Regularization behavior in a non-linguistic domain

Vanessa Ferdinand (v.a.ferdinand@sms.ed.ac.uk), Bill Thompson (bill@ling.ed.ac.uk),
Simon Kirby (simon@ling.ed.ac.uk), Kenny Smith (kenny@ling.ed.ac.uk)
Language Evolution and Computation Research Unit
School of Philosophy, Psychology & Language Sciences, University of Edinburgh
Dugald Stewart Building, 3 Charles Street, Edinburgh, EH8 9AD, UK

Abstract

Language learners tend to regularize unpredictable variation and some claim that is due to a language-specific regularization bias. We investigate the role of task difficulty on regularization behavior in a non-linguistic frequency learning task and show that adults regularize variable input when tracking multiple frequencies concurrently, but reliably reproduce the variation they have observed when tracking one frequency. These results suggest that regularization behavior may be due to domain-general factors, such as memory limitations.

Keywords: frequency learning; regularization; probability matching; Bayesian models;

Introduction

Languages contain very little unpredictable variation (Chambers et al., 2003) and language learners tend to regularize the inconsistent input they encounter (Reali & Griffiths, 2009; Hudson Kam & Newport, 2009, Smith & Wonnacott, 2010). For example, English contains two forms of the indefinite article *a* and *an*, but a deterministic rule (based on the initial phoneme of the following noun) governs the use of these two variants. Why are languages regular, and what drives learners to eliminate free variation in language? Some have suggested that we come to the task of language learning with the expectation that languages are regular and that this expectation takes the form of a language-specific innate bias (Bickerton, 1984; DeGraaff, 1999; Lumsden, 1999; Becker & Veenstra, 2003). Others claim that linguistic regularization can be explained by domain-general learning mechanisms, such as the effects of memory limitations on the type of variation that learners produce (Hudson Kam & Newport, 2005, 2009; Hudson Kam & Chang, 2009). Hudson Kam and Newport (2005, 2009) have shown that children tend to regularize free variation, whereas adults maintain it by probability matching, and attribute this difference to children having lower working memory capacity than adults. Newport (1990) demonstrated that children have more of a limited ability to learn from inconsistent input and Hudson Kam and Chang (2009) showed that adults probability matched more when word retrieval was made easier and regularized more when it was difficult, further corroborating their claim that memory limitations can lead to regularization, although see Perfors (2012) for an account of restricted memory encoding that does not lead to regularization.

A similar effect of memory limitations can be found in a non-linguistic tasks. In a study with adults, Kareev et al. (1997) reported an effect of individual differences in working memory capacity (as determined by a digit-span test) on participants' perception of the correlation of two probabilistic variables. Participants with lower capacity overproduced the most common variant, whereas participants with higher capacity did not. Regularization is also modulated by the number of variables in a task; adults regularized slightly more when predicting which of three lights will flash next than when predicting for two lights (Gardner, 1957).

In this paper, we explore the effect of tracking single versus multiple frequencies on the regularization behavior of adults in a non-linguistic task. We show that participants probability match when tracking a single frequency, but regularize when tracking six frequencies concurrently. Because concurrent frequency learning is a prominent aspect of language learning (Saffran, Alin & Newport, 1996), and also elicits regularization in a non-linguistic task, this is consistent with a domain-general account of the observed regularization bias in language, possibly attributable to limited working memory.

Frequency learning experiment

Participants 381 participants were recruited via Amazon's Mechanical Turk crowdsourcing platform and completed our experiment online. 37 participants were excluded on the basis of the following criteria: failing a color vision test (2), self-reporting the use of a pen or pencil during the task (14), not reporting their sex or age (2), or having previously participated in any of our experiments, as determined by their user ID with MTurk (19). More participants were recruited than necessary with the expectation that many would be excluded by these criteria. Once the predetermined number of participants per condition was met, data from the last participants was excluded, totaling 24 participants across all conditions and tasks. All excluded participants received the full monetary reward for the task. The average monetary reward per participant, converted to an hourly rate, was $2.64. Of the final 320 participants, 184 are female, and the mean age is 36 (min = 18, max = 69), with a standard deviation of 12 years.

