Low Entropy Facilitates Word Segmentation in Adult Learners

Ori Lavi-Rotbain (oriedit.lavi@mail.huji.ac.il)
Edmond & Lily Safra Center for Brain Sciences, Edmond J. Safra Campus,
The Hebrew University Jerusalem, 9190401, Israel

Inbal Arnon (inbal.arnon@mail.huji.ac.il)
Department of Psychology, The Hebrew University of Jerusalem,
Mount Scopus, Jerusalem 91905, Israel

Abstract

Do language learners benefit from exposure to input that is more predictable and has lower entropy? Frequency is known to facilitate learning (more frequent words acquired earlier). However, frequency is only one measure of the distributional structure of the linguistic input. Here, we show that entropy also impacts language learning: adults show better word segmentation in an artificial language when the sequence has lower entropy (created by making one word more frequent). Segmentation improved both for the language as a whole, and for the less frequent words, despite appearing half the number of times. These results illustrate the facilitative effect of entropy reduction on language learning. Theoretically, they show that the effect of frequency is relative, not absolute, and that language learners are sensitive to more complex measures of the environment. Methodologically, they suggest that the prevalent use of uniform distributions in word segmentation studies may underestimate learners' abilities.

Keywords: Statistical learning; Word segmentation; Language Learning; Information.

Introduction

Frequency effects are prevalent across many aspects of language learning and processing. More frequent sounds, words and constructions are acquired earlier (Diessel, 2007; Goodman, Dale, & Li, 2008), and more frequent words are easier to recognize and produce (Jescheniak & Levelt, 1994). These effects are not restricted to single words: more frequent multiword phrases are also processed faster by adults (Arnon & Priva, 2013; Arnon & Snider, 2010), and produced more accurately by children (Bannard & Matthews, 2008). Frequency also impacts the structure of the lexicon: more frequent words tend to be phonologically shorter (Zipf, 1936). While frequency affects many domains in language, it captures only one aspect of the distributional structure of the linguistic environment. Frequency alone does not tell us about the co-occurrence patterns of words; the contexts in which words tend to appear; or how predictable the input is overall. In order to quantify such aspects of the linguistic input, other measures are required. Here, we focus on one such measure, Shannon's Entropy (Shannon, 1948). Shannon's entropy quantifies how unpredictable a variable is, with higher entropy assigned to less predictable variables. For instance, a toss of a fair coin has higher entropy than a toss of an unfair coin. Entropy tells us something about the entire distribution of words, beyond the properties of each individual word.

In the past decade, there has been growing interest in applying more complex measures like entropy to the study of language, and growing evidence for their impact on language structure and use. For example, information content is a better predictor of word length than frequency, with less predictable words tending to have longer lexical forms (Piantadosi, Tily, & Gibson, 2011). Similar effects are found in online processing: reading times are affected by entropy (Linzen & Jaeger, 2015), and speakers' production of less predictable words is slower (Cohen Priva, 2017) and less contracted (Frank & Jaeger, 2008). Children are also sensitive to such measures: two-year-olds show better repetition of unfamiliar four-words sequences when the final word "slot" has higher entropy (Matthews & Bannard, 2010). Complex measures have been shown to impact naturalistic language learning: Words with greater contextual diversity (appearing with more unique words) are acquired earlier (Hills, Maouene, Riordan, & Smith, 2010); as are words that used in more predictable temporal, spatial or linguistic settings (Roy, Frank, DeCamp, Miller, & Roy, 2015). More generally, speech directed to infants seems to be more predictable than adult-to-adult speech: it is more associative, repetitive and consistent than adult-to-adult speech (Hills, 2013). That is, caregivers talk to infants in ways that reduce the entropy of their input.

