# Frequency-dependent regularization arises from a noisy-channel processing system

Zachary Houghton[1] & Emily Morgan[1]

[1] University of California, Davis

Language often has different ways to express the same or similar meanings. Despite this, however, people seem to have preferences for some ways over others. For example, people overwhelmingly prefer *bread and butter* to *butter and bread*. Previous research has demonstrated that these ordering preferences grow stronger with frequency (i.e., frequency-dependent regularization). In this paper we demonstrate that this frequency-dependent regularization can be accounted for by Gibson, Bergen, and Piantadosi (2013)'s noisy-channel processing model. We also show that this regularization can only be accounted for if the listener infers more noise than the speaker produces. Finally, we show that the model can account for the language distribution of binomial ordering preferences.

*Keywords:* Frequency-dependent regularization, Noisy-channel processing, Psycholinguistics

## Introduction

Speakers often have great flexibility in their choices to convey a meaning. For example, speakers are often confronted with many different ways to express the same meaning. A customer might ask whether a store sells "radios and televisions", but they could have just as naturally asked whether the store sells "televisions and radios." However, despite conveying the same meaning, speakers sometimes have strong preferences for one choice over competing choices (e.g., preference for *men and women* over *women and men*, Benor & Levy, 2006; Morgan & Levy, 2016a). These preferences are driven to some extent by generative preferences (e.g., preference for short words before long words), however they are sometimes violated by idiosyncratic preferences (e.g., *ladies and gentlemen* preferred despite a general men-before-women generative preference, Morgan & Levy, 2016b).

Interestingly, ordering preferences are often more extreme for higher frequency items (e.g., *bread and butter*). That is, higher-frequency items typically have more polarized preferences (Liu & Morgan, 2020, 2021; Morgan & Levy, 2015, 2016b, 2016a). However, it is still unclear what processeses this phenomenon, called frequency-dependent regularization, is driven by. In the present paper we examine whether a noisy-channel processing model (Gibson et al., 2013) combined with transmission across generations (Reali & Griffiths, 2009) can account for frequency-dependent regularization.

## Frequency-dependent regularization

In the last decade, frequency-dependent regularization – the phenomena of more frequent items having more extreme ordering preferences – has been documented for a variety of different constructions in English (Liu & Morgan, 2020, 2021; Morgan & Levy, 2015, 2016b). For example, Morgan and Levy (2015) demonstrated that more frequent binomial expressions (e.g., *bread and butter*) are more strongly regularized (i.e., are preferred in one order overwhelmingly more than the alternative). These ordering preferences are also not simply a result of abstract ordering preferences (e.g., short words before long words, Morgan & Levy, 2016a).

Additionally, Liu and Morgan (2020) demonstrated this effect holds true for the dative alternation in English (e.g., *give the ball to him* vs *give him the ball*). Specifically, they demonstrated higher frequency verbs have more polarized preferences with respect to the dative alternation. Similarly, Liu and Morgan (2021) showed that Adjective-Adjective-Noun orderings also show frequency-dependent regularization. That is, adjective-adjective-Nouns with higher head noun frequencies show stronger ordering preferences, even after taking into account generative preferences of adjective orderings.

How does this polarization for high-frequency items arise? One possibility is that it occurs as a consequence of imperfect transmission between generations. For example, Morgan and Levy (2016b) demonstrated that in an iterated-learning paradigm (Reali & Griffiths, 2009), this frequency-dependent regularization can arise from an interaction between a frequency-independent bias and transmission across generations. Specifically, they used an iterated learning paradigm (following Reali & Griffiths, 2009) and demonstrated that by introducing a frequency-independent regularization bias, after several generations the model predicted

Correspondence concerning this article should be addressed to Zachary Houghton, . E-mail: znhoughton@ucdavis.ed

frequency-*dependent* regularization. However, it is unclear what process in language is analogous to the frequency-independent bias.

### Noisy-channel Processing

One possibility is that frequency-dependent regularization arises as a product of noisy-channel processing (Gibson et al., 2013). That is, listeners are confronted with a great deal of noise in the form of perception errors (e.g., a noisy environment) and even production errors (speakers don't always say what they intended to, Gibson et al., 2013). In order to overcome these errors, a processing system must take into account the noise of the system.

