The learnability consequences of Zipfian distributions:

Word Segmentation is Facilitated in More Predictable Distributions

Ori Lavi-Rotbain[a1,2], Inbal Arnon[b3]

[a]The Edmond and Lilly Safra Center for Brain Sciences, Hebrew University.

[b]Department of Psychology, Hebrew University.

[1]Correspondence author: orilavirotbain@gmail.com

[2]ORCID ID: 0000-0002-0461-0717; [3]ORCID ID: 0000-0001-8934-718X

Keywords

Language acquisition; Statistical learning; Information theory; Zipf's law; Word segmentation

Author Contributions

O.L-R. and I.A. conceptualized the study. O.L-R designed the experiments, collected and analyzed the data, and conducted the corpus analyses. I. A. wrote the original draft of the manuscript, with O.L-R providing extensive feedback and editing.

Abstract

One of the striking commonalities between languages is the way word frequencies are distributed. Across languages, word frequencies follow a Zipfian distribution, showing a power law relation between a word's frequency and its rank (Zipf, 1949). Intuitively, this means that languages have relatively few high-frequency words and many low-frequency ones. While studied extensively, little work has explored the learnability consequences of the greater predictability of words in such distributions. Here, we propose such distributions confer a learnability advantage for word segmentation, a foundational aspect of language acquisition. We capture the greater predictability of words using the information-theoretic notion of efficiency, which tells us how predictable a distribution is relative to a uniform one. We first use corpus analyses to show that child-directed speech is similarly predictable across fifteen different languages. We then experimentally investigate the impact of distribution predictability on children and adults. We show that word segmentation is uniquely facilitated at the predictability levels found in language, compared both with uniform distributions and with skewed distributions that are less predictable than those of natural language. We further show that distribution predictability impacts learning more than distribution shape, and that learning is not improved further in distributions more predictable than natural language. These novel findings illustrate learners' sensitivity to the overall predictability of the linguistic environment; suggest that the predictability levels found in language provide an optimal environment for learning; and point to the possible role of cognitive pressures in the emergence and propensity of such distributions in language.

Significance Statement

While the world's languages differ in many respects, they share certain commonalities: these can provide crucial insight on our shared cognition and how it impacts language structure. We offer a

novel learnability perspective on one of the striking commonalities between languages: the Zipfian distribution of word frequencies. We show that languages have similarly predictable distributions, and that these predictability levels are uniquely facilitative for word segmentation in children and adults. We explain their lower and upper bounds as a trade-off between the competing pressures of learnability and expressivity. These findings have far-reaching empirical and theoretical implications, illustrating learners' sensitivity to distribution predictability and pointing to role of cognitive biases in the recurrence of such skewed distributions in language.

Introduction

Over 6,500 different languages are spoken in the world today. These languages vary along many dimensions, but they also share certain similarities. These similarities can provide a window onto our shared cognition and the way it impacts language structure. One of the striking commonalities between languages has to do with the way word frequencies are distributed. Across languages, the frequency of words follows a Zipfian distribution, showing a power law relation between a word's frequency and its rank (*1*, *2,* see Equation 1[*]). Intuitively, this reflects the fact that languages have relatively few high frequency words and many low frequency ones, and that the decrease in frequency is not linear: the most frequent word is twice as frequent as the second most frequent word and so on. This pattern was first noted by the American linguist Zipf in the 1930's (1), and is often called Zipf's law. Zipfian, or near-Zipfian (2) distributions are repeatedly found across languages, for different parts of speech (including nouns, verbs and adjectives).

Equation 1: $f(r) \propto \frac{1}{(r+\beta)^{\alpha}}$

There are many different explanations for the origin of Zipfian distributions in language, with ongoing controversy about the significance of this law and whether it tells us something fundamental about language. On the one hand, such distributions are found across the physical world, where they are thought to reflect general mathematical principles not unique to language (e.g., scale-invariance (3)). However, their recurrence in language - a human creation - may nevertheless reflect foundational properties of language and/or human communication. While there is no agreed account of their source, their presence has been argued to be a form of optimal coding (4), and to create an optimal trade-off between speaker and listener effort (5).

Interestingly, less work has examined Zipfian distributions from a learnability perspective. Independent of their source, the presence and propensity of Zipfian distributions in language may have consequences for learning because of their greater predictability: Words are more predictable

[*] The formula for Zipfian distributions, as extended by Mandelbrot (40), appears in Equation 1. It shows the relation between word's frequency - f(r) and its rank - r, with two constants that determine the shape of the distribution: α sets the steepness of the curve, and β introduces a skew which enables a better fit to natural language (1, 2).

in Zipfian distributions than in less skewed distributions. Intuitively, guessing the next word is easier in a skewed distribution where words differ in frequency compared to a uniform distribution where all words appear equally often. This increased predictability could provide a better environment for learning by making it easier to learn the high frequency elements and use them as stepping stones for subsequent learning. Importantly, despite their more limited vocabulary, word frequencies in speech directed to infants and young children (child-directed speech) also follow a Zipfian distribution (6). The objects that infants see also have a skewed distribution, with few objects appearing very often and many objects appearing infrequently (7). That is, from early on, both the words children hear and the objects they see are skewed in a certain way.

Such skewed distributions may be particularly helpful for segmenting words, a crucial first step in breaking into language: High frequency words could serve as anchors for segmenting less frequent ones, as seen in infants' use of their own name to segment adjacent words (8). Interestingly, the skewed nature of children's input is not reflected in lab-based investigations of word segmentation: While studied extensively using artificial language learning paradigms (see (9) for a review), almost all such investigations expose learners to uniform distributions, where all words appear equally often. Such uniform distributions are less predictable than the ones children are naturally exposed to: using them may underestimate learning and overlook learners' sensitivity to the overall predictability of the language. Very few studies have examined the impact of skewed distributions on learning. Only one previous study asked whether word segmentation is facilitated in a Zipfian distribution compared to a uniform one (10). Using a standard forced-choice task, and an online measure of segmentation (where participants had to mark word boundaries on a sequence of written syllables, a paradigm developed by Frank et al. 2010 (11)), they examined performance across varying lexicon sizes in the two distributions. They found contextual facilitation effects in the Zipfian distribution, but no overall advantage: Performance was not better overall in the Zipfian condition, but words were segmented more accurately when they appeared next to more frequent words. That is, higher frequency words in the Zipfian condition served as an aid for the online segmentation of lower frequency words. While these findings highlight a potentially facilitative aspect of Zipfian distributions (contextual facilitation), they provide limited evidence for a general learnability advantage in such distributions. For starters, accuracy was not significantly higher in the Zipfian condition in either measure (forced-choice or online segmentation). Moreover, accuracy measures were calculated over all the words in the Zipfian distribution, including the most frequent one: we do not know how well participants fared on the lower frequency words, which serve as the crucial comparison. The highest frequency word in the Zipfian condition is expected to be learned better because of its' higher token frequency (relative to the words in the uniform condition). However, if Zipfian distributions confer a learnability advantage, it should be present for the lower frequency words as well. An additional study examined the impact of Zipfian distributions

on word learning using a cross-situational learning paradigm (6), where participants learn novel object-label associations by aggregating statistics over trials (12). Participants showed better learning of frequency-matched items in the Zipfian distribution compared to the uniform one, but this effect disappeared when there was no ambiguity (when each label was only presented simultaneously with one object). Based on this, Zipfian distributions were predicted to improve learning only in ambiguous tasks (where there are multiple options for segmentation or object-label mapping).

Taken together, these studies suggest that learning is impacted by distribution type, but they do not provide compelling evidence for a general learnability advantage in Zipfian distributions. More importantly, they do not tell us *what* about Zipfian distributions impacts learning: One intriguing possibility is that it is not the particular shape of the distribution that impacts learning, but the greater predictability it confers relative to less skewed distributions. That is, the relevant factor for learning may be distribution predictability rather than distribution shape, with certain predictability levels leading to improved learning. Such individual biases, if they exist, could be amplified over time to impact language structure (13), creating a cognitive pressure to maintain similarly predictable distributions. Here, we offer a novel theoretical account for **why** Zipfian distributions facilitate learning that makes concrete predictions about when learning will be facilitated and when it will not. In a nutshell, we propose that languages have similar distribution predictability values; that these predictability values enhance learning; and that their recurrence in language reflects a trade-off between learnability pressures on the one hand, and expressivity pressures on the other.

