28 items of interest. Nine people were excluded in total, two from the UNIFORM condition, three from the ZIPFIANFREQUENCY condition, and three from the ZIPFIANLENGTH condition. An additional participant in the ZIPFIANLENGTH condition who verbally reported writing things down was also excluded.

Conditions. In all conditions people had to learn 28 word-object mappings, distributed across 71 screens each containing four word-object pairs (280 word presentations in total after excluding the four check items). The conditions differed only in the distribution of word and referent frequencies. In the UNIFORM condition, all 28 words had the same frequency during training (ten times each). In both of the two Zipfian conditions, the distribution was highly skewed, varying in frequency from one to 62 presentations per word with a best fitting power law of $x^{-0.99}$. Appendix B shows the distribution of word frequencies in the Zipfian conditions.

The Zipfian conditions differed from each other by whether the length of the words followed the law of abbreviation. In the UNIFORM and ZIPFIANFREQUENCY conditions, words were assigned randomly regardless of frequency. This means that in the ZIPFIANFREQUENCY condition, the most frequent (or least frequent) words could be of any length. (All words were equally frequent in the UNIFORM condition). By contrast, in the ZIPFIANLENGTH condition, assignment of words followed the law of abbreviation. The twelve one-syllable words were also the twelve most frequent ones, although within that subset the assignment of words to frequencies varied randomly across people. Similarly, the six three-syllable words were constrained to be the least frequent.

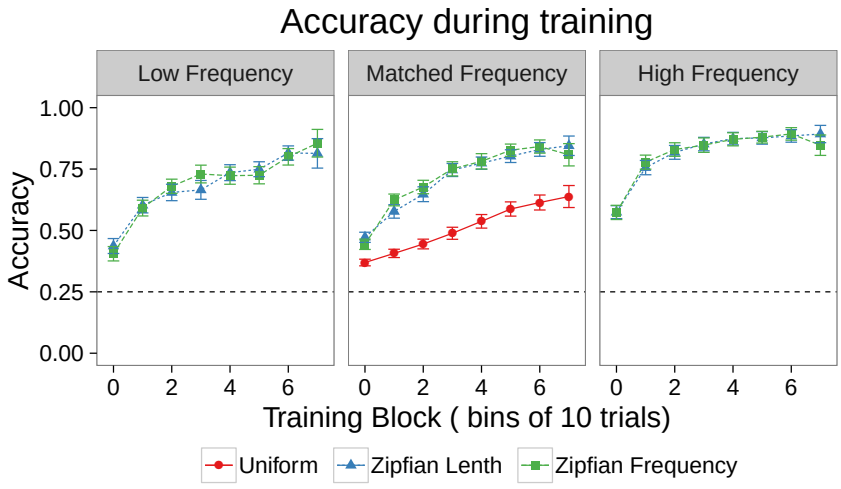

Figure 4. Mean accuracy in the training phase of Experiment 2. The three graphs are split according to the frequency of the words involved: low frequency words occurred one to five times during training (there were 12 of these in the two Zipfian conditions and none in the UNIFORM condition) while high-frequency words occurred 19 to 62 times (there were four of these in the two Zipfian conditions and none in the UNIFORM condition). The matched-frequency words are those that occurred between eight and eleven times during training, as did all of the words in the UNIFORM condition and twelve in the two Zipfian conditions. The x axis for each graph shows accuracy over the course of training, binned into blocks of ten screens each. Since the task involved picking one of four pictures in response to the label, chance is 25% (shown as a light dotted line). Accuracy in the two Zipfian conditions was far higher than in the UNIFORM condition, even comparing low-frequency words to the matched-frequency ones, although the Zipfian ones did not differ from one another. Error bars reflect standard errors.

Results

Training Phase. Over the course of training, were people able to learn the 28 words? Was there a difference in learning depending on condition? Figure 4 shows the accuracy in each condition over the course of the 71 training screens (each of which had four words). In all conditions mean accuracy is far above chance, but it is much higher in the Zipfian conditions, even when looking only at low-frequency words. Overall accuracy in each condition was 49.3% (sd: 17.7) in the UNIFORM condition, 75.0% (sd: 19.0) in the ZIPFIANFREQUENCY condition, and 74.4% (sd: 19.2) in the ZIPFIANLENGTH condition; a Bayesian linear regression with only distribution type predicting accuracy finds very strong evidence of a difference by distribution type ($BF > 10^{14}$) relative to an intercept-only model. Posterior estimates from the regression model of the differences between conditions show reliably lower accuracy in the UNIFORM condition relative to the ZIPFIANFREQUENCY condition (25.0 lower, 95% CI 18.0 to 31.9) and the ZIPFIANLENGTH condition (24.4 lower, 95% CI: 17.4 to 31.3) but no difference between the Zipfian conditions (95% CI on the difference spans zero, ranging from -7.4 to 6.2). Qualitatively similar results are obtained with the analogous frequentist tests, finding significant differences between the UNIFORM condition and both Zipfian conditions but not between the two Zipfian conditions. The ANOVA over distribution type is significant ($F(2,227) = 46.9, p < 0.0005$) and the the resulting post-hoc t-tests are significant when comparing UNIFORM to ZIPFIANFREQUENCY ($t(150) = 8.61, p < 0.0005$) and to ZIPFI-

ANLENGTH ($t(152) = 8.39, p < 0.0005$) but not the two Zipfian conditions ($t(152) < 1, p = 0.84$).