Indeed, there is evidence that our processing system does take noise into account. For example, Ganong (1980) presented people with word-nonword continuums, where the VOT of the initial consonant was manipulated (e.g., *task*-*dask*). He found that when people are presented with a word in which the initial phoneme was close to the VOT boundary, people will process the segment as being the word, rather than the non-word, even when the VOT was incongruous with the interpretation. Further, Albert Felty, Buchwald, Gruenenfelder, and Pisoni (2013) demonstrated that when listeners do misperceive a word, the word that they believe to have heard tends to be higher frequency than the target word. This suggests that misperceptions may sometimes actually be a consequence of noisy-channel processing (rather than a failure of our perceptual system).

In order to account for findings like these, Gibson et al. (2013) developed a computational model that demonstrated how a system might take into account noise. Specifically, their model operationalizes noisy-channel processing as a Bayesian process where a listener estimates the probability that their perception matches the speaker's intended utterance. Specifically, this is operationlized as being proportional to the likelihood of the intended utterance multiplied by the probability of the intended utterance being corrupted to the perceived utterance (See Equation (1)):

$$P(S_i|s_p) \propto P(S_i)P(S_i \rightarrow S_p) \tag{1}$$

Gibson et al. (2013)'s model made a variety of interesting predictions. For example, the model predicted that when people are presented with an implausible sentence (e.g., *the mother gave the candle the daughter*), they should be more likely to interpret the plausible version of the sentence (e.g., *the mother gave the candle to the daughter*) if there is increased noise (e.g., by adding errors to the filler items). Their model also predicted that increasing the likelihood of implausible events should increase the rate of implausible interpretations of the sentence. Interestingly both of these results were born out in their experimental data, suggesting that humans do utilize a noisy-channel system in processing.

### Present Study

Given the evidence of noisy-channel processing, it is possible that the frequency-dependent regularization that Morgan and Levy (2016b) saw is a product of listeners' noisy-channel processing. That is, perhaps the regularization bias responsible for the regularization across generations is a consequence of noisy-channel processing. Thus, the present study examines whether Gibson et al. (2013)'s noisy-channel processing model can also predict frequency-dependent regularization across generations of language transmission.

## Dataset

Following Morgan and Levy (2016b), we use Morgan and Levy (2015)'s corpus of 594 binomial expressions. This corpus has been annotated for various phonological, semantic, and lexical constraints that are known to affect binomial ordering preferences. The corpus also includes estimated generative preferences for each binomial (i.e., compositional ordering preferences, estimated from the above constraints) and observed binomial orderings (the proportion of the binomials that occur in alphabetical form). The observed binomial orderings are the number of times the binomial occurred in alphabetical form divided by the total times the binomial occurred in both alphabetical and non-alphabetical form. A visualization of the observed preferences and compositional preferences is included below in Figure , on the left and right respectively.

## Model

Following Reali and Griffiths (2009), we use a 2-alternative iterated learning paradigm. A learner hears N tokens of a binomial expression and then produces N tokens for the next generation. After hearing N tokens, they infer the probability, $\theta$, of the alphabetical form of the binomial. Using that probability, they then produce N tokens for the next generation, and this process continues iteratively.

The prior probability of a binomial's ordering is operationalized using the beta distribution (following Morgan & Levy, 2016b). Specifically, we treat the generative preference for a given binomial as the prior for our model. This is operationalized using the beta distribution with $\mu_{prior}$, which is the generative preference for a given binomial, and $\nu$, which determines the strength of the belief of the prior probability. Thus the prior probability for a given binomial in its alphabetical form is estimated as:

$$P(\theta_{prior}) = \frac{\mu_{prior} \cdot \nu}{(\mu_{prior} \cdot \nu) + (1 - \mu_{prior}) \cdot \nu} \tag{2}$$

