



# The learnability consequences of Zipfian distributions in language

Ori Lavi-Rotbain<sup>a,\*</sup>, Inbal Arnon<sup>b</sup>

<sup>a</sup> The Edmond and Lilly Safra Center for Brain Sciences, Hebrew University, Israel

<sup>b</sup> Department of Psychology, Hebrew University, Israel

## ARTICLE INFO

### Keywords:

Language acquisition  
Distributional learning  
Information theory  
Zipf's law  
Word segmentation

## ABSTRACT

While the languages of the world differ in many respects, they share certain commonalities, which can provide insight on our shared cognition. Here, we explore the learnability consequences of one of the striking commonalities between languages. Across languages, word frequencies follow a Zipfian distribution, showing a power law relation between a word's frequency and its rank. While their source in language has been studied extensively, less work has explored the learnability consequences of such distributions for language learners. We propose that the greater predictability of words in this distribution (relative to less skewed distributions) can facilitate word segmentation, a crucial aspect of early language acquisition. To explore this, we quantify word predictability using unigram entropy, assess it across languages using naturalistic corpora of child-directed speech and then ask whether similar unigram predictability facilitates word segmentation in the lab. We find similar unigram entropy in child-directed speech across 15 languages. We then use an auditory word segmentation task to show that the unigram predictability levels found in natural language are uniquely facilitative for word segmentation for both children and adults. These findings illustrate the facilitative impact of skewed input distributions on learning and raise questions about the possible role of cognitive pressures in the prevalence of Zipfian distributions in language.

## 1. Introduction

One of the striking commonalities between languages is 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 (Piantadosi, 2014; Zipf, 1949; see Eq. (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). First noted by Zipf in the 1930's (Zipf, 1949), Zipfian, or near-Zipfian (Piantadosi, 2014) distributions are repeatedly found across languages, and for different parts of speech (including nouns, verbs and adjectives). Eq. (1), which is an extension of Zipf's law formulated by Mandelbrot (Mandelbrot, 1953), shows the relation between a word's frequency -  $f(r)$  and its rank -  $r$ . Two constants determine the shape of the distribution:  $\alpha$  sets the steepness of the curve, and  $\beta$  introduces a skew which enables a better fit to natural language (Piantadosi, 2014; see also discussion in Dębowski, 2006). Frequency and rank show a power-law relation when looking at raw frequencies and a linear relation in log space.

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

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, Chater & Brown, 1999). However, their recurrence in language - a human creation - may nevertheless reflect foundational properties of human cognition and communication. While there is no agreed account of their source, their presence has been argued to be a form of optimal coding (Ferrer-i-Cancho, Bentz, & Seguin, 2020), to create an optimal trade-off between speaker and listener effort (Ferrer i Cancho & Sole, 2003), and to facilitate the hierarchical organization of word meanings (Manin, 2008). Zipfian distributions have also been proposed to reflect a trade-off between learnability pressures on the one hand and expressivity pressures on the other: Having a lot of words is needed for speakers to be able to communicate clearly and fully, yet acquiring many words is challenging from the learners' perspective. The particular shape of the Zipfian

\* Corresponding author.

E-mail address: [orilavirotbain@gmail.com](mailto:orilavirotbain@gmail.com) (O. Lavi-Rotbain).

distribution – with its' graded frequency - may offer a balance between those two pressures (Bentz, Alikaniotis, Cysouw, & Ferrer-i-Cancho, 2017; Coupé, Oh, Dediu, & Pellegrino, 2019; Lavi-Rotbain & Arnon, 2019a, 2019b).

Interestingly, less work has examined Zipfian distributions from a learnability perspective to ask whether they indeed impact language learning. Regardless of their source, their presence and propensity in language may have advantages for learning: In particular, words are more predictable in Zipfian distributions than in less skewed distributions. This increased word predictability could provide a facilitative environment for learning by making it easier to predict upcoming elements and easier to learn high frequency elements and use them as stepping stones for subsequent learning (as was suggested previously in Kurumada, Meylan, & Frank, 2013). Despite their more limited vocabulary, word frequencies in speech directed to infants and young children (child-directed speech) also follow a Zipfian distribution (Lavi-Rotbain & Arnon, under review; Hendrickson & Perfors, 2019), as do the objects (Clerkin, Hart, Reh, Yu, & Smith, 2017), and object combinations that infants see (Lavi-Rotbain & Arnon, 2021). 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 and challenging first step in breaking into language. Since words are not clearly separated in spoken language, infants need to discover their boundaries. Their ability to do so has been studied extensively, revealing infants' ability to use sophisticated distributional information to learn higher order structure (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005; Saffran, Aslin, & Newport, 1996). Skewed distributions could make this task easier by making word transitions easier to predict and allowing high frequency words to serve as anchors for segmenting less frequent ones, as seen in infants' use of their own name to segment adjacent words (Bortfeld et al., 2005).

Even though children's linguistic environment is skewed, this natural skew is rarely reflected in lab-based investigations (but see Kurumada et al., 2013; Meylan, Kurumada, Börschinger, Johnson, & Frank, 2012 for exceptions, also discussed below). Word segmentation in the lab is usually studied using artificial language learning paradigms (Saffran et al., 1996). Almost all such investigations expose learners to uniform distributions, where all novel words appear equally often. Such uniform distributions are useful for isolating the impact of specific factors on learning, but they are less predictable than children's real-world learning environment. Only a handful of studies have asked whether exposure to skewed distributions facilitates language learning (Hendrickson & Perfors, 2019; Kurumada et al., 2013; Meylan et al., 2012; Schuler, Reeder, Newport, & Aslin, 2017), with somewhat mixed results. In the only study to compare word segmentation in a Zipfian and uniform distribution, words were segmented more accurately in the Zipfian condition when they appeared more often next to the most frequent word (contextual facilitation), but accuracy was not better overall compared to a uniform distribution (Kurumada et al., 2013). An additional study (Meylan et al., 2012) showed that word segmentation was improved when the dependencies between words were asymmetrical (when some words appeared more often with others, resulting in a skewed distribution of dependencies). However, learning was better overall when there were no dependencies between words at all (as in the standard use of artificial word segmentation tasks, and unlike natural language). This study also did not find an overall advantage for skewed distributions. A similar pattern was found in a category formation task: participants assigned words to categories with similar success rates after being exposed to a Zipfian or a uniform distributions (Schuler et al., 2017). More facilitatory effects were found using a cross-situational word learning paradigm (Hendrickson & Perfors, 2019): Participants showed better learning of frequency-matched items in a Zipfian distribution compared to a uniform one, but only when there was ambiguity (when each object was presented simultaneously with two labels). A recent study showed better visual statistical learning - learning which triplet images appear together - when the triplets were presented in a

Zipfian distribution compared to a uniform, suggesting that learning relations between elements improves in a skewed distribution (Lavi-Rotbain & Arnon, 2021).

The few studies that have examined learning from Zipfian distributions suggest they are beneficial for learning, but the extent and generality of this effect is unclear. More importantly, we do not know *what* about Zipfian distributions impacts learning: which properties of the distribution facilitate learning? Here, we propose and test the hypothesis that Zipfian distributions are facilitative because of their greater word predictability. We predict that different languages will have similarly predictable word distributions and that these predictability values will enhance learners' word segmentation. We test these predictions using a combination of corpus-based and experimental studies where the corpus data serves both to assess the similarity between languages, and to generate predictions for the experimental investigation. Specifically, we first quantify word predictability in child-directed speech across 15 different languages from eight language families using unigram entropy (Study 1) and then test the impact of those values on learning using a classic artificial word segmentation paradigm (Studies 2–3). We start from the corpora investigation for two reasons: (1) to test whether languages have similar unigram predictability; and (2) to use the values we find to create experimental conditions for testing their impact on learning using artificial languages.

We find that different languages have similar unigram predictability, and that these predictability levels are uniquely facilitative for word segmentation. Comparing three levels of unigram predictability, we find that word segmentation accuracy is higher in languages that are as predictable as natural language, compared to uniform distributions, but also compared to skewed distributions less predictable than those found in natural language (Study 2). That is, increasing unigram predictability to a level lower than that of natural language did not facilitate learning, even though predictability was increased relative to a uniform distribution. We then show that the facilitation at language-like predictability, and the lack of facilitation for a less skewed distribution, is found for two additional skewed distributions that have similar unigram predictability but a different distribution shape (binary vs. Zipfian, Study 3). These findings highlight the importance of unigram predictability for learning and suggest that increasing unigram predictability does not impact learning in a linear fashion: It is not the case that any increase in predictability leads to an increase in accuracy. More generally, the findings deepen our understanding of the distributional factors impacting language acquisition, and offer a theoretical account for why (and when) Zipfian distributions facilitate learning. In the discussion, we relate these findings to broader question of how (and when) individual learning biases can impact language structure (e.g., Kirby, Cornish, & Smith, 2008).

## 2. Study 1: word distributions have similar unigram predictability across languages in child-directed speech

In this study, we examine the unigram predictability of word distributions in child-directed speech across different languages. Recent work suggests that words carry a similar amount of information across languages (Bentz et al., 2017; Takahira, Tanaka-Ishii, & Dębowski, 2016). A corpus study of 1000 languages using parallel corpora found that unigram entropy was similar across different languages (Bentz et al., 2017). Taken from information theory, Shannon's entropy (Shannon, 1948) quantifies the information content of a random variable (the amount of uncertainty) and has been shown to impact a range

of linguistic phenomena (Gibson et al., 2019). Unigram entropy refers to the average amount of information carried by all the words in the sample (where a word is defined by its orthographic form). Bentz et al. found that the average amount of information carried by individual words is similar across different languages: It is as easy to guess which word will appear next out of the entire lexicon.<sup>1</sup> This finding is consistent with our prediction that languages have similar unigram predictability, but needs to be expanded in several ways. First, since the study used parallel corpora (translations of the same text), the similarity between 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 only analyzed adult speech, which differs in many respects from child-directed speech, and may not give us an accurate picture of children's learning environment. To address these limitations, we conduct an investigation of word distributions in child-directed speech across 15 languages.