Of course, accuracy over all items may be somewhat misleading, because in the Zipfian condition the items occurred at very different frequencies. It would not be surprising if people were more accurate on the high-frequency item (which appeared 62 times across training) and less accurate on the low-frequency items (two of which appeared only once). How did frequency influence accuracy during learning within the Zipfian conditions and how does it compare to the UNIFORM condition? To address this we group the 28 items from the Zipfian distribution into three groups: items which are roughly matched in frequency to those in the UNIFORM condition (in the UNIFORM condition all items appeared ten times, and the Zipfian items occurred eight to eleven times, with an overall average of 9.2 times), items with low frequency below the matched group, and items with high frequency above the matched group. Considering only the data from the Zipfian conditions at first, a Bayesian linear regression with a random intercept for each participant and an effect of frequency was strongly preferred over an intercept-only model ($BF > 10^{10}$) as well as a model that also included the word frequency and word length conditions ($BF > 10^8$). The model parameters suggest there was a reliable difference between the high (mean: 80.9, sd: 21.2) and matched (mean: 69.9, sd: 10.3) frequency items (10.7 higher, 95% CI 9.7 to 20.0) but not between the matched and low (mean: 65.6, sd: 19.0) frequency items (4.2 higher, 95% CI -0.9 to 9.2). However, the low frequency items from the two Zipfian distribution conditions were reliably more accurate than all items in the UNIFORM condition (mean: 49.3, sd: 17.7, t-test $BF > 10^6$).

These findings are evidence of improved learning during training in the two Zipfian conditions compared to the UNIFORM one, though no evidence of differences between the two Zipfian conditions. The highly accurate performance during training for low- and matched- frequency words in the Zipfian conditions suggests that this advantage occurs because of improved disambiguation in the Zipfian conditions: people are good at figuring out the meanings of words they don't know by elimination, because they have already learned the high-frequency words. But does the ability in the Zipfian conditions to use mutual exclusivity to narrow down the true referent during learning lead to longer-term performance effects? Are the words actually learned better, or does the ability to disambiguate only help in the moment of disambiguation? To answer this, we turn to an analysis of test performance.

Test Phase. Did the improved disambiguation during training in the Zipfian conditions carry over to improved performance on the test items? We address this by examining overall accuracy rates at test. Overall accuracy indeed differed by condition, with UNIFORM attaining the lowest (mean: 34.3%, sd: 24.3), followed by ZIPFIANLENGTH (mean: 43.0%, sd: 22.5) and ZIPFIANFREQUENCY (mean: 45.4%, sd: 23.0). A preregistered Bayesian linear regression predicting average accuracy found support for a model that includes a difference between distribution conditions ($BF = 3.22$) relative to the intercept-only model. The analogous frequentist ANOVA also showed a significant effect of distribution condition ($F(2,227) = 4.82, p = 0.0089$).

As in the analysis of the training accuracy, overall accuracy can be misleading due to the difference in frequency across items. How did performance compare between conditions even among items that were similar in frequency? As before we address this by comparing the matched in frequency items from the Zipfian distributions to those in the UNIFORM condition. As shown in Figure 5, there is a strong difference in accuracy between conditions on those items: accuracy is 34.3% (sd: 24.3) in the UNIFORM condition, but 52.7% (sd: 29.1) in the ZIPFIANLENGTH condition and 49.9% (sd: 30.3) in the ZIPFIANFREQUENCY condition. The preregistered Bayesian linear regression found that this was very strong evidence in support of the model that includes the

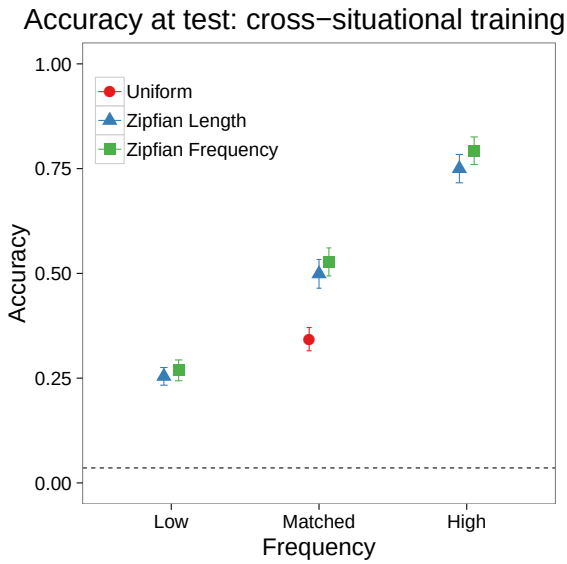

Figure 5. Accuracy during the testing phase in Experiment 2. Test items were split into bins based on their frequency during training (each of the 28 occurred once during test). The matched frequency bin occurred roughly the same amount (8 to 11 times) as all items in the UNIFORM condition (ten times), with those in the high frequency bin occurring more often than that and those in the low occurring less, sometimes as little as once. Accuracy in the two Zipfian conditions is higher for the key comparison, the matched frequency items. Chance is 3.1%, indicated by the dashed line; error bars reflect standard errors.

distribution ($BF = 192:1$). As before, the analogous frequentist ANOVA also showed a significant effect of condition ($F(2,227) = 9.54, p = 0.00011$).[5]