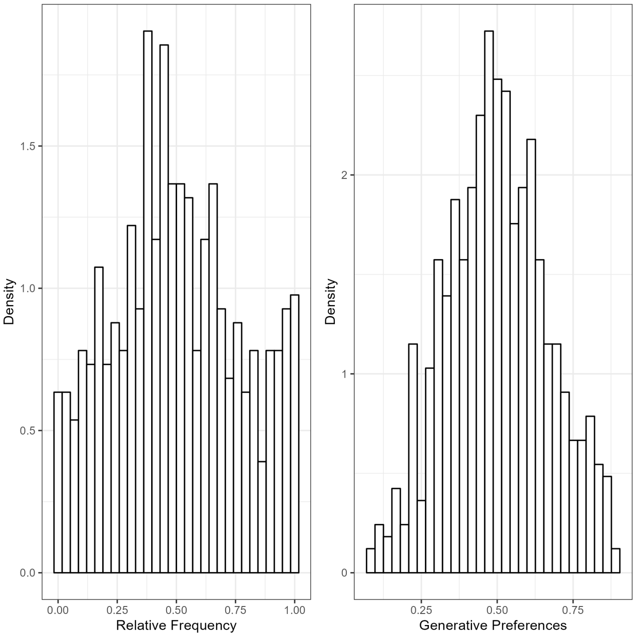

*Figure 1.* The left plot is a plot of the relative orderings of binomials in the corpus data from Morgan and Levy (2015), the right is the plot of the generative preferences of binomials in the same corpus. The x-axis is proportion of occurrences in alphabetical order and the y-axis is the probability density.

and the probability of hearing it in non-alphabetical form is $1 - P(\theta_{prior})$.

This is then used as the prior in Gibson et al. (2013)'s noisy-channel processing model, such that if a listener hears the alphabetical form of a binomial they update the probability of hearing it in alphabetical form according to the following parametrization,

$$\hat{p}(\alpha) \propto \frac{\mu_{prior} \cdot \nu}{(\mu_{prior} \cdot \nu) + (1 - \mu_{prior}) \cdot \nu} \cdot p_{noise} \tag{3}$$

where $1 - p_{noise}$ is a fixed parameter for every binomial. They also update the probability of hearing the non-alphabetical form:

$$\hat{p}(\neg\alpha) \propto 1 - \frac{\mu_{prior} \cdot \nu}{(\mu_{prior} \cdot \nu) + (1 - \mu_{prior}) \cdot \nu} \cdot (1 - p_{noise}) \tag{4}$$

If the listener hears the non-alphabetical form of the binomial, they update using the same equations, but instead the likelihood of hearing the alphabetical form is rate of noise, and the likelihood of hearing the non-alphabetical form is 1 minus the rate of noise (since hearing the non-alphabetical form is infact being faithful to the data in this context). Thus the estimated probability of the alphabetical form of the binomial is updated according to,

$$\hat{p}(\alpha) \propto \frac{\mu_{prior} \cdot \nu}{(\mu_{prior} \cdot \nu) + (1 - \mu_{prior}) \cdot \nu} \cdot p_{noise} \tag{5}$$

and the estimated probability of the non-alphabetical form is updated according to:

$$\hat{p}(\neg\alpha) \propto 1 - \frac{\mu_{prior} \cdot \nu}{(\mu_{prior} \cdot \nu) + (1 - \mu_{prior}) \cdot \nu} \cdot (1 - p_{noise}) \tag{6}$$

Finally, before $\hat{p}_{\alpha}$ and $\hat{p}_{\neg\alpha}$ are updated, the un-normalized estimates of the alphabetical and non-alphabetical above are normalized such that they sum to 1.[1]

Once the learner finishes hearing N tokens, they then produce N tokens generated binomially, where $\theta_1$ is their inferred probability of the alphabetical form of the given binomial (which is set to 0.5 for the first generation):

$$P(x_1|\theta_1) = \binom{N}{x_1} \theta^{x_1} (1 - \theta_1)^{N - x_1} \tag{7}$$

When the learner produces the N tokens, there is also a possibility the speaker will make an error. This is also generated binomially, with $\theta_1$ being a fixed parameter, which is the probability that the learner makes an error. In our model, if the learner makes an error, the opposite binomial form is produced. For example, if the learner intends to produce the alphabetical form and makes an error, the non-alphabetical form is produced.

## Results

### Speaker vs Listener Noise

First we demonstrate that frequency-dependent regularization does not arise when there is no listener or speaker noise.[2] Instead we see convergence to the prior, which is expected. That is, Griffiths and Kalish (2007) demonstrated that when learners sample from the posterior, as the number of iterations increases, the stationary distribution converges to the prior. In other words, without any noise, each generation of learners produces data that is more and more similar to the prior, until convergence is reached.

However, when we introduce noise (Figure 3, we see that the model can predict frequency-dependent regularization across generations.

[1] The full computational implementation can be found in the `iterated_learning.R` script in the github repository.