Rather than only computing unigram entropy, we operationalize the greater predictability of words in Zipfian distributions using the information-theoretic notion of efficiency (Eq. (2)), which is calculated using unigram entropy. Efficiency is the ratio between the observed entropy and the maximal entropy, which is entropy under a uniform distribution (see Eq. (2)). This measure has been used in the past to study human cognition (Pryluk, Kfir, Gelbard-Sagiv, Fried, & Paz, 2019) and is useful for us since it normalizes entropy by set size (see explanation below). In our use, efficiency is the ratio between the observed unigram entropy and unigram entropy under a uniform distribution with the same set size. Efficiency moves in the same direction as entropy: Given the same set size, lower entropy (a more predictable distribution) leads to lower efficiency. Efficiency has a range between zero (minimal efficiency, achieved when entropy is zero and the input is completely predictable), and one (maximal efficiency, achieved when the input has maximal entropy and is least predictable, as in the uniform distribution).<sup>2</sup>

$$\text{Efficiency} = \eta(X) = \frac{\text{observed entropy}}{\text{maximal entropy}} = - \frac{\sum_{i=1}^N p(x_i) \log_2(p(x_i))}{\log_2 N} \quad (2)$$

Efficiency is impacted by the parameters of the Zipfian distribution (as described in Eq. (1)): An increase in  $\alpha$  leads to a more skewed distribution, which will have lower unigram entropy and consequently lower efficiency. Zipfian distributions can differ from one another in their  $\alpha$ , and as a result, can also differ in their observed unigram entropy and efficiency (see Appendix 2 for a simulation illustrating the relation between changes in  $\alpha$  and efficiency, under a Zipfian distribution).

Using efficiency (instead of only unigram entropy) allows us to normalize entropy by set size and compare distribution predictability

<sup>1</sup> Unigram entropy does not take linguistic context into account: It doesn't quantify the predictability of a word given the preceding linguistic context, even though context is clearly predictive in natural language (e.g., Bell et al., 2003). This is a simplifying assumption, also made in prior corpus-based investigations of Zipfian distributions. We discuss its limitations in the discussion.

<sup>2</sup> Efficiency is an established mathematical term that is complementary to the mathematical term of redundancy (calculated as:  $1 - \text{efficiency}$ ). While efficiency measures the amount of "space" that is used in order to transmit a certain amount of information, redundancy measures the amount of "space" not used. We opted to use the term efficiency rather than redundancy for several reasons. First, it has the same direction as entropy – meaning that lower efficiency corresponds to greater predictability (the relation is the opposite with redundancy) – making it easier to understand the relation between the two. Second, the concept of redundancy is discussed within the study of language universals and language complexity, where it is used in many formulations and with no agreed upon mathematical or conceptual definition (see Tal, 2020 for a review and discussion).

across different experimental paradigms, and across languages whose lexicon size may differ and for which we have differently sized samples (as is the case for child-directed speech).

## 2.1. Methods

We used all the languages in the CHILDES database (MacWhinney, 2000) that had corpora for typically developing monolingual children with at least 150,000 tokens. We set this restriction following recent estimates of unigram entropy across languages and varying corpora sizes that showed that entropy calculations are reliable and stable when the corpus used includes at least 50,000 tokens (Bentz et al., 2017). This left us with the following languages: English (British and North-American), German, French, Japanese, Dutch, Polish, Spanish, Swedish, Portuguese, Hebrew, Mandarin, Estonian, Danish, Catalan and Norwegian, containing over 110 child-parent dyads. For each language, we collapsed over the different dyads to create one large corpus. We counted the number of appearances of each word (defined by orthographic form, as is done in prior studies of Zipfian distributions, e.g., (Piantadosi, 2014)) and calculated the unigram entropy for the observed frequency distribution. This is the observed unigram entropy. We then calculated the maximal unigram entropy, which is the unigram entropy under a uniform distribution for the same number of types (e.g., if the corpus had 1000 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 Eq. (1)).

To further explore the stability of the efficiency values, we divided the larger corpora (French, Japanese, German, North-American English and British English) into smaller samples of about 500,000 tokens each. The division procedure for each language was as follows: Each transcription file was read from beginning to end until a database of at least 500,000 tokens was created. This created samples that were conversationally continuous which is important for two reasons: (1) to better mimic the conversational continuity that is a property of actual linguistic input; (2) to avoid entropy inflation that could happen by mixing unrelated conversations: mixing words from different conversations could increase the number of distinct word types (e.g., by mixing words from a conversation around the dinner table with words from bath time), leading to an increased number of types which would lead to increased unigram entropy. For each sample, we counted the number of appearances of each word (defined by its orthographic form). We then calculated the efficiency of each sample in the same way described above.

## 2.2. Results

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 (see Table 1): all values were between 0.59 and 0.7 (average 0.64, SD = 0.03, Table 1). To ensure that the relatively stable deviation from the uniform we found is not dependent on the particular measure we used (efficiency), we also calculated the difference between the observed word distribution and a uniform one using Kullback–Leibler divergence (DKL) – a more commonly used measure for estimating the distance between two distributions. We found similar results: DKL values were similar across languages and spanned a relatively narrow range, meaning the distributions had a similar distance from the uniform distribution (mean  $D_{KL} = 5.05$ , SD = 0.70; for details see Appendix 1 and Supplementary Table 1). To further probe the stability of these efficiency 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 the Methods section). The efficiency values for the smaller bins ( $n = 36$ ) were still in the same range: the average efficiency value was 0.67 (SD = 0.01), with a range of 0.65–0.69 (Table 2). Interestingly, these values are lower than those of the Zipfian

**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	8611	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	8853	1–9217 (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	4924	1–10,741 (5.13–55,148)	7.71	12.27	0.63
Catalan	4	189,844	7970	1–12,318 (5.27–64,884)	8.71	12.96	0.67
Norwegian	2	184,676	8342	1–9196 (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**  
Corpora details and efficiency measures for samples ≈500,000 tokens.

Language	No. sample	No. tokens	No. types	Observed entropy [bits]	Maximal entropy [bits]	Efficiency
French	1	501,179	10,541	8.64	13.36	0.647
	2	502,135	10,915	8.77	13.41	0.65
	3	503,063	11,718	8.86	13.52	0.66
	Mean across samples	502,126	11,058	8.76	13.43	0.65
Japanese	1	502,147	19,085	9.27	14.22	0.65
	2	500,015	17,302	9.22	14.08	0.65
	3	500,260	16,543	9.22	14.01	0.66
	Mean across samples	500,807	17,643	9.24	14.10	0.65 (SD = 0.002)
German	1	501,663	14,709	8.83	13.37	0.66
	2	502,827	15,216	8.78	13.41	0.65
	3	502,260	15,322	8.83	13.42	0.66
	4	502,532	13,247	8.8	13.22	0.67
	Mean across samples	502,321	14,624	8.81	13.355	0.66 (SD = 0.005)
North American English	1	501,354	7082	8.67	12.79	0.68
	2	500,356	10,862	9.16	13.41	0.68
	3	500,910	9492	8.8	13.21	0.67
	4	500,338	7041	8.53	12.78	0.67
	5	500,258	8367	8.78	13.03	0.67
	6	500,398	9318	8.79	13.19	0.67
	7	511,638	9081	8.66	13.15	0.66
	8	500,653	10,230	8.94	13.32	0.67
	9	500,172	9315	8.81	13.19	0.67
	10	500,137	10,712	8.82	13.39	0.66
	11	505,194	8254	8.67	13.01	0.66
	12	501,693	9179	9.08	13.16	0.69
	Mean across samples	501,925	9078	8.81	13.14	0.67 (SD = 0.01)
British English	1	500,318	12,659	8.84	13.63	0.65
	2	502,646	7181	8.64	12.81	0.67
	3	501,751	6894	8.75	12.75	0.69
	4	501,688	7927	8.81	12.95	0.68
	5	503,121	8794	8.84	13.10	0.68
	6	500,141	9973	8.87	13.28	0.67
	7	502,413	7205	8.56	12.81	0.67
	8	500,536	6771	8.44	12.73	0.66
	9	501,870	8022	8.51	12.97	0.66
	10	503,255	7352	8.62	12.84	0.67
	11	501,043	6677	8.49	12.70	0.67
	12	501,263	6892	8.51	12.75	0.67
	13	500,976	6450	8.37	12.66	0.66
	14	503,049	6345	8.37	12.63	0.66
	Mean across samples	501,719	7796	8.61	12.90	0.67 (SD = 0.01)
Summary	Mean across all languages and samples	501,882	9219	8.84 (SD = 0.23)	13.3 (SD = 0.4)	0.67 (SD = 0.01)

distribution used in Kurumada et al., 2013 (mean = 0.84, SD = 0.03, range: 0.867–0.808, measures were calculated according to the word frequencies provided in Fig. 2 on page 443 of their article), raising the possibility that those languages were not learned better because they

were not as predictable as natural language.

However, how can we tell if the range of efficiency values we found is a narrow one? If efficiency values are limited to the range we found - if efficiency for Zipfian distributions cannot be higher than 0.7 or lower

than 0.6 - then it would not be surprising to repeatedly find those values in the corpora. To test this, and to see how efficiency changes as a function of alpha and lexicon size (assuming a Zipfian distribution), we ran a simulation calculating efficiency for artificial lexicons of different sizes and varying  $\alpha$  values, assuming a fixed corpus size of three million tokens (by artificial lexicon we mean that we generated a list of variables differing in frequency according to Eq. (1)). We looked at artificial lexicons with sizes of 30,000, 50,000 and 70,000 types (to reflect estimates of lexicon size in speakers, (Brybaert, Stevens, Mander, & Keuleers, 2016)), and at alpha values between zero and five (in increments of 0.1). For each lexicon size and each alpha value, we assigned frequency to the words according to the Zipfian equation (Eq. (1), under a corpus of three, five and seven million tokens respectively). That is, the lexicons were simply a list of frequencies, generated according to the Zipfian distribution with a limitation on lexicon size and total number of tokens. We then calculated the efficiency of all the resulting distributions. Efficiency values spanned the full possible range between zero to one for all lexicon sizes. More importantly, the efficiency values we found in natural language lie within the steepest part of the possible efficiency distribution, suggesting that their recurrence is not trivial: Randomly sampling from the distribution of efficiency values would not have given us a concentration of values between 0.6 and 0.7 (see Appendix 2 for full details).