Research overview

In the present paper, we investigate the impact of distribution predictability on word segmentation to test the novel proposal that Zipfian distributions are facilitative because of their greater predictability. We test three theoretical predictions: (1) that different languages have similarly predictable distributions, (2) that learners are sensitive to distribution predictability and show enhanced learning at the predictability levels found in language, and (3) that the facilitation is driven by distribution predictability, and not distribution shape. That is, distributions that are similarly predictable should be similarly facilitative, even if they differ in their shape, but not the other way around. To test these predictions, we draw on information theory and its' recent applications to language (see (14) for a recent review). We operationalize the greater predictability of words in Zipfian distributions using the information-theoretic notion of efficiency (Equation 2), which is calculated using Shannon's entropy (15). Entropy quantifies the information content of a random variable (the amount of uncertainty) and has been shown to impact a range of linguistic phenomena, from the structure of the lexicon (16) and the distribution of color terms in the worlds'

languages (17, 18), through historical sound change (19), to online processing (20–22). Efficiency – which is the ratio between the observed unigram entropy and unigram entropy under a uniform distribution (see Equation 2) - tells us how predictable a distribution is relative to a uniform one with the same set size, allowing us to normalize entropy by set size. This is crucial if we want to compare distribution predictability across languages whose lexicon size may differ, and across different experimental paradigms.

In the first study, we provide a systematic characterization of the efficiency of word distributions in child-directed speech across fifteen languages (from eight language families) and show that languages are similarly predictable: they deviate from the uniform to a similar extent. In study 2, we test the effect of efficiency on learning in adults and children using a classic artificial word segmentation paradigm (23). We manipulate efficiency by changing the frequency distribution to make some words more frequent than others. By comparing performance on languages with varying levels of efficiency, we show that word segmentation is uniquely facilitated in language-like predictability in both children and adults: accuracy is higher in language-like efficiency compared both to a uniform distribution and to a skewed distribution less predictable than found in language. In Study 3, we further investigate the beneficial effect of language-like efficiency on segmentation in two ways. In Study 3a we test the prediction that distribution predictability matters more than distribution shape by comparing performance on distributions with similar efficiency values, but a different shape: a Zipfian distribution - where word frequency followed a power law - and a binary distribution - where one word was more frequent, and the rest had the same low frequency. In study 3b, we look at an efficiency value lower than that of natural language (an even more predictable distribution), to see whether it will lead to increased learning compared to language-like efficiency. We find improved performance in language-like efficiency, regardless of distribution shape, and no added facilitation when efficiency was reduced to be lower than language. These findings point to the role of additional pressures, like expressivity, in determining the consistent predictability levels and distribution shape found in natural languages.

Study 1: Word distributions are similarly predictable across languages in child-directed speech

We begin by asking whether word distributions in child-directed speech are similarly predictable (relative to a uniform distribution) across different languages. A recent corpus study of adult speech provides some support for this prediction (23). The authors examined unigram entropy and entropy rates across 1,000 languages using three parallel corpora (Bible translations, translations of European Parliament discussions, and translations of the Human Rights Declaration). They found similar and narrow unigram entropy values across different languages: Despite structural and morphological differences, words carry a similar amount of information across languages. While consistent with our prediction that languages have similar distribution predictability, these findings

are limited in several important respects. First, the study used parallel corpora (translations of the same text) to make the samples more comparable. This strength is also a limitation: the similarity between the languages could have been driven in part by the identical content conveyed. Second, most of the data comes from written, rather than spoken language, and may not reflect the information theoretic properties of day-to-day spoken interaction. Finally, and most importantly from our perspective, the study analyses adult speech, which differs in many respects from the linguistic environment children are exposed to: The existing findings do not reflect distribution predictability in speech directed to young learners.

To address these limitations, and to generate a measure of distribution predictability that normalizes for set-size, we calculated and compared the efficiency values of words in child-directed speech in 15 different languages from eight language families (English, German, French, Japanese, Dutch, Polish, Spanish, Swedish, Portuguese, Hebrew, Mandarin, Estonian, Danish, Catalan and Norwegian). We focus on child-directed speech because this is what infants and young children actually hear. The data was taken from multiple corpora from the CHILDES database (24) (Table 1). The criterion for inclusion was that the corpus contained at least 150,000 tokens. This was done to increase the reliability of the entropy estimates and the efficiency values based on them. Previous work has shown that entropy calculations are reliable with a corpus of 50,000 tokens and above (25), which means that all our samples are above the required lower bound. We calculated efficiency for each corpus using Equation 2, which calculates the ratio between the observed unigram entropy and the maximal entropy. In equation 2, *N* is the number of types and $p(x_i)$ is the probability of the i'th word.

Equation 2: $$Efficiency = \frac{observed\ entropy}{maximal\ entropy} = \frac{\sum_{i=1}^{N} p(x_i) * \log_2(p(x_i))}{log_2 N}$$

Theoretically, efficiency, like entropy, captures the overall predictability of the linguistic environment. However, efficiency allows us to normalize entropy by set size and compare predictability across corpora and experimental paradigms. This is crucial since the existing child-directed corpora vary in size across languages. In addition, efficiency is a bounded measure (ranging from zero to one) with a fixed comparison point: it tells us how predictable a distribution is relative to a uniform baseline. Intuitively, it is easier to interpret than entropy, which is unbounded in nature. This measure has been used in the past to study human cognition, for example, for normalizing the entropy of neuronal firing rate across brain regions (26).

Even though the corpora varied both in overall size (number of tokens) and in the size of the lexicon (number of types), efficiency values spanned a relatively narrow range: all values were between 0.59 and 0.7 (average 0.64, SD=0.03, Table 1). To further probe the stability of these values, we repeated the calculation, but this time divided each of the five larger corpora (American

English, British English, German, French and Japanese) into bins containing 500,000 words (see Materials and Methods). The efficiency values for the smaller bins were still in the same range: for the 36 samples we now have, the average efficiency value was 0.67 (SD=0.01), with a range of 0.65-0.69 (Table S1). That is, efficiency values span a narrow range across languages[†]. Interestingly, these values are lower than those of the artificial languages used in Kurumada et al. (2013, mean=0.84, SD=0.03, range: 0.867-0.808), raising the possibility that those languages were not learned better in the Zipfian distribution because they had less predictable distributions[‡].

In the next two studies, we test the impact of efficiency values and distribution shape using the classic word segmentation paradigm developed by Saffran et al. (23). In this paradigm, learners are exposed to recurrent tri-syllabic 'words' where the only cue to word boundary is the higher transitional probabilities between syllables within words (compared to across word boundaries). While used extensively, almost all existing studies expose learners to uniform distributions, where each word appears equally often.

Study 2: Language-like efficiency facilitates word segmentation in adults and children

We ask three questions in this study: (1) Is word segmentation improved in more predictable distributions? (2) Is the facilitation greater at the predictability values found in natural language? And (3) will we see a similar effect in children? To ask these questions, we examine the impact of efficiency on learners' ability to segment a four-word unsegmented artificial language. We manipulated efficiency by varying the frequency of the four words to create a skewed distribution with varying efficiency values, where one word is more frequent than the other three. We started with this binary distribution (which differs from the Zipfian one found in language, see Study 3), because it allows us to better examine the impact of efficiency on learning low-frequency words without having to control for frequency effects (since they all had the same frequency). We compared performance at three efficiency levels, based on the efficiency values we found in natural language: (1) *Maximal efficiency*: a uniform distribution where each of the four words appear equally often, as is the norm in this paradigm (efficiency=1). This is the least predictable distribution; (2) *Reduced efficiency*: a skewed distribution more predictable than the uniform, but less predictable than natural language, with higher efficiency than what we found in natural language (efficiency=0.85). This efficiency value is similar to that of the languages used in Kurumada et al. (2013), where no overall advantage was found in the Zipfian distribution; and (3) *Language-like*

[†] To ensure that the relatively stable and narrow deviation from the uniform we found is not dependent on the particular measure we used (efficiency), we also calculated it using Kullback–Leibler divergence (DKL) - a more commonly used measure for estimating the distance between two distributions – and found similar results (Table S2).

[‡] We calculated the efficiency of the artificial languages in Kurumada et. al (2013) according to the word frequencies provided in Figure 2 (page 443) of their article,

efficiency: a skewed distribution with efficiency values similar to those found in natural language, (efficiency=0.54, see Table 2 for full details). By comparing performance across the three levels, we can ask whether any increase in predictability leads to improved learning, or whether segmentation is more facilitated in language-like efficiency.

