Prior Expectations in Linguistic Learning: A Stochastic Model of Individual Differences

R. Alexander Schumacher (robertschumacher2016@u.northwestern.edu)
Department of Linguistics, Northwestern University, 20160 Sheridan Rd.
Evanston, IL 60208 USA

Janet Pierrehumbert (janet.pierrehumbert@oerc.ox.ac.uk)
Oxford e-Research Centre, 7 Keble Rd.
Oxford, UK OX1 3QG

Abstract

When learners are exposed to inconsistent input, do they reproduce the probabilities in the input (*probability matching*), or produce some variants disproportionately often (*regularization*)? Laboratory results and computational models of artificial language learning both argue that the learning mechanism is basically probability matching, with regularization arising from additional factors. However, these models were fit to aggregated experimental data, which can exhibit probability matching even if all individuals regularize. To assess whether learning can be accurately characterized as basically probability matching or systematizing at the individual level, we ran a large-scale experiment. We found substantial individual variation. The structure of this variation is not predicted by recent beta-binomial models. We introduce a new model, the Double Scaling Sigmoid (DSS) model, fit its parameters on a by-participant basis, and show that it captures the patterns in the data. Prior expectations in the DSS are abstract, and do not entirely represent previous experience.

Keywords: artificial language learning; regularization; structural bias; substantive bias; individual differences; probability matching

Theoretical Background

In artificial language learning experiments, individuals learn a mock language that manipulates some property of interest. The extent to which learners acquire and generalize the linguistic patterns, and how their generalizations deviate from the input, shed light on the cognitive factors that shape language systems. Previous work has highlighted two important issues. One is regularization: faced with inconsistent input, do learners output the same probabilities as in the input? Or is the output more regular than the input, where one of the variants occurs disproportionately often? In other words, do learners have a prior expectation of systematicity in language systems, and the predisposition to enforce systematicity in their output?

The other factor is the expectations that learners bring in the form of substantive biases. Faced with patterns that deviate from known patterns, or patterns that are typologically unusual, how are learners influenced? Do their outputs favor the more familiar or unmarked patterns? These factors are potential sources of regularization.

There is a paradox at the heart of the theoretical understanding of regularization behavior. At historical time scales, languages tend to regularize variation. However, laboratory work has reported that language learning in adults is basically probability matching (Hudson Kam & Newport 2005). If learners fundamentally probability match, how do languages become regular over time? This challenge has been addressed by proposals in which a probability matching learning mechanism has a twist that can generate regularization at long time scales. Reali and Griffiths (2009) propose a beta-binomial Bayesian model in which a slight bias towards systematicity can create structure over multiple generations. Culbertson and Smolensky (2012) propose a related Bayesian model in which domain-specific constraints bias learning.

Problematically, these models have only been fit to aggregated data. Aggregated data have the potential to be very misleading about the cognitive mechanisms of individual learners. Consider input evenly split between variant A and variant B: P(A) = P(B) = 0.5. If half the learners have P(A) = 1.0 in their output, and half have P(B) = 1.0, the aggregated data would appear to show probability matching. However, each individual learner would have regularized the data. Motivated by this issue, we carried out a large-scale experiment that would make it possible to assess individual learning patterns. The experiment manipulated the consistency and the familiarity/markedness of the input in a four by two design. Here, we summarize the results of the experiment. These results are not consistent with any of the probability matching learning mechanisms proposed in the prior literature. We therefore present a model that is basically systematizing, and show how this model can capture the effects of individual cognitive style and bias for familiar patterns.

The Experimental Results

Schumacher and Pierrehumbert (submitted) taught 632 English-speaking participants an artificial language over Amazon Mechanical Turk (AMT) using the game-like protocol in Schumacher, Pierrehumbert, and LaShell (2014). The language exposed learners to two complementary number marking systems. The English-like singular/plural (Plural) has bare singular stems and adds a suffix on the plural (*brick/brick-s*). The Welsh-like singulative/collective (Singulative) uses a bare form for the collective and adds a suffix on the singulative (*brics-en/brics*). Because the Plural system is typologically common and familiar to English speakers, whereas the Singulative system is not, we expected that learners would be biased in favor of the Plural.