[2] All code and results can be found publicly available here: https://github.com/znhoughton/Noisy-Channel-Iterated-Learning

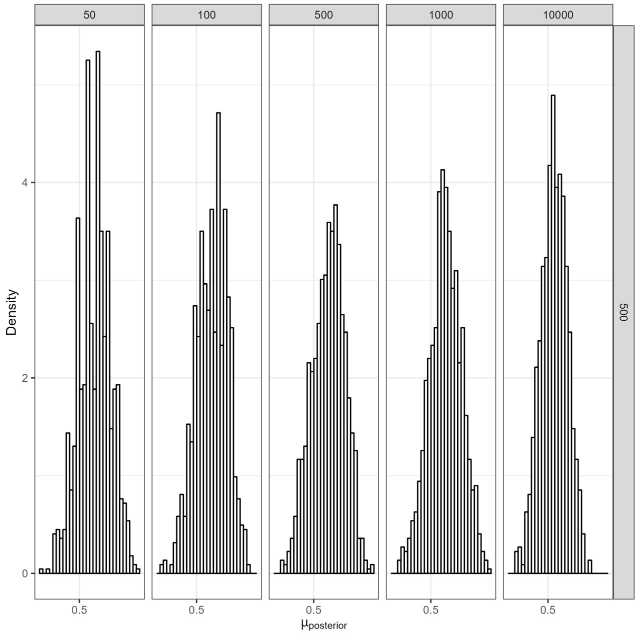

*Figure 2*. A plot of the distribution of simulated binomials at the 500th generation, varying in frequency. The top value represents N. On the x-axis is the predicted probability of producing the binomial in alphabetical form. On the y-axis is probability density. Speaker and listener noise was set to 0. The generative preference was 0.6, and nu was set to 10. 1000 chains were run. Note that there is no frequency-dependent regularization apparent.

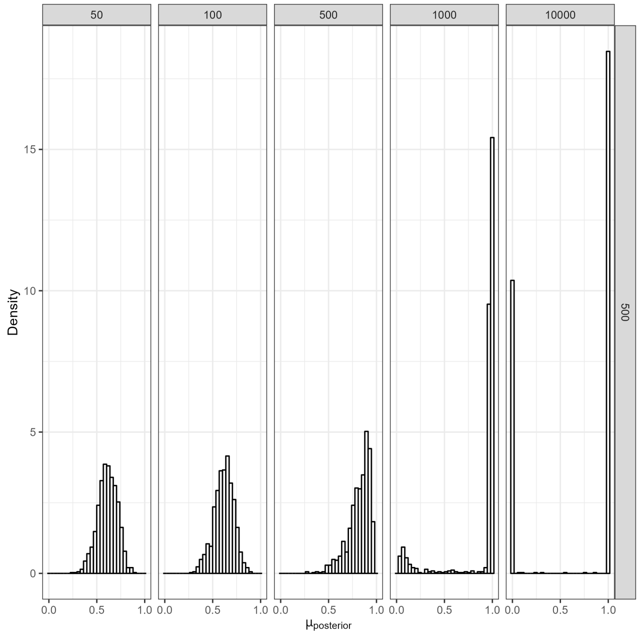

*Figure 3*. A plot of the distribution of simulated binomials at the 500th generation, varying in frequency. The top value represents N. On the x-axis is the predicted probability of producing the binomial in alphabetical form. On the y-axis is probability density. Speaker noise was set to 0.001, listener noise was set to 0.01, the generative preference was 0.6, and nu was set to 10. 1000 chains were run. Note how for the binomials with large N, the ordering preferences tend to be more extreme.

Further, the disparity of the noise affects the rate of regularization. Increased noise results in weaker regularization (i.e., less regularization for lower frequency items, see (**ref?**)(fig:absoluteDiffMatters)), however a larger relative difference between the speaker and listener noise parameters increases both the strength and the speed of the regularization (see (**ref?**)(fig:fasterSlowerReg)).

Interestingly this regularization disappears if the listener's noise parameter is less than or equal to the speaker's noise parameter (Figure 6).

It is useful to revisit here what the speaker and listener noise parameters represent. The speaker noise parameter is how often the speaker produces an error and the listener noise parameter is the listeners' belief of how noisy the environment is. Framed this way, it is perhaps unsurprising that we do not see regularization when the parameters equal eachother, since they essentially cancel eachother out (everytime a speaker makes an error, the listener is accounting for it, thus we get convergence to the prior).

Thus our model makes a novel prediction: In order to account for frequency-dependent regularization, listeners must be inferring more noise than speakers are actually producing (according to our model).

### Corpus Data

Finally, we now demonstrate that our model also predicts the language-wide distribution of binomial preference strengths seen in the corpus data. Specifically, we show that with $\nu$ set to 10, listener noise set to 0.02, and speaker noise set to 0.005, our model does a pretty good job of approximating the distribution in the corpus data (See Figure 7).