To further assess learning of the low frequency words, which are the crucial test for the impact of efficiency on learning, we added a second uniform condition ("uniform-short") where each of the words appeared only 19 times, as often as the lower frequency words in the reduced-efficiency condition. This additional condition allowed us to compare learning of words that appear the same number of times at different efficiency levels. Following exposure, participants in all conditions completed 16 two-alternative forced-choice trials where they heard a real triplet and a foil and had to say which was a real word in the language. If any increase in predictability improves learning, then performance should be similarly improved in the two skewed conditions compared to the uniform one. If, alternatively, learning is uniquely facilitated at certain predictability levels, then performance should be better in the language-like condition compared to the other two. We used the same conditions and procedure with adults (Study 2a) and children (Study 2b) to see if they are similarly impacted by distribution predictability.

Study 2a: Adults

Participants (N=142, mean age 24;0, 108 females) were randomly allocated to one of the four conditions ($N_{uniform}$=31; $N_{reduced}$=41; $N_{language\text{-}like}$=40; $N_{uniform\text{-}short}$=30). In the two skewed distributions, there were four different exposure streams, each with a different word serving as the frequent one (this did not impact the results, see supplementary information (SI) of Study 2a, so we collapsed over the four exposure streams in the analyses). As predicted, word segmentation was better in language-like efficiency compared to both the uniform and the reduced-efficiency conditions ($M_{langauge\text{-}like}$=81.66%, SD=15.9%; $M_{redueced}$=67.5%, SD=15.9%; $M_{uniform}$=65.7%, SD=12.5%, Fig. 1A). A mixed-effect model comparing these three conditions (with the uniform one as the baseline) showed that segmentation was significantly improved in language-like efficiency (β=1.27, SE=0.23, p<0.001), but not in the reduced efficiency condition (β=0.19, SE=0.2, p>0.1, Fig. 1A, Table S4, SI of Study 2a). Despite being more predictable, reduced-efficiency did not facilitate learning relative to the uniform condition. To further explore the lack of difference between the uniform and the reduced efficiency conditions, we conducted a Bayes factor analysis on a mixed-effect model comparing the two conditions, using the 'BayesFactor' package in R, (for more details: SI of Study 2a). We found strong support for the null hypothesis (no difference in accuracy, BayesFactor=18.28), strengthening the claim that the effect of efficiency is discontinuous. For comparison, when performing a similar analysis comparing the language-like and the uniform conditions, we found very strong support for the alternative hypothesis, stating there is a difference

in accuracy (BayesFactor=9,643). In an additional mixed-effect model, we asked whether language-like efficiency resulted in better performance than reduced-efficiency, and it did (β=1.12, SE=0.23, p<0.001, Fig. 1C, Table S5). This comparison also shows that frequency affects performance in a relative rather than an absolute way: Raw frequency predicted performance within conditions, but not across conditions. To give an example, a word that appeared only 9 times (the low frequency word in the language-like condition) was learned numerically better than a word appearing 71 times (the frequent word in the reduced condition, see Fig. 1C).

Importantly, the effect was not driven only by performance on the higher frequency word. We used a mixed-effect model to compare performance on the lower frequency words in the two skewed conditions to the words in the shorter-uniform condition (which had similar frequency). Here also, accuracy was better in language-like efficiency compared to the shorter-uniform condition ($M_{language\text{-}like\text{-}infrequent}$=78.8% vs. $M_{shorter\text{-}uniform}$=65.0%), despite the words appearing half the number of times (9 vs. 19 times, β=0.78, SE=0.22, p<0.001). There was no such facilitation in the reduced efficiency condition ($M_{reduced\text{-}infrequent}$=64.8%), even though each low frequency word appeared more often than in the language-like condition (19 times, as in the uniform-short condition, β= -0.002, SE=0.21, p=0.99, see Fig. 1B, Table S6). A Bayes factor analysis showed support for the null hypothesis (BayesFactor=33.7). The opposite pattern was found when comparing the language-like and the uniform-short conditions (support for the alternative hypothesis, BayesFactor=705.34).

The increased facilitation in language-like efficiency could not have been driven only by an anchoring effect (where the frequent word is learned early on and used to segment the lower frequency words (10)), since that should have led to improvement also in the reduced efficiency condition. The reduced efficiency condition still provided ample opportunity for anchoring: the low frequency words appeared next to the frequent word often (12-15 times, between 63-78% of the time, a proportion similar to that found in Kurumada et al. (10), see Table S3). Moreover, the frequent word was learned well allowing it to be an anchor. However, despite providing more anchoring opportunities than the uniform distribution (which provided no anchoring since all words were equally frequent), accuracy was not higher in the reduced-efficiency condition.

Study 2b: Children

Children (N=147, mean age=10;1 years; range: 9;0 to 12;0 years, 68 females) completed the same four conditions ($N_{uniform}$=30; $N_{reduced}$=47; $N_{language\text{-}like}$=40; $N_{uniform\text{-}short}$=30). Like adults, children showed better learning in language-like efficiency: they were more accurate in this condition compared to both the reduced efficiency and the uniform conditions ($M_{langauge\text{-}like}$=66.1%, SD=16.6%; $M_{uniform}$=60.2%, SD=11.5%; $M_{reduced}$=58.0%, SD=13.8%; Fig. 2A). A mixed-effect model comparing the three conditions (with the uniform as the baseline) showed that segmentation was significantly improved in language-like efficiency relative to the uniform condition (β=0.53, SE=0.17, p<0.01). No such facilitation was found in the reduced efficiency condition (β= -0.01, SE=0.15,

p>0.9, Fig. 2A, Table S8, SI for Study 2b). A Bayes factor analysis showed strong support for the null hypothesis (no difference in accuracy between the reduced and the uniform conditions, BayesFactor=28.44). A separate mixed-effect model, showed that language-like efficiency led to significantly better performance than reduced efficiency (β=0.44, SE=0.16, *p*<0.01, Fig. 2C, Table S9). Together, these results strengthen the finding that the effect of distribution predictability on learning is discontinuous.

The facilitation at language-like efficiency is also seen when looking only at the low frequency words. Children did not manage to learn words appearing 19 times in the reduced efficiency condition ($M_{reduced\text{-}infrequent}$=52.5%, SD=16.0%; not higher than chance, t(46)=1.06, *p*>0.1), and learned them relatively poorly in the uniform-short condition ($M_{uniform\text{-}short}$=54.17%, SD=12.5%; t(29)=1.8, *p*=0.07). However, they managed to learn words appearing half the number of times (only nine times) when presented in language-like efficiency ($M_{language\text{-}like\text{-}infrequent}$=62.5%, SD=17.3%; t(39)=4.57, p<0.001 compared to chance). A mixed-effect model comparing the low frequency words in the language-like condition to words in the shorter-uniform condition confirmed that segmentation was better in the language-like condition, despite the lower frequency (9 vs. 19 times, β=0.34, SE=0.15, *p*<0.05; Fig. 2B, Table S10). That is, children's word segmentation is also facilitated in language-like efficiency.

Study 3: Further investigation of the effect of efficiency on word segmentation

Study 2 showed that word segmentation is facilitated in children and adults at language-like efficiency. Here, we address two open questions. The first is whether this facilitation is driven by distribution shape or distribution predictability. In study 2, efficiency was reduced by making one word more frequent than the other three (hereon, a binary distribution). This design allowed us to better control for the effect of word frequency (since all the low frequency words within each condition appeared an equal number of times). However, this distribution does not resemble the Zipfian one found in natural language. It is possible that the effect we found is limited to the specific distribution used. In Study 3a, we use the same word segmentation paradigm to compare adults' performance on two skewed distributions (binary vs. Zipfian) in two efficiency levels (reduced vs. language-like). We predict similar performance in languages that have the same efficiency level, but differ in shape (and not the other way around): That is, we expect better performance in the two language-like conditions, regardless of distribution shape. This study also serves to replicate the greater facilitative effect of language-like efficiency compared to reduced efficiency (we will sample three additional efficiency values, one higher than language, and two in the language-like range). The second question is what happens at efficiency values that are lower than the ones found in natural languages? Does increasing predictability more lead to even better learning? If so, this would suggest that the absence of lower efficiency values in natural language is not driven by learnability constrains, but by other pressures, such as expressivity (e.g., in languages with lower

efficiency a greater mass of the distribution will be taken up by fewer words, see discussion). If, however, learning is not improved more at lower efficiency values, this would suggest that learning is optimal at language-like efficiency: learnability pressures on their own may explain the range of values found in language. In Study 3b, we explore this by comparing performance on a language with lower efficiency to the language-like condition.