Are these differences in performance at test a result of simply being better able to eliminate the high-frequency items and guess better on the remaining ones? The difference in performance on the matched and low frequency items within the Zipfian conditions suggests that it is not. However, we can assure ourselves that this is the case by estimating how much benefit disambiguating the high-frequency items would provide. There were four high-frequency items that appeared more than 11 times during training. Even if we assume that participants in the Zipfian conditions learned all of them, this would leave 28 (i.e., 32-4) items to guess from at test for all of the other words. This implies that chance performance in the Zipfian conditions would effectively be 3.6% (i.e., $\frac{1}{28}$) rather than 3.1% (i.e., $\frac{1}{32}$). Thus, the improvement in accuracy we should expect to see if all improvement was all due to disambiguation during the test phase is 0.5% (i.e., 3.6% - 3.1%), far less that the observed improvement in accuracy relative to the UNIFORM condition of 18.4% for the

[5]In addition to these analyses, a preregistered mixed effects logistic regression on individual test item accuracy match the previous results. Specifically, a frequentist mixed effects logistic regression ($acc \sim frequency + syllables + distribution + (1|ID)$) found a significant effect of training frequency ($\chi^2(1) = 595.4, p < 0.001$), no effect of the number of syllables ($\chi^2(1) = 0.194, p = 0.66$), as well as a significant effect of the distribution ($\chi^2(2) = 10.211, p = 0.006$). The parameter estimates for the distribution conditions show a difference between the ZIPFIANFREQUENCY condition and the UNIFORM condition ($\beta = -0.71, SE = 0.30, p = 0.002$) but no difference between the ZIPFIANFREQUENCY condition and the ZIPFIANLENGTH condition ($\beta = -0.15, SE = 0.22, p = 0.52$).

ZIPFIANLENGTH condition or 15.6% for the ZIPFIANFREQUENCY condition .

Overall, Experiment 2 provides strong evidence that when words and referents follow a more Zipfian distribution, learning improves. This advantage occurs even for words that are equally matched in frequency and is not explained by improved guessing during test. Our analysis of learning during training suggests that the reason for this is that the high-frequency words in the Zipfian conditions allowed participants to eliminate referential uncertainty as they learned, thus succeeding in using mutual exclusivity to figure out the referents for the lower-frequency items.

If this is indeed what is happening, there is one strong test of this prediction. Presenting the items one-by-one rather than in a cross-situational context would eliminate any differences between conditions due to the amount of disambiguation provided. Indeed, there would be no need for disambiguation at all, because each item would appear paired with only one label. If the Zipfian advantage is due to improved ability to overcome referential uncertainty, this manipulation should remove the Zipfian advantage entirely. It is due to something else, it should not. We test this prediction in the next section.

Experiment 3: How does distribution shape affect learning when it's not cross-situational?

Experiment 3 is an exact replication of Experiment 2 except that each item occurred individually on its own screen during training. This eliminated the need for learning the mapping between words and objects, since each word was always presented with the correct object.

Method

Participants. 240 participants were recruited from Amazon's Mechanical Turk and paid US$3 for the ten minute task. Complete data was collected from 236 participants. 57% were female, with ages ranging from 18 to 70 years old (mean: 36.4). 93.9% were from the US. Of these, six were excluded for failing the checks described below, leaving 230 in the final analysis. Participants were randomly assigned to the UNIFORM condition (71 people), the ZIPFIANFREQUENCY condition (84 people), and the ZIPFIANLENGTH condition (75 people).

Materials. The exact same images and words as in Experiment 2 were used in this study. The frequency distribution of items in the Zipfian conditions was also identical. As before, there were three conditions: UNIFORM, ZIPFIANFREQUENCY, and ZIPFIANLENGTH.

Procedure. The test phase was identical to Experiment 2, but the training phase differed. In this experiment, people were shown only one object on the screen while hearing the corresponding word. Each word was played 1000ms after the object appeared, and each object was visible for a total of 2000ms. There was a blank screen lasting 500ms between each object presentation. Because the same total objects and words were shown in this experiment, people saw 284 screens of one item each rather than 71 screens of four each. After every 40 screens participants took a short break. The four check items were the same as in Experiment 2 and occurred once each. The exclusion criteria were also the same. Overall, six people were excluded, three from the UNIFORM condition, one from the ZIPFIANFREQUENCY condition, and two from the ZIPFIANLENGTH condition.

Results

Because training involved only one item per screen, participants made no selections during training and there was no training data to analyze. We therefore concentrate on test performance. Overall accuracy is very similar across conditions: 54.8% (sd: 30.2) in the UNIFORM condition,

