ITERATED LEARNING PREDICTS FREQUENCY-DEPENDENT REGULARIZATION OF WORD ORDER PREFERENCES

ANONYMOUS AUTHOR 1
University Department, University Name
City, Country
email@university

Binomial expression ordering preferences (e.g. "bread and butter" versus "butter and bread") are more regularized at higher frequencies (Morgan & Levy, 2015). Although regularization of stochastic linguistic items has been demonstrated previously, this tendency is not known to be frequency-dependent. By augmenting a 2-Alternative Iterated Learning Model with a frequency-*independent* regularization bias, we demonstrate that frequency-*dependence* emerges through a combination of this regularization bias and the bottleneck of cultural transmission. Our model predicts the distribution of binomial expression preferences seen in corpus data.

1. Introduction

Languages are shaped both by the cognitive architectures of individual speakers and by the process of cultural transmission that acts across generations. In this paper we ask how these two factors jointly contribute to a key dichotomy in language structure: the trade-off between broadly-applicable compositional knowledge and knowledge of item-specific idiosyncrasies. Specifically, we take up the case of frequency-dependent regularization in word order preferences for binomial expressions (e.g. "bread and butter" versus "butter and bread"). Although regularization is a well-attested phenomenon in statistical learning, *frequency-dependent* regularization is not. We demonstrate that the frequency-dependence seen in corpus data can arise as an emergent property of a frequency-*independent* regularization bias combined with the bottleneck effect of cultural transmission.

Item-specific idiosyncrasies (i.e exceptions to the rules) are well known to be frequency-dependent. For example, more frequent verbs are more likely to have irregular conjugations (Lieberman, Michel, Jackson, Tang, & Nowak, 2007). More recently, Morgan and Levy (2015) have demonstrated a different type of frequency-dependent idiosyncrasy at the level of multi-word phrases, specifically *binomial expressions* of the form "X and Y" (Cooper & Ross, 1975; Benor & Levy, 2006). Word order preferences for these expressions are gradient; for example, "radio and television" is preferred to "television and radio" in a 63/37 ratio, while "bread and butter" is preferred to "butter and bread" in a 99/1 ratio

(Lin et al., 2012). These ordering preferences are partially determined by violable constraints operating during compositional generation of the expression, e.g. a constraint to put shorter words before longer words. But these expressions are also subject to learned item-specific idiosyncrasies, e.g. despite a generally strong constraint to put men before women, "ladies and gentlemen" is preferred over "gentlemen and ladies". In addition to the possibility of the complete reversal of compositionally determined preferences, item-specific idiosyncrasies can also be gradient, e.g. a binomial whose compositional preference predicts a 60/40 distribution might instead be used in a 90/10 ratio. As in the case of irregular verbs, the distribution of idiosyncrasies in binomial ordering preference is frequency-dependent: more frequent binomial expressions deviate more from compositional preferences. Moreover, this deviation takes on a particular form, with more frequent expressions becoming more *regularized*, i.e. preferred more strongly in a given order. (We note that this is different from the notion of "regular" items as those that conform to compositional rules.)

Regularization is a well-established phenomenon in statistical learning. In a variety of tasks, both linguistic and non-linguistic, in which participants learn and reproduce probabilisty distributions over alternates, both children and adults tend to regularize their productions (Hudson Kam & Newport, 2005; Reali & Griffiths, 2009; Ferdinand, Kirby, & Smith, 2014). For example, Reali and Griffiths (2009) found that when exposed to two labels for a novel object, subjects reproduced the more frequent label *even more frequently* than that label was seen in training. Although this tendency was weak, they demonstrated that even such a small bias towards regularization can have significant long-term impacts, as the bias acts across successive generations to shape language over time. Bickerton (1981), Hudson Kam and Newport (2005), and others have argued that childrens' tendency to regularize is an important mechanism of language change, e.g. for forming more consistent languages out of pidgins.

However, to the best of our knowledge, there has been no previous demonstration of regularization being dependent upon the frequency of an item (not to be confused with the frequency of the forms used to label the item)—e.g. in Reali and Griffith's experiments, regularization was not dependent upon the frequency of occurrence of the novel object—nor is it immediately predicted by any existing theories that it should be. Thus Morgan and Levy's finding of frequency-dependent regularization of binomial ordering preferences is unexpected. Here, we will demonstrate that a frequency-*independent* regularization bias in learners, combined with properties of cultural transmission, can produce the pattern of frequency-*dependent* regularization of binomial ordering preferences seen in corpus data. We demonstrate this using iterated learning models, which allow us to explore how individual cognitive biases acting at each generation, combined with the bottleneck effect of transmission across generations, can yield the patterns we see in natural language data.