Materials The experiment was coded up as a java applet that ran in the participant's web browser in a 600x800-pixel field. Photographs of 6 different containers (a box, pouch, jar, bowl, bucket, and basket) and graphically generated images of marbles in 12 different colors (blue, orange, brown, grey, black, yellow, red, teal, olive, pink, purple, and lime) served as stimuli.

One-item task This experiment consisted of a training phase in which participants observed a series of 10 marble draws from a bag, and a testing phase in which participants were asked to produce another several likely draws from the

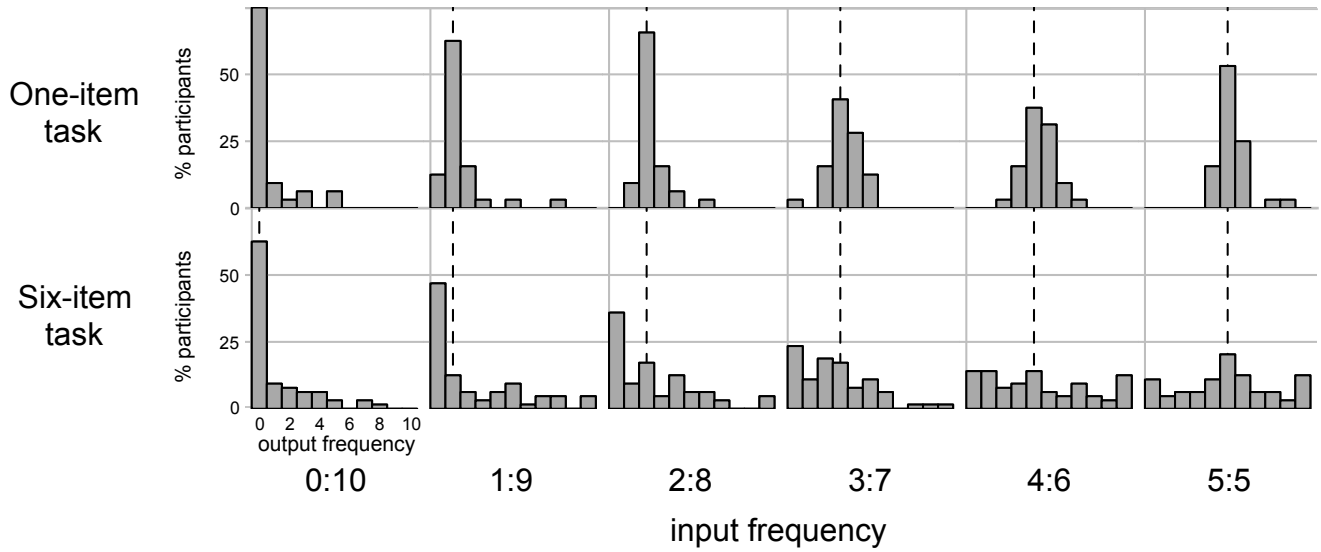

Figure 1: Each pane displays the percentage of participants that responded with a given output frequency of the minority marble (m) during testing. Columns are the input ratio of m:M during training. Dashed lines mark the input frequency of m. In the one-item task, participants probability matched, reproducing the input ratio with high fidelity. This task was between-subjects; each participant was trained on one input ratio only. In the six-item task, participants were more likely to regularize than to reproduce the input ratio. This task was within-subjects; each participant was trained on all six input ratios concurrently.

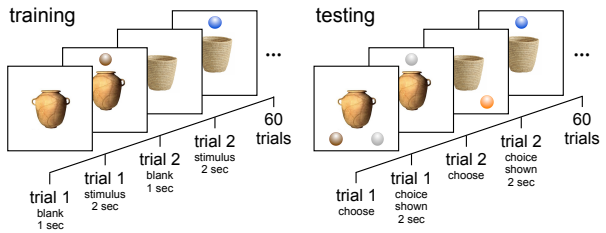

Figure 2: Training and testing trials for the six-item task.

same bag. In each training trial, a picture of the bag was displayed for 1000 milliseconds and then a marble (blue or orange) appeared over the bag for 2000 milliseconds. There were 10 training trials, with no break between trials. In each testing trial, the bag was displayed with the two marble colors below. Participants mouse clicked on a marble to make their choice of one draw from the bag. Their choice was displayed above the bag for 2000 milliseconds and then the next testing trial began. There were 10 testing trials with no breaks between trials. Locations (left or right) of the blue and orange marbles were held constant across test trials for each participant, but counterbalanced across participants.