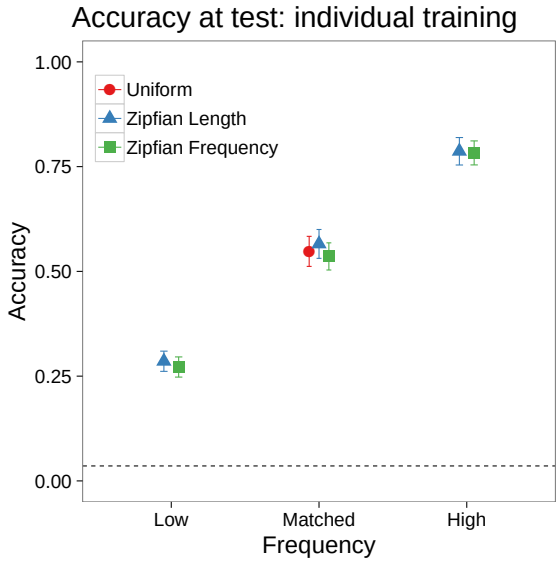

Figure 6. Accuracy during the testing phase in Experiment 3. Test items were split into bins based on their frequency during training (each of the 28 occurred once during test). The matched frequency bin occurred at roughly the same amount (9 to 11 times) as all items in the UNIFORM condition (ten times), with those in the high frequency bin occurring more often than that and those in the low occurring less, sometimes as little as once. Accuracy in all three conditions is identical for the key comparison, the matched frequency items. Chance is 3.1%, indicated by the dashed line; error bars reflect standard errors.

45.8% (sd: 23.2) in ZIPFIANFREQUENCY, and 47.7% (sd: 22.9) in ZIPFIANLENGTH. A Bayesian ANOVA finds that this is weak evidence in favor of the null hypothesis that there is no effect of distribution ($BF = 2.22$). The analogous frequentist ANOVA did not show a significant effect of condition ($F(2,227) = 2.58, p = 0.078$).

As before, overall accuracy might be somewhat misleading given that the items differed so markedly in frequency. We therefore also compare performance on the same items as in Experiment 2 that were similar in frequency. As shown in Figure 6, there is no difference in accuracy between conditions on these items: accuracy on the matched-frequency items is 54.8% (sd: 30.2) in the UNIFORM condition, but 56.6% (sd: 29.8) in the ZIPFIANLENGTH condition and 53.6% (sd: 29.7) in the ZIPFIANFREQUENCY condition. A Bayesian linear regression finds that this is very strong evidence in support of the null hypothesis that there is no effect of distribution ($BF = 18.3$), and the analogous frequentist ANOVA also does not show a significant effect of condition ($F(2,227) < 1, p = 0.82$). [6]

Taken together, Experiment 3 supports the conclusion that the Zipfian advantage in cross-situational word learning occurs because such a distribution supports learning through disambiguation. The high-frequency words can be learned quickly, thus helping people to narrow down the

[6] In addition to these analyses, a preregistered mixed effects logistic regression on individual test item accuracy match the previous results. Specifically, a frequentist mixed effects logistic regression ($acc \sim frequency + syllables + distribution + (1|ID)$) found a significant effect of training frequency ($\chi^2(1) = 591.6, p < 0.001$), a significant effect of the number of syllables ($\chi^2(1) = 17.9, p < 0.001$), but no significant effect of the distribution ($\chi^2(2) = 3.1, p = 0.22$).

meanings of the other words by eliminating some of the potential referents in each context. When this disambiguation advantage is removed by presenting each word and object individually, the Zipfian advantage is eliminated. Interestingly, there is no additional advantage when the words follow the law of abbreviation (with the shortest being the most highly frequent), even though this regularity is also ubiquitous in human languages. This, too, suggests that the Zipfian advantage occurs because of the disambiguation provided by high-frequency words and thus occurs in both the ZIPFIANFREQUENCY and ZIPFIANLENGTH conditions, regardless of whether the length of words correlated with their frequency.

Our results indicating a Zipfian learning advantage are interesting, but an open question remains: how to reconcile them with the mathematical analyses suggesting that at best learning in a Zipfian environment is equally difficult and in many circumstances should be more difficult (Blythe et al., 2010; Vogt, 2012; Reisenauer et al., 2013; Blythe et al., 2016). We suggest that people's performance reflects something that was not integrated into these mathematical models: memory limitations. Although some of the models assumed imperfect learners of various sorts, none modelled learners who could remember the most *recent* and *frequent* items better. Perhaps Zipfian distributions are well-tailored to the structure of human memory because the existence of high-frequency items provides an "in" for the system to gain a foothold. Indeed, this explanation is consistent with the reasoning suggested in Reisenauer et al. (2013) to explain why effective learning is so dependent on referential uncertainty.

General Discussion

This paper explored word learning in both cross-situational and non-cross-situational contexts. We began by exploring word learning cross-situationally, leaving the task of disambiguation over the course of many such screens up to participants. In all conditions, performance was above chance, even for words that were very low in frequency. This suggests that despite the inherent ambiguity of the learning environment, people were able to make informed guesses about the correct object for those words. These results are consistent with any computational model that can use disambiguating information to make informed guesses, including associationist (Fazly, Alishahi, & Stevenson, 2010; Kachergis et al., 2016, 2012; Yu & Smith, 2007; McMurray, Horst, & Samuelson, 2012; Räsänen & Rasilo, 2015), hypothesis testing (Frank et al., 2009; Trueswell, Medina, Hafri, & Gleitman, 2013; Kachergis & Yu, 2018), and hybrid models (K. Smith et al., 2011; Yurovsky, Fricker, Yu, & Smith, 2014; Yu & Smith, 2012; Yurovsky & Frank, 2015; Tilles & Fontanari, 2013) of cross-situational word learning. Since these experiments were not designed to distinguish amongst these models, but rather to address a different question, this is not surprising.