2. Morgan and Levy's binomial expression corpus

In this work, we will take advantage of a uniquely appropriate real-world data set: Morgan and Levy (2015) collected a corpus of 594 binomial expression types. The corpus was hand-annotated for the semantic, phonological, and lexical constraints that affect binomial ordering preferences, and occurrence frequencies for the binomials' two possible orders were gathered from the Google Books corpus (Lin et al., 2012). Moreover, Morgan and Levy developed a model that combines these constraints using logistic regression to provide an estimate of people's combinatorial knowledge of ordering preferences for binomial expressions, independent of their actual experience with an item in a given order. We thus have knowledge of three independent measures with respect to these expressions:

- The *overall (unordered) frequency* of an expression: freq("X and Y")+freq("Y and X")
- The *relative frequency* of occurrence of a given order, expressed as a number between 0 and 1: freq("X and Y")/(freq("X and Y")+freq("Y and X"))
- The *compositional preference* for occurrence in a given order, expressed as a number between 0 and 1, given by Morgan and Levy's model.

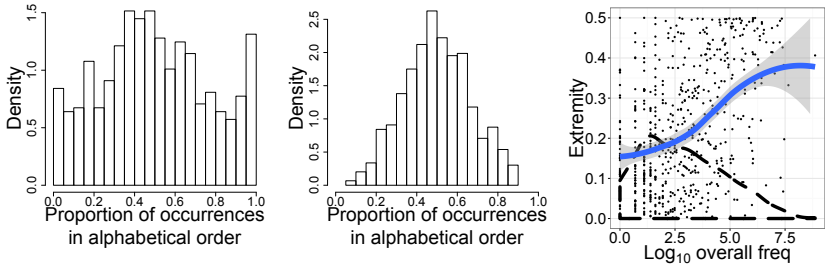
Morgan and Levy demonstrate that relative frequencies in corpus data are multimodally distributed, with modes at the extremes as well as around 0.5 (Fig. 1a). Crucially, this pattern is not predicted by compositional preferences, which predict only a single mode (Fig. 1b). It is the expressions with higher overall frequency that diverge most from compositional preferences to become more regularized (Fig. 1c).

We ask what combination of individual cognitive biases and properties of cultural transmission predict the patterns seen in Figs. 1a and 1c. We will see that a standard 2-Alternative Iterated Learning Model does not correctly predict these patterns, but an augmented model with an explicit regularization bias does.

3. Standard 2-Alternative Iterated Learning Model

Following Reali and Griffiths (2009), we use iterated learning models to simulate the evolution of binomial expressions over generations of speakers. A learner hears N tokens of a binomial expression, with x_1 of them in a given order—we will use alphabetical order as neutral reference order—and then infers a hypothesis $\theta_1 \in [0, 1]$ which is the proportion of time a binomial should be produced in alphabetical order. The learner then generates new data using θ_1.

The prior probability $P(\theta_1)$ of a binomial being preferred in a given order can be expressed using the beta distribution. We can treat the compositional preference as a form of prior knowledge of ordering preferences for a binomial. To incorporate this prior knowledge, we use a parameterization of the beta distribution with a parameter μ that determines the mean of draws and a concentration parameter ν that determines how tightly clustered around the mean those draws



(a) (b) (c)

Figure 1. Results from Morgan and Levy (2015). (a) Histogram of binomial types' relative frequencies. (b) Histogram of binomial types' compositional preferences. (c) We define an expression's *extremity* as the absolute difference between its relative frequency and 0.5. More frequent expressions have more extreme/regularized preferences (solid blue line). Density of overall frequency counts (scaled as described in Section 4.2) shown by dotted line. The distribution is non-Zipfian because the corpus is restricted to binomial types with at least 1000 occurrences in the Google Books corpus.

are. (ν can also be thought of as reflecting how confident in the prior we are, e.g. $\nu = 10$ would indicate confidence equivalent to having seen ten instances of a given binomial expression type before.) Under this parameterization:

$$P(\theta_1) = \frac{\theta_1^{\mu\nu-1}(1-\theta_1)^{(1-\mu)\nu-1}}{\mathrm{B}(\mu\nu,(1-\mu)\nu)} \tag{1}$$

where B is the beta function. Because μ represents compositional ordering preferences, it varies for each binomial, and is set according to Morgan and Levy's model. All learners are assumed to have the same μ value for a given binomial. ν is constant for all binomial expressions for all learners, and is a free parameter. Given θ_1, data is generated binomially:

$$P(x_1|\theta_1) = \binom{N}{x_1}\theta_1^{x_1}(1-\theta_1)^{N-x_1} \tag{2}$$

We define a chain of learners under this model by initializing a single learner with some hypothesis. (For all the simulations in this paper, we will initialize our first generation with $\theta_1 = 0.5$.) This first generation produces N utterances according to the distribution defined in Eq. 2. The learner in the next generation applies Bayes rule and choose a hypothesis from the resulting posterior distribution over hypotheses. (For the sake of space, all results presented in this paper will choose the MAP estimate, but all results are qualitatively the same if posterior sampling is used instead.) This process continues iteratively.