A fixed ratio of blue to orange marbles was shown in the training phase. Each participant was randomly assigned to one of 6 training conditions based on this ratio. The color of the training ratio's minority marble (m) and majority marble (M) was counterbalanced across participants. All possible ratios of m:M were tested and will be referred to as the 0:10, 1:9, 2:8, 3:7, 4:6, and 5:5 conditions. 192 participants took part in this task, with 32 in each condition.

Six-item task This task is based on the word frequency learning task from Reali and Griffiths (2009). Participants observed 10 marble draws each from six different containers, totaling 60 marble draws (see Figure 2). Each container was associated with 2 unique marble colors (12 unique marble colors were therefore used). Training and testing trials were identical to the one-item task. Each container was uniquely associated with one of the possible ratios specified by condition 0:10, 1:9, 2:8, 3:7, 4:6, and 5:5 above. Thus, the six-item task is a within-subject version of the one-item task, with the addition that training and testing trials from all six conditions are interleaved. Assignments of a ratio and marble colors (in predefined color pairs) to each container was randomized per participant. 64 participants took part in this task. Two additional versions of this experiment were also run; one where all 6 bags were in condition 0:10 (each container was mapped to one color only) and one where all 6 containers were in condition 5:5. Each of these versions was completed by 32 new participants.

Experiment results

Participants in the six-item task were more likely to regularize their responses per container than participants in the one-item task. Here, we refer to regularization as the production of a more extreme ratio than that observed during training,

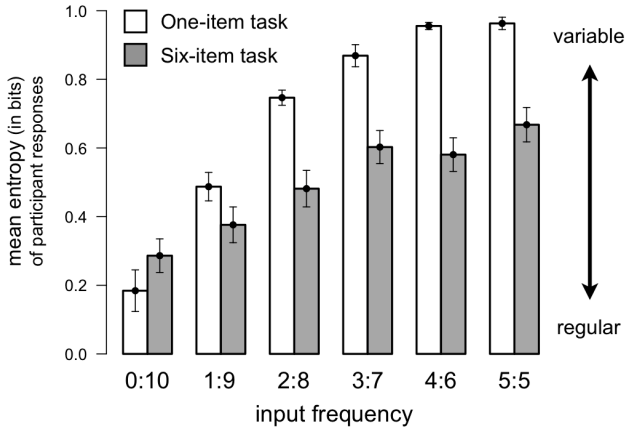

Figure 3: Difference in mean entropy scores between tasks, for each input ratio. Each participant's sequence of marble draws during testing was converted into an entropy score. Lower scores denote greater regularity within a response. Participant responses were significantly more regular in the six-item task than in the one-item task for input ratios 3:7, 4:6, and 5:5. Error bars show the standard error of the mean.

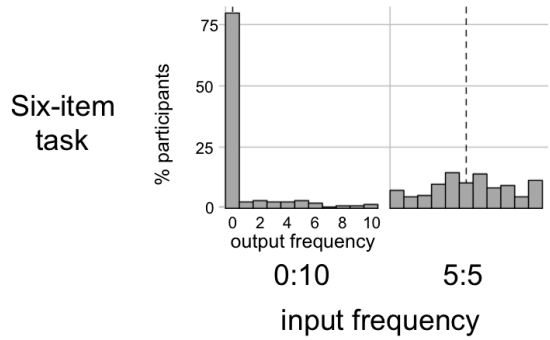

Figure 4: Distribution of participant responses for two additional versions of the six-item task, where all items contained the same input ratio of *m*:*M*. One group of participants was trained on all 0:10 ratios and another group was trained on all 5:5 ratios.

where 0:10 and 10:0 are the most extreme ratios and 5:5 is the least extreme. The distributions of participant responses are shown in Figure 1. Each pane displays the percentage of participants that responded with a given output frequency of *m*, per input frequency and per task. In the one-item task, participants probability matched; the mode of the population is on the input frequency of *m*, meaning that the most common response was perfect reproduction of the ratio observed during training. In the six-item task, visual inspection suggests that participants did not reproduce the training ratios with as high fidelity. Most participants regularized by overproducing the majority marble (all mass in the bars to the left of the dotted line) and a large number of responses are fully regular, meaning the output frequency of *m* is 0 or 10.