The question we were focused on was how word learning is impacted by Zipfian environments like those found in natural language in both child-directed and adult-directed speech. To explore this, in Experiment 1 we varied the distribution of words people were presented with, and people were trained until they learned the full set of words. In the UNIFORM condition, all of words occurred equally often. In both Zipfian conditions, the frequency of the words followed a power law. We found no difference across the distributions in the number of training blocks it took until people learned all of the words. These results challenge the predictions from the simulations of learning words that follow a Zipfian distribution (Blythe et al., 2010; Vogt, 2012; Reisenauer et al., 2013; Blythe et al., 2016), though the small vocabulary and low referential uncertainty may have disguised any effect of distribution. Nevertheless, our results are interesting because one of the main keys to the Zipfian disadvantage in the mathematical analyses was the problem of learning the low-

frequency words. Experiment 1 showed that even with this difficulty taken into account (i.e., even when all words were learned), there was still no advantage for the UNIFORM word distribution.

In Experiment 2, we increase the difficulty of the learning task by increasing the number of words to learn as well as the ambiguity during training. In the ZIPFIANFREQUENCY condition, this was the extent of the manipulation, and the length of the words was randomized relative to word frequency. In the ZIPFIANLENGTH condition, the length of the words followed the law of abbreviation observed in natural language, so that the shortest words were most frequent and the longest words were least frequent. We found that although people were able to learn words at an above-chance level in all conditions, performance was considerably better in the two Zipfian conditions than in the UNIFORM one. This is especially true when comparing words that were equally frequent in the UNIFORM and ZIPFIAN frequency distributions. These results are interesting in light of mathematical analyses suggesting that cross-situational learning should be easier (or, at most, no harder) in a uniform environment. Even the analyses that suggest mutual exclusivity should help (e.g., Reisenauer et al., 2013) learning do not suggest that Zipfian environments should make learning easier than uniform environments.

What is going on? Our hypothesis is that Zipfian environments are easier because they promote disambiguation based on mutual exclusivity. It might help during training because learning the high-frequency items quickly might mean there are fewer potential referents to disambiguate during learning, thus making it easier to figure out the meanings of the other words. Or it might help during test because having learned a few of the higher-frequency options means there are now fewer distractors to choose from when guessing the meanings of the others.

Why do we believe that this is driving our results? One factor is that performance in the ZIPFIANLENGTH and ZIPFIANFREQUENCY conditions was identical – and in both cases better than the UNIFORM condition – suggesting that it was the mere presence of a skewed frequency distribution of words that drove the effect. Since the disambiguation effect relies on precisely that, this makes sense. Second, and more tellingly, the Zipfian advantage was eliminated in Experiment 3, when learning was no longer cross-situational. Since it is only in the cross-situational contexts that disambiguation is necessary, this strongly suggests that the benefit provided by the Zipfian environment was because of the increase in disambiguation during training. This benefit was so strong that accuracy in Zipfian contexts with four items each (shown in Figure 5) was of a similar magnitude as accuracy when words were presented alone (shown in Figure 6). Given that words in the real world do not obligingly come one-by-one and nicely labeled, this is a substantial advantage indeed.

Why do Zipfian distributions provide this benefit? Although at this point we must speculate, it seems reasonable that it emerges at least in part due to the structure of human memory, which was not incorporated into the mathematical analyses that predicted poorer (or at best equivalent) learning in Zipfian environments. Perhaps the advantage for learning words from a Zipfian distribution comes from the fact that it is the high-frequency words – the very ones that are also most useful for disambiguation – that are the easiest for people to remember. Indeed, many researchers have suggested that human memory follows power-law relationships in memory or retrieval (see, e.g., Wickelgren, 1974; Wixted & Ebbesen, 1991; Wixted, 2004; Baronchelli, Ferrer-i-Cancho, Pastor-Satorras, Chater, & Christiansen, 2013). Why memory should follow this form is unclear, but some rational analyses suggest that it is because our learning system is well-adapted for recalling things that occur with a power-law distribution in the real world (Shepard, 1987; Anderson, 1990; Anderson & Schooler, 1991). A memory explanation of our findings is also consistent with another

otherwise difficult-to-explain finding reported by Piantadosi (2014) in which participants were asked to write a story about eight aliens about whom they knew nothing other than their names. The use of these novel eight words in the resulting 2000-word texts followed a Zipfian frequency distribution, a finding which Piantadosi (2014) suggests is due to "properties of human memory since it is hard to think of other pressures in this experiment that would lead people into power-law use of words."

It is quite interesting in light of this that there was no additional benefit for distributions that also followed the law of abbreviation, in which the more frequent words were also shorter. Given the prevalence of the law of abbreviation in human languages, plus the well-known fact that shorter words are easier to remember (e.g., Baddeley, Thompson, & Buchanan, 1975), one might have expected – indeed, we did expect – that both aspects would be helpful. One possibility is that the highly-frequent words in our study were so frequent that the additional benefit due to being shorter was not necessary. Another possibility is that the law of abbreviation emerges for reasons other than its benefit in cross-situational word learning. The origin of Zipfian distributions is widely researched, with most explanations pointing out that Zipfian distributions are optimal in an information-theoretic sense. Many of these information-theoretic explanations focus on the process of communication. For instance, there is evidence that Zipfian distributions are actually better explained by word-to-word predictability, which suggests that they occur because they allow for efficient information transfer during communication (Piantadosi et al., 2011b). Still other research finds that word length is correlated with aspects of meaning: shorter words are more ambiguous (Piantadosi, Tily, & Gibson, 2011a), speakers and listeners both tend to assume that longer formulations are associated with less predictable meanings (Horn, 1984), and people judge longer words to be more conceptually complex (Lewis & Frank, 2016). All of these effects tend to be explained based on communicative factors – language with these characteristics is more predictable and efficiently transmitted. Since our experiments did not incorporate communication, it is perhaps not a surprise that we did not see improved learning when the law of abbreviation was followed.