The systems were taught using the same suffix in four con-

sistency conditions. In the 1.00 baseline conditions, the learners saw only one marking system (Singulative or Plural) that all nouns used. In the 0.875, 0.75 and 0.625 conditions, the suffix marked one of the marking systems on a random subset of the thirty-two training items. The same suffix marked the opposite system on the remaining items. Thus, in the 0.875, 0.75 and 0.625 conditions, learners were exposed to both marking systems in the training, one that they were familiar with (Plural) and one that they were not (Singulative). Training trials were two-alternative forced choice with immediate feedback. Participants advanced towards a goal by providing correct answers. Items answered incorrectly were repeated. After training, learners entered the test phase. In the test phase, learners provided answers to thirty-two novel generalization items, interspersed with the training items. Trials were two-alternative forced-choice with no feedback. The data of central interest are the proportion of responses consistent with the dominant system on the generalization items. We are particularly interested in the extent to which this proportion differs from the proportion in the input. A probability matching mechanism predicts no reliable difference. A positive deviation indicates regularization of the dominant pattern, and a negative deviation means that the participant "irregularized": or – in the extreme – regularized the minority pattern. This comparison is shown in Figure 1.

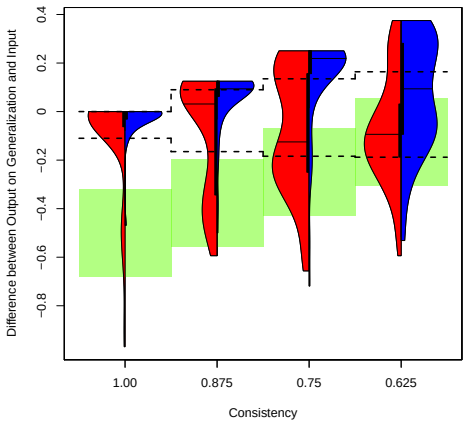

Figure 1: Violin plots of the difference between the generalization proportions and input. Values > 0 represent regularization. Singulative conditions are in red, Plural are in blue. The dashed line is the 95% confidence interval around the input proportion. The shaded area is the 95% confidence interval around a 0.5 generalization rate (chance).

The variability across individuals is substantial. Several distributions appear bimodal, an outcome that probability matching does not predict. Because the upper modes are near ceiling while the lower modes fall within the 95% confidence interval for random guessing ($P = 0.5$), it appears that learners either succeeded or failed in forming a productive generalization. Prior proposals along these lines include Mikheev (1997), Albright and Hayes (2003) and Yang (2005).

Systematic differences between the Singulative and Plural conditions indicate that the unexpectedness of the Singulative affects learning. However, there is not an across-the-board shift to the Plural. Any model must capture the following:

- In the 1.00 conditions, the Singulative and Plural distributions are both near ceiling.
- In the 0.875 and 0.75 conditions, more people regularize in the Plural than in the Singulative.
- The difference between the Singulative and the Plural is attenuated in the 0.625 conditions, where even in the Plural many participants exhibit random behavior.

Double Sigmoid Scaling

The predisposition for systematizing rather than reproducing the input variability indicates a nonlinear input-output relationship. Many prior efforts to capture systematization have used a sigmoid function (Ashby & Maddox, 1993; Mandelshtam & Komarova, 2014; Pierrehumbert et al., 2014). The inflection point in the middle of the sigmoid, interpretable as decision threshold, pushes the outputs toward more extreme probabilities than the input. Figure 1 shows that this behavior is problematic for explaining the results. In the Singulative and Plural 0.625 conditions, the majority of participants deviated towards 0.5 rather than away from it. This indicates the existence of a flat region in the middle of the nonlinear function. For less severe inconsistency, in contrast, there are tendencies towards regularization. These observations inspire us to posit a double sigmoid function containing two inflection points. The challenge is to capture the interaction of the propensity to regularize with the bias towards a familiar or unmarked pattern.

The Double Sigmoid Scaling model (DSS) is based on the logit of the natural logarithm (1). The term p is the input proportion. We only discuss the simple binary case here.

$$logit(p) = \ln\left(\frac{p}{1-p}\right) \tag{1}$$

The logit is interpretable as log odds. The inverse of the logit is the logistic; the composition of the logit with the logistic is the function is the line $y = x$, which describes probability matching (2):

$$p = \frac{1}{1+e^{-\left(\ln\left(\frac{p}{1-p}\right)\right)}} \tag{2}$$

We describe regularization by adding a scaling factor $c \geq 1$ to (1) before transforming back to probabilities.

$$f(p) = \ln\left(\frac{p}{1-p}\right)^{c} \tag{3}$$

This introduces nonlinearity into the model. As shown by the black lines in Figure 2, the composition of (3) with the logistic is a double-sigmoid. A flat region centered at 0.5 is introduced, which is dramatically different from the behavior of the sigmoid functions mentioned above. The region becomes

flatter and wider as c increases. Consequently, changing the value in the middle of the range of p causes little change in the output. The flat region is interpretable as the input range where there is insufficient evidence for the learner to infer a regularization-inducing rule. Uncertainty leads the learner to guess, a behavior seen with learners that fall within the confidence interval around chance in Figure 1.

