Segmentation Can Drive the Cultural Evolution of the Statistical Properties of Language

Lucie Wolters (lucia.wolters@mail.huji.ac.il)
The Hebrew University of Jerusalem, Jerusalem, Israel.

Inbal Arnon (inbal.arnon@mail.huji.ac.il)
The Hebrew University of Jerusalem, Jerusalem, Israel.

Simon Kirby (simon.kirby@ed.ac.uk)
The University of Edinburgh, Edinburgh, United Kingdom.

Abstract

Language is passed from one generation of learners to the next via cultural transmission. This process has been shown to give rise to core properties of language that enhance its learnability. Recent experimental work shows that statistical properties of language can also emerge through cultural transmission: specifically, the statistical coherence of words, and the Zipfian distribution of word frequencies. It has been proposed that these properties emerge because they facilitate segmentation. However, it is not clear whether segmentation is necessary for their emergence. We use a computational iterated learning model to simulate the cultural transmission of unsegmented sequences under different assumptions about the nature of learning. We show that segmentation indeed promotes the emergence of these statistical properties, whereas tracking of unigram statistics does not. In addition, we show that tracking sequential statistics alone can also promote their emergence.

Keywords: language evolution; cultural transmission; segmentation; statistical properties; Zipfian distribution; agent-based computational model; iterated learning; statistical learning

Introduction

Languages are repeatedly passed from one generation to the next via cultural transmission. A large number of studies have shown that this can constrain which linguistic forms are more likely to be transmitted (e.g. Christiansen & Chater, 2008; Kirby et al., 2007). One such constraint is that language has to be learned by each generation, creating a pressure for the system to be learnable. This model of language evolution has been used to explain why certain linguistic forms are prominent across languages, as well as the emergence of core properties of language (Culbertson et al., 2012; Kirby et al., 2008; Smith et al., 2017; Verhoef et al., 2014).

Recent work suggests that cultural transmission can also explain the emergence of two characteristic statistical properties of language (Arnon & Kirby, 2024; Cornish et al., 2017; Shufaniya & Arnon, 2022): (1) language consists of statistically coherent units, meaning that the transitional probabilities between adjacent syllables tend to be higher when both syllables belong to the same word, as opposed to when they cross a word-boundary (Saksida et al., 2017; Stärk et al., 2022); and (2) word frequencies follow a highly skewed distribution that is consistent with Zipf's power law (Piantadosi, 2014; Zipf, 1949). Language learners can use the statistical coherence of words to discover word boundaries in continuous speech, a crucial ability in language learning (for review: Saffran & Kirkham, 2018). Moreover, speech segmentation is facilitated when the relevant words follow a Zipfian distribution (Kurumada et al., 2013; Lavi-Rotbain & Arnon, 2022). Based on these findings, Arnon and Kirby (2024), hypothesised that the learnability advantages of statistically coherent words and Zipfian frequency distributions can drive their emergence in language over the course of cultural transmission. They used a non-linguistic iterated learning experiment, simulating the process of cultural transmission, in which participants reproduced sets of unsegmented colour sequences that were produced by a previous participant. To analyse the statistical properties of the sets, sequences were segmented into units (sub-sequences) based on the transitional probabilities of the colours in the set, using a segmentation method that was motivated by the infant statistical-learning literature. Indeed, sets evolved to have statistically coherent units with a frequency distribution approaching a power-law, and the emergence of these properties was correlated with higher transmission accuracy.

Arnon and Kirby (2024) hypothesised that the emergence of these statistical properties arises from the repeated segmentation and learning of a set of units by each generation. In other words, sequences end up having the statistical properties that make them segmentable and learnable because they are transmitted by learners who segment those sequences into units when reproducing them. However, although this hypothesis aligns with previous work on iterated learning in which languages tend to evolve to optimise their own learnability, it is not clear whether segmentation is required to explain their findings. Could the frequency distribution found in language arise from a simpler process in which learners at each generation are tracking lower-level statistics in their input? In other words, is segmentation actually required for the emergence of statistically coherent units that exhibit a Zipfian distribution? To address these questions, we use a computational iterated learning model to simulate the cultural evolution of sequences under different assumptions about the nature of learning.

