[Proceedings of the Society for Computation in Linguistics](https://scholarworks.umass.edu/scil)

Volume 3 Article 52

2020

Frequency-(in)dependent Regularization in Language Production and Cultural Transmission

Emily Morgan
University of California, Davis, eimorgan@ucdavis.edu

Roger P. Levy
Massachusetts Institute of Technology, rplevy@mit.edu

Frequency-(in)dependent regularization in language production and cultural transmission

Emily Morgan (UC Davis; eimorgan@ucdavis.edu) and Roger Levy (MIT; rplevy@mit.edu)

In cases of variation in language, how do people learn and reproduce probabilistic distributions over linguistic forms? Given a certain amount of variation in their linguistic input, speakers could aim to reproduce the variation exactly (i.e. to *probability match*) or could instead aim to *regularize*—to make their productions more consistent by reproducing the most frequent variant *even more frequently* than it was heard in the input. While we know that people retain detailed statistics about their linguistic input (Levy, 2008; Arnon & Snider, 2010), there is also evidence for regularization in language learning (Hudson Kam & Newport, 2005; Reali & Griffiths, 2009), although the circumstances that lead to regularization versus probability matching are not yet well understood. Morgan and Levy (2015) found evidence in corpus data that *binomial expressions* of the form "X and Y" are more regularized the higher their frequency—i.e. their ordering preferences (e.g. "bread and butter" vs. "butter and bread") are more extreme when the two words ("bread" and "butter") co-occur in a binomial more frequently, regardless of order. This finding is puzzling because previous experimental research does not suggest that regularization should be frequency-dependent. However, when we find systematic patterns in corpus data, we would like to be able to attribute them to motivated preferences (based on language learning and/or production; Hawkins, 2004). Does this corpus data in fact provide evidence for regularization in online language processing, and if so, does speakers' regularization behavior depend on an item's frequency, contrary to previous claims?

We demonstrate that frequency-dependent regularization can arise diachronically through a combination of a frequency-independent synchronic regularization bias and the bottleneck effect of cultural transmission. We simulate diachronic language change using an Iterated Learning Model (Smith, 2009) in which speakers in successive generations iteratively learn binomial expression preferences from the previous generations' productions and then generate their own productions. We augment the standard model with a regularization bias that applies during production. Although the bias itself is frequency-*independent*, we demonstrate that frequency-*dependent* regularization emerges from the iterated learning process. For lower frequency items, a tighter bottleneck (fewer productions per generation) favors convergence to the prior. Because prior preferences depend only on the words in the binomial—not on its frequency—the bottleneck thus prevents the regularization bias from having a strong effect. With increasing frequency, a wider bottleneck (more productions per generation) increasingly transmits the effects of the regularization bias across generations. Our model thus correctly predicts the qualitative pattern of frequency-dependent regularization (Fig 1).

Moreover, our model correctly predicts the observed language-wide distribution of ordering preferences in Morgan and Levy's (2015) binomial corpus. For each binomial expression in the corpus, we predict its ordering preference based on its frequency of occurrence (as well as other word-level properties). Our model correctly predicts the multimodal distribution of ordering preferences found in the corpus (Fig 2).

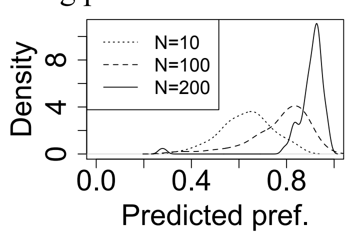

Fig 1. Model-predicted preferences for a hypothetical binomial, from 0 (always one order) to 1 (always the other), are more extreme (closer to 0 and/or 1) with increasing number of productions per generation N.

Proceedings of the Society for Computation in Linguistics (SCiL) 2020, pages 466-467.
New Orleans, Louisiana, January 2-5, 2020

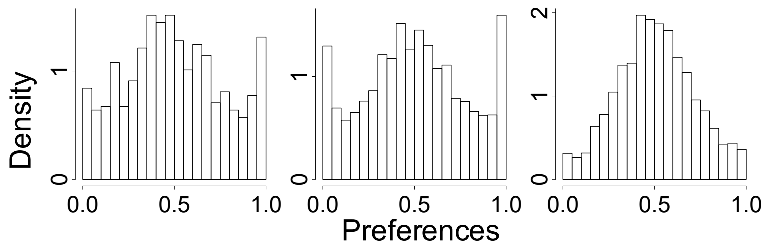

Fig 2. Language-wide distribution of preferences in corpus data (left) and as predicted by a model with (center) and without (right) a regularization bias.

Our model thus confirms previous demonstrations of a regularization bias in language learning and/or production, but demonstrates that frequency-dependent regularization in a corpus distribution does not imply that frequency influences regularization at the level of individual speakers. Rather, the pattern of frequency-dependent regularization seen in corpus data can arise from the interaction of a frequency-independent bias in online language processing and the bottleneck effect of cultural transmission.

We conclude by questioning why language learning and/or production might include a regularization bias. One hypothesis is that an online regularization bias promotes efficiency in language processing by reducing the choices that must be made, hence reducing the cost of online utterance planning. Another hypothesis relates to difficulty during early learning rather than during online production. Focusing on one variant during learning may reduce cognitive load, and therefore regularization may be both particularly prevalent and particularly beneficial during early language learning when cognitive resources are more limited than in adulthood.

This work was previously published in: Morgan, E., & Levy, R. (2016). Frequency-Dependent Regularization in Iterated Learning. In S. G. Roberts, C. Cuskley, L. McCrohon, L. Barceló-Coblijn, O. Fehér, & T. Verhoef (Eds.), *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG 11)*. http://evolang.org/neworleans/papers/193.html

References

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

Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford University Press.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change. *Language Learning and Development*, *1*(2), 151–195.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Morgan, E., & Levy, R. (2015). Modeling idiosyncratic preferences: How generative knowledge and expression frequency jointly determine language structure. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *37th Annual Meeting of the Cognitive Science Society* (pp. 1649–1654). Austin, TX: Cognitive Science Society.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, 111(3), 317–328.

Smith, K. (2009). Iterated learning in populations of Bayesian agents. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 697–702.