The corpus investigation suggests that languages indeed have similar unigram predictability and provides us with a range of efficiency values whose impact on learning we can test experimentally: We can now ask whether exposure to distributions with the same efficiency values we found in natural language will facilitate word segmentation in the lab.

### 3. Study 2: language-like efficiency facilitates word segmentation in children and adults

Having found that languages have similar efficiency values, we can ask whether those values facilitate word segmentation. We ask three questions: (1) Is word segmentation improved in more predictable distributions? (2) Is the facilitation greater at the predictability values found in natural language? And (3) is the effect found in both children and adults? The child sample serves as a replication and extension: given that children and adults do not always show the same learning biases, we want to see whether they will be similarly affected by distribution predictability (see more detailed discussion below).

We examine the impact of efficiency on learners' ability to segment a four-word unsegmented artificial language using a classic artificial word segmentation paradigm (Saffran et al., 1996). In this paradigm, learners are exposed to recurrent tri-syllabic 'words' where the only cue to word boundary is that the transitional probabilities between syllables are higher within words than across word boundaries. While used extensively, almost all existing studies expose learners to uniform distributions, where each word appears equally often (but see Kurumada et al., 2013; Meylan et al., 2012 for exceptions). We manipulated efficiency by varying the frequency of the four words to create a skewed distribution where one word appears more often than the other three (we call this a binary distribution). We start with this distribution (which differs from the Zipfian one found in language, see Study 3), because it allows us to examine the impact of efficiency without having to control for frequency effects (since all low frequency words are equally frequent).

We compared performance at three efficiency levels (presented here in order from highest to lowest), based on what 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. The efficiency of this distribution (0.85) is lower than that of the uniform, but higher than that of natural language providing an intermediate value between the two. Importantly, the efficiency of this distribution is similar to that of the skewed

distribution used in Kurumada et al., (2013), allowing us to ask whether the lack of facilitation was driven by having a distribution not predictable enough; and (3) *Language-like efficiency*: a skewed distribution with efficiency values similar to those we found in natural language in Study 1, (see Table 3 for full details). The efficiency of this distribution (0.54) is lower than the other two, making it the most predictable distribution of the three, and the one we predict to facilitate learning the most. By comparing performance across the three conditions, we can ask whether any increase in unigram predictability leads to improved learning, or whether accuracy will be improved more (or only) in language-like efficiency. If any increase in predictability improves learning, then performance should be improved in the two skewed conditions compared to the uniform one. If, alternatively, learning is uniquely facilitated at certain efficiency 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 whether a similar effect is found in younger learners. The child sample included children aged 9–12 years and serves as a replication in another population of learners, and as an extension. We predict that children at this age will behave similarly to adults but this is not a trivial prediction since they still often perform worse than adults in artificial language learning studies (Arciuli & Simpson, 2012; Raviv & Arnon, 2018), and do not always show the same learning biases as adults (Jennifer Culbertson & Schuler, 2019; Lavi-Rotbain & Arnon, 2017). Given this literature, we wanted to ask whether the effect of efficiency will be similar, as we expect if it reflects a fundamental aspect of the linguistic environment. Moreover, the collection of child data provides an extension and comparison to recent work showing that children's visual SL improves in a skewed distribution (Lavi-Rotbain & Arnon, 2021). The visual study only compared a uniform to a skewed distribution and did not examine the impact of efficiency. By collecting a child sample of similar ages, we can compare the impact of skew in the visual and auditory domains.

#### 3.1. Method

##### 3.1.1. Participants

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

*Adult participants*: 142 participants took part in study 2a (mean age 24;0; 108 females, 34 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.

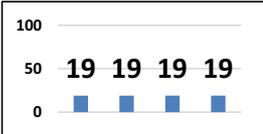
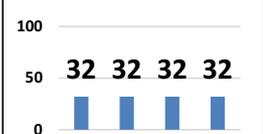
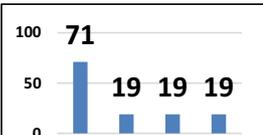
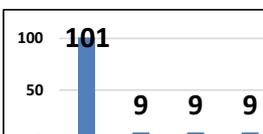
*Child participants*: 147 children took part in Study 2b (age range: from 9;0 to 12;0 years, mean age: 10;1 years; 68 girls, 79 boys). Children's ages did not differ across conditions ( $F(3) = 1.69, p > 0.1$ ). All children were visitors at the Bloomfield Science Museum in Jerusalem and were recruited 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.

#### 3.2. Materials

##### 3.2.1. 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 (Glicksohn & Cohen, 2013). They were created using the PRAAT synthesizer (Boersma & van Heuven, 2001) 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

**Table 3**  
Study 2 experimental conditions.

Condition	Length [min]	Total no. tokens	Frequency distribution of the four words	Unigram entropy [bits]	Efficiency	Average no. appearances after the frequent word (% of total appearances)
Shorter-uniform	1:05	76		2	1	–
Uniform	1:50	128		2	1	–
Reduced efficiency	1:50	128		1.7	0.85	13 (68.4%)
Language-like efficiency	1:50	128		1.1	0.54	8 (88.9%)

ensure that there were no co-articulation cues to word boundary. The words were matched for length (mean word length = 860 ms, range = 845–888 ms). The words were then concatenated together using MATLAB in a randomized order to create the auditory familiarization streams, with the only constraint being that words could not repeat themselves in the uniform condition (to keep the transitional probabilities between words constant). Importantly, there were no breaks between words or syllables and no prosodic or co-articulation cues in the stream to indicate word boundaries.

### 3.2.2. Experimental conditions

We used the same four words to create all our experimental conditions (see Table 3 for full details). The two skewed conditions and the uniform condition had the same exposure length. For the two skewed conditions, we created four different exposure streams where the frequent word was different (to ensure that the effect of efficiency is not driven by one particular word being easier to learn). This did not impact the results (see supplementary information of Study 2 and 3) so we collapsed over the four exposure streams in the analyses.

### 3.2.3. 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: 860 ms; range 854–868 ms). 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 discriminate words from part-words (a finding shown extensively), but to see how efficiency affects this ability.

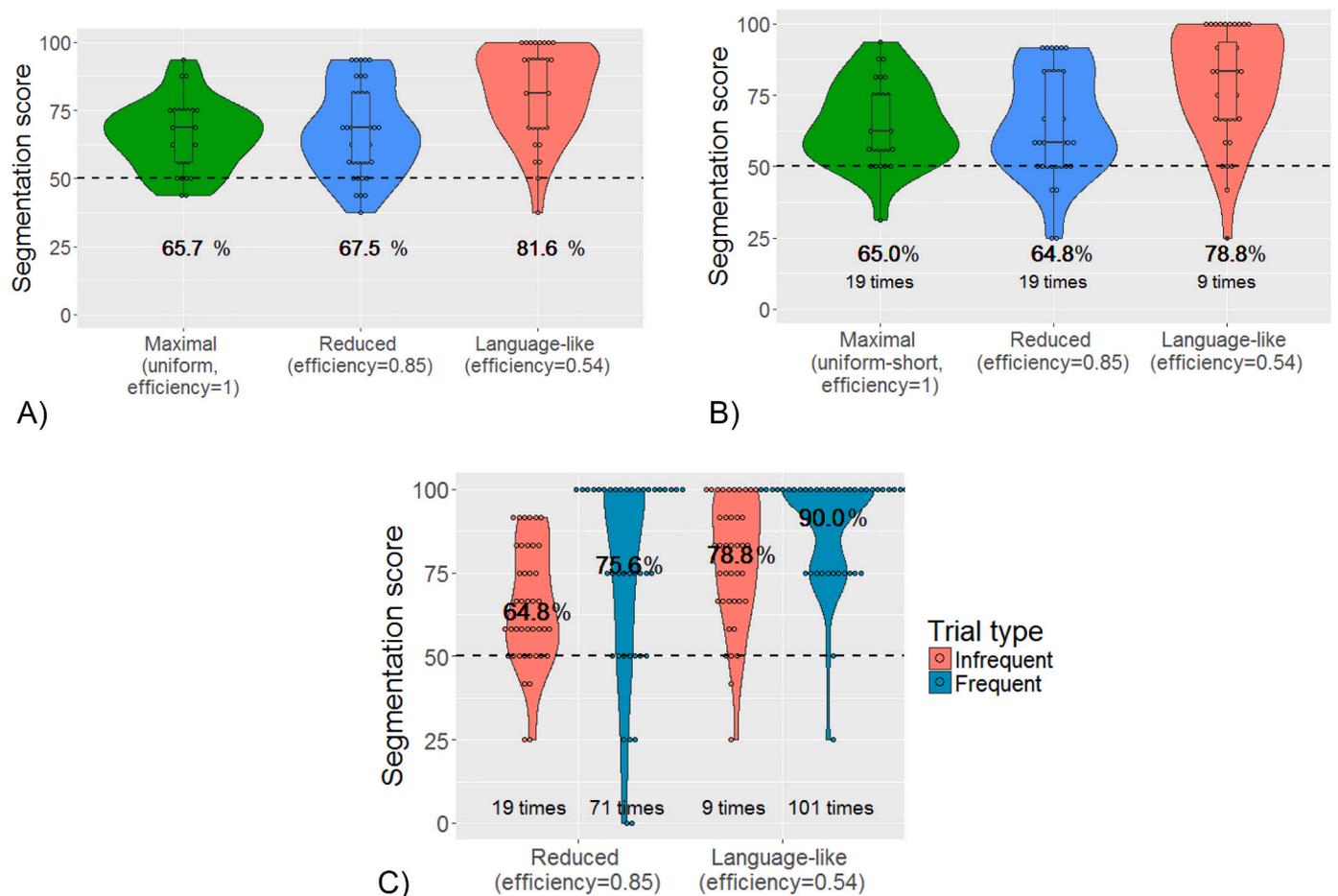
## 3.3. Results

### 3.3.1. Study 2a: adults

Participants were randomly allocated to one of the four conditions ( $N_{\text{uniform}} = 31$ ;  $N_{\text{reduced}} = 41$ ;  $N_{\text{language-like}} = 40$ ;  $N_{\text{shorter-uniform}} = 30$ ). Participants showed learning (were above chance, which was 50%) in all conditions (language-like efficiency:  $t(39) = 12.57$ ,  $p < 0.001$ ; reduced efficiency:  $t(40) = 7.0$ ,  $p < 0.001$ ; uniform:  $t(30) = 7.0$ ,  $p < 0.001$ ; shorter-uniform:  $t(29) = 5.8$ ,  $p < 0.001$ ). As predicted, word segmentation was better in language-like efficiency compared both to the uniform and the reduced-efficiency conditions ( $M_{\text{language-like}} = 81.66\%$ ,  $SD = 15.9\%$ ;  $M_{\text{reduced}} = 67.5\%$ ,  $SD = 15.9\%$ ;  $M_{\text{uniform}} = 65.7\%$ ,  $SD = 12.5\%$ , Fig. 1A).