The scaling factor alone allows the model to produce regularization. However, the model is still rotationally symmetric by 180° around $(0.5, 0.5)$. In order to produce different rates of regularization for Plural versus Singulative conditions in the 0.875 and 0.75 consistencies, the model must be asymmetric. The addition of a substantive bias parameter to (3), b in (4), shifts the center of the flat region to the right or left.

$$f(p) = \left(\ln\left(\frac{p}{1-p}\right) + b\right)^c \tag{4}$$

Equation (5) is thus the final model.

$$Output(p|b,c) = \frac{1}{1+e^{-\left(\ln\left(\frac{p}{1-p}\right)+b\right)^c}} \tag{5}$$

The model is symmetric when $b = 0$. $b > 0$ moves the output left (red lines), and $b < 0$ moves it right (blue lines). Thus, b shifts how much evidence is needed to form a rule in a particular direction, describing the preference for one system over another. An extreme value of $b = 2$, for example, can push a learner in the Singulative 0.875 condition (= Plural 0.125) down to guessing. The model also has the ability to favor a minority system by manipulating b, an outcome observed for a few participants. This method of formalizing substantive bias is neutral about its source, and is consistent with a number of interpretations. A positive bias favors the Plural, which is consistent with a preference for a familiar system, or even a cognitive bias for Plural marking. A negative bias favoring the Singulative could indicate a preference for novelty.

Using scaling factor c, the width of the flat region can be manipulated to explain observed differences in regularization versus uncertainty. The similarity between the Plural and Singulative 0.625 conditions follows from the fact that 0.625 falls within the flat region for more individual combinations of b and c than more extreme input proportions do. For the 1.00 conditions, the model analytically predicts that all participants will perform at ceiling. Individual biases have no opportunity to express themselves. This prediction is in qualitative agreement with the observation that the outcomes for the Singulative 1.00 and Plural 1.00 were extremely similar and displayed little variability.

Figure 2 illustrates the behavior of the model. The model takes a probability as input and outputs a probability. It is guaranteed to have fixed points at (0,0) and (1,1). The DSS captures both systematizing and substantive biases in a framework that can also capture regions of uncertainty. We now turn to validating the DSS against the experimental results.

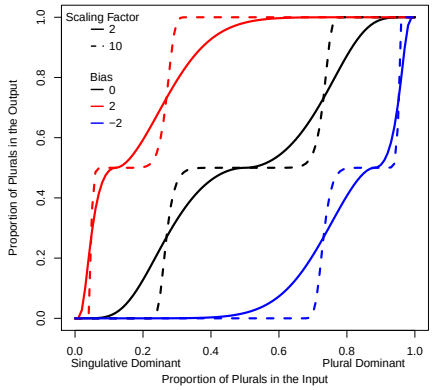

Figure 2: Behavior of the DSS for various parameter combinations. $b < 0$ moves the function rightwards from 0.5 and $b > 0$ moves it leftwards. Increasing c widens the flat area in the middle and increases the speed of scaling towards the asymptotes.

Fitting the DSS Model

Fitting Procedure

Model parameters cannot be estimated for individuals in the 1.00 conditions because the predicted outputs in this condition are always at ceiling. For the inconsistent conditions, there is enough variability to allow the model to be fit to the training data for each participant separately. Parameters were estimated by fitting the models to the training responses made on each unique training trial using the Levenberg-Marquardt algorithm (using *nlsLM* in R).

Criteria

Self-consistency Because participants are a random sample from the same population, and the parameters represent their mental state prior to training, there is no reason for them to vary *across groups* in their prior expectations for Singulative versus Plural marking systems. Consequently, the parameter estimates in each group should be reasonably similar. Initial parameter estimates which anticipate the presentation condition prior to any exposure are not plausible, since they imply that learners were precognizant of the input.

Generalization performance The degree to which the post-training model output predicts generalization performance demonstrates success. Generalization performance here is the proportion of Plural responses produced on generalization items. We evaluate generalization performance with the Pearson correlation coefficient R of the predicted values of the DSS against the output. The relative success of the DSS is evaluated against a baseline of $R = 0.735$, which is the correlation between the input proportion and output.