Study 3a: Distribution predictability vs. distribution shape

To test the impact of distribution shape on segmentation, we created a two-by-two array, with efficiency (reduced versus language-like) and distribution shape (binary versus Zipfian) as the variables. In the binary conditions, one word was more frequent, while the other three had the same low frequency (as in Study 2). In the Zipfian conditions, word frequency followed a power law distribution (see Equation 1, with higher exponents for lower efficiency values). We compared the following four conditions: (1) a binary condition with reduced efficiency (the same condition as in Study 2a, efficiency =0.85); (2) a Zipfian condition with reduced efficiency (efficiency=0.83); (3) a binary condition with language-like efficiency (efficiency=0.65); and (4) a Zipfian condition with language-like efficiency (efficiency=0.61, see Table 3 for full details). We used the existing sample from Study 2a for condition 1, but collected new data for the three other conditions. We collected a new sample for the binary-language-like condition (even though we had a similar condition in Study 2a) for two reasons: to ensure a similar difference in efficiency between the two distribution shapes, and to show that the facilitative effect we found is replicated in an additional sample.

120 additional adult participants completed the three conditions (for all four conditions: N=161, age=23;7; 115 females). Figure 3A shows segmentation scores in all conditions for the most frequent word and the lower frequency words separately. Since the most frequent word is expected to be learned well in all conditions and might inflate the total segmentation score, it is important to look at success on low frequency words as well. As predicted, accuracy was higher in the two language-like conditions compared to the reduced conditions. This difference is found when looking at total segmentation score ($M_{language\text{-}like\text{-}binary}$=78.4%, SD=16.5%; $M_{language\text{-}like\text{-}Zipfian}$=78.4%, SD=16.1%; $M_{intemediate\text{-}Zipfian}$=74.7%, SD=12.8%; $M_{reduced\text{-}binary}$=67.5%, SD=15.9%), and becomes clearer when looking at performance on the low frequency words : Participants showed better learning of the low frequency words in the language-like conditions compared to the reduced efficiency ones ($M_{language\text{-}like\text{-}binary\text{-}infrequent}$=77.1%, SD=19.1%; $M_{language\text{-}like\text{-}Zipfian\text{-}infrequent}$=76.4%, SD=18.9%; $M_{reduced\text{-}Zipfian\text{-}infrequent}$=69.8%, SD=16.5%; $M_{reduced\text{-}binary\text{-}infrequent}$=64.8, SD=17.9%).

We used a mixed-effect model with frequency (frequent vs. infrequent, binary coded, see details in SI for Study 3a and Table S12), efficiency level and distribution type as variables, along with the triple interaction between them to see whether efficiency affected segmentation differently for the two distribution shapes, and whether the effect of frequency differed across efficiency levels

and distribution types. As predicted, performance was better at language-like efficiency (β=0.62, SE=0.15, p<0.001), but the effect of distribution shape, and the interaction between them were not significant (β=0.23, SE=0.14, p=0.096; β= -0.27, SE=0.221 p>0.1, respectively). That is, the effect of efficiency is greater than that of distribution shape. The interaction between frequency and distribution shape was significant: the frequent word in the Zipfian condition was learned better than expected compared to the binary condition (β=0.78, SE=0.35, p<0.05), which could suggest an impact of distribution shape on learning of the most frequent word. However, further analyses showed that this effect was only found in the reduced-efficiency condition (as can be seen in Fig. 3A and in the SI for Study 3a), suggesting it is not a robust one. These results support the facilitative effect of language-like efficiency (for two additional efficiency values), and indicate that in this experimental paradigm, distribution predictability impacts word segmentation, but not distribution shape.

As in Study 2a, anchoring cannot explain the full pattern of results, and does not provide an alternative explanation to efficiency. Anchoring predicts that the low frequency words in the two Zipfian conditions (reduced efficiency vs. language-like) would be learned equally well. The frequent word could serve as an anchor in both conditions (it was learned very well), and the lower frequency words appeared next to it a similar number of times (10.66 times on average in the Zipfian-reduced and 10.33 in the Zipfian-language-like, Table S3). Moreover, the low frequency words in the Zipfian-reduced condition appeared more times on average than in the Zipfian-language-like condition (19.3 versus 11.67 respectively), providing more opportunities to learn them. However, the low frequency words were learned significantly worse in the reduced condition compared to the language-like one (69.8% versus 76.5%, in mixed effect model reported in Table S13 in the SI of Study 3a: β= 0.35, SE=0.149 p<0.05), a pattern predicted by efficiency, but not anchoring.

Study 3b: What happens in a lower efficiency value?

Here, we wanted to see how a further reduction in efficiency would affect segmentation. We created a new binary language that was more predictable than natural language (efficiency=0.4, see Table 4) and compared it to the binary language-like condition from Study 3a. The new language had the same total exposure as the previous ones (N=128 tokens). We chose this value so that the change in efficiency is similar to the one between the language-like and the reduced efficiency conditions (binary-reduced=0.85, binary-language-like=0.65, binary-low=0.4). We did not use an even lower efficiency value since we wanted to have enough repetitions of the low frequency words. 40 additional adult participants completed this condition (age=23;1, 28 females): they showed accuracy similar to that found in the language-like condition (M_{low}=83.9%, SD=15.8%). A mixed effect model with frequency (frequent vs. infrequent, binary coded), efficiency level and the

interaction between them showed no difference between these conditions (β=0.3, SE=0.3 p>0.1): learning was not better in the lower-efficiency condition compared to the language-like one. The interaction between frequency and condition was significant, with the frequent word learned better than expected in the low efficiency condition (β=1.1, SE=0.44, p<0.05, Table S14). We do not have an explanation for this effect. Importantly, learning of the low frequency words was also not better in the low efficiency condition ($M_{low\text{-}infrequent}$=80.6%, SD=19.4%). A separate mixed-effect model showed no increase in performance compared to the language-like condition (β=0.28, SE=0.16, p>0.1, Table S15). That is, reducing efficiency to a level lower than found in language did not result in an additional increase in accuracy. We discuss the implications of this below.

Discussion

In the current paper, we set out to explore the possible learnability consequences of Zipfian distributions in language: while much work has debated their origin (3, 5), less research has examined their impact on learning. Given their propensity in language, do they provide a beneficial, or even optimal, environment for learners? Specifically, we ask whether the greater predictability of words in such distributions can facilitate word segmentation – a critical first step in language acquisition. This prediction receives some support from a previous study showing that Zipfian distributions provide more contextual facilitation for word segmentation compared to uniform ones (10). However, this study did not find an overall advantage for the Zipfian distribution, and did not provide separate analyses of how low frequency words were learned. Here, we go beyond existing findings to provide a novel theoretical account about when (and why) Zipfian distributions facilitate word segmentation. We quantify distribution predictability using the information-theoretic measure of efficiency, which captures how predictable a distribution is relative to a uniform one and provides a way to normalize entropy by set size. In the first study, we show that distribution predictability has very similar values in child-directed speech across fifteen different languages. Studies 2 and 3 use a classic artificial word segmentation paradigm to test the impact of distribution predictability and shape on learning by shifting the distribution away from the uniform one used in most SL studies. Study 2 shows that word segmentation is uniquely facilitated in language-like efficiency in both children and adults. Study 3 shows that the facilitation is driven by distribution predictability (efficiency) and not distribution shape, and that performance is not improved more when efficiency is lower than the lower bound found in natural language. That is, accuracy was higher in language-like efficiency compared both to a uniform distribution, and to skewed distributions with efficiency values higher than those found in natural language (see Fig. S1). Importantly, these findings cannot be explained by anchoring (where the frequent word is used to segment the lower frequency ones): while anchoring undoubtedly plays a role in word segmentation, it cannot explain the full range of results. The impact of frequency was also modulated by efficiency: While the frequent word was learned better than the low frequency words in all conditions, words in language-like efficiency were

learned better than words appearing more times in reduced efficiency. That is, frequency affected performance in a relative, rather than an absolute way.