*Language-like efficiency led to better segmentation compared to a uniform distribution, while reduced efficiency did not.* To test the effect of efficiency on segmentation accuracy, we compared the three conditions using a mixed-effect linear regression model (Model 1). Our dependent binomial variable was success on a single trial of the segmentation test. Our fixed effects were: experimental condition (dummy coded, meaning that reduced efficiency and language-like efficiency were compared to the uniform condition); log word frequency (centered); trial number (centered); and order of appearance in the test (word-first trials vs. foil-first trials). The model had random intercepts for participants and for items (see Supplementary Table 2). To examine the overall effect of experimental condition and word frequency, we used model comparisons.

Experimental condition had a significant effect on performance ( $\chi^2(2) = 36.09$ ,  $p < 0.001$  in model comparisons): Participants showed better learning in language-like efficiency compared to the uniform



**Fig. 1.** Adult segmentation scores from Study 2a. (A) Accuracy across conditions. (B) Accuracy only for the lower frequency words. (C) Comparing low frequency words and high frequency words in the two skewed conditions. Dashed lines represent chance level. Boxes indicate quartiles. Points represents individual. Numbers indicate frequency during exposure and percentages indicate mean segmentation score.

distribution ( $\beta = 1.27$ ,  $SE = 0.23$ ,  $p < 0.001$ ). Despite being more predictable, performance in the reduced efficiency condition did not differ from the uniform condition ( $\beta = 0.19$ ,  $SE = 0.20$ ,  $p > 0.1$ ). Frequency also had a significant effect on segmentation, with higher accuracy for more frequent words ( $\beta = 0.41$ ,  $SE = 0.1$ ,  $p < 0.001$ ,  $\chi(1) = 19.26$ ,  $p < 0.001$ ). The effect of trial number was significant, with worse accuracy over time (possibly reflecting the repetition of both foils and words,  $\beta = -0.04$ ,  $SE = 0.01$ ,  $p < 0.001$ ). Accuracy was higher when the real word was presented first ( $\beta = 0.61$ ,  $SE = 0.11$ ,  $p < 0.001$ ), as has been found in other studies using the same paradigm (Lavi-Rotbain & Arnon, 2017; Raviv & Arnon, 2018; Shufaniya & Arnon, 2018).

*Language-like efficiency leads to better segmentation compared to reduced efficiency.* To directly examine the difference between the two skewed conditions, we compared them using an additional model (Model 2). We used a mixed-effect linear regression model with success on a single trial of the segmentation test as our dependent binominal variable. Our fixed effects were: experimental condition (language-like efficiency vs. reduced efficiency); log word frequency (centered); the interaction between word frequency and condition; trial number (centered); and order of appearance in the test (word-first trials vs. foil-first trials). The model had random intercepts for participants and for items. The model shows that language-like efficiency resulted in better performance than reduced-efficiency ( $\beta = 1.11$ ,  $SE = 0.23$ ,  $p < 0.001$ , Fig. 1C, Supplementary Table 3).

To further support the difference between the language-like and the uniform condition, and the lack of difference between the uniform and the reduced efficiency, we conducted Bayes factor analyses. A Bayes

factor analysis on the mixed-effect model comparing the uniform and reduced conditions found strong support for the null hypothesis predicting no difference between them (BayesFactor = 18.28, details in Supplementary). In contrast, a similar analysis comparing the language-like and the uniform conditions found very strong support for the prediction that they differ in accuracy (BayesFactor = 9643). That is, accuracy was improved only in the language-like condition. This facilitation 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 (Kurumada et al., 2013), since the reduced efficiency condition also provided anchoring (the low frequency words appeared next to the frequent word often between 63 and 78% of the time, a proportion similar to that found in Kurumada et al. (2013), and much more than the uniform, see Table 3), but did not improve accuracy. 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.

*Language-like efficiency leads to better segmentation of low frequency words.* To test the prediction that the facilitation is not driven only by the frequent word and that low-frequency words are also learned better under language-like efficiency, we compared accuracy of low frequency words across three efficiency levels. We used a mixed-effect linear regression model (Model 3) with success on a single trial of the segmentation test as our dependent binominal variable. We included all trials from the shorter-uniform condition (16 trials) and only the low frequency trials from the reduced and language-like efficiency conditions (12 trials per condition). Our fixed effects were: experimental

condition (dummy coded, shorter-uniform as baseline); trial number (centered); and order of appearance in the test (word-first trials vs. foil-first trials). The model had random intercepts for participants and for items. Importantly, this model shows that the increased accuracy in the language-like condition was not driven only by performance on the higher frequency word: Accuracy on the low frequency words was better in language-like efficiency compared to the shorter-uniform condition ( $M_{\text{language-like-infrequent}} = 78.8\%$  vs.  $M_{\text{shorter-uniform}} = 65.0\%$ ), despite the words appearing half the number of times (9 vs. 19 times,  $\beta = 0.78$ ,  $SE = 0.22$ ,  $p < 0.001$ ). There was no such facilitation in the reduced efficiency condition ( $M_{\text{reduced-infrequent}} = 64.8\%$ ), even though each low frequency word appeared more often than in the language-like condition (19 times, as in the uniform-short condition,  $\beta = -0.001$ ,  $SE = 0.21$ ,  $p = 0.99$ , see Fig. 1B, Supplementary Table 4). 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).

To summarize, adults showed better segmentation in the language-like efficiency condition compared to the uniform and to the reduced efficiency conditions, even for the low frequency words.

### 3.3.2. Study 2b: children

Children completed the same four conditions ( $N_{\text{uniform}} = 30$ ;  $N_{\text{reduced}} = 47$ ;  $N_{\text{language-like}} = 40$ ;  $N_{\text{shorter-uniform}} = 30$ ). Children showed learning (were above chance) in language-like efficiency:  $t(39) = 6.12$ ,  $p < 0.001$ ; reduced efficiency:  $t(46) = 3.97$ ,  $p < 0.001$ ; and the uniform condition:  $t(29) = 4.84$ ,  $p < 0.001$ . Performance in the shorter-uniform condition was not significantly higher than chance ( $t(29) = 1.82$ ,  $p = 0.08$ ). Like adults, children showed better learning in language-like efficiency: They

were more accurate in this condition compared to the reduced efficiency and the uniform conditions ( $M_{\text{language-like}} = 66.1\%$ ,  $SD = 16.6\%$ ;  $M_{\text{uniform}} = 60.2\%$ ,  $SD = 11.5\%$ ;  $M_{\text{reduced}} = 58.0\%$ ,  $SD = 13.8\%$ ; Fig. 2A).

Language-like efficiency leads to better segmentation compared to a uniform distribution, while reduced efficiency does not. To test the effect of efficiency on segmentation accuracy, we compared the three conditions using a mixed-effect linear regression model (Model 4). Our dependent binomial variable was success on a single trial of the segmentation test. Our fixed effects were: experimental condition (dummy coded, meaning that reduced efficiency and language-like efficiency were compared to the uniform condition); age (centered); log frequency of the word (centered); trial number (centered); and order of appearance in the test (word-first trials vs. foil-first trials). The model had random intercepts for participants and for items (see Supplementary Table 6 for full model). To examine the overall effect of experimental condition, word frequency and age, we used model comparisons. Experimental condition had a significant effect on performance ( $\chi(2) = 16.02$ ,  $p < 0.001$ ): Children showed better learning in language-like efficiency compared to the uniform condition ( $\beta = 0.53$ ,  $SE = 0.17$ ,  $p < 0.01$ ). Performance in the reduced efficiency condition did not differ from the uniform condition ( $\beta = -0.03$ ,  $SE = 0.15$ ,  $p > 0.8$ ). 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). Frequency impacted segmentation ( $\chi(1) = 35.76$ ,  $p < 0.001$ ): Children showed higher accuracy for more frequent words ( $\beta = 0.45$ ,  $SE = 0.08$ ,  $p < 0.001$ ). In addition, age had a significant effect on performance ( $\chi(1) = 6.89$ ,  $p < 0.001$ ): older children showed higher accuracy ( $\beta = 0.16$ ,  $SE = 0.06$ ,  $p < 0.05$ ), as has been found in previous studies (Raviv & Arnon, 2018).