To better assess the different degrees of regularization between tasks, we calculated the entropy of each participant's sequence of test choices. This quantifies the amount of variation (in bits) with a value between 0 and 1; where 0 denotes a completely regular sequence (i.e. a series of all blue marble draws) and 1 denotes a maximally variable sequence (i.e. a series of 5 blue and 5 orange draws, in any order). This allows us to refine our definition of regularization as the overproduction of one marble, such that the entropy of the participant's testing choices is lower than that of their training observations. The mean entropy scores of participant responses per input frequency are shown in Figure 3.

A linear mixed effects regression analysis showed a significant effect of task on entropy scores, $t(34) = -7.226, p < .001$, and a significant effect of input frequency on entropy scores, $t(34) = -10.832, p < .001$. This means the two tasks elicited different amounts of regularity within participants' responses and that participants' responses were modulated by training frequencies; they noticed differences in the input frequencies and this affected their responses. A significant interaction of task and input frequency on entropy scores was also obtained, $t(34) = 4.570, p < .001$; participants responded differently to different input frequencies, and this pattern of responses also differed by task.

There was a significant difference in mean entropy scores between tasks for input frequencies 3:7, 4:6, and 5:5 ($W = 1427.5, p = .001; W = 1714, p < .001; W = 1585.5, p < .001$), respectively.[1] The difference in mean entropy between tasks was not significant for input frequencies 0:10, 1:9, and 2:8 ($W = 894, p = .228; W = 1184.5, p = .192; W = 1264, p = .054$[2]), respectively.

Two additional experiments were conducted to explore the possibility that regularization in the six-item task is due to interference between containers, such that ratios learned for one container get confused with ratios learned for another container. We eliminated this type of interference by training participants on 6 containers with identical ratios. Figure 4 shows participant responses when trained on all 0:10 ratios (left) and all 5:5 ratios (right). The average entropy for the all 0:10 task is significantly lower than that of the 0:10 condition in the six-item task ($W = 5061, p = .004$), but not significantly different than the 0:10 condition in the one-item task ($W = 2900.5, p = .466$). Tracking multiple 0:10 ratios is no different than tracking one 0:10 ratio, but it is different from tracking one 0:10 ratio concurrently with other ratios. This means interference may account for the errors participants make in the original six-item task when producing draws for the container they observed as 0:10. However, for the all 5:5 task, the average entropy was not significantly different from

[1] These were determined with a non-parametric t-test, the Whitney-Mann U-test, since the distributions of entropy scores are non-normal.

[2] After correction for multiple comparisons, this is not approaching significance.

the 5:5 condition in the six-item task ($W = 5892.5, p = .617$). Participants still produced 0:10 and 10:0 responses in the absence of observing these ratios during training. Therefore, interference may account for some of the differences between the one-item and six-item tasks, but this isn't the sole cause of the regularization behavior observed in the six-item task.

Frequency learning models

What cognitive processes cause regularization? So far our analyses have quantified the difference in regularity between participants' training and testing responses. In this section, we turn our focus to an internal force that can affect a learner's behavior; an inductive bias favoring certain ratios of marbles.

Bayesian model

Bayesian models provide a way to quantify inductive biases and understand their effect on behavior. We fit a beta-binomial Bayesian sampler model to participants' responses, following Reali and Griffiths (2009), and ask what prior expectation for regularity a Bayesian rational learner would need to have in order to produce the data that our participants produced.

A Bayesian rational learner uses Bayes' rule, $P(h|d) \propto P(d|h)P(h)$, to infer what proportion of marbles generated the draws that they observed. Here, each proportion is a hypothesis and the observed draws are the data. Bayes rule combines the prior probability of a hypothesis, $P(h)$, with the likelihood of the data under that hypothesis, $P(d|h)$, to arrive at a posterior probability of that hypothesis given the data, $P(h|d)$. The prior is a beta distribution over all hypotheses, Beta$(\frac{\alpha}{2}, \frac{\alpha}{2})$, where the parameter α determines whether the learner expects to see regular draws or variable draws. A learner with $\alpha < 2$ will tend to regularize their productions, a learner with $\alpha = 2$ is unbiased toward any particular proportion of draws, and a learner with $\alpha > 2$ is biased towards variability in draws. The likelihood of drawing N marbles in ratio $k : (N-k)$ from a container of marbles in proportions $p : (1-p)$ follows a binomial distribution (Equation 1).

$$P(k|p,N) = \binom{N}{k} p^k (1-p)^{N-k} \quad (1)$$

Once the posterior probability over all hypotheses has been determined, the learner must choose a hypothesis to generate testing responses from. We take the case where learners sample a hypothesis from the posterior distribution, and then sample data from this hypothesis according to its likelihood (as if the learner were randomly drawing marbles from the hypothesized proportion, with replacement, as in Equation 1).