Our findings show that learners are sensitive to the overall predictability of the linguistic environment, and that this sensitivity does not follow a continuous trajectory (see Fig. S1). Despite being more predictable than a uniform distribution, word segmentation was not improved when efficiency values were higher than those found in natural languages. It seems that there is a minimal increase in predictability that has to happen before learning is facilitated: the distribution needs to be skewed enough. This could explain the lack of overall advantage when word segmentation was previously assessed in a Zipfian distribution that did not have language-like efficiency (10): in this previous study, the Zipfian distribution was not predictable enough (it had reduced efficiency and not language-like efficiency), and consequently did not enhance learning. A similarly discontinuous effect is found in the animal learning literature where there are stronger neural responses to highly deviant (infrequent) stimuli (stimulus-specific adaptation (27)). While not identical, this literature highlights the benefit that low frequency items can receive in certain distributional environments. The lack of improvement in the reduced efficiency conditions also challenges a recent proposal suggesting that Zipfian distributions are beneficial only in ambiguous learning environments. A recent study found better cross-situational learning from a Zipfian distribution, but only when the task contained ambiguity (6): Learning was not improved in a Zipfian distribution when each object was only presented with one label. The authors propose that the facilitation stems from using the frequent word to reduce ambiguity quickly, and predict it will not be found in unambiguous learning settings (i.e., when learning already segmented object-label associations). However, since all our conditions involved ambiguity (there are many possible ways to segment the novel speech stream), such an explanation cannot explain the lack of facilitation in the reduced efficiency conditions. While ambiguity may play an important factor in how facilitative a skewed distribution is, it cannot explain the range of data explained by efficiency.

Our findings also suggest that segmentation does not continue to improve when efficiency decreases more: Accuracy was not greater in a language with lower efficiency (more predictable) than found in language. One interpretation of this is that the predictability values found in language are *optimal* for learning, so that increasing predictability more will not facilitate performance further. However, since the low frequency words in the low-efficiency condition appeared only six times (compared to nine or twelve in the language-like conditions), the smaller number of repetitions could have masked the effect of the lower efficiency. It is possible that increasing distribution predictability to levels higher than language can facilitate word segmentation under certain conditions. Importantly, the current study is only a first step in investigating the efficiency values that languages display and their impact on learning. More work is needed to track how changes in

efficiency impact learning, systematically investigate whether the effect is indeed discontinuous, and ask how learning is impacted by lower efficiency values.

Why do languages display a certain range of distribution predictability, as captured by similar efficiency values? One possibility is that learnability constraints alone drive both the lower and the upper bound of the efficiency range: Languages do not have higher (or lower) efficiency values because those are less optimal for word segmentation. However, in such a scenario, we would expect *worse* learning at lower efficiency levels, which is both unmotivated given the generally positive effect of increased predictability on learning (e.g., 29), and is unlike what we found in Study 3b. Alternatively, and more likely, the observed predictability values may reflect the impact of competing pressures on language structure (13, 29–31). Specifically, the narrow range of efficiency values may be shaped by two competing pressures: a learnability pressure on the one hand, and an expressivity pressure on the other. From a cognitive perspective, learners benefit from languages that are more predictable, creating a pressure for lower efficiency values. At the same time, languages with lower efficiency values are ineffectual from a communicative perspective. Having very low efficiency values would result in a language that is not expressive: such values can be obtained only if very few words take up a disproportionate part of the distribution. The noise present in any communication channel could create an additional need to push efficiency away from the two extremes (so they don't accidently fall into the two ineffectual boundaries). The middle region of the curve keeps languages as far as possible from these too extremes, ensuring that they that are not too surprising or inadequate for communication. The idea that learnability and expressivity both impact language structure is not new (e.g., 25, 32). However, existing studies estimate expressivity (using corpus data), but do not measure (or test) learnability. Their claims are based on an empirical assessment of one pressure (expressivity), but not the other (learnability). Here, we focus on that second pressure: we quantify learnability (using efficiency) and experimentally test its' effect on human learning. That is, we provide an empirical test of how (and when) it impacts learning.

However, an open question remains: If the relevant factor is distribution predictability, then why do languages consistently have Zipfian distributions? Since there are many different distributions that could have the same efficiency values, this alone would not explain the recurrence of Zipfian (or near-Zipfian) distributions in language. Several communicative and cognitive pressures may converge to make Zipfian distributions particularly advantageous for language learning and use. First, it is possible that for larger lexicon sizes, distribution shape will impact learning beyond the effect of distribution predictability. One limitation of the current findings is that our conclusions are based on learning an artificial language with only four words, a long stretch from the large lexicons in natural language. With numerous words to learn, as in natural language, the graded difference in frequency, which is a hallmark of Zipfian distributions, could facilitate

learning by making each higher frequency word an anchor for learning less frequent words. The high frequency words can serve as anchors to learn mid frequency words, while these mid frequency words can in turn help segment low frequency words. The graded difference in frequency may also be beneficial from a lexical access perspective - making each word more distinguishable from its' lexical neighbors. Moreover, the Zipfian distribution - with its graded frequency and particular slope - may be optimal for maintaining the facilitative language-like predictability levels for a large number of samples and sample sizes: when words vary in frequency, we need to use words from different regions of the frequency distribution to form an utterance. This means that the contrast between high and low frequency remains even within an utterance, which could make the utterance itself easier to segment. That is, the particular shape of the Zipfian distribution may confer a unique learnability advantage with large enough lexicons, by making words more distinguishable and allowing for stable predictability values for varying samples and sample sizes. These cognitive benefits are joined by communicative pressures: such distributions are claimed to create an optimal trade-off between speaker and listener effort (33). We are currently investigating these possibilities using computational simulations, mathematical modelling, and expanded word segmentation paradigms.

The improved segmentation we found in both children and adults also has implications for the statistical learning literature. Our results highlight the importance of using linguistic environments that resemble those of actual language, and the danger of experimental paradigms that strip away the multiple cues present in real-world learning environments. Using uniform distributions is useful for assessing the impact of one particular cue (e.g., transitional probabilities) on learning. However, presenting learners with environments that are less informative than natural language may limit our understanding of learning in the wild and lead us to underestimate learners' abilities (34, 35). This is especially risky when asking questions about what can and cannot be learned, as is often the case in developmental research. For instance, from two uniform conditions alone, we could have concluded that children (at the tested age) cannot use TPs to segment novel words when they appear only 19 times. This conclusion is not warranted given their performance in the language-like condition where less frequent words were learned well. Manipulating distribution predictability could similarly impact learning in other domains that have been studied in the lab. Language-like efficiency also seems to facilitates learning novel word-object associations (36), but its' effect on learning grammatical relations has not yet been examined.

We set out to assess the impact of distribution predictability and shape on language learning in children and adults. Our broader goal, however, is to study learning biases at the level of the individual as a way to understand both the consequences and the source of the kind of skewed word distributions found in language: Can individual learning biases explain the propensity of Zipfian (or near-Zipfian) distributions in language? This question is inspired by research

highlighting the way individual biases can be amplified over time to impact language structure (13). Our experimental results suggest that certain efficiency levels are better for learning than others. If substantiated, such individual learning biases may in turn shape language structure over time, by creating or maintaining similarly skewed distributions[§]. This proposal makes several testable predictions. The first is that languages will maintain stable efficiency values over time, even as they change and even when new words are added. Such stable patterns have been reported for the transfer of information across languages (32), as well as for the ratio between word and sequence entropy (37). The second, and harder to test prediction, is that efficiency values become more language-like in the process of emergence, for example in the development of new sign languages. We are currently testing both predictions using historical and diachronic corpora as well as iterated learning paradigms to see whether the individual learning biases we saw in the lab can emerge through the process of cultural transmission.

Conclusion

In this paper, we investigated the possible learnability advantage of one of the most striking commonalities between languages: the way words are distributed. We characterize the predictability of word distributions in natural language corpora using the information-theoretic measure of efficiency, which tells us how predictable a distribution is relative to a uniform one. We find that similar distribution predictability across languages. We then show that word segmentation is uniquely facilitated in language-like predictability for both children and adults. These findings illustrate the impact of a novel information-theoretical measure on learning; show that learners are sensitive to the structure of the environment as a whole; and point to distribution predictability as an important factor in learning. More importantly, the findings suggest that Zipfian distributions confer a learnability advantage, because of their predictability levels, and open up new directions in explaining their source.