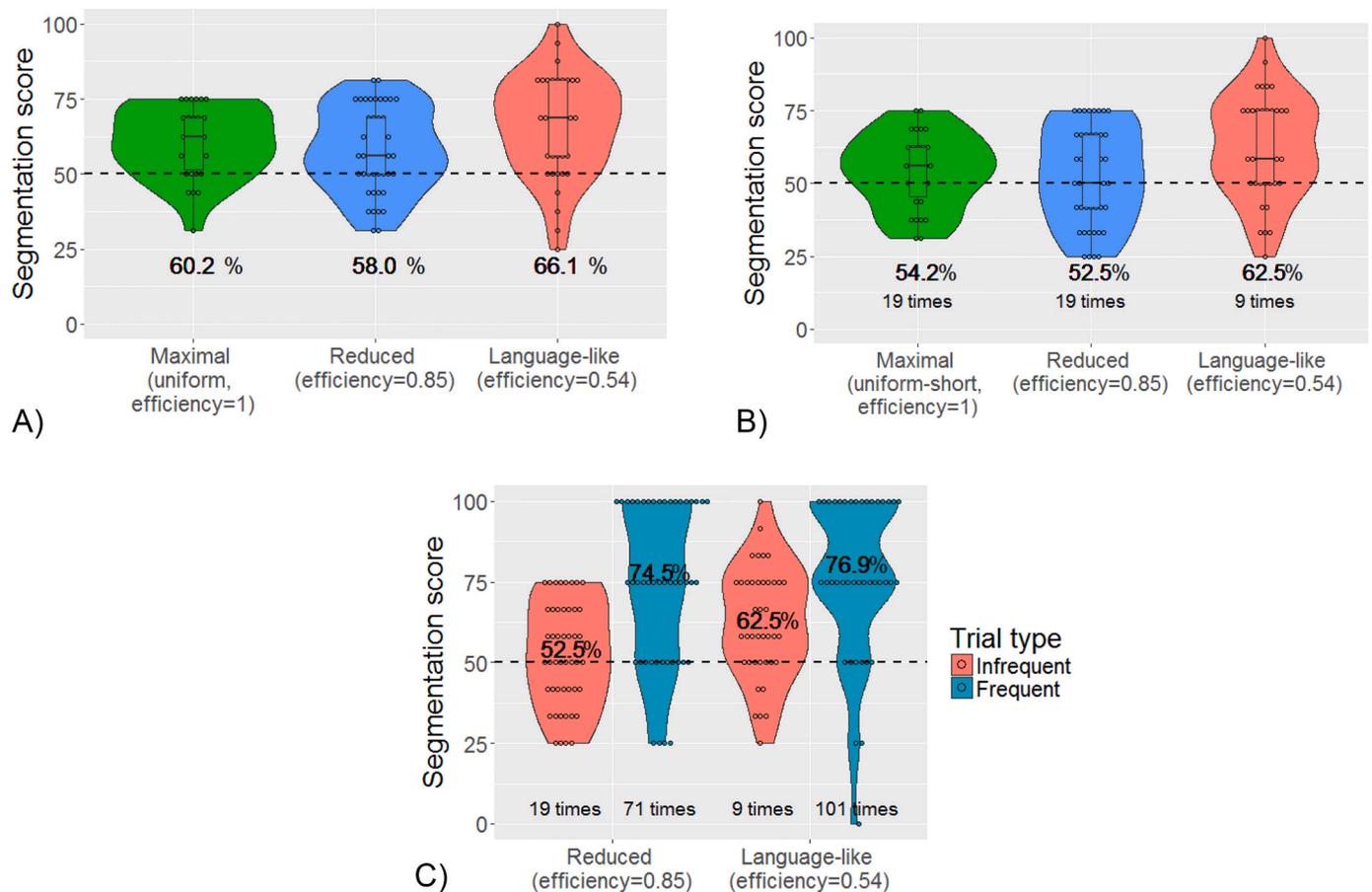


Fig. 2. Children's segmentation scores from Study 2b. (A) Accuracy across conditions. (B) Accuracy on the lower frequency words. (C) Comparing low frequency words and high frequency words in the two skewed conditions. Dashed lines represent chance level. Boxes indicate quartiles. Points represents individual. Numbers indicate frequency during exposure and percentages indicate mean segmentation score.

*Language-like efficiency leads to better segmentation compared to reduced efficiency.* To see if there is a difference between the two skewed conditions, as we found for adults, we compared them using an additional model (Model 5). We used a mixed-effect linear regression model with success on a single trial of the segmentation test as our dependent binomial variable. Our fixed effects were: experimental condition (language-like efficiency compared to reduced efficiency); log word frequency (centered); the interaction between them; age (centered); trial number (centered); and order of appearance in the test (word-first trials vs. foil-first trials). The model had random intercepts for participants. This model showed that language-like efficiency led to significantly better performance than reduced efficiency ( $\beta = 0.44$ ,  $SE = 0.16$ ,  $p < 0.01$ , Fig. 2C, Supplementary Table 7). Together, these results suggest that the effect of unigram predictability on learning is not linear: it is not the case that any increase in predictability leads to an increase in accuracy. Instead, it seems that a large enough reduction in efficiency is needed to facilitate learning.

*Language-like efficiency leads to better segmentation of low frequency words.* Finally, we wanted to test if 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 ( $M_{\text{reduced-infrequent}} = 52.5\%$ ,  $SD = 16.0\%$ ; not higher than chance,  $t(46) = 1.06$ ,  $p > 0.1$ ), and learned them relatively poorly in the uniform-short condition ( $M_{\text{uniform-short}} = 54.17\%$ ,  $SD = 12.5\%$ ;  $t(29) = 1.8$ ,  $p = 0.07$ ). However, they did manage to learn words appearing half the number of times (only nine times) when presented in language-like efficiency ( $M_{\text{language-like-infrequent}} = 62.5\%$ ,  $SD = 17.3\%$ ;  $t(39) = 4.57$ ,  $p < 0.001$  compared to chance). To test whether low frequency words were learned better in the language-like condition compared to the uniform-short, we used a mixed-effect linear regression model (Model 6). Our dependent binomial variable was success on a single trial of the segmentation test. We included all trials from the shorter-uniform condition (16 trials) and only low frequency trials from the language-like efficiency condition (12 trials). Our fixed effects were: experimental condition (dummy coded, shorter-uniform as baseline); trial number (centered); and order of appearance in the test (word-first trials vs. foil-first trials). The model had random intercepts for participants and for items. The model confirmed that segmentation was better in the language-like condition, despite the lower frequency (9 vs. 19 times,  $\beta = 0.34$ ,  $SE = 0.15$ ,  $p < 0.05$ ; Fig. 2B, Supplementary Table 8). That is, exposure to a skewed distribution with language-like efficiency facilitated children's segmentation and enabled them to learn low frequency words that were not learned in a uniform distribution.

To summarize, children showed a similar pattern as adults: segmentation in the language-like condition was better than in the uniform condition or in the reduced condition, including for low-frequency words.

### 3.3.3. Study 3: the relative impact of distribution shape and unigram predictability

Study 2 showed that word segmentation is facilitated in children and adults at language-like efficiency. However, efficiency in Study 2 was reduced by making one word more frequent than the other three (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), but was limited in using a distribution that does not resemble the Zipfian one found in natural language. In Study 3, we want to ask two additional questions: (1) Does the impact of efficiency on learning extend to a more language-like distribution? And (2) is the facilitation impacted more by distribution shape or unigram predictability? Since different distributions can have the same efficiency level the way to examine this is to compare distributions that have the same shape (e.g., binary) but differ in efficiency. 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 with similar efficiency levels: We expect better performance in the two language-like conditions, regardless of distribution shape. This study also serves to replicate the facilitative effect of language-like efficiency compared to reduced efficiency: We will sample three additional efficiency values to see if we again observe better performance in language-like efficiency compared to reduced efficiency.

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 Eq. (1), lower efficiency values were obtained by higher a higher exponent). 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 4 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) to ensure a similar difference in efficiency for the two skewed distributions (0.63 vs. 0.85 for the binary distribution and 0.61 vs. 0.83 for the Zipfian one), and to show that the facilitative effect of language-like efficiency is replicated in an additional sample.

## 3.4. Method

### 3.4.1. Participants

Adult participants: 120 additional adult participants completed the three new conditions (for all four conditions:  $N = 161$ , mean age = 23;7; 115 females). 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.

### 3.4.2. Materials

Auditory stimuli were identical to the ones used in Study 2. The experimental conditions were created in the same manner as the ones in Study 2. Full details on each condition are listed in Table 4. The procedure was identical to that of Study 2.

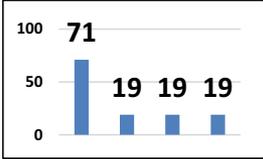
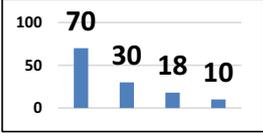
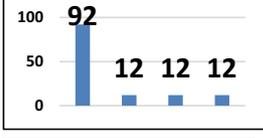
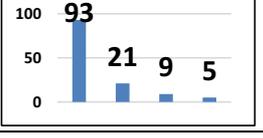
## 3.5. Results

Participants were randomly allocated to one of the three new conditions ( $N_{\text{Zipfian-reduced}} = 40$ ;  $N_{\text{binary-language-like}} = 40$ ;  $N_{\text{language-like}} = 40$ ;  $N_{\text{Zipfian-language-like}} = 40$ ). Participants in all conditions showed learning (were above chance level of 50%): Zipfian-reduced:  $t(39) = 12.19$ ,  $p < 0.001$ ; binary-language-like:  $t(39) = 10.9$ ,  $p < 0.001$ ; Zipfian-language-like:  $t(39) = 11.15$ ,  $p < 0.001$ . As predicted, accuracy was higher in language-like efficiency compared to the reduced efficiency conditions ( $M_{\text{language-like-binary}} = 78.4\%$ ,  $SD = 16.5\%$ ;  $M_{\text{language-like-Zipfian}} = 78.4\%$ ,  $SD = 16.1\%$ ;  $M_{\text{intermediate-Zipfian}} = 74.7\%$ ,  $SD = 12.8\%$ ;  $M_{\text{reduced-binary}} = 67.5\%$ ,  $SD = 15.9\%$ ).