Other information-theoretic explanations focus on optimality in terms of compressibility: Zipf's law of abbreviation (Ferrer-i-Cancho et al., 2015) and word frequency distribution (Ferrer-i-Cancho, 2016) are coding schemes with optimal compressibility. This approach is also consistent with Zipf's principle of least effort (Zipf, 1949), which he put forward to explain the ubiquity of such distributions. Our work provides empirical support for this theoretical explanation. We found that when words follow a Zipfian distribution, they are learned better (perhaps because they are more information-theoretically optimal and have compressibility properties that human memory takes advantage of). When distributions are better learned, they are more likely to be passed on; over time, language might be Zipfian because it was shaped that way by the brain (Christiansen & Chater, 2008). This is speculative and much more empirical work is necessary to determine whether this process is plausible. But our work provides some compelling evidence that it might be.

Of course, it is quite likely that all of these explanations play some role, since language is shaped both by memory and the process of transmission. Moreover, all of these explanations emerge naturally from considering the cognitive pressures faced by human learners and communicators (e.g., Christiansen & Chater, 2016). Indeed, it may well be that language is overdetermined to be Zipfian. We are not trying to argue that transmission and communication play no role; however, our work is some of the first experimental research on human learning in Zipfian contexts that we are aware of that incorporates no communicative context at all. It therefore suggests that Zipfian environments may provide a word-learning advantage for compression or memory reasons alone.

It is also quite likely that the Zipfian advantage we observed here complements Zipfian ad-

vantages observed in different situations. For instance, Kurumada et al. (2013) found that word segmentation was improved when the words followed a Zipfian frequency distribution (although the word lengths in their task did not also follow the law of abbreviation). They suggest that this is because the high-frequency words in Zipfian environments were more easily learned and remembered, and these helped the learner to identify and segment the other words. For instance, if a learner hears the sequence *abcjkl* and already knows the word *abc*, that facilitates the identification of *jkl* in the future. The high-frequency words in our tasks appeared to serve an analogous disambiguating role. Just as the presence of *abc* makes it easier to segment *jkl*, if there are four objects in a context but you know object *a*, it is indeed easier to figure out the referents of objects *b*, *c*, and *d*.

These results provide converging evidence that the Zipfian distribution of words children hear does not hinder word learning and instead supports faster learning. Most importantly, it seems the advantage people gain from learning from a Zipfian distribution comes from a decrease in uncertainty about the referent for new words when seen in the presence of high frequency words that are very likely to have already been learned. This mechanism was not captured by previous simulation studies that predicted Zipfian distributions would hurt learning but is analogous to advantages in word segmentation when frequencies are skewed (Kurumada et al., 2013). Finally, we found no evidence that preserving the correlation between word frequency and word length had any impact on people's proficiency at learning words. The importance of the law of abbreviation for language learning is unclear more generally, but our work suggests that it does not seem to be critical for learning word meanings in a cross-situational context.

Acknowledgments

AFP received salary support from ARC grant DE120102378. Research costs and salary support for ATH were funded through ARC grants DP110104949 and DP180103600.

References

Anderson, J. (1990). *The adaptive character of thought*. Psychology Press.

Anderson, J., & Schooler, L. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

Aussems, S., & Vogt, P. (2018). Adults use distributional statistics for word learning in a conservative way. *IEEE Transactions on Cognitive and Developmental Systems*.

Baddeley, A., Thompson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 575-589.

Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in Cognitive Sciences*, *13*(7), 348–360.

Bentz, C., & Ferrer-i-Cancho, R. (2015). Zipf's law of abbreviation as a language universal. In *Capturing phylogenetic algorithms for linguistics.* Lorentz Center Workshop, Leiden.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Blythe, R., Smith, A., & Smith, K. (2016). Word learning under infinite uncertainty. *Cognition*, *151*, 18-27.

Blythe, R., Smith, K., & Smith, A. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, *34*(4), 620–642.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489 –558.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*(6), 1017–1063.

Fellbaum, C. (1998). *Wordnet*. Wiley Online Library.

Ferrer-i-Cancho, R. (2016). Compression and the origins of zipf's law for word frequencies. *Complexity*, *21*(S2), 409–411.

Ferrer-i-Cancho, R., Bentz, C., & Seguin, C. (2015). Compression and the origins of Zipf's law of abbreviation. *arXiv preprint arXiv:1504.04884*.

Frank, M. C., Goodman, N., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578–585.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*, 1–55.

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*(1), B23–B34.

Horn, L. (1984). Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. *Meaning, Form, and Use in Context*, *42*, 11–42.

Horst, J., & Hout, M. (2015). The novel object and unusual name (NOUN) database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*(4), 1393-1409.

Kachergis, G. (2018). Word learning: Associations or hypothesis testing? *Current Biology*, *28*(9), R555–R557.