However, little work to date has looked at the impact of entropy on learning novel linguistic information: will entropy reduction lead to better learning? Here, we examine this question by looking at statistical learning, and in particular, at the classic word segmentation task of Saffran et al., (1996). Statistical learning (SL) has been studied extensively over the past 20 years, demonstrating human's ability to use distributional information to learn about various aspects of language structure (Romberg & Saffran, 2010). One of the first demonstrations of SL was for word segmentation, where infants were shown to use the lower transitional probabilities (TP's) between words as a cue to word boundaries (Saffran, Aslin, & Newport, 1996). Research since has shown that humans can also make use of such distributional information to learn more complex relations such as non-adjacent dependencies (Gomez, 2002) or multimodal associations (Cunillera, Laine, Càmara, & Rodríguez-Fornells, 2010; Lavi-Rotbain & Arnon, 2017).

Interestingly, even though word segmentation has been studied extensively, almost all such studies present learners with a uniform distribution where all elements appear an equal number of times (e.g., each of the words in the Saffran segmentation task appear equally often, but see Kurumada, Meylan & Frank, 2013, which we discuss in detail below). The uniform distribution differs from that of natural language where words follow a Zipfian distribution (Zipf, 1936). In this skewed distribution there is a small number of very frequent words, and a large number of low frequency words. Zipfian distribution is found across language, for both adult-to-adult speech (Zipf, 1936; Piantadosi, 2014) and child directed speech (Hendrickson & Perfors, 2019; Lavi-Rotbain & Arnon, submitted). That is, unlike word segmentation studies, words in natural language do not have a uniform distribution. More importantly from our perspective, uniform distributions are less predictable than Zipfian ones: Since each word is equally likely to appear, no guess is better than the other. Skewed distributions, such as the Zipfian distribution, are more predictable: when only a small number of words are highly frequent, they make a better guess than the rest. The difference in predictability between uniform and Zipfian distributions can be captured using entropy: the uniform distribution has maximal unigram entropy, while the Zipfian distribution has lower unigram entropy. That is, the uniform distributions used in word segmentation experiments differ from those of natural language in ways that may impede learning.

A recent word segmentation study provides some mixed results about the impact of non-uniform distributions on learning. Kurumada, Meylan & Frank (2013) compared performance on a word segmentation task in a uniform and a Zipfian distribution. While they found no advantage overall in the Zipfian condition, they did find strong frequency effects (more frequent words were segmented better), as well as contextual predictability effects (words that appeared more often next to the frequent word were segmented better). Stronger support for a Zipfian advantage can be found in a recent cross-situational word learning study where adults showed significantly better learning of novel word-object mappings when they were exposed to a Zipfian, rather than to a uniform, distribution (Hendrickson & Perfors, 2019). How are we to understand the lack of overall facilitation in Kurumada et al. (2013)? Non-uniform distributions could be facilitative for several different reasons. First, using non-uniform distributions leads to having words that are much more frequent than others. These frequent words can be learned early on and used as anchors for detecting additional word boundaries, similar to the way that presenting words in isolation can facilitate subsequent segmentation (Cunillera, Càmara, Laine, & Rodríguez-Fornells, 2010). In addition, TP's between words in such a distribution can be lower compared to TP's between words in a uniform distribution, making word boundaries more salient. More importantly, non-uniform distributions have lower entropy than uniform ones. The kind of Zipfian word distributions found in natural languages seem to have a particularly low unigram entropy. Using nine corpora of child-directed and adult-to-adult speech from five languages, we calculated the degree to which the unigram word entropy in the corpus deviates from the unigram entropy of a uniform distribution (assuming all words appear equally often). This deviation value was surprisingly consistent across languages, speech types (child-directed vs. adult), and corpus sizes, and had the average value of 0.63 (SD=0.03, see Lavi-Rotbain & Arnon, submitted, for full details). Note that we do not think there is something magical about this value, but rather are using it as a proxy for how unigram entropy is manifested in natural language. Interestingly, while the Zipfian condition in Kurumada et al. provided an anchor (a very frequent word), it deviated less from the uniform than natural languages (M=0.83). It is possible that learning is impacted by the degree of entropy such that it is facilitated only when there is a large enough reduction in entropy (compared to the entropy of a uniform distribution).