Results

First consider the distributions of the scaling factor c across conditions in Figure 3.

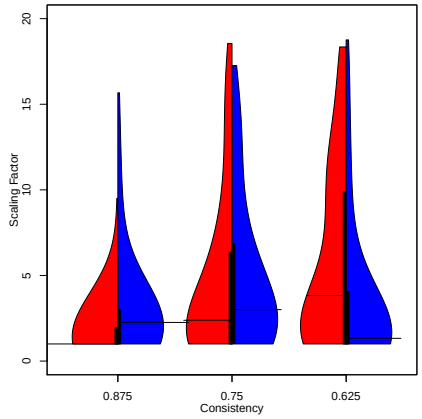

Figure 3: Violin plots of the scaling factor c broken down by condition. Sides in red are Singulative dominant conditions, sides in blue are Plural dominant, and they are arrayed in decreasing order by consistency of presentation.

In general their shapes and modes are close, although there is some tendency for more variable scaling factors in more variable conditions. Changes in the scaling factor parameter thus seem to be more dependent on the input proportion than on whether the dominant system was Singulative or Plural. This is a desirable outcome, since participants were drawn from the same population and a substantial number regularized in both cases.

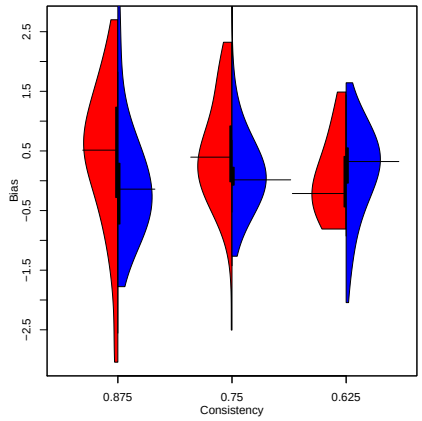

Figure 4: Violin plots of the bias parameter broken down by condition. Sides in red are Singulative dominant conditions, sides in blue are Plural dominant, and they are arrayed in decreasing order by consistency of presentation.

Moving on to the bias parameter b, the conditions are again generally similar. The most salient difference is greater variability in the less variable conditions, where the scaling factor is more variable. This suggests an interaction between the two factors in the fitting procedure. The means also differ somewhat by system. However, these differences are relatively inconsequential, due to the low sensitivity of the output mean to the bias parameter: a change of 0.5 in b results in a change of no more than 0.1 in the average output.

We now move on to the generalization performance criterion. Generalization performance is shown in Figure 5.

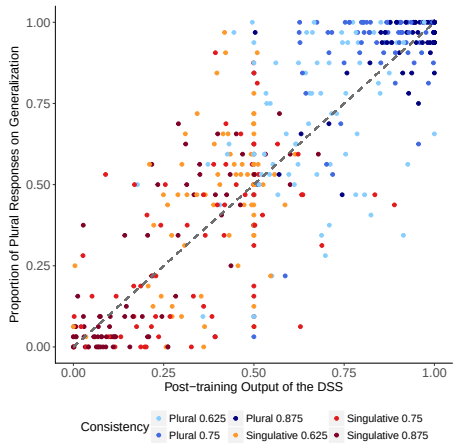

Figure 5: Scatter plot of the last output of the DSS against generalization performance, where the y-axis is the proportion of Plural responses. Red points are Singulative dominant, blue points are Plural dominant, with the intensity of the color corresponding to higher input consistencies. The dashed line is $y = x$.

Finally, consider generalization performance, shown in Figure 5. There is a clear, apparently linear relationship between the post-training output of the DSS and generalization performance. There is a salient vertical bar at 0.5, indicating learners who were considered to be guessing. The correlation is $R = 0.859$, which is a significant improvement over correlation of $R = 0.735$ for a probability matching baseline. In terms of improving prediction of generalization performance, then, the DSS is successful.

To demonstrate that the model captures the key observations about the data, we used the model to generate fake data, as follows: 1) we pooled the distributions of fitted bias and scaling factors across all conditions. 2) For each condition, we took a random sample of values from the pooled distribution 3) We generated the predicted outcomes for those parameter values for those conditions. The result (Figure 6) has the same overall structure as the real data (Figure 1.)