Since the most frequent word is expected to be learned well in all conditions (and may disproportionately impact overall accuracy), we focus on performance on the low frequency words. Fig. 3 shows segmentation scores in all conditions for the most frequent word and the lower frequency words separately. As predicted, participants showed better learning of the low frequency words in the language-like conditions compared to the reduced efficiency ones ( $M_{\text{language-like-binary-infrequent}} = 77.1\%$ ,  $SD = 19.1\%$ ;  $M_{\text{language-like-Zipfian-infrequent}} = 76.4\%$ ,  $SD = 18.9\%$ ;  $M_{\text{reduced-Zipfian-infrequent}} = 69.8\%$ ,  $SD = 16.5\%$ ;  $M_{\text{reduced-binary-infrequent}} = 64.8$ ,  $SD = 17.9\%$ ). To test the significance of these effects, we used mixed-effects models.

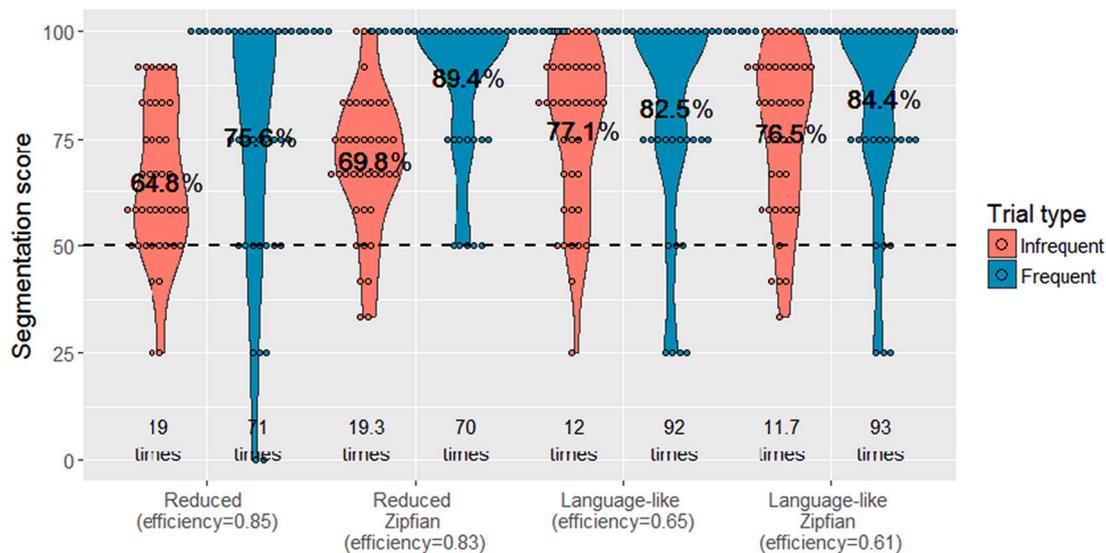
*Segmentation is facilitated at language-like efficiency, but is not affected by distribution shape.* In order to test the effect of efficiency versus

**Table 4**  
Study 3 experimental conditions.

	Distribution type	Frequency distribution of the four words	Unigram entropy [bits]	Efficiency	Average no. appearances after the frequent word (% of total appearances)
Reduced efficiency	Binary*		1.7	0.85	13 (68.4%)
	Zipfian		1.65	0.83	10.66 (55.1%)
Language-like efficiency	Binary		1.3	0.65	9.66 (80.1%)
	Zipfian		1.21	0.61	10.33 (88.5%)

All conditions in Study 3 had length of 1:50 min and 128 tokens.

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



**Fig. 3.** Comparing adult segmentation scores on low frequency words and high frequency words across conditions (note that the sample for the binary-reduced efficiency is the one from Study 2a). Dashed lines represent chance level. Boxes indicate quartiles. Points represents individual. Numbers indicate frequency during exposure (in the Zipfian conditions this is the average of the low frequency words) and percentages indicate mean segmentation score.

distribution shape, we compared all four conditions using the same model (Model 7). We used a mixed-effect linear regression model with success on a single trial of the segmentation test as our dependent binominal variable. Our fixed effects were: efficiency level (reduced versus language-like, reduced as baseline), distribution shape (binary versus Zipfian, binary as baseline), word frequency (binary coded: frequent trials versus infrequent ones, infrequent as baseline) as well as all the interaction between distribution shape and frequency; trial number (centered); and order of appearance in the test (word-first trials vs. foil-first trials). The model had random intercepts for item (See Supplementary Table 10 for full model). Efficiency had a significant

effect on performance: Participants showed better learning in language-like efficiency compared to reduced efficiency ( $\beta = 0.62$ ,  $SE = 0.15$ ,  $p < 0.001$ ). Distribution shape did not affect performance ( $\beta = 0.23$ ,  $SE = 0.14$ ,  $p = 0.1$ ), suggesting that in this task, efficiency impacted learning more than shape. The interaction between frequency and distribution shape was significant ( $\beta = 0.79$ ,  $SE = 0.35$ ,  $p < 0.05$ ), suggesting better learning of the frequent word in the Zipfian distribution. However, this effect was only found in the reduced efficiency conditions: In the reduced efficiency conditions, the frequent word was learned better in the Zipfian distribution than in the binary one ( $t(65.56) = 2.53$ ,  $p < 0.05$ ). No such difference was found when comparing the two language-

like conditions ( $t(77.6) = -0.35, p > 0.7$ , as can be seen in Fig. 3), suggesting it is not a robust one.

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) will be learned equally well since the most frequent word was learned very well in both conditions 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 4). Despite this, the low frequency words were learned significantly worse in the reduced condition compared to the language-like one (69.8% versus 76.5%). To test the significance of this difference, we used a mixed-effect model comparing learning of the low frequency words between the two Zipfian conditions (Model 8). Our fixed effects were: efficiency level (reduced versus language-like, reduced as baseline); trial number (centered); and order of appearance in the test (word-first trials vs. foil-first trials). The model had random intercepts for item (Supplementary Table 11). The low frequency words were learned significantly worse in the reduced condition compared to the language-like one ( $\beta = 0.35, SE = 0.149, p < 0.05$ ). This pattern is predicted by efficiency, but not anchoring.

These results support the facilitative effect of language-like efficiency on learning (for two additional efficiency values), and indicate that in this experimental paradigm, unigram predictability impacts word segmentation more than distribution shape.

#### 4. 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 (Chater & Brown, 1999; Ferrer i Cancho & Sole, 2003), less research has examined their impact on learning. Given their prevalence 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 previous studies showing that Zipfian distributions provide more contextual facilitation for word segmentation compared to uniform ones (Kurumada et al., 2013) and that they can improve learning in other linguistic and non-linguistic domains (Hendrickson & Perfors, 2019; Meylan et al., 2012; Lavi-Rotbain & Arnon, 2020). However, previous work did not identify what about the distribution facilitates learning and under what conditions.

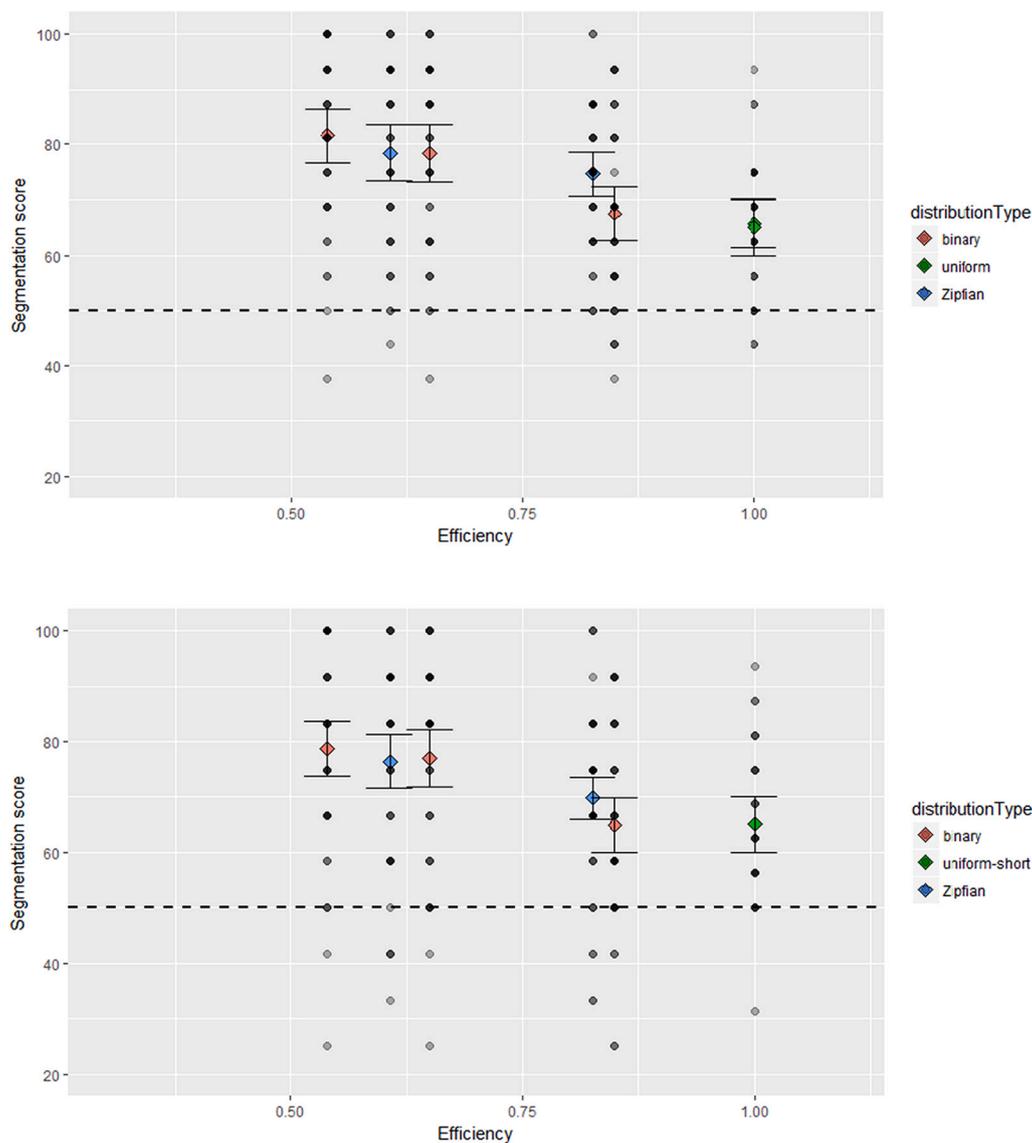
Here, we go beyond existing findings to provide a comprehensive theoretical account about when (and why) Zipfian distributions facilitate word segmentation. Specifically, we propose that segmentation is aided by the greater predictability of words in such distributions. We quantify distribution predictability 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 unigram predictability has very similar values in child-directed speech across fifteen different languages. We then use the predictability values we found to create experimental conditions asking whether word segmentation in the lab is facilitated in distributions with similar efficiency. Studies 2 and 3 use a classic artificial word segmentation paradigm to test the impact of unigram predictability and shape on learning by shifting the distribution away from the uniform one used in most SL studies. In line with our predictions, we find that word segmentation is facilitated in language-like efficiency relative to a uniform distribution, and also relative to a skewed distribution less predictable than language (with higher efficiency). This finding holds across two skewed distributions: a binary one where one word is more frequent than the other three and a Zipfian one where word frequency follows a power law distribution. Importantly, these findings cannot be explained by anchoring alone: while anchoring – where the frequent word is learned earlier and used to segment lower