Kachergis, G., & Yu, C. (2013). More naturalistic cross-situational word learning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the Cognitive Science Society* (pp. 710–715). Austin, TX: Cognitive Science Society.

Kachergis, G., & Yu, C. (2018). Observing and modeling developing knowledge and uncertainty during cross-situational word learning. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(2), 227–236.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009). Temporal contiguity in cross-situational statistical learning. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 1704–1709). Austin, TX: Cognitive Science Society.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012). Cross-situational word learning is better modeled by associations than hypotheses. *IEEE Conference on Development and Learning*, 1–6.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2016). A bootstrapping model of frequency and context effects in word learning. *Cognitive Science*, 1-33.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773–795.

Kurumada, C., Meylan, S., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, *127*, 439–453.

Lewis, M., & Frank, M. C. (2016). The length of words reflects their conceptual complexity. *Cognition*, *153*, 182–195.

Li, P., & Shirai, Y. (2000). *The acquisition of lexical and grammatical aspect*. Berlin and New York: Mouton de Gruyter.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold jeffreys's default bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32.

MacWhinney, B. (2000). The childes project (3rd ed.) [Computer software manual]. Mahwah, NJ.

Markman, E., & Wachtel, G. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121-157.

McMurray, B., Horst, J., & Samuelson, L. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*, 831–877.

Piantadosi, S. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin and Review*, *21*, 1112–1130.

Piantadosi, S., Tily, H., & Gibson, E. (2011a). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280-291.

Piantadosi, S., Tily, H., & Gibson, E. (2011b). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526-3527.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Quine, W. (1960). *Word and object*. Cambridge, MA: MIT Press.

Räsänen, O., & Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological Review*, *122*(4), 792–829.

Reisenauer, R., Smith, K., & Blythe, R. (2013). Stochastic dynamics of lexicon learning in an uncertain and nonuniform world. *Physics Review Letters*, *110*(258701).

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225–237.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Sigurd, B., Eeg-Olofsson, M., & van Wiejer, J. (2004). Word length, sentence length and frequency - Zipf revisited. *Issue Studia Linguistica*, *58*(1), 37–52.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39–91.

Smith, K., Smith, A., & Blythe, R. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*, 480–498.

Smith, L., & Yu, C. (2008). Infants rapidly learn word/referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.

Strauss, U., Grzybek, P., & Altmann, G. (2006). Word length and word frequency. In P. Grzybek (Ed.), *Contributions to the science of text and language: Text, speech, and language technology* (Vol. 31, p. 15-90). Springer.

Suanda, S. H., & Namy, L. L. (2012). Detailed behavioral analysis as a window into cross-situational word learning. *Cognitive Science*, *36*(3), 545–559.

Tilles, P. F., & Fontanari, J. F. (2013). Reinforcement and inference in cross-situational word learning. *Frontiers in behavioral neuroscience*, *7*, 163.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. (2013). Propose but verify: Fast mapping meets cross-situational learning. *Cognitive Psychology*, *66*, 126–156.

Vlach, H. A., & Johnson, S. P. (2013). Memory constraints on infants' cross-situational statistical learning. *Cognition*, *127*(3), 375–382.

Vogt, P. (2012). Exploring the robustness of cross-situational learning under Zipfian distributions. *Cognitive Science*, *36*, 726–739.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*, 729–742.

Wickelgren, W. (1974). Single-trace fragility theory of memory dynamics. *Memory & Cognition*, *2*(4), 775-780.

Wixted, J. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, *55*, 235-269.

Wixted, J., & Ebbesen, E. (1991). On the form of forgetting. *Psychological Science*, *2*, 409–415.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414–420.

Yu, C., & Smith, L. (2012). Modeling cross-situational word-referent learning: Prior questions. *Psychological Review*, *119*(1), 21–39.

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. *Cognition*, *145*, 53–62.

Yurovsky, D., Fricker, D., Yu, C., & Smith, L. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin and Review*, *21*, 1–22.

Yurovsky, D., & Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 30, pp. 715–720).

Yurovsky, D., Yu, C., & Smith, L. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, *37*(5), 891–921.

Zipf, G. (1949). *Human behavior and the principle of least effort*. New York, NY: Addison-Wesley.

Appendix A: Is child-directed speech Zipfian?

Given that one of the main motivations of this research lies in the implications it has for the acquisition of language by children, it would be helpful to first establish that the distribution of words children experience is actually Zipfian. We are specifically interested in how children resolve the problem of referential uncertainty in cross-situational contexts, which – especially as implemented in a typical experimental paradigm – translates most naturally to a problem resolving noun-label mappings. Thus, here we investigate whether the distribution of nouns in child-directed speech follows a Zipfian pattern.

