Word-like units emerge through iterated sequence learning

Lucie Wolters[*1], Simon Kirby[2], and Inbal Arnon[1]

[*]Corresponding Author: lucia.wolters@mail.huji.ac.il
[1]Department of Psychology, The Hebrew University, Jerusalem, Israel
[2] School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, Edinburgh, UK

Language exhibits systematicity in the way sets of discrete units, such as words or syllables, are used and reused to form linguistic sequences. This kind of systematicity is reflected in the distributional statistics of language, which in turn provides cues that can help learners discover the building blocks of language – a crucial milestone in language learning. Previous work has shown that a difference in within-unit and between-unit transitional probabilities (e.g. Fiser & Aslin, 2002; Saffran et al., 1996) and a skewed frequency distributions of these units (e.g. Lavi-Rotbain & Arnon, 2021, 2022) can facilitate segmentation in both linguistic and non-linguistic learning domains. However, it is still unexplained how linguistic units and their distributional properties arise in language in the first place. Here, we investigate experimentally whether their emergence may be driven by domain-general constraints on sequence learning over the course of language being repeatedly learned and transmitted.

We conducted an online non-linguistic iterated sequence learning experiment based on Cornish, Smith, and Kirby (2013) in which participants observed and reproduced sets of color sequences that were produced by a previous participant. In each trial, a participant was shown a sequence, made up of four possible colors (red, yellow, green, and blue), and asked to immediately reproduce it. Each participant reproduced a set of 30 sequences, which was transmitted to the next participant. We collected data for 10 transmission chains of 10 generations. The sequences in the initial sets had a length of 12 and were randomly generated.

As is typical in iterated learning experiments, we found a decrease in reproduction error over generations (see figure 1), indicating that the sets of sequences evolved to become easier to reproduce. This was found after accounting for sequence length. To extract units (sub-sequences) from the

sequence sets, we used a segmentation method developed by Arnon and Kirby (2024) that segments individual sequences based on the transitional probabilities of the colors the sequence sets. Unit boundaries were created when there was a drop in probability (see figure 1) – similar to how word-boundaries in natural language are often found where the probability of syllable transitions are low (shown in e.g. Stärk et al., 2022). We found that statistically coherent units emerge over iterations and that the distribution of these units became increasingly more skewed, reflected by a decrease in unit entropy (see figure 1). Moreover, the distribution of units showed an increasingly better fit to a Zipfian distribution over time, the typical distribution of word frequencies across languages (Mehri & Jamaati, 2017). Importantly, unit entropy was highly correlated with reproduction error, indicating that the distributional structure of the sequence sets increased their learnability. In addition to these results, I will explore different methods to extract units from sequences and in doing so, contrast the outcomes of transition- and chunking-based learning strategies.

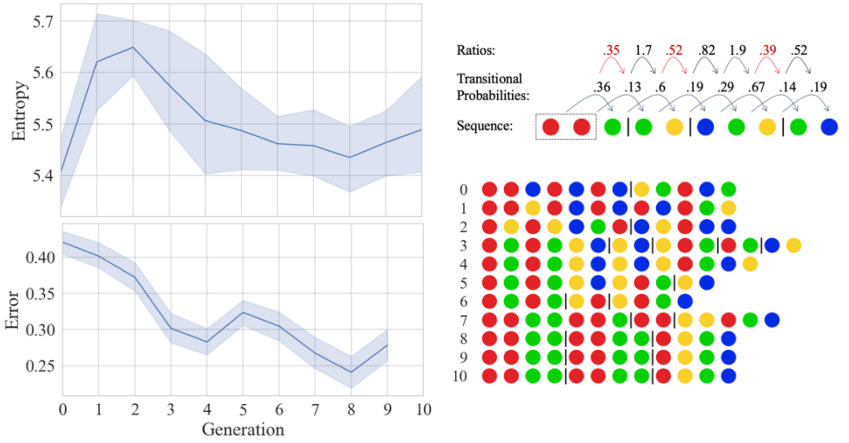

Figure 1. Left-top: Drop in mean unit entropy of chains over generations. Left-bottom: Drop in mean transmission error of chains over generations Right-top: Visualization of the segmentation method used to segment the sequences into units. Right-bottom: The change of a single sequence (out of a set of 30), over the 10 generations. The figure shows sequence 5 from chain D.

Taken together, our findings suggest that domain-general learning pressures during cultural transmission can shape the distributional structure of language and with that, that aspects of linguistic structure may emerge independently of the structure in the meanings that are being conveyed. These results lead to interesting predictions on the emergence of the distributional structure in other culturally transmitted behaviors, such as music.

References

Arnon, I., & Kirby, S. (2024). Cultural evolution creates the statistical structure of language. Scientific reports, 14(1), 5255. https://doi.org/10.1038/s41598-024-56152-9

Cornish, H., Smith, K., & Kirby, S. (2013). Systems from Sequences: An Iterated Learning Account of the Emergence of Systematic Structure in a Non-Linguistic Task. Proceedings of the Annual Meeting of the Cognitive Science Society, 35(35).

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. Journal of Experimental Psychology: Learning, Memory, and Cognition, 28(3), 458–467.

Lavi-Rotbain, O., & Arnon, I. (2021). Visual statistical learning is facilitated in Zipfian distributions. Cognition, 206, 104492.

Lavi-Rotbain, O., & Arnon, I. (2022). The learnability consequences of Zipfian distributions in language. Cognition, 223.

Mehri, A., & Jamaati, M. (2017). Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations. Physics Letters A, 381(31), 2470–2477.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. Science, 274(5294), 1926–1928.

Stärk, K., Kidd, E., & Frost, R. L. A. (2022). Word Segmentation Cues in German Child-Directed Speech: A Corpus Analysis. Language and Speech, 65(1), 3–27.