We will proceed as follows: First, in Study 1, we present a set of simulations comparing two types of learners: a frequency learner that reproduces the unigram statistics of the input, and a lexicon[1] learner that segments the input into units and reproduces the frequency distribution of the lexicon of units. Next, in Study 2, we ask whether reproducing the sequential statistics of the input can lead to the emergence of the statistical properties as well. To do so, we compare a learner that reproduces the sequential statistics of the input but does not use them to segment the input into units, to the lexicon learner from Study 1. We evaluate each learner by asking whether cultural transmission leads to the emergence of statistically coherent units and a Zipfian frequency distribution. However, first, we will provide a brief description of the segmentation method developed by Arnon and Kirby (2024), which we use in the simulations and analyses.

Segmentation Method

To analyse the emergence of statistical properties in sequence sets over transmission, Arnon and Kirby (2024) segmented the sequences into units (sub-sequences) using a segmentation method that was motivated by the infant statistical-learning literature. This method uses drops in the transitional probabilities between colours as indicators of unit boundaries, similar to how infants use drops in the transitional probabilities between syllables to discover word boundaries in continuous speech (e.g. Saffran et al., 1996). The probability of a transition was calculated as the probability of a colour to follow the two preceding colours, based on their frequency in the whole set. The segmentation method posits a boundary when the probability of a transition is low compared to the preceding one by looking at the ratio between the current and previous transition. A set of randomly generated sequence sets was used to estimate when a drop in probability was large enough to justify segmentation: the transitional-probability ratio distributions of the sets were aggregated, and the 5% lower tail of this distribution was set as the segmentation threshold: sequences were segmented where a ratio was lower than that (ratio = 0.425).

Agent-based Segmentation Method

We use an agent-based iterated learning model to simulate the evolution of sequence sets. At each generation, a single learner observes a sequence set produced by a previous agent, extracts frequency information about the set, and then uses this information to produce a set of sequences that serves as input to the next learner.

The setup of the simulations is chosen to match the experimental design in Arnon and Kirby (2024): sequence sets consist of 60 sequences made up of four colours (red, green, yellow, and blue); each simulation starts with a set of randomly generated sequences with a length of 12 – resulting in a near-uniform distribution of the four colours in the set; each sequence set is transmitted over ten generations of learners; and the length of the sequences is constrained to a range of 8 to 16 during transmission. Sequences are represented as strings of letters (e.g. ‘RYBGGBY’), where each letter represents a colour. Our simulations differ from the experimental design in that the agents are not required to accurately reproduce individual sequences. Instead, they reproduce a set of new sequences by reproducing the statistics of the input. We ran 1000 simulations for each type of learner.

Study 1: Unigram Learning and Lexicon Learning

In the first set of simulations, we simulate the evolution of sequence sets with two types of learners: (1) unigram learners, that learn the frequency distribution of the four colour tokens in the set, and (2) lexicon learners, that segment the individual sequences to extract a set of units of varying lengths and token frequencies. If segmentation introduces a pressure for the emergence of statistically coherent units and their Zipfian frequency distribution, then we expect to see these properties emerge with lexicon learners. If reproducing low-level statistics is sufficient for their emergence, then we should see similar results with unigram learners.

To look at the emergence of statistical coherence and Zipfian distributions, all simulated sequence sets, irrespective of the type of learner, are segmented using the method described in the introduction. The resulting units and their respective frequency distributions are analysed for their statistical coherence and fit to a Zipfian distribution using the same measures as in Arnon and Kirby (2024), which are described in the Results section of Study 1.

Unigram Learner

At each generation, a learner extracts the frequencies of the four colours in the observed sequence set. A single sequence is produced by consecutively sampling a colour from the observed frequency distribution. The N of sampled colours is set by randomly sampling a number from the predetermined range of 8 to 16. Each learner produces 60 sequences.

Lexicon Learner

At each generation, a learner applies the segmentation method described in the introduction to the observed sequence set and extracts the frequencies of the segmented units. A single sequence is produced by sampling and concatenating N units from the observed frequency distribution of units. The N of sampled units is set as a random number between 1 and 16. Only sequences with a length that falls within the predetermined range of 8 to 16 are considered valid. A learner produces sequences until it has a set of 60 valid sequences.

[1] By 'lexicon' we refer to a collection of segmented units extracted from the sequence set, which, unlike the traditional use of the term, do not carry any inherent meaning.

