Why are languages skewed? A Bayesian account for how skew and type count, but not entropy, facilitate rule generalisation

Elizabeth Pankratz[*1], Simon Kirby[1], and Jennifer Culbertson[1]

[*]Corresponding Author: e.c.pankratz@sms.ed.ac.uk
[1]Centre for Language Evolution, University of Edinburgh, Edinburgh, UK

One reason that languages come to have the properties they do is because those properties make a language easier to learn, its rules easier to generalise (Kirby, Griffiths, & Smith, 2014). Here, we ask which statistical properties of a language help people to decide whether a linguistic rule can be extended to new instances.

Previous research has generated seemingly conflicting hypotheses regarding the role of *Shannon entropy*. Segmentation and generalisation (closely related processes; Frost & Monaghan, 2016) are facilitated when items the rule applies to follow a skewed frequency distribution (e.g., Kurumada, Meylan, & Frank, 2013; Lavi-Rotbain & Arnon, 2019b, 2019a, 2021, 2022; Casenhiser & Goldberg, 2005; Goldberg, Casenhiser, & Sethuraman, 2004). Since skewed distributions have lower entropy than uniform distributions over the same number of items, this finding has been explained as a facilitatory effect of low entropy (e.g., Lavi-Rotbain & Arnon, 2022). At the same time, rules are more readily generalised when they apply to a large number of distinct types (e.g., Gómez, 2002; Tamminen, Davis, & Rastle, 2015; Valian & Coulson, 1988; Radulescu, Wijnen, & Avrutin, 2020). This result has also been explained in terms of entropy—but now, since a distribution over more types has higher entropy than a distribution over fewer, the prediction is that *high* entropy prompts generalisation. How do these seemingly contradictory findings fit together?

In this preregistered artificial language learning experiment (`osf.io/5keh9`), we disentangle how skew, type count, and entropy influence generalisation. Participants learned two different plural suffixes, each occurring with stems that followed one of two frequency distributions (Figure 1A). Then at test, they were asked to choose which of the two suffixes to use with novel stems. For half of the participants, the distributions that were contrasted were a uniform distribution over four types (Unif4) vs. a skewed distribution over four types (Skew4). For the other half, the distributions were, again, Unif4 vs. a uniform distribution over twice as many types (Unif8). Including Unif4 in both groups gives a baseline from which we can evaluate the individual effects of skew (in Group 1) and type count (in Group 2).

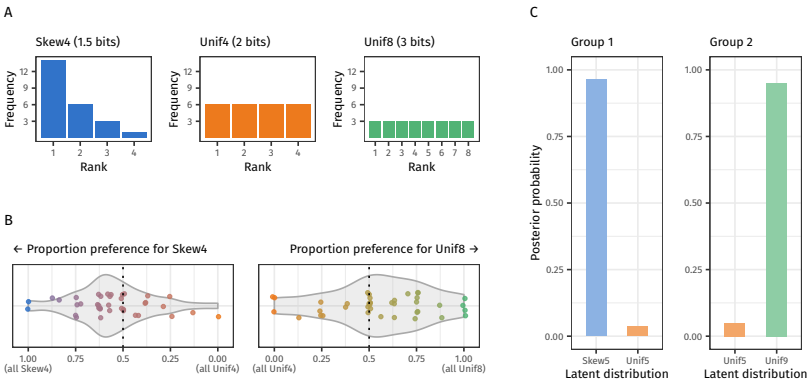

Figure 1. (A) The frequency distributions compared (Group 1 saw Skew4 and Unif4; Group 2 saw Unif4 and Unif8). (B) The suffixes that perfect learners generalised with tend to be the ones that appear with more variable stems (Skew4 in Group 1; Unif8 in Group 2). (C) This finding aligns with the posterior probabilities of missing a type when sampling from posited latent distributions.

We analyse data for participants who perfectly learned the language (N = 77 of 100, split 38–39 between groups). We found that Group 1 preferred to generalise with Skew4, while Group 2 preferred Unif8 (Figure 1B). A Bayesian linear model estimated the probability of generalising with the non-baseline suffix as 56.3% (95% CrI: [49.8%, 62.8%]). The same model estimated no difference between the two participant groups (β = 0 log-odds, 95% CrI [–0.49, 0.51]); skew and a greater type count provide comparable evidence that the suffix can be generalised.

These results are consistent with the empirical findings of both sets of studies summarised above, and thus *not* with an explanation based on entropy. We propose instead that the explanation follows from participants reasoning in a probabilistic, Bayesian way. In particular, participants in our task must essentially guess which suffix is more likely to appear with additional types beyond the ones they've already seen. If Unif4 were only a sample from some larger latent distribution, say Unif5, then observing only Unif4 and missing that fifth type is relatively unlikely. But missing a type when sampling from a latent Skew5 or Unif9 would be much more likely because of the greater number of low-frequency types overall.

Figure 1C shows each group's posterior probabilities of failing to encounter one or more types when sampling from each latent distribution. These probabilities heavily favour the distribution that learners in our experiment preferred. The greater generalisability of rules with these features could be part of the explanation for why languages come to have properties such as skew; ultimately, probabilistic reasoning of the kind we observe may shape the statistical structure of language.

Acknowledgements

EP gratefully acknowledges funding from the Economic and Social Research Council (award number ES/P000681/1) and the Social Sciences and Humanities Research Council of Canada (award number 752-2021-0366).

References

Casenhiser, D. M., & Goldberg, A. E. (2005). Fast mapping of a phrasal form and meaning. *Developmental Science*, *8*(6), 500–508.

Frost, R. L., & Monaghan, P. (2016). Simultaneous segmentation and generalisation of non-adjacent dependencies from continuous speech. *Cognition*, *147*, 70–74.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics*, *15*(3).

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 6.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114.

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. (2019a). Children learn words better in low entropy. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 631–637).

Lavi-Rotbain, O., & Arnon, I. (2019b). Low entropy facilitates word segmentation in adult learners. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 2092–2097).

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*, 105038.

Radulescu, S., Wijnen, F., & Avrutin, S. (2020). Patterns bit by bit: An entropy model for rule induction. *Language Learning and Development*, *16*(2), 109–140.

Tamminen, J., Davis, M. H., & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive Psychology*, *79*, 1–39.

Valian, V., & Coulson, S. (1988). Anchor points in language learning: The role of marker frequency. *Journal of Memory and Language*, *27*(1), 71–86.