## Conclusion

Our results demonstrate the frequency-dependent regularization emerges from a noisy-channel processing model (Gibson et al., 2013) in an iterative-learning paradigm (Reali & Griffiths, 2009) when listeners assume more noise in the environment than the speakers actually produce.

-noisy channel processing

-speaker vs listener noise

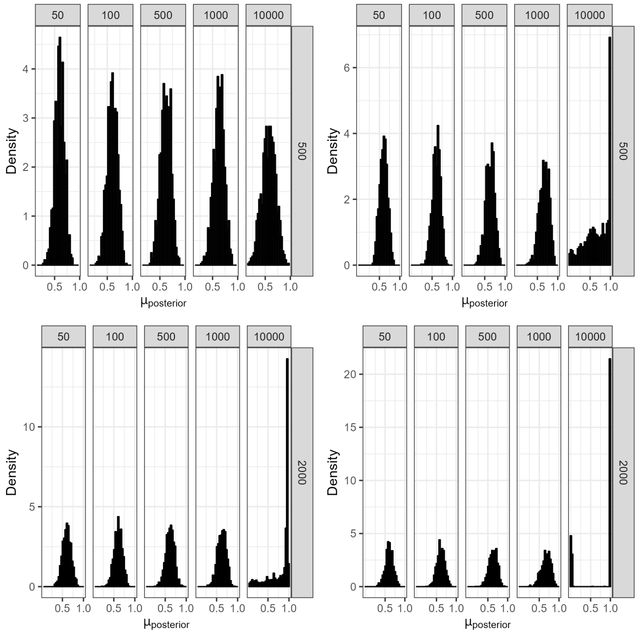

*Figure 4.* A plot of simulations with different noise parameters at 500 (top plots) and 2000 (bottom plots) generations. For the left plots, the speaker noise was set to 0.009 and the listener noise parameter was set to 0.01. For the right plots, the speaker noise was set to 0.0075 and the listener noise parameter was set to 0.01. For both plots, the generative preference was set to 0.6 and nu was set to 10.

## References

Albert Felty, R., Buchwald, A., Gruenenfelder, T. M., & Pisoni, D. B. (2013). Misperceptions of spoken words: Data from a random sample of american english words. *The Journal of the Acoustical Society of America*, *134*(1), 572–585.

Benor, S. B., & Levy, R. (2006). The chicken or the egg? A probabilistic analysis of english binomials. *Language*, 233278.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(1), 110. Retrieved from https://psycnet.apa.org/record/1981-07020-001

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056. https://doi.org/10.1073/pnas.1216438110

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*(3), 441–480. https://doi.org/10.1080/15326900701326576

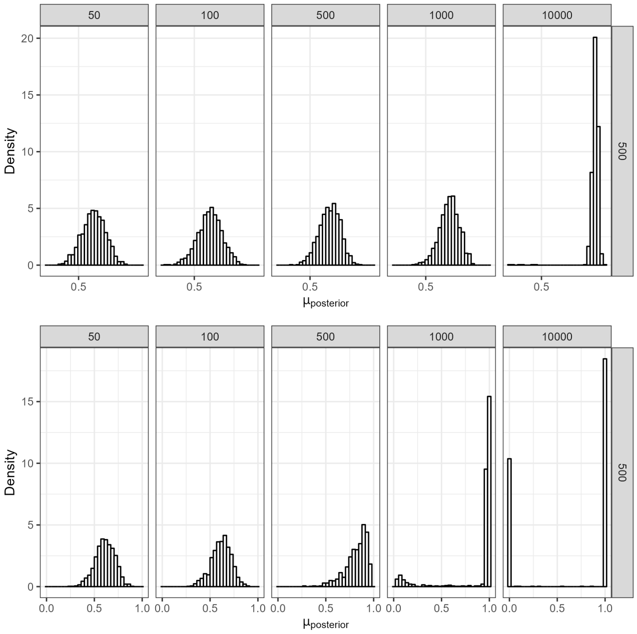

*Figure 5.* A plot of simulations with different noise parameters, but the same relative difference between the speaker and listener noise parameters. The top plot For the top plot, the speaker noise was set to 0.091 and listener noise was set to 0.1. For the bottom plot, the speaker noise was set to 0.001 and listener noise was set to 0.01. Note that the relative difference between the listener and speaker noise parameters for both plots was the same (0.009).