frequency ones – undoubtedly plays a role in word segmentation, and occurs more in more skewed distributions, it cannot explain the full range of results since the reduced efficiency condition also had high anchoring but did not improve segmentation.

Our findings show that learners are sensitive to the overall predictability of the linguistic environment, and that the effect on learning is not linear: Accuracy does not increase with any increase in unigram predictability. Fig. 4 shows accuracy from all experimental conditions ordered by efficiency. We can see that (a) performance improves (compared to the uniform) only in language-like efficiency, and (b) that this happens regardless of distribution shape: Accuracy was higher at language-like efficiency for a binary and Zipfian distribution, and was not improved at reduced-efficiency for both distribution shapes. That is, 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 unigram predictability that has to happen before learning is facilitated: the distribution needs to be skewed enough. This effect has parallels in the animal learning literature where there are stronger neural responses to highly deviant (infrequent) stimuli (stimulus-specific adaptation, Taaseh, Yaron, & Nelken, 2011). While not identical, this literature highlights the benefit that low frequency items can receive in highly skewed distributional environments. The lack of improvement in the reduced efficiency conditions 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 (Hendrickson & Perfors, 2019): Learning was not improved 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 when skewed distribution are facilitative, it cannot explain the range of data explained by efficiency.

The non-linear effect of increased unigram predictability on learning can also explain the lack of overall advantage when word segmentation was previously assessed in a Zipfian distribution that did not have language-like efficiency (Kurumada et al., 2013): in this previous study, the Zipfian distribution was not predictable enough, and consequently did not enhance learning. While there are several differences in the design of the two studies, they cannot explain away the effect of efficiency we found. Kurumada et al. had larger vocabularies than in the current study (between 6 and 36, compared to our four) and used partwords as foils whereas we used nonwords (partwords include a two-syllable sequence from a real word while nonwords are made up of syllables from real words that didn't appear together as a sequence). While having a smaller vocabulary and using nonwords could have made our task overall easier compared to that of Kurumada et al., this cannot explain why accuracy was higher in the language-like condition in our study compared to other two (given that nonwords were used as foils in all conditions). Moreover, our results replicate aspects of Kurumada's: We also found no overall facilitation in the reduced condition – which is parallel in terms of efficiency to their Zipfian condition. That is, differences in efficiency can explain both the lack of overall improvement in Kurumada et al., 2013, and the improvement in the language-like condition in our study.

The current study shows that certain efficiency values facilitate word segmentation more than others, but why then, do languages display the certain range of efficiency values we observed? 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. In



**Fig. 4.** Adult segmentation accuracy by efficiency (Study 2a and Study 3). (A) Accuracy across conditions. (B) Accuracy only for the lower frequency words. Dashed lines represent chance level. Error bars represents confidence intervals of 95%. Points represents individual scores with greater darkness for more individual participants.

such a scenario, lowering efficiency levels more will not further enhance word segmentation in the lab, which may be unmotivated given the generally positive effect of increased predictability on learning (Christiansen & Chater, 2008). Alternatively, and more likely, the observed unigram predictability values may reflect the impact of competing pressures on language structure (Christiansen & Chater, 2008; Fedzechkina, Jaeger, & Newport, 2012; Kirby et al., 2008; MacWhinney, 1987). 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 (Bentz et al., 2017; MacWhinney, 1987; Smith & Kirby, 2008). 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 accidentally 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 are predictable enough for learning while still being adequate for communication.

Under this explanation, word segmentation will not be further improved in distributions with efficiency values that are lower than what we found in natural languages (under 0.6). To test this, we had run an additional study using the same word segmentation paradigm with a distribution with a lower efficiency value than natural language (efficiency = 0.4). We used a binary distribution where the frequent word appeared 110 times and each of the lower frequency words appeared six times. We found that accuracy in this condition was similar to accuracy in the language-like condition ( $N = 40$ ;  $M = 83.9$ ;  $SD = 0.16$ ): Performance didn't improve more when efficiency was lower. This could suggest that the lower bound found in natural language is not driven by learnability pressures: Learning did not improve when efficiency was decreased further. However, while suggestive, these findings should be interpreted with caution (which is why we did not include them in the result section). The low frequency words in the lower efficiency condition were less frequent than those in the language-like efficiency (appearing six times vs. nine times, see Appendix 4 for details). This was

done to maintain the same exposure length between conditions, but may have masked the effect of lowering efficiency: additional work is needed to compare conditions that differ in efficiency but where the low frequency words are matched for frequency.

Importantly, the current study is only a first step in investigating the efficiency values that languages display and their impact on learning. It raises many additional questions that need to be addressed. The first is to evaluate how gradual changes in efficiency impact learning. We examined three levels of efficiency: maximal, reduced and language-like, and found a non-linear effect of efficiency reduction on accuracy. However, a more systematic investigation is needed to see what happens at additional values, within and outside the range we found in our corpus investigation, and whether learning improves at lower efficiency levels (after controlling for token frequency). Such an investigation is needed to better understand what underlies the facilitation and what drives the range we observed in natural languages. Related to this, it is important to assess the range of efficiency values in other linguistic and non-linguistic corpora to see how similar they are to the ones we found in child-directed speech. A comparison across domains may also illuminate the different pressures that impact efficiency values: if learnability plays a role in setting the upper bound, then we should find a similar upper bound in other domains where a set of elements and the distinction between them needs to be learned (e.g., in learning to recognize visual objects). A preliminary investigation indicates that efficiency values have the same range in adult-to-adult speech: Using the EuroParl corpus (a corpus of spoken language created from the European Parliament Proceeding in 21 European languages, (Koehn, 2008)), we found that efficiency spans a very similar range in adult speech (mean efficiency = 0.62, SD = 0.002, range: 0.58–0.68, (Shor, Reichart, & Arnon, 2022)). In future work we plan to examine efficiency values in sign languages: If learning is optimal within a certain range, we would expect to find similar values in signed language corpora (unfortunately, there are few such corpora that are of sufficient size to reliably estimate sign entropy). A second question is whether younger learners, including infants, show improved word segmentation in language-like efficiency. If the skewed nature of the environment plays a facilitative role in language learning, as we predict, then infants and younger children should also show such effects. There are some indications in the literature that this will be the case. Infants are sensitive to frequency distributions in the lab, and can utilize differences in sound distributions to determine whether to create one or two phonemic categories (Maye, Werker, & Gerken, 2002). Two- and three-year-olds' ability to produce unfamiliar four-word sequences is affected by the frequency distribution of the fourth word (referred to as "slot entropy" (Matthews & Bannard, 2010)). These examples indicate that younger learners are sensitive to the frequency distribution of their linguistic input in ways that affect their learning outcomes. We are currently in the process of testing 8-month-old infants using a similar (infant-adapted) design to see if they also show improved word segmentation in language-like efficiency (compared to a uniform and less skewed distribution).

Our findings give rise to another important question: If distribution predictability impacts word segmentation more than distribution shape, as Study 3 implies, then why do languages consistently have Zipfian distributions? Many other distributions could provide the same efficiency values. We think several communicative and cognitive pressures 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, an utterance will include words from different regions of the frequency distribution. This means that the contrast between higher and lower frequency words will hold even within a single 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 (Ferrer-i-Cancho et al., 2020) and to enable a better semantic space, by allowing different levels of specificity to be represented by different levels of frequency (Lestrade, 2017; Manin, 2008). That is, a mixture of pressures (that may be weighted differently during learning and processing) could lead recurrence of the Zipfian distribution (with its particular shape) in language. 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 broader implications for the study of how humans use distributional information to learn higher-order structure (this includes the literatures on statistical learning, sequence learning, and distributional learning). 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 predictable than natural language may limit our understanding of learning in the wild and lead us to underestimate learners' abilities (Erickson & Thiessen, 2015; Frost, Armstrong, & Christiansen, 2019). 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 transitional probabilities 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 facilitate learning novel word-object associations (Lavi-Rotbain & Arnon, 2019a), but its effect on learning grammatical relations has not yet been examined.

