WORD SEGMENTATION IS FACILITATED IN ZIPFIAN DISTRIBUTIONS

Lavi-Rotbain, O. ^{*1}, Arnon, I. ^{1}

*Corresponding Author: orilavirotbain@gmail.com
^{1}Psychology Department, Hebrew University, Jerusalem, ISRAEL

While languages differ from one another in many respects, they share certain commonalties: these can provide insight on our shared cognition and the way it may impact language structure. Here, we focus on one of the striking commonalities between languages, the way word frequencies are distributed. Across languages, words follow a Zipfian (or near-Zipfian) distribution, showing a power law relation between a word's frequency and its rank (Piantadosi, 2014; Zipf, 1949). The source of such distributions in language is debated (e.g., Chater & Brown, 1999), but they have been proposed to reflect foundational aspects of human cognition and/or communication (e.g., Ferrer i Cancho & Sole, 2003; Ferrer i Cancho et al. 2020; Manin, 2008). Regardless of their source, their recurrence in language may have advantages for learning: skewed distributions may facilitate various aspects of language. Such individual biases could be amplified over time, creating pressure to maintain skewed distributions (as has been argued for other linguistic domains, e.g., Kirby et al. 2008).

In this paper, we explore the learnability consequences of Zipfian distributions for word segmentation, a crucial aspect of early language acquisition. Word segmentation has been studied extensively in the lab using artificial language learning tasks (Saffran et al. 1996), but learners are usually presented with a uniform learning environment where each novel word appears equally often. The few studies that examined learning from Zipfian distributions suggest they are beneficial for word segmentation (Kurumada et al. 2013; Meylan et al. 2012), but the extent and generality of this effect is unclear. More importantly, we do not know *what* about Zipfian distributions impacts learning.

Here, we propose and test the prediction that Zipfian distributions are facilitative because of their lower unigram entropy. We start by quantifying unigram entropy (using efficiency, a normalized entropy measure, see Eq. 1) in child-directed speech – children’s actual learning environment - across 15 languages (following Bentz et al. 2017 who did so for adult-to-adult speech). We find that efficiency spans a surprisingly narrow range across languages (range=0.6-0.7, mean=0.63, SD=0.03). We then test the impact of those values on learning by manipulating unigram entropy and distribution shape using a classic artificial word segmentation paradigm. We compare performance at three efficiency levels: maximal (uniform distribution), reduced (skewed distribution, with lower unigram entropy than the uniform but higher than natural language), and language-like (within the range found for natural language). We find that word segmentation in adults is uniquely facilitated in language-like efficiency overall (see Fig. 1a), and for the low frequency words (see Fig. 1b), but does not improve in reduced efficiency. We find similar results in a second study using a differently skewed distribution with similar efficiency.

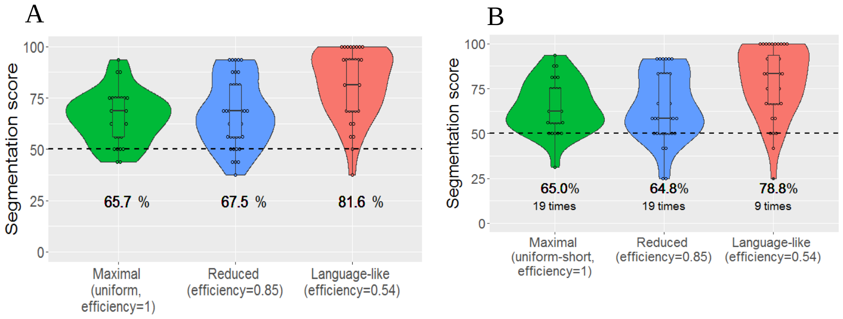

Figure 1. Adults' segmentation scores across conditions (a), and for low frequency words (b).

These findings show that child-directed speech has similar unigram entropy across languages, and that these values are uniquely facilitative for word segmentation. We discuss the possible role of learnability pressures in the emergence of such distributions.

References

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.

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*, *105*(31), 10681-10686.

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. Retrieved from

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.

Zipf, G. K. (1949). Human behavior and the principle of least effort. In *Human* *behavior and the principle of least effort.* Oxford, England