Liu, Z., & Morgan, E. (2020). *Frequency-dependent regularization in constituent ordering preferences.* Retrieved from https://www.cognitivesciencesociety.org/cogsci20/papers/0751/0751.pdf

Liu, Z., & Morgan, E. (2021). *Frequency-dependent regularization in syntactic constructions*. 387389. Retrieved from https://aclanthology.org/2021.scil-1.41.pdf

Morgan, E., & Levy, R. (2015). *Modeling idiosyncratic preferences : How generative knowledge and expression frequency jointly determine language structure*. 1649–1654.

Morgan, E., & Levy, R. (2016a). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, *157*, 384–402. https://doi.org/10.1016/j.cognition.2016.09.011

Morgan, E., & Levy, R. (2016b). Frequency-dependent regularization in iterated learning. *The Evolution of Language: Proceedings of the 11th International Conference*.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317–328. https://doi.org/10.1016/j.cognition.2009.02.012

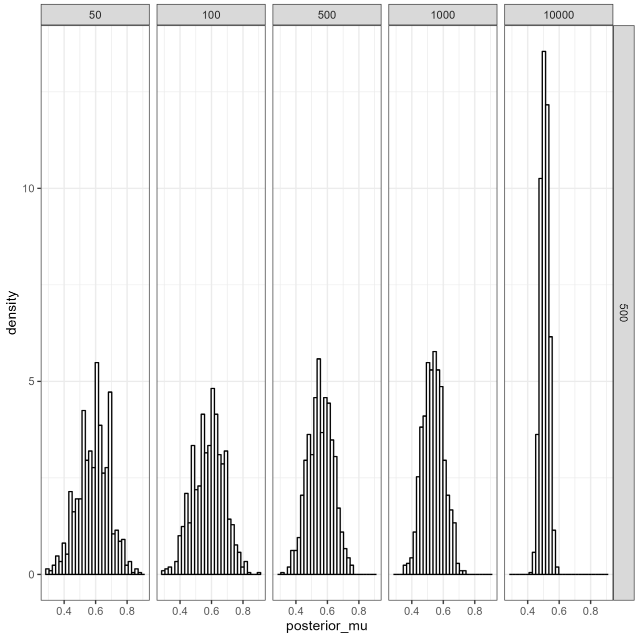

*Figure 6*. A plot of the distribution of simulated binomials at the 500th generation, varying in frequency. The top value represents N. On the x-axis is the predicted probability of producing the binomial in alphabetical form. On the y-axis is probability density. Speaker noise was set to 0.01, listener noise was set to 0.001, the generative preference was 0.6, and nu was set to 10. 1000 chains were run. Note how regularization does not appear to be present in this graph.

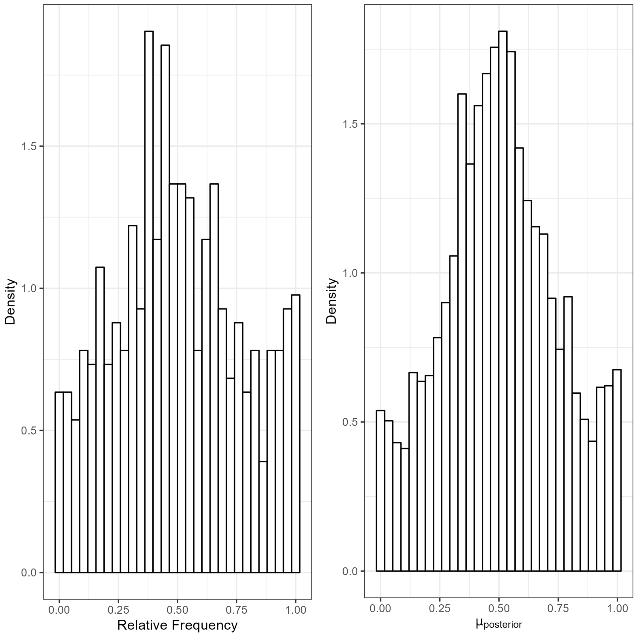

*Figure 7*. A plot of the distribution of ordering preferences after 500 generations of our iterated learning model (left) and the distribution of ordering preferences in the corpus data from Morgan and Levy (2015). For our simulations, the binomial frequencies and generative preferences were matched with the corpus data. $\nu$ was set to 10, listener noise was set to 0.02, and speaker noise was set to 0.005.