The current study

In the current study, we explore the impact of entropy reduction on word segmentation to test the hypothesis that entropy reduction is beneficial for learning when there is a large enough reduction in entropy. That is, we predict that entropy reduction will facilitate word segmentation beyond the effect of anchoring. To explore this hypothesis, we compare performance on a Saffran-style word segmentation task in three levels of entropy: high, medium and low. Entropy was reduced by making one word more frequent than the rest. The high entropy condition had a uniform distribution (all words appeared equally often). In the medium entropy condition, one word appeared 55% of the time: importantly, the entropy deviation of this condition was similar to that of Kurumada et al. (2013), and higher than what we found for words in natural language. In the low entropy condition, one word appeared 80% of the time, resulting in even lower entropy. Importantly, facilitation due to anchoring and lower TP's should happen similarly in both the medium and the low entropy conditions compared to the uniform one (the infrequent words appear almost always next to the frequent one in both). However, facilitation due to low entropy, if indeed a greater reduction in entropy is needed, should be present only in the low entropy condition.

We examine the effect of entropy reduction in two ways: by asking whether it is beneficial for words segmentation (1) in general, and (2) of infrequent words. We examine the first prediction by looking at adults' segmentation across the three entropy levels with the same exposure durations. If language learners are mostly sensitive to frequency, performance on the segmentation test should be affected by word frequency rather than entropy level. However, if learners are sensitive to more than mere frequency, e.g. to unigram entropy, then segmentation score in the low entropy condition should be better than in the high entropy condition. We examine the second prediction by comparing segmentation of items with the same low frequency, across different levels of entropy. We expect that low entropy will boost learning of low frequency items, such that low frequency words will be learned better when they appear in a sequence with lower unigram entropy, compared to when they appear in a uniform distribution (with high unigram entropy). If we will see facilitation only in the low entropy condition, but not in the medium entropy condition, this will show that this effect is not driven only by anchoring and lower TP's, but due to entropy of the input.

Method

Participants

142 undergraduate students at the Hebrew University of Jerusalem participated in the study (108 females, 34 males, mean age 24;0). Participants were randomly assigned to one of the four experimental conditions. All of the participants were native Hebrew speakers without learning disabilities or attention deficits. Participants received 10 NIS or course credit in return for their participation.

Materials

Auditory stimuli

The task was modelled on the audio-only condition from Lavi-Rotbain & Arnon (2017). Participants were exposed to a familiarization stream corresponding to the condition they were assigned to. All streams were composed of the same four unique tri-syllabic synthesized words: "dukame", "nalubi", "kibeto", and "genodi". We used only four words since we plan to use the same paradigm with children and needed to ensure that the task was usable also with young learners. The syllables making up the words were taken from 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 ensure that there were no co-articulation cues to word boundary. The words were matched for length (mean word length=860ms, range=845-888ms). The words were then concatenated together using MATLAB in a semi-randomized order to create the auditory familiarization streams. Importantly, there were no breaks between words and no prosodic or co-articulation cues in the stream to indicate word boundaries. The only cue for word boundaries was transitional probabilities (TP's): TP's between words were lower compared to TP's within words.

Experimental conditions

We created auditory sequences with three levels of entropy: high, medium and low, but with the same number of tokens (128) and length (1:50 minutes), in order to see if reduced entropy can facilitate segmentation. In the high entropy level, words followed a uniform distribution with each word appearing 32 times in a semi-randomized order (no word appeared twice in a row). TP's within a word were 1, and TPs between words were 0.333. In the medium entropy level, words appeared with a skewed distribution: one word appeared 55% of the time (71 appearances) while each of the other three words appeared 15% of the time (19 appearances for each word). In the low entropy level, words appeared with an even more skewed distribution: one word appeared 80% of the time (101 appearances) while each of the other three words appeared only 7% of the time (9 appearances for each word). In both the low and medium entropy conditions, the identity of the frequent word was counterbalanced across subjects. In addition, in both conditions the TP's within a word were 1, but the TP's between words varied depending on whether the next word was a frequent or infrequent one (see Table 1 for all the TPs). These conditions were used to examine the effect of entropy on the general segmentation score.