Materials and Methods

Study 1: Corpus Analyses

We used all the available corpora for typically developing monolingual children in CHILDES (24) for the following languages: English, German, French, Japanese, Dutch, Polish, Spanish, Swedish, Portuguese, Hebrew, Mandarin, Estonian, Danish, Catalan and Norwegian. The data was taken from multiple child-parent dyads (over 110): we collapsed over the different dyads to create one larger corpus for each language. For each language, we counted the number of appearances of

[§] As discussed above, it may be possible to derive their propensity from the combination of these individual biases, and other structural and communicative constraints (such that Zipfian distributions will emerge as a byproduct of interacting pressures). We are currently exploring this idea using mathematical and computational tools.

each word (defined by orthographic form) and calculated the unigram entropy for the observed frequency distribution. We call this the observed unigram entropy. We then calculated the maximal entropy for this set size, which is the unigram entropy under a uniform distribution for the same number of types (e.g., if the corpus had 1,000 distinct word forms, we assumed each appeared the same number of times). The last step was to calculate efficiency for each corpus: the ratio between the observed unigram entropy and the maximal unigram entropy (see Equation 1). As can be seen from Table 1, efficiency values are similar across languages and corpus size.

Studies 2 and 3: Word Segmentation

Participants

All studies were approved by the IRB committee at the relevant university.

Adult participants: 302 participants took part in study 2a and Study 3 together (mean age 23;8; 219 females, 83 males). All were undergraduate students. All participants were native Hebrew speakers without learning or language disabilities. Adults read and signed a consent form prior to participating. They received 10 NIS or course credit in return for their participation.

Child participants: 147 children took part in Study 2b (age range: from 9;0 to 12;0 years, mean age: 10;1 years; 68 girls, 79 boys). Children's ages did not differ across conditions ($F(3)=1.69$, $p>0.1$). All children were visitors at the Bloomfield Science Museum in Jerusalem and were recruited for this study as part of their visit to the Living Lab. Parental consent was obtained for all children. All children were native Hebrew speakers without learning disabilities or attention deficits. Children received a small prize in return for their participation.

Materials

Auditory stimuli

Participants were exposed to one of the familiarization streams according to the experimental condition they were assigned to. All streams consisted of the same four tri-syllabic words: "dukame", "nalubi", "kibeto", and "genodi". We used only four words because we wanted to compare child and adult performance on a language that will be learnable for both. The 12 unique syllables were taken from Glicksohn & Cohen (38). They were created using the PRAAT synthesizer (39) and were matched on pitch (~76 Hz), volume (~60 dB), and duration (250–350 ms). The four words were created by concatenating the syllables using MATLAB to ensure that there were no co-articulation cues to word boundary. The words were matched for length (mean word length=860ms, range=845-888ms). The words were then concatenated together using MATLAB in a semi-randomized order to create the auditory familiarization streams. Importantly, there were no breaks between words or syllables and no prosodic or co-articulation cues in the stream to indicate word boundaries.

Experimental conditions

We used the same four words to create all our experimental conditions described in Tables 2-4. All

the skewed conditions had the same exposure length as each other, and as the uniform condition. For each of the skewed conditions, we created four different exposure streams: in each, the frequent word was different (to ensure the effect of efficiency is not driven by one particular word being easier to learn). Tables 2-4 and Table S3 show all the parameters of each condition.

Procedure

Children and adults wore noise-cancelling headphones while sitting in front of a computer. All participants were told they are going to listen to an alien language and that they need to pay attention and try to learn it as best as they can. The instructions were identical in all conditions. During exposure, a check-board image was displayed on the screen. After the familiarization phase, participants performed a segmentation test. On each trial, they heard two words and were asked to decide which belonged to the language they heard. They were told to guess if they were not sure. Each of the four words appeared once with each of the four foils to create 16 two-alternative-forced-choice trials. The trials appeared in a semi-randomized order, with the constraint that the same word/foil did not appear in two consecutive trials. The order of words and foils was counter-balanced so that in half the trials, the real word appeared first and in the other half, the foil appeared first. Our foils were non-words created by combining three syllables from three different words while maintaining their position ("dunobi", "nabedi", "kilume", and "gekato", average length: 860ms; range 854-868ms). Non-words, as opposed to part-words, never appeared together during exposure, making it easier to distinguish between them and real words. We used the "easier" non-words (rather than part words) to ensure that children will be able to complete the task, and because we did not set out to show that learners can be discriminate words from part-words (a finding shown extensively), but to see how efficiency affects this ability.

Data Availability

All raw data for Study 2 and 3 is available at http://dx.doi.org/10.23668/psycharchives.3009

Acknowledgments

We thank Zohar Aizenbud and Rana Abu-Zhaya for help with running the studies. We thank Israel Nelken, Roi Reichart and Yuval Hart for helpful comments and discussions. We thank Damian Blasi, Ram Frost, Noam Siegelman, and Shira Tal for feedback on previous versions of the paper. We thank the Living Lab staff and the Bloomfield Science Museum in Jerusalem, as well as the parents and children who participated. The research was funded by the Israeli Science Foundation grant number 584/16 awarded to the second author.

References

1. G. K. Zipf, *Human behavior and the principle of least effort.* (Addison-Wesley Press, 1949).
2. S. T. Piantadosi, Zipf's word frequency law in natural language: A critical review and future directions. *Psychon. Bull. Rev.* **21**, 1112–1130 (2014).
3. N. Chater, G. D. A. Brown, Scale-invariance as a unifying psychological principle. *Cognition* **69**, 17–24 (1999).
4. R. Ferrer-i-Cancho, C. Bentz, Optimal coding and the origins of Zipfian laws. *arXiv Prepr. arXiv* (2019).
5. R. Ferrer i Cancho, R. V Sole, Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci.* **100**, 788–791 (2003).
6. A. T. Hendrickson, A. Perfors, Cross-situational learning in a Zipfian environment. **189**, 11–22 (2019).
7. E. M. Clerkin, E. Hart, J. M. Rehg, C. Yu, L. B. Smith, Real-world visual statistics and infants' first-learned object names. *Philos. Trans. R. Soc. B Biol. Sci.* **372** (2017).
8. H. Bortfeld, J. L. Morgan, R. M. Golinkoff, K. Rathbun, Mommy and Me. *Psychol. Sci.* **16**, 298–304 (2005).
9. A. Romberg, J. R. Saffran, Statistical learning and language acquisition. *Wiley Interdiscip. Rev. Cogn. Sci.* **1**, 906–914 (2010).
10. C. Kurumada, S. C. Meylan, M. C. Frank, Zipfian frequency distributions facilitate word segmentation in context. *Cognition* **127**, 439–453 (2013).
11. M. C. Frank, H. Tily, I. Arnon, S. Goldwater, Beyond Transitional Probabilities : Human Learners Impose a Parsimony Bias in Statistical Word Segmentation. *Proc. 32nd Annu. Meet. Cogn. Sci. Soc.* (2010).
12. C. Yu, L. B. Smith, Rapid word learning under uncertainty via cross-situational statistics: Research article. *Psychol. Sci.* **18**, 414–420 (2007).
13. S. Kirby, H. Cornish, K. Smith, Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 10681–10686 (2008).
14. E. Gibson, *et al.*, How Efficiency Shapes Human Language. *Trends Cogn. Sci.* **23**, 389–407 (2019).
15. C. E. Shannon, A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
16. S. T. Piantadosi, H. Tily, E. Gibson, Word lengths are optimized for efficient communication. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 3526–3529 (2011).
17. E. Gibson, *et al.*, Color naming across languages reflects color use. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 10785–10790 (2017).
18. N. Zaslavsky, C. Kemp, T. Regier, N. Tishby, Efficient compression in color naming and its evolution. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 7937–7942 (2018).
19. U. Cohen Priva, Informativity and the actuation of lenition. *Language (Baltim).* **93**, 569–597 (2017).
20. T. F. Jaeger, Redundancy and reduction: Speakers manage syntactic information density. *Cogn. Psychol.* **61**, 23–62 (2010).

21. T. Linzen, T. F. Jaeger, Uncertainty and Expectation in Sentence Processing: Evidence From Subcategorization Distributions. *Cogn. Sci.*, n/a-n/a (2015).

22. R. Levy, Expectation-based syntactic comprehension. *Cognition* **106**, 1126–1177 (2008).

23. J. R. Saffran, R. N. Aslin, E. L. Newport, Statistical learning by 8-month-old infants. *Science* **274**, 1926–1928 (1996).