This model defines the probability of generating all testing proportions (output states) from all training proportions (input states) and can be visualized as a transition matrix between all possible states in the system. Because our experiment covers all possible training proportions for 10 draws from a bag, we can also construct an empirical transition matrix from participant responses in each task. From here on, we switch to visualizing our data in terms of marble 1 (m_1) and marble 2 (m_2)[3]. Figure 5 (top row) shows the two empirical transition matrices and three model matrices for different values of the prior parameter α. Each value of α defines a unique transition matrix, and thus a unique pattern of behavior. For example, if a Bayesian learner observes 1 draw of m_1 and 9 of m_2, and if their prior is $\alpha = 0.01$, they are most likely to produce 0 draws of m_1 and 10 of m_2, regularizing their productions. If their prior is $\alpha = 2$, they are most likely to produce 1 draw of m_1 and 9 of m_2, probability matching their productions. And if their prior is $\alpha = 10$, they are most likely to produce 3 draws of m_1 and 7 of m_2, increasing variation in their productions. Thus, the prior used here intuitively captures a range of human behaviors in frequency learning.

The model fitting task at hand is to determine which model transition matrix most resembles the empirical transition matrix, by assigning the most likelihood to the empirical data. The prior associated with the best-fit model is the one that best explains participant behavior and gives us an idea of what biases our participants may have.

The best-fit bias in the one-item task is $\alpha = 1.55$ with a log likelihood of -413, which is equivalent to correctly predicting 20% of participant responses in this task [4]. This prior shows an expectation for a slight amount of regularity in the data set. For the six-item task, the best-fit bias is $\alpha = 1.21$ with a log likelihood of -1186, equivalent to 9% response prediction. This prior shows a stronger bias toward regularity in the six-item task than in the one-item task.

Prediction percentages are lower for the six-item task because participant responses are more variable in the this task than in the one-item task. Only deterministic processes (with one output per input) can be predicted with 100% accuracy. The ceiling on model prediction for each task was determined by fitting each data set to itself, yielding a maximum of 32% accuracy for the one-item task and 16% accuracy for the six-item task. Relative to these ceilings, the best-fit models account for 61% and 56% of participant responses in the one-item and six-item tasks, respectively.

Bootstrap model

An input-based random sampling model was also fit to the data. This model defines the transition matrix that would be obtained if participants produced their testing responses by randomly sampling 10 draws from their training observations, with replacement. In this case, each row would be a binomial where p equals the training proportion of m_1. It is important to note that this transition matrix defines the dynamics of drift in one generation and may be used as a baseline for the loss of variation that can occur in the absence of a regularization bias.

[3] marble 1 (m_1) refers to the blue marble in the one-item task, and to the blue, brown, black, red, olive, and purple marbles in the six-item task.

[4] The raw log likelihoods should not be compared between tasks, because there are a different number of observations per task. This is corrected for in the prediction percentages, which are comparable between tasks.