In order to look at the segmentation of the low frequency items, we added a uniform condition with high entropy but with shorter length (uniform-short). In this condition, each word appeared 19 times (76 tokens, lasting 1:05 minutes). The frequency of each word in this condition was matched to that of the infrequent words from the medium entropy condition (19 times), and was twice as frequent as the infrequent words from the low entropy condition (nine times). By comparing these infrequent words, we can examine the impact of entropy on words with low frequency. If the infrequent words in the low entropy condition will be learned better than the words in the uniform condition (despite appearing half the number of

Table 1: Different experimental conditions

	Uniform-short	Uniform	Medium entropy	Low entropy
Exposure length [minutes]	1:05	1:50	1:50	1:50
Number of tokens	76	128	128	128
Tokens per word	19	32	Frequent: 71 Infrequent: 19	Frequent: 101 Infrequent: 9
Unigram entropy [bits]	2	2	1.7	1.1
TP's within words	1	1	1	1
TP's between words	0.33	0.33	For the frequent word: 0.43 For infrequent words: 0.18	For the frequent word: 0.75 For infrequent words: 0.08

times), this will serve as strong evidence in favor of the facilitative nature of low entropy. Since this comparison resulted in the predicted effect, we did not run an additional uniform condition where each word appeared even fewer times. See Table 1 for full details of conditions.

Segmentation test

16 two alternative forced choice trials appeared in a random order, with the constraint that the same word/foil did not appear in two consecutive trials. Participants heard two words and were asked to decide which belonged to the language they heard. We used non-words as foils ("dunobi", "nabedi", "kilume", and "gekato", average length: 860ms; range 854-868ms), created by taking three syllables from three different words, while keeping their original position. Non-words, as opposed to part-words, never appeared together during exposure, making it easier to distinguish between them and real words. Since our goal was not to show that adults can discriminate words from part-words (a finding shown extensively), but to see how entropy affects this ability, we chose to focus only on the “easier” non-word versus word distinction (this was again motivated by our plan to run the same task with young children). In the test, each of the four words appeared once with each of the four foils to create 16 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.

Procedure

Participants completed the experiment on a computer while seated in a quiet room. They were told that they are going to listen to an alien language and will then be asked about it. A check-board image was displayed while they listened to the familiarization stream. After the exposure phase, participants completed the segmentation test.

Results

Participants were divided as follows between the four conditions: uniform, N=31; uniform-short, N=30; medium entropy, N=41; low entropy, N=40. In the medium and low conditions, each of the four words was the frequent one for ten subjects. A one way ANOVA (on each entropy rate separately) revealed that segmentation did not differ due to which word was the frequent one (for the medium entropy condition: F(3)=0.72, p=0.55; for the low entropy condition: F(3)=1.7, p=0.18). Consequently, in all subsequent analyses we collapsed the data across the different frequent words, for each of these conditions. Participants showed learning (were above chance) in all four conditions (low entropy condition: t(39)=12.57, p<0.001; medium entropy condition: t(40)=7.0, p<0.001; uniform condition: t(30)=7.0, p<0.001; uniform-short condition: t(29)=5.8, p<0.001) (see Fig. 1).