General Discussion

The DSS can explain the main observations in the experiment: ceiling performance on both Singulative and Plural 1.00, more random performance for 0.625 and more regularization for Plural than Singulative in 0.75 and 0.875. The

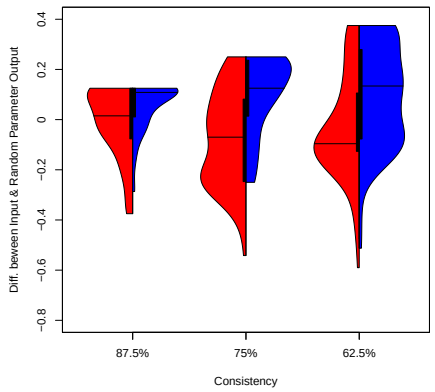

Figure 6: Violin plots comparing the output of the model from randomly sampled parameters of the global parameter distributions against the input proportion.

by-participant fitting of the DSS to training data yields reasonably consistent results for both scaling factor and bias parameters. It also predicts generalization performance better than a probability matching baseline model.

There are implications of the success of the DSS for the understanding of structural and substantive biases and their relationship to regularization. The DSS has an explicit tradeoff between generalization and guessing. The evidence has to be sufficiently high before a learner will generalize, but critically, the amount of evidence required by any particular learner may vary. The requisite amount of evidence for regularization is affected by a preference for more regular systems (scaling factor) or for particular systems (bias). When they favor the same direction (as in the Plural), regularization results. When the biases pull in opposite directions, as in the Singulative conditions, their effects interact. Depending on what system is preferred and how quickly scaling occurs, a learner will either generalize or guess at chance. Consequently, individual differences manifested more greatly in the Singulative, leading to guessing in some cases and generalization in others.

The success of the DSS is not matched by other classes of models and approaches to regularization. At their most basic, neither naive probability matching mechanisms nor naive systematizing mechanisms prove adequate. They both struggle to explain the individual variability across conditions.

An obvious mechanism for capturing individual differences in Bayesian models is the prior. Different classes of Bayesian models have different formulations of the prior. One influential class of Bayesian models which have gained traction in artificial language learning recently is the beta-binomial class (Reali & Griffiths 2009; Culbertson & Smolensky 2012), where the prior has the force of some previous experience. A beta-binomial model can produce regularization through variation in a prior with hyperparameters α and β, where the prior is equivalent to $\frac{\alpha}{\alpha+\beta}$. One of the hyperparameters is added to the observed counts of the Plural, and the other is added to the counts of the Singulative. Consequently, the hyperparameters function as previous experience of the systems. The expected value of the posterior after training is thus (6), where A is counts of the Plural and B is counts of the Singulative:

$$\frac{\alpha+A}{\alpha+A+\beta+B} \tag{6}$$

High initial differences in the hyperparameters can cause different rates of regularization. Simulated results of (6) for the task are presented in Figure 7. It is calculated on the assumption that the prior represents previous knowledge and individuals differ in the effective strength of this knowledge. The shaded area represents possible outputs for initial biases favoring the Plural, from no bias (the dashed line) to extreme ($\alpha = 2000$, $\beta = 1$).

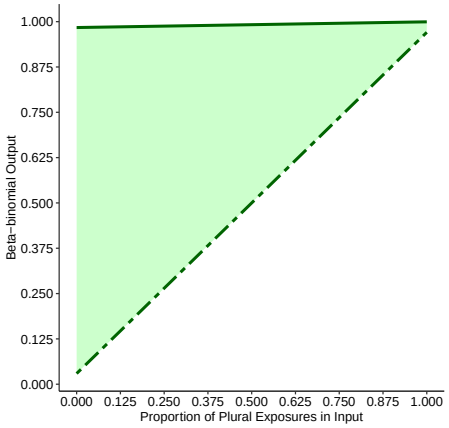

Figure 7: Possible outputs of a beta-binomial Bayesian model ranging from no bias (the dashed line) to extreme Plural bias ($\alpha = 2000$, $\beta = 1$).

Figure 7 shows that outcomes should be most variable in the Singulative 1.00 condition, since any difference in the prior will produce values that spread out in the shaded area. However, performance in the Singulative 1.00 was at ceiling.

This particular problem is indicative of a more general problem with applying beta-binomial models to regularization behavior. No assignment of priors to the population could produce ceiling in the Singulative 1.00 as well as the variability in the Singulative 0.875 and Singulative 0.75 without large and systematic initial differences in the prior across conditions. While individuals are permitted to differ in their priors on a Bayesian perspective, they should not do so according to the condition they were about to be exposed to in the task. Nevertheless, this is the approach that must be taken if the prior is the ultimately the source of the variability. For example, in order to capture cross-condition variability with a hyperprior, no participants in the Singulative 1.00 could have a bias against the Singulative, while roughly half in the Singulative 0.75 must. Given that all participants were recruited

from AMT at the same time, this is highly implausible.