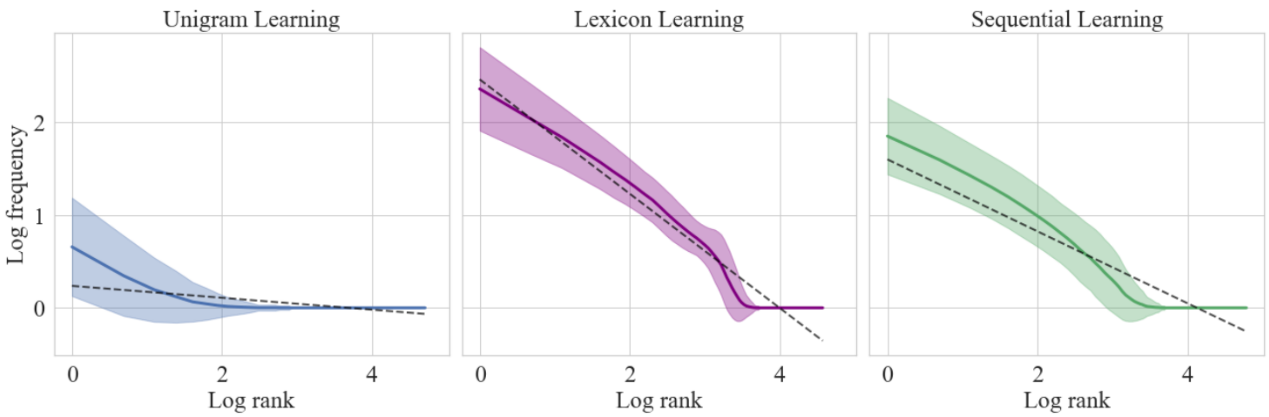

Figure 1: Distributions of units of the 1000 simulated chains after 10 generations. Each is plotted as the frequency of each unit against its rank on a log-log scale. The standard deviation across chains is shown as a shaded region around the mean. The reference linear regression for each distribution is shown as a dashed black line.

Results

To measure the statistical coherence of segmented units, we look at the transitional probabilities of colours within units, where high transitional probability indicates high statistical coherence. Figure 2 (top) shows the distributions of mean within-unit transitional probability (TP) for unigram- and lexicon learners after ten generations of learning. Lexicon learning results in units with higher within-unit TP, suggesting that segmentation promotes the emergence of statistically coherent units. In contrast, unigram learning does not promote higher within-unit TPs, suggesting that reproducing the unigram distribution of the sets is not sufficient for the emergence of statistically coherent units. To determine whether the difference in within-unit TP between unigram- and lexicon learners is statistically significant after ten generations, we generated a distribution of differences by randomly pairing unigram learners with lexicon learners and calculating the difference in their values. We then calculated the probability that the difference between two learners is less than zero. We found that the mean within-unit TP values resulting from lexicon learning are significantly higher than those resulting from unigram learning ($\mu = 0.18$, $\sigma = 0.03$, $p < .001$).

Figure 1 (panel 1 and 2) shows the frequency distributions of units for unigram- and lexicon learners after ten generations of learning. To test the fit of these distributions to a Zipfian power-law, we look at the R^2 fit of the frequency distributions to a linear regression on a log-log plot of frequency and rank. If a distribution is Zipfian, showing a power law relation between a unit's frequency and rank, then we expect a high R^2 fit. Figure 2 (bottom) shows the distributions of R^2 values for unigram- and lexicon learners after ten generations of learning. Lexicon learning results in frequency distributions with higher R^2 values, suggesting that learning segmented units promotes distributions with a better fit to a Zipfian distribution. In contrast, unigram learning does not lead to high R^2 values, suggesting that reproducing the unigram distribution of sequence sets alone, results in frequency distributions with a poor fit to a Zipfian distribution. To determine whether the difference in R^2 values between unigram- and lexicon learners is statistically significant, we calculated a distribution of differences in R^2 values. We found that the R^2 values resulting from lexicon learning are significantly higher than those resulting from unigram learning ($\mu = 0.68$, $\sigma = 0.18$, $p < .001$).