We used mixed-effect linear regression models to examine the effect of condition on performance. Following Barr et al. 2013, the models had the maximal random effect structure justified by the data that would converge. Our dependent binominal variable was success on a single trial of the segmentation test. We had experimental condition (dummy coded, meaning that each condition is compared to the uniform condition) as a fixed effect, as well as: log frequency of the word (centered); gender; 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 (Table 2). To examine the overall effect of experimental condition and word's frequency, we used two model comparisons. As predicted, experimental condition had a significant effect on performance (chi(3)=42.07, p<0.001). Participants showed better learning in the low entropy condition compared to the uniform condition (β=1.25, SE=0.22, p<0.001). However, performance in the medium entropy condition did not differ from the uniform condition (β=0.19, SE=0.19, p>0.1), suggesting that a larger reduction in entropy is needed to facilitate word segmentation. Performance on the uniform-short condition did not differ from that in the uniform condition, suggesting accuracy does not increase linearly with increased exposure (β=0.19, SE=0.2, p>0.1) (see also Siegelman, Bogaerts, Kronenfeld, & Frost, 2018).

Frequency also had a significant effect on segmentation (chi(1)=18.9, p<0.001). Participants showed higher accuracy for more frequent words (β=0.4, SE=0.09, p<0.001). Trial number significantly affected performance, with better accuracy in the beginning of the test (β= -0.03, SE=0.01, p<0.01). Accuracy was higher on trials where the word appeared before the foil (β=0.59, SE=0.1, p<0.001), as has been found in previous studies (Lavi-Rotbain & Arnon, 2017; Raviv & Arnon, 2018). Since the order of presentation of words and foils was counter-balanced this could not reflect a preference for pressing 1 or 2. Gender did not affect performance (β= -0.02, SE=0.16, p>0.1).

In order to examine the effect of entropy on low frequency words, we compared accuracy in learning: (1) the words in the uniform-short condition; (2) infrequent words from the medium entropy condition; and (3) infrequent words from the low entropy condition. The first two sets of words appeared 19 times during exposure, while the third set appeared only nine times. We used all trials (16 per subject) from the uniform-short condition (since they all had

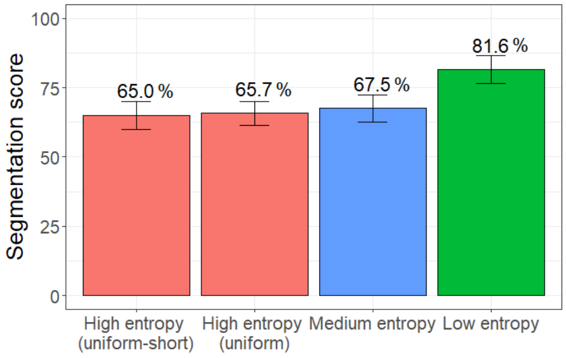

Fig. 1: Mean segmentation score by condition with 95% confidence intervals

the same frequency). However, for the medium and low entropy conditions, we included only trials in which the correct answer was one of the infrequent words (denoted as 'infrequent trials'). In these conditions, there were 12 infrequent trials for each subject. Participants showed learning of infrequent items (above chance) in all conditions (low entropy condition: M=78.8%, t(39)=9.59, *p*<0.001; medium entropy condition: M=64.8%, t(40)=5.3, *p*<0.001).

We used a mixed-effect linear regression model to look at the effect of entropy level on learning infrequent words. Our dependent binominal variable was success on a single trial. We had experimental condition as a fixed effect (each condition was compared to the uniform-short condition) as well as: gender, trial number (centered); and order of appearance in the test. The model had random intercepts for participants and for items. To examine the overall effect of condition, we used model comparisons. As predicted, experimental condition had a significant effect on learning infrequent words (chi(2)=16.9, *p*<0.001). Low frequency words were learned better in the low entropy condition than in the uniform-short condition (M=65%, β=0.78, SE=0.22, *p*<0.001). This effect is opposite to what would be expected based on mere frequency: these words appeared only nine times in the low entropy condition as opposed to 19 times in the uniform-short condition. Performance on infrequent trials in the medium entropy condition did not differ from the uniform-short condition (β=0.0, SE=0.2, *p*>0.1), suggesting again that a smaller reduction in entropy is not facilitative. Trial number affected performance, with better accuracy in the beginning of the test (β= -0.03, SE=0.01, *p*<0.05). Order of appearance in the test also affected performance, with better accuracy on trials where the word appeared before the foil (β=0.53, SE=0.1, *p*<0.001). Gender did not affect performance (β=0.06, SE=0.2, *p*>0.1).