Reali and Griffiths (2009) have demonstrated that regularization occurs in iterated learning models with sparse priors (i.e. those that favor hypothesis close to

0 and 1); given our parameterization of the beta distribution, these are hypothesis with $\nu < 2$. However, this regularization is not dependent on the expression's overall frequency. We demonstrate this by modeling chains of learners with different values of N. We model a single binomial expression with prior probability $\mu = 0.6$. We explore different values of ν, specifically $\nu = 1$ (a sparse prior) and $\nu = 10$ (a dense prior). We explore values of $N = 10, 100, 200, 500$. For each combination of ν and N, we approximate the distribution over expression preferences by running 500 chains of learners for 500 generations each and taking the hypothesis of the final generation in each chain. As seen in Fig. 2, the distribution of preferences in the resulting distribution is sensitive to ν but not to N. In other words, the number of times a binomial expressions is seen in each generation does not affect how regularized it becomes.

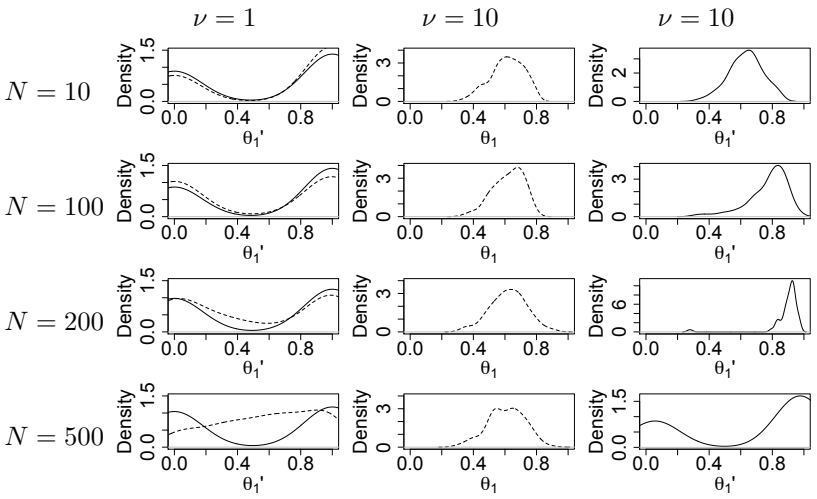


Figure 2. Distribution of binomial ordering preferences in a standard 2-Alternative Iterated Learning Model (dotted lines) and one with an explicit regularization bias (solid lines)

4. Modeling frequency-dependent regularization

The standard 2-Alternative Iterated Learning Model does not predict frequency-dependent regularization. We will now demonstrate that we can predict frequency-dependent regularization by introducing a frequency-*independent* regularization bias into our model. Under this model, frequency-dependent regularization is an emergent property of the interaction of the frequency-independent regularization bias with the bottleneck effect of cultural transmission.

We augment the learning process as follows. After hearing data, the learner

chooses a hypothesis θ_1 as before, then applies a regularization function to produce a new hypothesis θ_1', then generates data from θ_1' rather than θ_1.

The regularization function is the regularized incomplete beta function (equivalently, the cumulative distribution function of the beta distribution), restricted to be symmetric such that it has a single free parameter α:

$$f(x;\alpha) = \frac{\int_0^x t^{\alpha-1}(1-t)^{\alpha-1}\,\mathrm{d}t}{\mathrm{B}(\alpha,\alpha)} \tag{3}$$

As shown in Fig. 3, the bias parameter α controls how strong the tendency to regularize is. When $\alpha = 1$, this is the identity function, i.e. no explicit regularization is added. As α increases, the regularization bias grows stronger.

4.1. *Results: Frequency-dependent regularization*

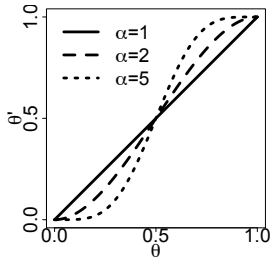


Figure 3. Regularization function with different values of α

When we repeat the simulations from above using a non-trivial regularization bias $\alpha = 1.1$, we see frequency-dependent regularization in the case with a dense prior (Fig. 2). Although the regularization bias itself is frequency-independent, frequency-dependence emerges from the interaction of the regularization bias with the process of cultural transmission: At lower frequencies, there is not sufficient data for the regularization bias to overcome the prior. At higher frequencies, the regularization bias becomes increasingly dominant as there is increasingly enough data for the effects of this bias to be carried across generations. However, the prior always continues to exert some influence; thus, even the highest frequency expressions do not become completely regularized.