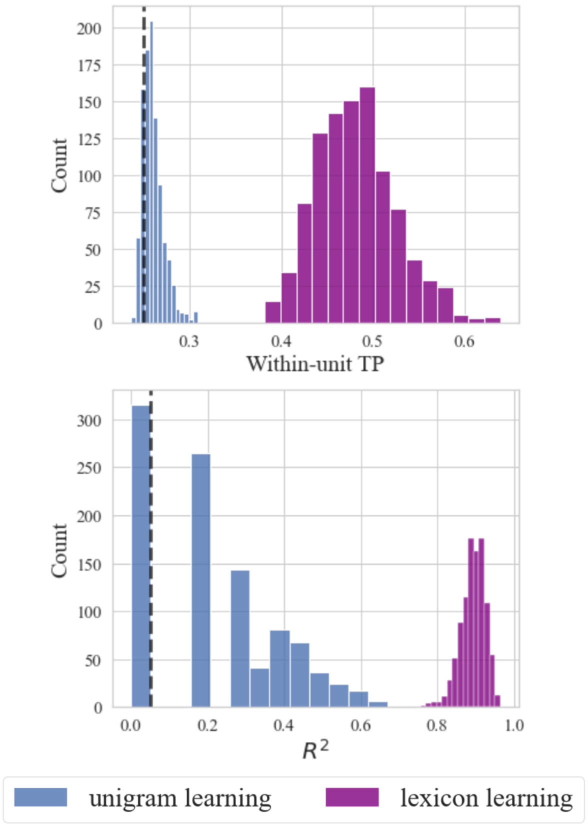

Figure 2: Distribution of mean within-unit transitional probability (top) and R^2 values (bottom), of the 1000 simulated chains after 10 generations. The mean value of generation zero (which is randomly generated**)** is indicated with the black dashed line.

Study 2: Sequential Learning

The two types of learners in Study 1 differed in two respects: first, the unigram learner had no knowledge of any sequential information in the input, whereas the lexicon learner tracks the sequential information in order to segment the sequences. Second, the lexicon learner builds a lexicon, whereas the unigram learner does not. It is possible that the emergence of the statistical properties by lexicon learning is largely driven by the transmission of sequential information, rather than by the segmentation process. If this is the case, then we may also see the emergence of these statistical properties when sets are transmitted by learners who reproduce the sequential statistics of sequence sets.

To test this, we simulate the evolution of sequence sets by a learner that reproduces the sequential statistics of the input but does not use these statistics to segment the input into units, and compare the results to that of the lexicon learner in Study 1, a learner that reproduces the lexicon of segmented units by using those same sequential statistics but does not seek to reproduce the sequential statistics directly. If sequential learning introduces a pressure for the emergence of statistically coherent units with a Zipfian frequency distribution, then we expect to see these properties emerge with sequential learners. If sequential learning promotes these properties to a lesser extent than lexicon learning, then we expect to find higher within-unit transitional probabilities and R^2 values with lexicon learners compared to sequential learners.

Sequential Learning

At each generation, a learner identifies the conditional statistics of how often each of the four colours follows any given pair of colours in the set. For example, it learns how often red, green, blue, and yellow occur after the pair blue-green in the set. A single sequence is produced by sampling and concatenating individual colours, like the unigram learner in Study 1, with the difference being that a colour is sampled based on its conditional probability given the two preceding colours. For example, if the preceding colours are red-green, and yellow follows red-green 40% of the time in the input, then yellow has a 40% chance of being sampled as the next colour in the sequence. The initial two colours of each sequence are randomly chosen. The length of each sequence is set by randomly sampling a number the predetermined range of 8 to 16. Each learner produces 60 sequences.

Results

Figure 3 (top) shows the distributions of mean within-unit TP for sequential- and lexicon learners after ten generations. The figure shows that sequential learning results in high within-unit TP, albeit to a lesser extent than lexicon learning. This suggests that sequential learning alone, without segmentation, can also promote the emergence of statistically coherent units. The mean within-unit TP values resulting from sequential learning are significantly higher than those resulting from unigram learning ($\mu = 0.15$, $\sigma = 0.03$, $p < .001$), however, they are not significantly different from those resulting from lexicon learning ($\mu = 0.03$, $\sigma = 0.04$, $p = .24$).

Figure 1 (third panel) shows the frequency distribution of units after ten generation of sequence learning. Sequential learning results in frequency distributions with higher R^2 values, albeit again to a lesser extent than lexicon learning (see Figure 3, bottom). This suggests that sequential learning alone, can also promote the emergence of Zipfian distributions. Analyses of the differences between the different types of learners reveals that while the R^2 values of sequential learners are significantly higher than those of unigram learners in Study 1 ($\mu = 0.55$, $\sigma = 0.2$, $p < .01$), there is no significant difference between the sequential learners and lexicon learners ($\mu = -0.13$, $\sigma = 0.1$, $p = .09$).