How can we reconcile the general effect of frequency with the finding that words that appeared only nine times were learned better than those appearing 19 times? Our data suggests that what matters is not absolute frequency, but relative frequency: within each condition, the more frequent words were learned better. This is best illustrated in Fig. 2 in which we plotted segmentation means by condition (medium and low entropy) and by trial type (infrequent versus frequent trials). Frequency affects performance within conditions: frequent words are learned better in both entropy levels. However, this does not hold across conditions. For example, infrequent trials from the low entropy condition are numerically better than frequent trials from the medium entropy condition, despite of the sharp difference in frequency in the opposite direction: only nine appearances compared to 71. That is, only the relative frequency within each condition affected performance.

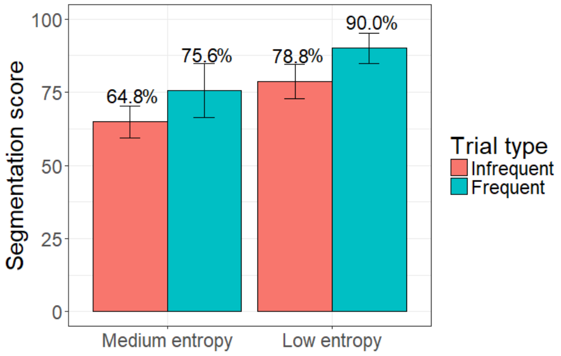

Fig. 2: Mean segmentation score by condition with 95% confidence intervals

One possible explanation for the entropy effect we found is that participants only learned the frequent word, and used it to rule out foils by elimination. If this is what they did, we should see a difference in segmentation scores across foils: foils that share a syllable with the frequent word should be easier to reject compared to foils that do not. For example, if the frequent word for a participant is 'nalubi', we should see better accuracy in rejecting 'nabedi' that shares the first syllable with 'nalubi', compared to rejecting 'gekato' that does not share a syllable with 'nalubi'. However, we saw no such effects (success when the foil shared a syllable with frequent word was 82% vs. 79% when it did not). We used a linear regression model with success on a single trial as the dependent binominal variable, and "is foil frequent" (assigned '1' for trials in which the foil shared any of its three syllables with the frequent word and ‘0’ when it didn’t) as a fixed effect, as well as log frequency (centered), gender, trial number (centered); and order of appearance in the test. The model had random intercepts for items. "Is foil frequent" was not a significant predictor of accuracy (β=0.26, SE=0.23, *p*>0.1), while log frequency, trial number

Table 2: Mixed-effect regression model for all four conditions. Variables in bold were significant. Significance obtained using the lmerTest function in R.

	Estimate	**Std. Error**	**z value**	**p-value**
(Intercept)	0.27331	0.17793	1.536	>.1
uniform-short condition	0.17777	0.20571	0.864	>.1
Medium entropy condition	0.18484	0.18791	0.984	>.1
Low entropy condition	1.25277	0.21789	5.750	**<.001 *****
Log frequency (centered)	0.40138	0.09691	4.142	**<.001 *****
Gender (male)	-0.01982	0.16113	-0.123	>.1
Trial number (centered)	-0.03469	0.01061	-3.271	**<.01 ****
Order of appearance (word)	0.59277	0.09781	6.061	**<.001 *****

and order of appearances remained significant predictors. That is, the boost for the infrequent words in the low entropy condition seems to reflect the better learning of those words.