24. B. MacWhinney, The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. *Mahwah, NJ Lawrence Erlbaum Assoc.* (2000) https:/doi.org/10.1162/089120100750105984.

25. C. Bentz, D. Alikaniotis, M. Cysouw, R. Ferrer-i-Cancho, The entropy of words-Learnability and expressivity across more than 1000 languages. *Entropy* **19**, 1–32 (2017).

26. R. Pryluk, Y. Kfir, H. Gelbard-Sagiv, I. Fried, R. Paz, A Tradeoff in the Neural Code across Regions and Species. *Cell* **176**, 597-609.e18 (2019).

27. N. Taaseh, A. Yaron, I. Nelken, Stimulus-Specific Adaptation and Deviance Detection in the Rat Auditory Cortex. **6** (2011).

28. K. M. M. J. Diederen, T. Spencer, M. D. D. Vestergaard, P. C. C. Fletcher, W. Schultz, Adaptive Prediction Error Coding in the Human Midbrain and Striatum Facilitates Behavioral Adaptation and Learning Efficiency. *Neuron* **90**, 1127–1138 (2016).

29. M. H. Christiansen, N. Chater, Language as shaped by the brain. *Behav. Brain Sci.* **31**, 489–508; discussion 509-558 (2008).

30. M. Fedzechkina, T. F. Jaeger, E. L. Newport, Language learners restructure their input to facilitate efficient communication. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17897–17902 (2012).

31. B. MacWhinney, "The Competition Model" in *Mechanisms of Language Acquisition*, (1987), pp. 249–308.

32. C. Coupé, Y. Oh, D. Dediu, F. Pellegrino, Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Sci. Adv.* **5**, eaaw2594 (2019).

33. R. Ferrer-i-cancho, C. Bentz, Optimal coding and the origins of Zipfian laws.

34. L. C. Erickson, E. D. Thiessen, Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Dev. Rev.* **37**, 66–108 (2015).

35. R. Frost, B. C. Armstrong, M. H. Christiansen, Statistical Learning Research: A Critical Review and Possible New Directions. *Psychol. Bull.*, 1–87 (2019).

36. O. Lavi-Rotbain, I. Arnon, Children Learn Words Better in Low Entropy. *Proc. 41thth Annu. Conf. Cogn. Sci. Soc.* (2019).

37. U. Cohen Priva, E. Gleason, Simpler structure for more informative words : a longitudinal study. *Proc. 38th Annu. Conf. Cogn. Sci. Soc.*, 1895–1900 (2016).

38. A. Glicksohn, A. Cohen, The role of cross-modal associations in statistical learning. *Psychon. Bull. Rev.* **20**, 1161–9 (2013).

39. P. Boersma, V. van Heuven, Speak and unSpeak with Praat. *Glot Int.* **5**, 341–347 (2001).

40. B. Mandelbrot, An Informational Theory of the Statistical Structure of Language. *Commun. Theory*, 486–502 (1953).

Figures and Tables

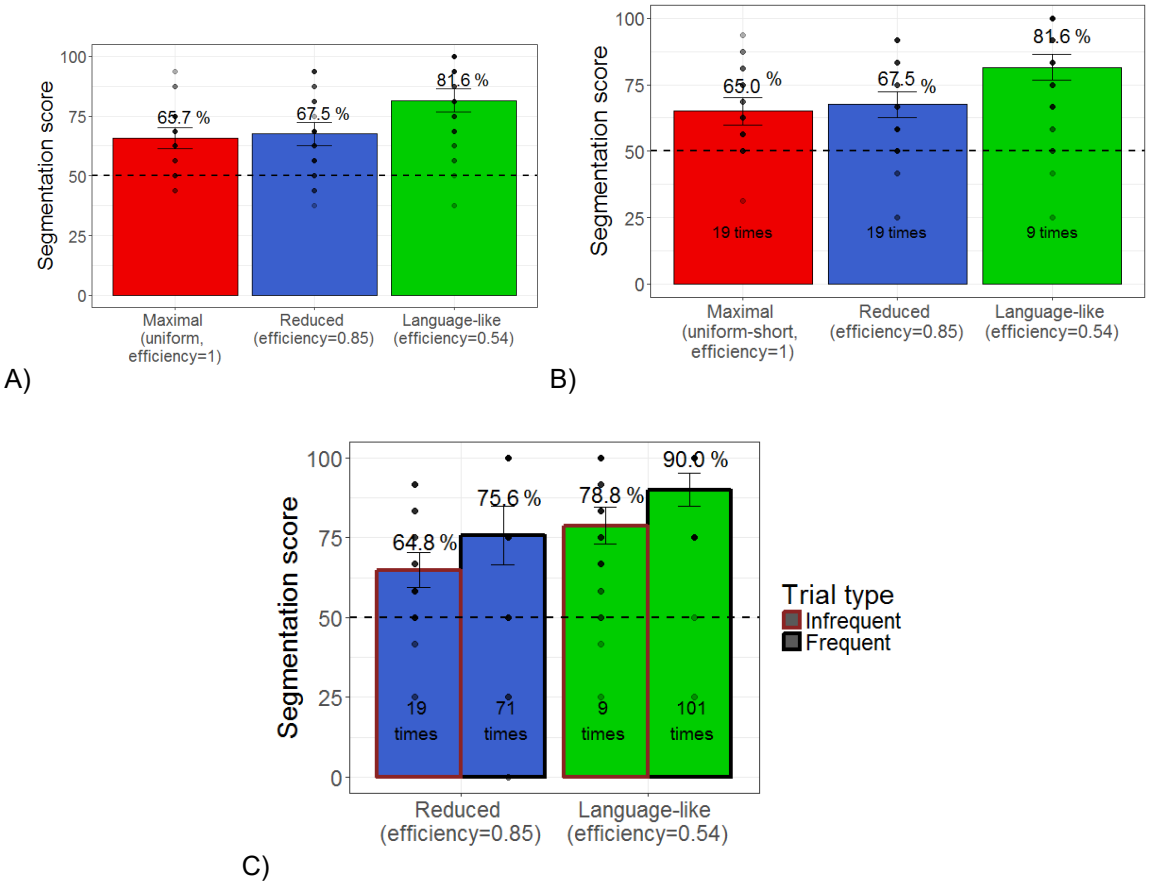
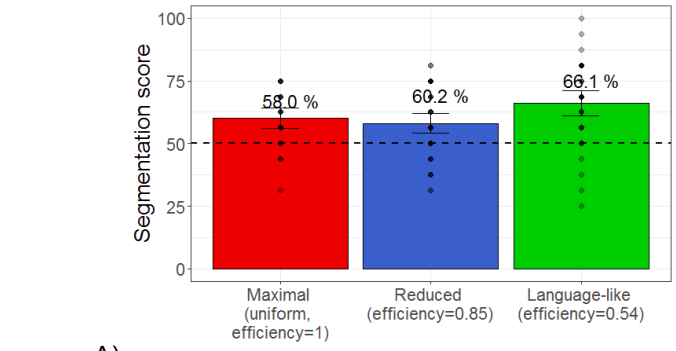
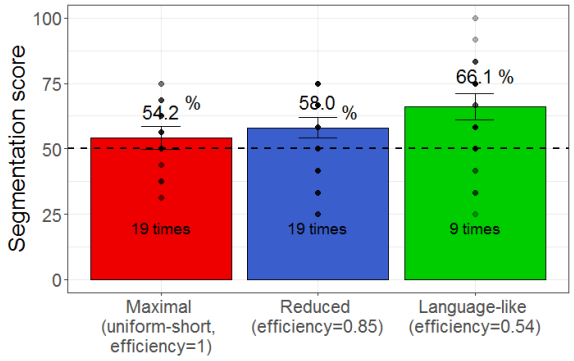
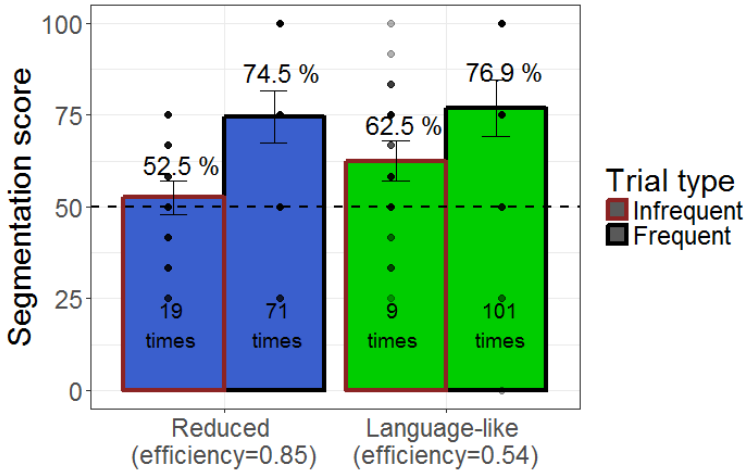


Figure 1. Adult segmentation scores from Study 2a. (A) Accuracy across conditions. (B) Accuracy only for the lower frequency words. (C) Comparing low frequency words (with brown borders) and high frequency words (with black borders) in the two skewed conditions. Dashed lines represent chance level. Error bars represents confidence intervals of 95%. Points represents individual scores with greater darkness for more individual participants. Numbers indicate frequency during exposure.