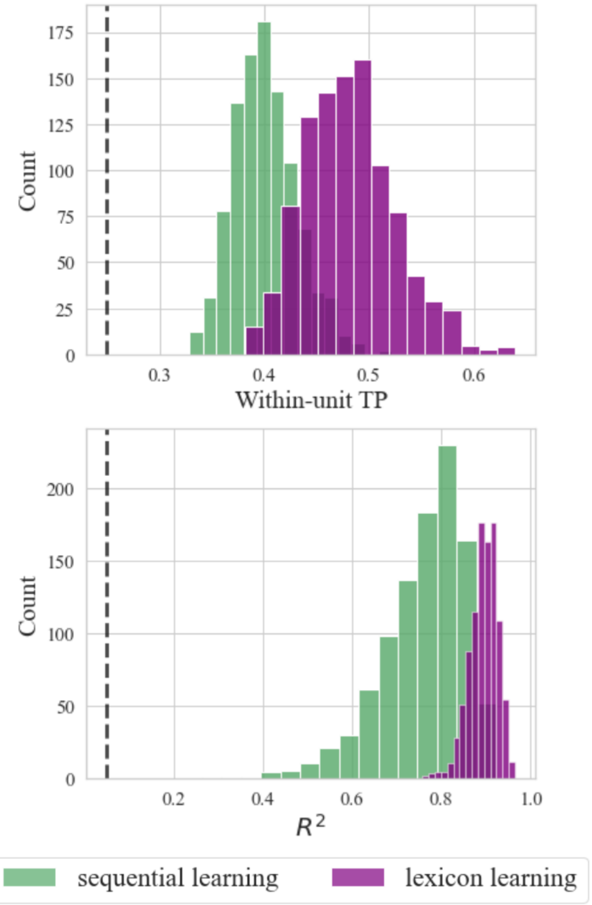

Figure 3: Distribution of mean within-unit transitional probability (top) and R^2 values (bottom), of the 1000 simulated chains after 10 generations. The mean value of generation zero (which is randomly generated) is indicated with the black dashed line.

Discussion

Recent work has proposed that certain characteristic statistical properties of language can emerge through cultural transmission, due to their facilitative effect on segmentation—the process of breaking continuous speech into smaller units, such as words. This was supported by an iterated learning experiment, showing that sets of unsegmented sequences evolve to have statistically coherent units that follow a Zipfian frequency distribution, under a pressure for learnability (Arnon & Kirby, 2024). However, it was not clear whether these properties emerged because learners segment the sequences into smaller units.

The simulations in Study 1 show that cultural transmission by learners who segment sequences into units drives the emergence of statistically coherent units that exhibit a Zipfian frequency distribution. These findings support the hypothesis put forth by Arnon and Kirby (2024), that segmentation introduces a pressure for the emergence of characteristic statistical properties of language. In contrast, simulations with learners who reproduced the unigram distribution of colours did not result in the emergence of these properties.

The unigram model closely parallels the process of drift. This is a kind of unbiased model of transmission where the frequencies of variants evolve through random copying, showing what change can be expected in the absence of selective biases. Using this approach, computational models of the transmission of word frequency distributions have shown that drift can drive the emergence of several cross-linguistic phenomena, including Zipfian frequency distributions (Bentley and Shennan, 2003; Hahn & Bentley, 2003; Reali & Griffiths, 2010). Our findings show that drift cannot explain the findings by Arnon and Kirby (2024), and that instead, a more sophisticated representation of the input is required for the emergence of statistically coherent units with a Zipfian frequency distribution in the cultural transmission of sequential information.

Interestingly, the simulations in Study 2 show that the two statistical properties can also emerge without segmentation. Cultural transmission by sequential learners, who tracked the conditional probabilities of individual colors without explicitly segmenting the sequences, also resulted in the emergence of statistically coherent units with a Zipfian frequency distribution. This suggest that the presence of these statistical properties in a culturally transmitted behaviour does not imply segmentation by its learners, and that instead there are multiple processes by which these statistical properties can emerge: by using the sequential statistics of the input to explicitly identify and reproduce a set of units (lexicon learners), and by reproducing the sequential frequency information of the input, which incidentally results in statistically coherent units and a Zipfian frequency distribution (sequential learners).