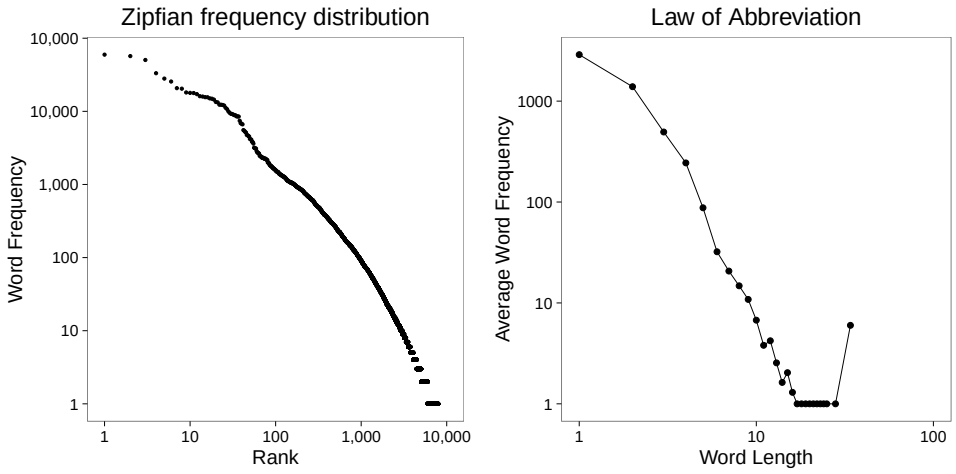

Figure 7. Distribution of nouns in child-directed speech in the CHILDES database. **Left:** Distribution of word frequencies, shown as a log-log plot with word frequency on the *y* axis and frequency rank on the *x* axis. The relationship is linear on the log-log plot, a hallmark of a Zipfian distribution. The most frequent items, with high frequency ranks, occur many thousands of times, while most items occur much less frequently and many occur only once. **Right:** Word length (*x* axis) and its relation to word frequency (*y* axis) follows the law of abbreviation in which shorter words are more frequent. The outlier is the word *supercalifragilousticexpialidocious*, a made-up children's word that is specifically interesting because it was designed to be abnormally long.

To explore this issue, we examined the distribution of nouns in child-directed speech in the CHILDES Parental corpus (MacWhinney, 2000; Li & Shirai, 2000). The frequencies of these items were calculated and made available by Ping Li and include utterances from parents, caregivers, and experimenters extracted from the original CHILDES database, which consists of corpora from 27 authors and contains about 2.6 million word tokens and 24,000 word types. We further identified the nouns in that database by including only those items which also occurred uniquely in WordNet (Fellbaum, 1998) as a noun (and not also a verb, adverb, or adjective), resulting in 1,255,235 word tokens and 7,950 word types.[7]

[7]Results are qualitatively similar, indicative of a Zipfian distribution, when the analysis included all nouns, regardless of whether they also occurred as other parts of speech, or when it includes all words in child-directed speech rather than only nouns.

We are interested in two aspects of the distribution of these nouns. First, do they follow a skewed distribution, with a few highly frequent items and many low-frequency ones? Second, if so, are the highly frequent items also shorter, as predicted by the law of abbreviation? The left panel of Figure 7 demonstrates that the answer to the first question is yes: the distribution of word frequencies is linear when graphed on a log-log plot, which is one of the hallmarks of a Zipfian distribution (Piantadosi, 2014). The most frequent nouns (pronouns or words like *boy* or *baby*) occur many thousands of times in the 2.6 million token corpus, while 1,969 items (24.8%) occur only once, like *prairie* or *deodorant*. Although this result only reflects the distribution of words children hear, not the referents they observe, it does suggest that their lexical input follows the kind of highly skewed distribution that occurs in adult speech as well.

The right panel of Figure 7 shows that the answer to the second question is also yes: the most frequent words also tend to be the shortest in child-directed speech. The sole outlier to that general trend is the proverbial exception that proves the rule: *supercalifragilousticexpialidocious*, a made-up children's word notable entirely because it is abnormally long. Other than that, the law of abbreviation holds: the most frequent words are short and there is a long tail of longer words that each occur very few times.

Overall, these results indicate that the linguistic environment children are learning nouns in is Zipfian, both in terms of the frequency distribution and in terms of the relationship of word length to word frequency. Children hear some nouns (which are usually short) very frequently and others (which are often longer) quite rarely.

Appendix B: Stimuli

Experiment 1: The frequency distribution of the 12 words in the Zipfian condition were: 11, 5, 3, 3, 3, 3, 3, 1, 1, 1, 1, and 1. The words used in this experiment were: boam, chave, glay, loid, naft, heef, plook, queed, rinch, shug, thorp, and zye.

Experiments 2 and 3: The frequency distribution of the 28 words in the Zipfian conditions were: 62, 33, 21, 19, 11, 11, 11, 11, 9, 9, 8, 8, 8, 8, 8, 8, 5, 5, 5, 4, 3, 3, 2, 2, 2, 2, 1, and 1. The four attention-check words were presented once each. The words used in these experiments were: boam, chave, glay, loid, naft, heef, plook, queed, rinch, shug, thorp, zye, ank, drib, mav, troz, vurl, wope, duppy, enzol, pazoo, impet, mandle, oidup, arturum, fobivan, kepata, and prebantik. The four attention-check words were: silodox, drezelist, yunkiter, and epichuf. In Experiment 2 the attention-check words appeared on the 52nd, 54th, 56th, and 58th training screen. In Experiment 3 the first item occurred between the 208th and 212th screens, the second between the 216th and 220th, the third between the 224th and 228th, and the fourth between the 232nd and 236th screen. This resulted in the the attention-check words appearing after the same number of words in both Experiments 2 and 3.

Figure 8. The 12 images used in Experiment 1, from NOUNS database (Horst & Hout, 2015)

Figure 9. The 32 images used in Experiments 2 and 3. The bottom row of objects were assigned to be attention-check referents for all participants. These images come from the NOUNS database (Horst & Hout, 2015).