Even a more sophisticated beta-binomial model like the Bayesian Mixture Model (BMM) (Culbertson & Smolensky 2012) requires significant heterogeneity within and across conditions to capture the results. In addition to α and β, the BMM employs weights that uniformly enforce a preference for unmarked or familiar patterns. Therefore, it predicts as strong a dispreference for the Singulative in the Singulative 1.00 as in the Singulative 0.875 and 0.75. Yet, the Singulative 1.00 was at ceiling. The Reali and Griffiths (2009) beta-binomial model uses the prior in a different way. Taking $\alpha = \beta < 1$, they provide a symmetric U-shaped prior distribution that weakly favors systematizing without favoring either competitor. Equal levels of regularization are predicted for both Singulative and Plural conditions, contrary to fact.

The contrast between the success of the DSS and the inadequacy of beta-binomial models relates to the larger debate between connectionist and Bayesian approaches. The DSS can be interpreted as a simple feed-forward two node neural network: (4) describes a non-linear (logit) input encoding subject to scaling and bias, (5) performs a logistic transformation on the output of (4). Zhang and Maloney (2012) reviews evidence that probabilities are represented by the cognitive system on a log odds (logit) scale. Related double sigmoid models are developed in Madhavan et al. (1995) and Lipovetsky (2015), but have more free parameters and different fixed points.

These observations suggest that a connectionist approach to regularization behavior might be able to incorporate the strengths of the DSS model. The reason for this concerns the characterization of what individuals bring to the task. In beta-binomial models, individuals begin the task with prior expectations, which are structured probabilistic representations that presuppose a fully-formed analysis of the input. In the DSS and connectionist approaches, the free parameters that could differ amongst individuals need not assume anything about the form of the input.

Further work is necessary to determine whether a connectionist interpretation of the DSS is the most appropriate. For now, we note that the success of the DSS argues that a more abstract representation of prior expectations in language learning is necessary, one that is not based entirely on previous experience. The mechanism for producing individual differences needs to be more potent than is commonly assumed, particularly in beta-binomial models. More work is needed that focuses on modeling on an individual level, so that the mechanisms can be evaluated by how they capture variation.

Conclusion

The results of Schumacher and Pierrehumbert (submitted) exhibit striking patterns of individual variation. To accurately capture the data, we presented the Double Sigmoid Scaling model. Through its shape, the DSS can explain the observed patterns of variation found in the experiment. The model was validated against the data, providing consistent parameter estimates and better predictive ability for generalization data. The comparative success of the DSS suggests that more abstract representations of prior expectations in language learning are needed to understand regularization behavior both at individual and group levels.

Acknowledgments

This project was made possible through a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

References

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*(2), 119–161.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *J. Math. Psychol.*, *37*(3), 372–400.

Culbertson, J., & Smolensky, P. (2012). A Bayesian Model of Biases in Artificial Language Learning: The Case of a Word-Order Universal. *Cognitive Sci.*, *36*(8), 1468–1498.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Lang. Learn. Dev.*, *1*(2), 151–195.

Lipovetsky, S. (2015). Trinomial response modeling in one logit regression. *Annals of Data Science*, *2*(2), 157–163.

Madhavan, P., Stephens, B., & Low, W. (1995). Tri-state neural network and analysis of its performance. *Intelligent Automation & Soft Computing*, *1*(3), 235–245.

Mandelshtam, Y., & Komarova, N. L. (2014). When learners surpass their models: mathematical modeling of learning from an inconsistent source. *B. Math. Biol.*, *76*(9), 2198–2216.

Mikheev, A. (1997). Automatic rule induction for unknown-word guessing. *Comput. Ling.*, *23*(3), 405–423.

Pierrehumbert, J. B., Stonedahl, F., & Daland, R. (2014). A model of grassroots changes in linguistic systems. *arXiv:1408.1985*.

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

Schumacher, R. A., & Pierrehumbert, J. B. (submitted). *Regularization of Expected versus Unexpected Marking Systems.*

Schumacher, R. A., Pierrehumbert, J. B., & LaShell, P. (2014). Reconciling Inconsistency in Encoded Morphological Distinctions in an Artificial Language. In *Cogsci.*

Yang, C. (2005). On productivity. *Linguistic Variation Yearbook*, *5*(1), 265–302.

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, *6*, 1.