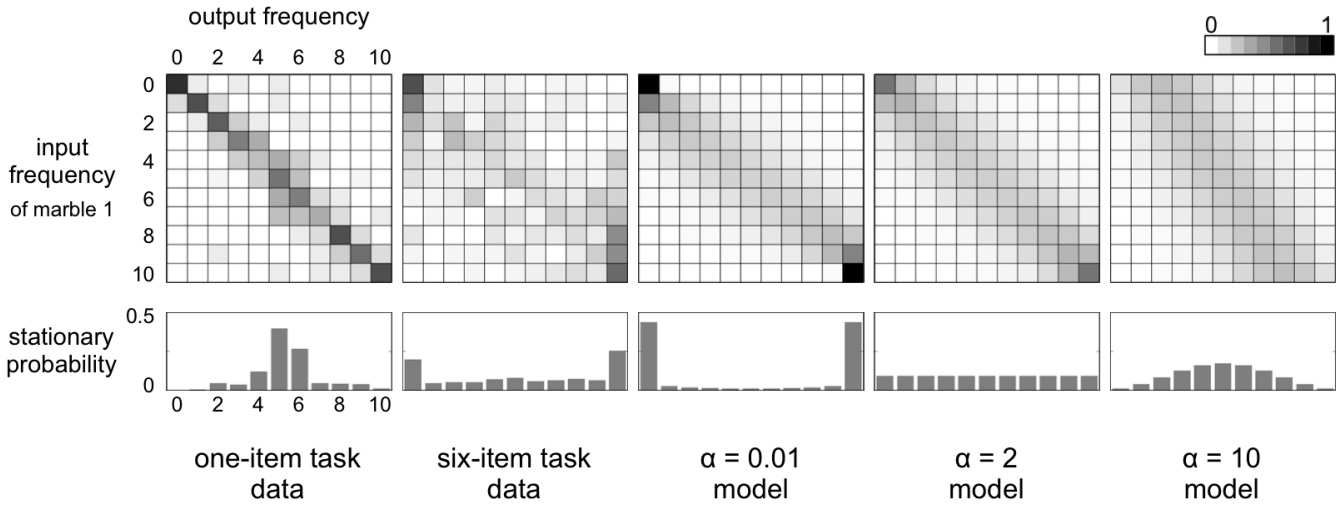

Figure 5: Transition matrices (top row) and their associated stationary distribution (bottom row) for the experimental results of the two frequency learning tasks, and for the Bayesian model showing three example bias strengths ($\alpha = 0.01, 2, 10$). Transition matrices give the probability of moving from each input frequency (the number of training trials showing marble 1) to each output frequency (the number of testing trials in which participants produced marble 1)[3]. The stationary distribution shows how often the transition matrix will produce each output frequency of marble 1.

For this model, the log likelihood of the one-item task data is -259, equivalent to 25% response prediction, and is a better fit than the best-fit beta-binomial sampler model[5]. Thus, of the models explored in this paper, drift provides the best account of our participants' probability matching behavior. However, a repeated measures Monte Carlo test shows that the standard deviation among participant output entropies in the one-item task data are significantly lower than that obtainable by drift: $p = .04$, $p = .03$, $p = .01$, $p = .003$, for conditions 2:8, 3:7, 4:6, 5:5, respectively. Although these data are well-accounted for by the drift model, they still show a quantitative difference in standard deviation, meaning that the forces behind probability matching are not truly isomorphic to drift. As for the six-item task, the log likelihood is -1076, equivalent to 6% response prediction. Here, the sampler model with a bias toward regularization is still the better fit.

Null model

This model is the transition matrix that would be obtained if participants were randomly sampling from the two testing choices each trial (i.e. not engaging in the task). Here, every row would be a binomial distribution where $p = 0.5$. For this model, the log likelihood of the one-item task data is -604, equivalent to 4% response prediction. For the six-item task, the log likelihood is -1630, equivalent to 1% response prediction. Of all models considered, this is the worst fit for both tasks, meaning that participants are not likely to be randomly sampling from their testing choices.

[5]This bootstrap model, which defines the dynamics of evolutionary drift, is equivalent to a Bayesian MAP model with $\alpha = 0$. See Reali & Griffiths (2010) for the proof.

The results of these model fits strongly suggest that participants in the six-item condition are not just performing poorly at reproducing their training proportion, but they are regularizing their responses in a way that can not be accounted for by random errors.

Learning biases and long-term behavior

In addition to comparing the transition matrices, which describe the behavior of one generation of learners, we can also look at the long-term behavior of the system, which is described by the stationary distribution of the transition matrix (Figure 5, bottom row). This distribution tells us what percent of the population we would expect to see in each state, after an arbitrarily large number of generations, if the output state of one learner served as the input state to another. Griffiths and Kalish (2007) have shown that the stationary distribution mirrors the prior distribution over hypotheses for the Bayesian sampler model utilized here. The stationary distributions of the empirical transition matrices are most interesting because these would be an estimate of our participants' regularization bias (the prior) if they were Bayesian sampler learners[6]. In line with this interpretation, the stationary distribution of the six-item task closely resembles that of its best-fit Bayesian model, which has a beta distribution Beta(0.605,0.605). However, the stationary distribution of the one-item task does not resemble that of its best-fit Bayesian model, which has a u-shaped beta distribution Beta(0.775,0.775). In general, the Bayesian model is a good fit to participant behavior in the six-item task, but does not account very well for participant behavior in the one-item task.