Another linguistically accurate property of this model is that for sufficiently high values of N, the distribution over hypotheses includes a mode on the opposite side of 0.5 from the prior. Thus the model correctly predicts that at high enough frequencies, an expression can become idiosyncratically preferred in the opposite of its compositionally predicted direction (as in "ladies and gentlemen").

4.2. *Results: Simulating corpus data*

Having demonstrated that our augmented model produces frequency-dependent regularization, we will now show that it additionally predicts the true language-wide distribution of binomial preference strengths seen in corpus data. Specifically, we simulate the distribution of relative frequencies seen in Fig. 1a.

We take the true properties of each binomial expression in the corpus: its compositional preference determines μ and its overall frequency determines N. We scale overall frequency counts based on a lifetime exposure to 300 million total words (based on an estimate from Levy, Fedorenko, Breen, and Gibson (2012)). The resulting distribution of values N is shown in Fig. 1c. For each binomial in the corpus, we model 10 chains of learners for 200 generations each and take the hypothesis θ_1' of the final generation of each chain.

Our model has two free paramters, ν and α. We model the corpus data as described above for a range of values of both of these parameters. As shown in Fig. 4, our model displays a trade-off between the prior and the regularization bias as a function of these parameters. At appropriate values, our model correctly predicts the multimodal distribution of corpus data as seen in Fig. 1a.

5. Discussion

We have demonstrated that a frequency-independent regularization bias combined with cultural transmission can produce the pattern of frequency-dependent regularization of binomial ordering preferences seen in corpus data. This finding sheds light on the origins of linguistic structure in two important ways: one, it confirms earlier demonstrations of a bias to regularize when learning stochastic linguistic items. Second, it shows that this bias can apply equally across all levels of frequency, but that combined with properties of cultural transmission it will produce frequency-dependent effects; thus, the distribution of idiosyncracy seen in the language emerges from the interaction of individuals' cognitive biases with the bottleneck effect of cultural transmission. This work additionally expands upon the existing iterated learning paradigm by demonstrating that it can be used to successfully model not just general patterns of language use but actual corpus data, by choosing test cases that are at once theoretically interesting and computationally tractable.

Acknowledgements

We gratefully acknowledge support from research grants NSF 0953870 and NICHD R01HD065829 and fellowships from the Alfred P. Sloan Foundation and the Center for Advanced Study in the Behavioral Sciences to Roger Levy.

References

Benor, S., & Levy, R. (2006). The Chicken or the Egg? A Probabilistic Analysis of English Binomials. *Language*, *82*(2), 233–278.

Bickerton, D. (1981). *Roots of language.* Ann Arbor, MI: Karoma.

Cooper, W. E., & Ross, J. R. (1975). World Order. In R. E. Grossman, L. J. San, & T. J. Vance (Eds.), *Papers from the parasession on functionalism* (pp. 63–111). Chicago: Chicago Linguistics Society.

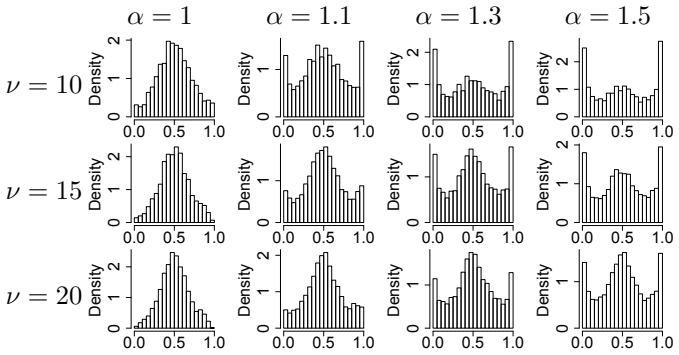


Figure 4. Predicted distribution of θ_1'. We see a trade-off between effects of the prior and the regularization bias. When the prior is stronger (high ν, low α), we see a unimodal distribution of preferences, similar to Fig. 1b. When the regularization bias is stronger (low ν, high α), we see too much regularization. At appropriate values of α and ν, we see the correct multimodal distribution of preferences as seen in corpus data (Fig. 1a).

Ferdinand, V., Kirby, S., & Smith, K. (2014). Regularization in language evolution. *Proceedings of the 10th International Evolution of Language Conference*.

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., Fedorenko, E., Breen, M., & Gibson, E. (2012). The processing of extraposed structures in English. *Cognition*, *122*(1), 12–36.

Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*(7163), 713–716.

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic Annotations for the Google Books Ngram Corpus. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 169–174.

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.