The current study documents a facilitative effect of skewed distributions on an individual level. Extrapolating from this, we can ask whether such individual learning biases could explain the prevalence 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 (Culbertson & Kirby, 2015; Kirby et al., 2008; Smith & Kirby, 2008). This has been demonstrated for various linguistic properties, among them compositionality (Kirby et al., 2008), regularization (Ferdinand, Kirby, & Smith, 2019), harmonic alignment (Jennifer Culbertson & Newport, 2015) and more. If Zipfian distributions help learners discover word boundaries, as our experimental findings indicate, this could create a cognitive pressure to maintain similarly skewed distributions across languages and time. 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 introduced. Such stability has been reported for the transfer of information across languages (Coupé et al., 2019), as

well as for the ratio between word and sequence entropy (Cohen Priva & Gleason, 2016). The second prediction is that learners will show a cognitive preference for skewed distributions, leading them to change their input to make it more skewed. We are exploring this prediction (Shufaniya & Arnon, under review) using iterated learning paradigms, which can be used to reveal weak individual biases that are amplified over time (e.g., Kirby et al., 2008). In these studies, we ask whether speakers are biased to produce skewed word distributions in telling a novel story and whether this bias leads learners to shift uniform distributions towards more skewed ones in re-telling a story containing six nonce words that appear equally often. Our results suggest that word distributions became more skewed (as measured by lower levels of entropy), suggesting a cognitive bias for a shift from the uniform. The third, 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 all three 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.

## 5. 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 use corpus analyses to show that child-directed speech is similarly skewed across languages, and that the unigram predictability values found in natural language are uniquely facilitative for word segmentation. These findings show that learners are sensitive to the structure of the environment as a whole; and point to unigram predictability as an important factor in learning. More broadly, the findings suggest that Zipfian distributions confer a learnability advantage and open up new directions in explaining the impact of learning biases on their recurrence in language.

## Data availability

All raw data for Study 2 and 3 is available at <https://osf.io/z58hr/>.

## 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.

## Declaration of Competing Interest

We do not have any interests that might be interpreted as influencing the research, and APA ethical standards were followed in the conduct of the study.

## 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.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.cognition.2022.105038>.

[org/10.1016/j.cognition.2022.105038](https://doi.org/10.1016/j.cognition.2022.105038).

## References

- Arciuli, J., & Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science*, 36(2), 286–304. <https://doi.org/10.1111/j.1551-6709.2011.01200.x>
- Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America*, 113(2), 1001–1024.
- Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i-Cancho, R. (2017). The entropy of words-learnability and expressivity across more than 1000 languages. *Entropy*, 19(6), 1–32. <https://doi.org/10.3390/e19060275>
- Boersma, P., & van Heuven, V. (2001). Speak and unSpeak with Praat. *Glott International*, 5(9–10), 341–347.
- Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me. *Psychological Science*, 16(4), 298–304. <https://doi.org/10.1111/j.0956-7976.2005.01531.x>
- Brysaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, 7(JUL), 1–11. <https://doi.org/10.3389/fpsyg.2016.01116>
- Chater, N., & Brown, G. D. A. (1999). Scale-invariance as a unifying psychological principle. *Cognition*, 69(3), 17–24. [https://doi.org/10.1016/S0010-0277\(98\)00066-3](https://doi.org/10.1016/S0010-0277(98)00066-3)
- Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *The Behavioral and Brain Sciences*, 31(5), 489–508. discussion 509–558 <https://doi.org/10.1017/S0140525X08004998>.
- Clerkin, E. M., Hart, E., Rehg, J. M., Yu, C., & Smith, L. B. (2017). Real-world visual statistics and infants' first-learned object names. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, 372(1711). <https://doi.org/10.1098/rstb.2016.0055>
- Cohen Priva, U., & Gleason, E. (2016). Simpler structure for more informative words: A longitudinal study. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society, 2012* (pp. 1895–1900).
- Coupé, C., Oh, Y., Dediu, D., & Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9). <https://doi.org/10.1126/sciadv.aaw2594>. eaaw2594.
- Culbertson, J., & Kirby, S. (2015). Simplicity and specificity in language: Domain general biases have domain specific effects. *Frontiers in Psychology*, 6(January), 1–11. <https://doi.org/10.3389/fpsyg.2015.01964>
- Culbertson, J., & Newport, E. L. (2015). Harmonic biases in child learners: In support of language universals. *Cognition*, 139, 71–82.
- Culbertson, J., & Schuler, K. (2019). Artificial language learning in children. *Annual Review of Linguistics*, 5, 353–373.
- Dębowski, L. (2006). On Hilberg's law and its links with Guiraud's law. *Journal of Quantitative Linguistics*, 13, 81–109.
- Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, 37, 66–108. <https://doi.org/10.1016/j.dr.2015.05.002>
- Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences of the United States of America*, 109(44), 17897–17902. <https://doi.org/10.1073/pnas.1215776109>
- Ferdinand, V., Kirby, S., & Smith, K. (2019). The cognitive roots of regularization in language. *Cognition*, 184(February 2017), 53–68. <https://doi.org/10.1016/j.cognition.2018.12.002>
- Ferrer i Cancho, R., & Sole, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3), 788–791.
- Ferrer-i-Cancho, R., Bentz, C., & Seguin, C. (2020). Optimal coding and the origins of Zipfian Laws. *Journal of Quantitative Linguistics*, 00(00), 1–30. <https://doi.org/10.1080/09296174.2020.1778387>
- Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, 1–87. <https://doi.org/10.1037/bul0000210>
- Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, 23(5), 389–407. <https://doi.org/10.1016/j.tics.2019.02.003>
- Glicksohn, A., & Cohen, A. (2013). The role of cross-modal associations in statistical learning. *Psychonomic Bulletin & Review*, 20(6), 1161–1169. <https://doi.org/10.3758/s13423-013-0458-4>
- Hendrickson, A. T., & Perfors, A. (2019). Cross-situational learning in a Zipfian environment (vol. 189(May 2017), pp. 11–22). <https://doi.org/10.1016/j.cognition.2019.03.005>
- Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences of the United States of America*, 105(31), 10681–10686.
- Koehn, P. (2008). Adquisición de una segunda lengua en estancias cortas en el extranjero: un análisis actitudinal. *Didáctica (Lengua y Literatura)*, 20(20), 117–134. <https://doi.org/10.5209/DIDA.19853>
- Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. *Cognition*, 127(3), 439–453. <https://doi.org/10.1016/j.cognition.2013.02.002>

- Lavi-Rotbain, O., & Arnon, I. (2017). *Developmental differences between children and adults in the use of visual cues for segmentation* (pp. 1–15). <https://doi.org/10.1111/cogs.12528>
- Lavi-Rotbain, O., & Arnon, I. (2019a). Children learn words better in low entropy. In *Proceedings of the 41th annual conference of the cognitive science society*.
- Lavi-Rotbain, O., & Arnon, I. (2019b). Low entropy facilitates word segmentation in adult learners. In *Proceedings of the 41th annual conference of the cognitive science society*.
- Lavi-Rotbain, O., & Arnon, I. (2021). Visual statistical learning is facilitated in Zipfian distributions. *Cognition*, 206.
- Lestrade, S. (2017). Unzipping Zipf's law. *PLoS One*, 12(8), 1–13. <https://doi.org/10.1371/journal.pone.0181987>
- MacWhinney, B. (1987). The competition model. In *Mechanisms of language acquisition* (pp. 249–308).
- MacWhinney, B. (2000). *The childes project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates. <https://doi.org/10.1162/089120100750105984>
- Mandelbrot, B. (1953). An informational theory of the statistical structure of language. *Communication Theory*, 2, 486–502.
- Manin, D. (2008). Zipf's law and avoidance of excessive synonymy. *Cognitive Science*, 32(7), 1075–1098. <https://doi.org/10.1080/03640210802020003>
- Matthews, D., & Bannard, C. (2010). Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child-directed speech. *Cognitive Science*, 34(3), 465–488. <https://doi.org/10.1111/j.1551-6709.2009.01091.x>
- Maye, J., Werker, J. F., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3), 101–111. [https://doi.org/10.1016/S0010-0277\(01\)00157-3](https://doi.org/10.1016/S0010-0277(01)00157-3)
- Meylan, S. C., Kurumada, C., Börschinger, B., Johnson, M., & Frank, M. C. (2012). Modeling online word segmentation performance in structured artificial languages. In *Proceedings of the 34th annual meeting of the cognitive science society*. <http://langcog.stanford.edu/papers/MKBJF-cogsci2012.pdf>.
- Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, 21(5), 1112–1130. <https://doi.org/10.3758/s13423-014-0585-6>
- Pryluk, R., Kfir, Y., Gelbard-Sagiv, H., Fried, I., & Paz, R. (2019). A tradeoff in the neural code across regions and species. *Cell*, 176(3), 597–609.e18. <https://doi.org/10.1016/j.cell.2018.12.032>
- Raviv, L., & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science*, 21(4), 1–13. <https://doi.org/10.1111/desc.12593>
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science (New York, N.Y.)*, 274(5294), 1926–1928. <https://doi.org/10.1126/science.274.5294.1926>
- Schuler, K. D., Reeder, P. A., Newport, E. L., & Aslin, R. N. (2017). The effect of Zipfian frequency variations on category formation in adult artificial language learning. *Language Learning and Development*, 13(4), 357–374. <https://doi.org/10.1080/15475441.2016.1263571>
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423. <http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5312191%5Cnhttp://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5312191&queryText=mathematicaltheoryofcommunication&newsearch=true>.
- Shor, Y., Reichart, R., & Arnon, I. (2022). *A cross-linguistic investigation of efficiency across speech genres*.
- Shufaniya, A., & Arnon, I. (2018). Statistical learning is not age-invariant during childhood: Performance improves with age across modality. *Cognitive Science*, 42, 3100–3115. <https://doi.org/10.1111/cogs.12692>
- Smith, K., & Kirby, S. (2008). Cultural evolution: Implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, 363(1509), 3591–3603.
- Taaseh, N., Yaron, A., & Nelken, I. (2011). Stimulus-specific adaptation and deviance detection in the rat auditory. *Cortex*, 6(8). <https://doi.org/10.1371/journal.pone.0023369>
- Takahira, R., Tanaka-Ishii, K., & Dębowski, L. (2016). Entropy rate estimates for natural language—a new extrapolation of compressed large-scale corpora. *Entropy*, 18(10), 5–8. <https://doi.org/10.3390/e18100364>
- Zipf, G. K. (1949). Human behavior and the principle of least effort. In *Human behavior and the principle of least effort*. Addison-Wesley Press.