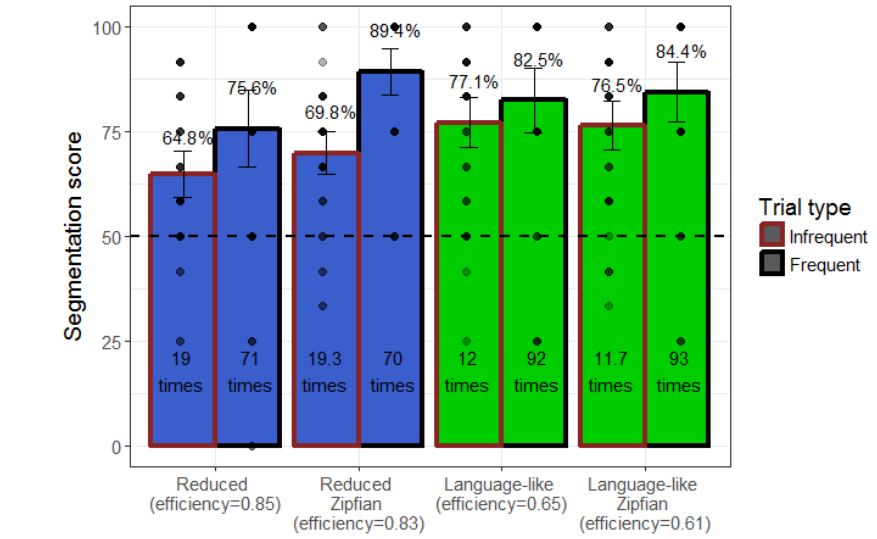
A) B)



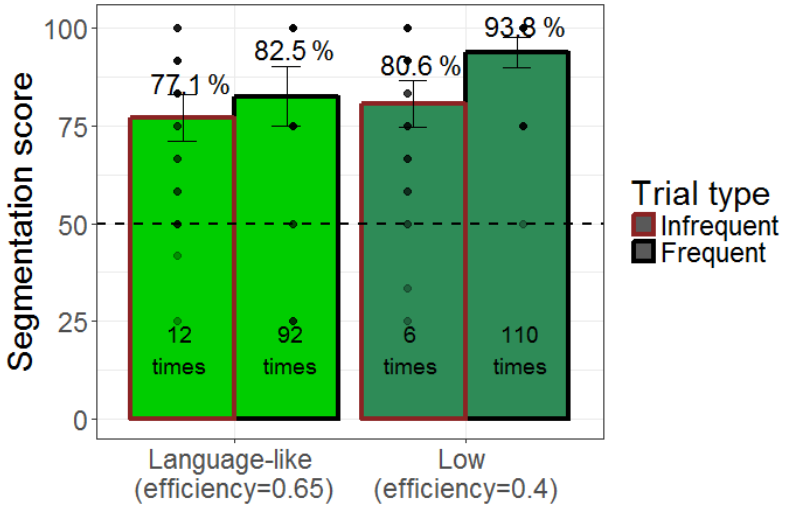


C)

Figure 2. Children's segmentation scores from Study 2b. (A) Accuracy across conditions. (B) Accuracy on the lower frequency words. (C) Comparing low frequency words (with brown borders) and high frequency words (with black borders) in the two skewed conditions. Dashed lines represent chance level. Error bars represents confidence intervals of 95%. Points represents individual scores with greater darkness for more individual participants. Numbers indicate frequency during exposure.



A)



B)

Figure 3. Comparing adult segmentation scores on low frequency words (with brown borders) and high frequency words (with black borders) across conditions: (A) Study 3a (note that the sample for the binary-reduced efficiency is the one from Study 2a); (B) Study 3b. Dashed lines represent chance level. Error bars represents confidence intervals of 95%. Points represents individual scores with greater darkness for more individual participants. Numbers indicate frequency during exposure (in the Zipfian conditions this is the average of the low frequency words).

Table 1. Summary of corpora measures across languages.

Language	No. Corpora	No. Tokens	No. Types	Word Frequency (per million)	Observed Entropy [bits]	Maximal Entropy [bits]	Efficiency
British English	12	7,066,980	30,470	1 - 333,802 (0.14 - 47,234)	8.83	14.9	**0.59**
North American English	34	6,404,744	34,593	1 - 294,493 (0.16 - 45,980)	8.99	15.08	**0.6**
German	7	2,177,584	37,236	1 - 71,561 (0.46 - 32,862)	9.16	15.18	**0.60**
French	8	1,938,055	21,776	1 - 75,134 (0.52 - 38,767)	8.86	14.41	**0.61**
Japanese	6	1,503,673	36,031	1 - 71,580 (0.67 - 47,603)	9.45	15.14	**0.62**
Dutch	5	1,149,781	19,870	1 - 45,815 (0.87 - 39,846)	8.67	14.28	**0.61**
Polish	8	794,232	44,555	1 - 32,172 (1.26 - 40,509)	10.35	15.44	**0.67**
Spanish	12	608,277	15,485	1 - 22,652 (1.64 - 37,239)	8.97	13.92	**0.64**
Swedish	2	376,879	10,035	1 - 19,259 (2.65 - 51,101)	8.4	13.29	**0.63**
Portuguese	2	352,661	8,611	1 - 23,828 (2.84 - 67,566)	8.23	13.07	**0.63**
Hebrew	6	306,765	14,095	1 - 16,372 (3.26 - 53,369)	9.37	13.78	**0.68**
Mandarin	2	216,715	8,853	1 - 9,217 (4.61 - 42,530)	8.66	13.11	**0.66**
Estonian	5	216,504	12,072	1 - 12,224 (4.62 - 56,460)	9.5	13.56	**0.70**
Danish	1	194,765	4,924	1 - 10,741 (5.13 - 55,148)	7.71	12.27	**0.63**
Catalan	4	189,844	7,970	1 - 12,318 (5.27 - 64,884)	8.71	12.96	**0.67**
Norwegian	2	184,676	8,342	1 - 9,196 (5.41 - 49,795)	8.75	13.03	**0.67**
Summary					**Mean = 8.91** (SD=0.6)	**Mean = 13.96** (SD=0.98)	**Mean= 0.64** (SD=0.033)

Table 2. Study 2 experimental conditions.

Condition	Length [min]	Total No. tokens	No. repetitions per word	Unigram entropy [bits]	Efficiency
Shorter-uniform	1:05	76	19	2	1
Uniform	1:50	128	32	2	1
Reduced efficiency	1:50	128	Frequent: 71 Infrequent: 19	1.7	0.85
Language-like efficiency	1:50	128	Frequent: 101 Infrequent: 9	1.1	0.54

Table 3. Study 3a experimental conditions.

Efficiency	Distribution type	Length [min]	Total No. tokens	No. repetitions per word	Unigram entropy [bits]	Efficiency
Reduced efficiency	Binary	1:50	128	Frequent: 71 Infrequent: 19	1.7	0.85
	Zipfian	1:50	128	Frequent: 70 Infrequent: 30, 18, 10	1.65	0.83
Language-like efficiency	Binary	1:50	128	Frequent: 92 Infrequent: 12	1.3	0.65
	Zipfian	1:50	128	Frequent: 93 Infrequent: 21, 9, 5	1.21	0.61

Table 4. Study 3b experimental conditions.

Condition	Length [min]	Total No. tokens	No. repetitions per word	Unigram entropy [bits]	Efficiency
Language-like efficiency*	1:50	128	Frequent: 92 Infrequent: 12	1.3	0.65
Low efficiency	1:50	128	Frequent: 110 Infrequent: 6	0.81	0.4

* This is the same sample as in Study 3a.