[6]Both of the empirical transition matrices are ergodic.

A close examination of the model's transition matrices and stationary distributions shows that probability matching behavior with a low standard deviation is not within this model's range of behavior.

Discussion

We have shown that learning a single versus multiple frequencies modulates participants' regularization behavior in a non-linguistic task. When participants tracked the frequency associated with a single item, they probability matched; reproducing the variation they had observed with high fidelity. However, when tracking multiple frequencies concurrently, participants regularized their responses, usually by overproducing the most common variant.

A beta-binomial Bayesian sampler model was fit to the results of each task and showed a stronger prior bias toward regularization in the six-item task than in the one-item task. Strictly speaking, the prior represents the inductive bias of the learner, and participants should come to a marble-drawing task with a particular expectation about the ratios of marbles in containers, regardless of the difficultly of the task. The fact that we find different best-fit priors according to different task demands means that we are not revealing the inductive bias of our participants, per se, but a composite picture that characterizes more than one cognitive constraint. At least one constraint that is sensitive to task demands should be added to the model, such as a memory constraint that disproportionally forgets lower-frequency observations. Such an addition could free up the prior to more accurately reflect participants' inductive bias. This raises a point of caution in comparing inductive biases across domains without controlling for task demands, since task demands can modulate bias strengths.

Our modeling results also suggest that human probability matching and regularization behavior do not lie on a simple continuum that can be captured by the prior alone. Although the Bayesian model accounted well for our participants' regularization behavior, it failed to account for the restricted variance of probability matching. Participants may be trying to produce a representative sample of draws, where the most likely response is the training ratio itself. Such a parameter might lead to high-fidelity reproduction of the training proportion under low memory constraints only.

If memory constraints are the cause of the regularization bias revealed when learning the frequencies of marbles in several containers, then this same domain-general factor may be the cause of regularization in tasks naturally characterized by concurrent frequency learning, such as language learning.

Acknowledgements

Special thanks to Tom Griffiths and Luke Maurits for feedback. This research was supported by the University of Edinburgh's College Studentship, the SORSAS award, and the Engineering and Physical Sciences Research Council.

References

Becker, A., & Veenstra, T. (2003). The survival of inflectional morphology in French-related creoles. *Studies in Second Language Acquisition, 25*, 283-306.

Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and Brain Sciences, 7*, 173-221.

Chambers, J., Trudgill, P., & Schilling-Estes, N. (2003). *The handbook of language variation and change*. Blackwell, Malden, MA.

DeGraaff, M. (1999). Creolization, language change, and language acquisition: an epilogue. In M. DeGraaf (Ed.) *Language creation and language change: creolization, diachrony, and development.* Cambridge, MA: MIT Press.

Gardner, A. (1957). Probability-learning with two and three choices. *The American Journal of Psychology, 70(2)*, 174-185.

Hudson, C., & Newport, E. (2005). Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Language Learning and Development, 1(2)*, 151195.

Hudson Kam, C., & Newport, E. (2009). Getting it right by getting it wrong: when learners change languages. *Cognitive Psychology, 59*, 30-66.

Hudson Kam, C., & Chang, A. (2009). Investigating the cause of language regularization in adults: memory constraints or learning effects? *Journal of Experimental Psychology, 35(3)*, 815-821.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science, 31*, 441480.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: sample size and the perception of correlation. *Journal of Experimental Psychology, 126(3)*, 278-287.

Lumsden, J. S. (1999). Language acquisition and creolization. In M. Degraaf (Ed.) *Language creation and language change: creolization, diachrony, and development.* Cambridge, MA: MIT Press.

Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science, 14*, 11-28.

Perfors, A. (2012). When do memory limitations lead to regularization? An experimental and computational investigation. *Preprint submitted to Elsevier*.

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

Reali, F., & Griffiths, T. L. (2010). Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift. *Proc. R. Soc. B, 277*, 429-436.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science, 274(5294)*, 1926-1928.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition 116*, 444-449.