Discussion

We set to ask if reduced unigram entropy can improve segmentation in a classic auditory SL task (1) in general, and (2) of infrequent words. In addition, we wanted to see if the effect of entropy reduction may be not linear, such that the lack of general facilitation when using a Zipfian distribution in a previous word segmentation study (Kurumada, Meylan, & Frank, 2013) was driven by a not large enough reduction of entropy. In addition, we wanted to explore the potential advantage of low entropy beyond the effect of anchoring driven by better learning of the more frequent word. To do so, we examined adults' word segmentation in an artificial language across three levels of entropy (high, medium and low). Entropy was reduced by making one word more frequent than the rest, so that it appeared 55% (medium) or 80% (low entropy) of the time.

Our results show that entropy reduction does facilitate learning, but only when entropy is low enough. As in the Zipfian condition in Kurumada (2013), reducing entropy to medium level did not facilitate segmentation. However, lower levels of entropy did facilitate learning compared to uniform conditions with the same length. This effect was not driven only by improved learning of the frequent words. The low frequency words were learned better in the low entropy condition compared to medium and high levels, despite appearing half the number of times (nine vs. 19). Further analyses ruled out alternative explanations: the facilitation cannot be explained by ruling out foils that share syllables with the frequent word. The facilitation in the low entropy conditions cannot be fully attributed to anchoring: even though the infrequent words appeared almost always next to a frequent word in the medium entropy condition, their segmentation was not facilitated relative to a uniform distribution. Instead, it seems that learners are sensitive to the overall entropy of the distribution, with better learning at lower entropy levels. Further research is needed to understand which entropy levels are facilitative and why: Is there an optimal entropy level for learning or is the effect of entropy reduction continuous? We are examining these questions in ongoing studies.

In addition to the effect of entropy, our findings highlight the importance of relative, rather than absolute frequency on learning. Frequency effects were present only within conditions and not across conditions. For example, infrequent words from the low entropy condition, that appeared only nine times, were learned better than the infrequent words in the medium entropy condition (appearing 19 times). Moreover, they were learned as well as the frequent word in the medium entropy condition despite appearing much less (nine vs. 71 times). Our results indicate that humans are sensitive to complex measures such as unigram entropy in the process of language learning, and that a more predictable distribution, more similar to the one found in natural language, can be beneficial for learning compared to a uniform one. In addition, we provide novel evidence showing that low frequency items can 'overcome' their frequency when appearing with higher frequency items, in a more predictable distribution.

These results have implications for artificial language experiments. The vast majority of artificial language experiments use a uniform distribution in which all items have equal frequency. Uniform distributions are not ecological since the natural language we are exposed to shows a Zipfian distribution (Zipf, 1936; Piantadosi, 2014) even in speech directed to infants at their first stages (Lavi-Rotbain & Arnon, submitted). Our results highlight an additional drawback of using uniform distributions in the lab: such distributions can impede performance compared to more skewed, low entropy distributions. That is, we may be significantly underestimating learners' abilities when using uniform distributions. This is of particular importance when such tasks are used to determine what learners can (or cannot) learn. We are currently investigating the impact of entropy on learning in children, and for other kinds of SL tasks. Specifically, we are investigating other distributions with low entropy to ensure that the effect observed holds for other distributions (e.g. Zipfian distributions with similar deviation from the uniform). In addition, further research should look if a similar effect can be found in even more ecologically environments. For example, although we used a non-uniform distribution, words were not predictable of one another. However, in natural language this is not the case: words are predictable of one another. This seem to be important for proper segmentation: a segmentation model that accounted only for a unigram model, ended up with under-segmentation of the corpus. However, a model that assumed that words are predictive of one another gained better results (Goldwater, Griffiths, & Johnson, 2009). Therefore, creating distributions that are non-uniform and with word dependencies, could lead to better results.

Beyond artificial language experiments, these results have implications for our understanding of the factors that impact language learning. While frequency effects on language learning have been studied extensively (Goodman et al., 2008; Jescheniak & Levelt, 1994), the effect of more complex measures remain relatively understudied. Our results highlight the role of entropy in learning and open up new research directions on the impact of entropy on real-life language learning. What is the informative structure of child-directed speech? Does variance in entropy predict the age of acquisition of words? Can we see similar effects of the environment words appear in on natural language learning? We are currently engaged in a series of studies investigating these questions, which can further deepen our understanding of infants' first steps into language and the formation of their vocabulary.