Although segmentation is not the sole mechanism that can drive the emergence of these statistical properties, existing work suggests that segmentation is often employed in sequence learning tasks like the one in Arnon and Kirby (2024): learners have a tendency to segment sequences into smaller units and segmentation has been shown to facilitate sequence encoding and recall (e.g. Bo & Seidler, 2009; Sakai et al., 2003; Verwey & Eikelboom, 2003).

One limitation of this study is that differently from the participants in Arnon and Kirby (2024), the simulated agents did not reproduce individual sequences but instead produced a new sequence set by reproducing the frequency information of the input. As a result, our simulations do not provide a direct comparison the experimental results. To address this, future research should explore models that integrate both distributional and sequence learning processes, and test to what extent sequential- and lexicon learning fits the sequence reproduction behaviour in Arnon and Kirby (2024).

To conclude, the findings of this study provide further insights into how cultural transmission can shape the statistical properties of language. Successful language learning requires the segmentation of linguistic input into meaningful units. In other words, language learners are inherently lexicon learners. Our results suggest that this introduces a pressure for the emergence of two characteristic statistical properties of language, as proposed by Arnon and Kirby (2024). Importantly, we also show that drift of unigram frequencies does not drive the emergence of these properties but that there is more than one pathway by which these properties can emerge. Further research is needed to fully understand the role of segmentation in shaping the statistical properties of language, as the simulations in our study necessarily simplify multiple aspects of speech segmentation. Notably, the sequence sets—and consequently, the segmented units—lacked semantic content. The referential nature of language is likely to influence how and to what extent learners segment their input, which in turn may impact the evolution of statistical properties.

References

Arnon, I., & Kirby, S. (2024). Cultural evolution creates the statistical structure of language. Scientific Reports, 14(1), 5255.

Bentley, R. A., & Shennan, S. J. (2003). Cultural Transmission and Stochastic Network Growth. American Antiquity, 68(3), 459–485.

Bo, J., & Seidler, R. D. (2009). Visuospatial Working Memory Capacity Predicts the Organization of Acquired Explicit Motor Sequences. *Journal of Neurophysiology*, *101*(6), 3116–3125.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. Behavioral and Brain Sciences, 31(5), 489–509. https://doi.org/10.1017/S0140525X08004998

Cornish, H., Dale, R., Kirby, S., & Christiansen, M. H. (2017). Sequence Memory Constraints Give Rise to Language-Like Structure through Iterated Learning. PLOS ONE, 12(1), e0168532.

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. Cognition, 122(3), 306–329.

Hahn, M. W., & Bentley, R. A. (2003). Drift as a mechanism for cultural change: An example from baby names. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(suppl_1), S120–S123.

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.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. Proceedings of the National Academy of Sciences, 104(12), 5241–5245.

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. (2022). The learnability consequences of Zipfian distributions in language. Cognition, 223.

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.

Reali, F., & Griffiths, T. L. (2010). Words as alleles: Connecting language evolution with Bayesian learners to models of genetic drift. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1680), 429–436.

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

Saffran., J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual review of psychology, 69*(1), 181-203.

Sakai, K., Kitaguchi, K., & Hikosaka, O. (2003). Chunking during human visuomotor sequence learning. *Experimental Brain Research*, *152*(2), 229–242.

Saksida, A., Langus, A., & Nespor, M. (2017). Co-occurrence statistics as a language-dependent cue for speech segmentation. Developmental Science, 20(3), e12390.

Shufaniya, A., & Arnon, I. (2021). A cognitive bias for Zipfian distributions? Uniform distributions become more skewed via cultural transmission. Proceedings of the Annual Meeting of the 43th Cognitive Science Society.

Smith, K., Perfors, A., Fehér, O., Samara, A., Swoboda, K., & Wonnacott, E. (2017). Language learning, language use and the evolution of linguistic variation. Philosophical Transactions of the Royal Society B: Biological Sciences, 372(1711), 20160051.

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.

Verhoef, T., Kirby, S., & de Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. Journal of Phonetics, 43, 57–68.

Verwey, W. B., & Eikelboom, T. (2003). Evidence for lasting sequence segmentation in the discrete sequence-production task. *Journal of Motor Behavior*, *35*(2), 171–182.

Zipf, G. K. (1949). Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology. Addison-Wesley Press.