Acknowledgments

We wish to thank Zohar Aizenbud for her help. The research was funded by the Israeli Science Foundation grant number 584/16 awarded to the second author.

References

Arnon, I., & Cohen Priva, U. (2013). More than words: The effect of multi-word frequency and constituency on phonetic duration. *Language and speech*, *56*(3), 349-371.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. Journal of Memory and Language, 62(1), 67–82.

Bannard, C., & Matthews, D. (2008). Stored Word Sequences in Language Learning. Psychological Science, 19(3), 241–248.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255–278.

Boersma, P., & van Heuven, V. (2001). Speak and unSpeak with Praat. Glot International, 5(9–10), 341–347.

Cohen Priva, U. (2017). Not so fast: Fast speech correlates with lower lexical and structural information. Cognition, 160, 27–34.

Cunillera, T., Càmara, E., Laine, M., & Rodríguez-Fornells, A. (2010). Words as anchors: Known words facilitate statistical learning. Experimental Psychology, 57(2), 134–141.

Cunillera, T., Laine, M., Càmara, E., & Rodríguez-Fornells, A. (2010). Bridging the gap between speech segmentation and word-to-world mappings: Evidence from an audiovisual statistical learning task. Journal of Memory and Language, 63(3), 295–305.

Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. New Ideas in Psychology, 25(2), 104–123.

Frank, A. F., & Jaeger, T. F. (2008). Speaking Rationally : Uniform Information Density as an Optimal Strategy for Language Production. The 30th Annual Meeting of the Cognitive Science Society (CogSci08), 939--944.

Glicksohn, A., & Cohen, A. (2013). The role of cross-modal associations in statistical learning. Psychonomic Bulletin & Review, 20(6), 1161–1169.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. Cognition, 112(1), 21–54.

Gomez, R. L. (2002). Variability and Detection of Invariant Structure. Psychological Science, 13(5), 431–436.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. Journal of Child Language, 35(3), 515–531.

Hendrickson, A. T., & Perfors, A. (2019). Cross-situational learning in a Zipfian environment. 189(May 2017),11–22.

Hills, T. (2013). The company that words keep: comparing the statistical structure of child- versus adult-directed language. Journal of Child Language, 586–604.

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. Journal of Memory and Language, 63(3), 259–273.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word Frequency Effects in Speech Production: Retrieval of Syntactic Information and of Phonological Form. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(4), 824–843.

Kurumada, C., Meylan, S. C., & Frank, M. C. (2013). Zipfian frequency distributions facilitate word segmentation in context. Cognition, 127(3), 439–453.

Lavi-Rotbain, O. & Arnon, I. (submitted). Zipf's Law in Child-Directed Speech.

Lavi-Rotbain, O. & Arnon, I. (submitted). Language-like Entropy Facilitates Word Segmentation in both Adults and Children.

Lavi-Rotbain, O., & Arnon, I. (2017). Developmental Differences Between Children and Adults in the Use of Visual Cues for Segmentation. Cognitive Science.

Linzen, T., & Jaeger, T. F. (2016). Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science*, *40*(6), 1382-1411.

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.

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.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences of the United States of America, 108(9), 3526–3529.

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, (May), 1–13.

Romberg, A., & Saffran, J. R. (2010). Statistical learning and language acquisition. Wiley Interdisciplinary Reviews: Cognitive Science, 1(6), 906–914.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. Proceedings of the National Academy of Sciences, 112(41), 12663–12668.

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.

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2018). Redefining “Learning” in Statistical Learning: What Does an Online Measure Reveal About the Assimilation of Visual Regularities? Cognitive Science, 42(1996), 692–727.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27, 379–423.

Zipf, G. (1936). The Psychobiology of Language. London: Routledge.