Cognition 111 (2009) 317–328

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

# The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning

Florencia Reali, Thomas L. Griffiths *

*University of California, Berkeley, Department of Psychology, 3210 Tolman Hall # 1650, Berkeley, CA 94720-1650, United States*

ARTICLE INFO

*Article history:*
Received 18 September 2008
Revised 13 January 2009
Accepted 15 February 2009

*Keywords:*
Iterated learning
Bayesian models
Frequency distributions
Word learning
Language acquisition

ABSTRACT

The regularization of linguistic structures by learners has played a key role in arguments for strong innate constraints on language acquisition, and has important implications for language evolution. However, relating the inductive biases of learners to regularization behavior in laboratory tasks can be challenging without a formal model. In this paper we explore how regular linguistic structures can emerge from language evolution by iterated learning, in which one person's linguistic output is used to generate the linguistic input provided to the next person. We use a model of iterated learning with Bayesian agents to show that this process can result in regularization when learners have the appropriate inductive biases. We then present three experiments demonstrating that simulating the process of language evolution in the laboratory can reveal biases towards regularization that might not otherwise be obvious, allowing weak biases to have strong effects. The results of these experiments suggest that people tend to regularize inconsistent word-meaning mappings, and that even a weak bias towards regularization can allow regular languages to be produced via language evolution by iterated learning.



## 1. Introduction

Languages are passed from one learner to the next via processes of cultural transmission. Such processes introduce linguistic variation, with the generalizations produced by each learner changing the prevalence of linguistic forms. A particular type of change occurs when components of language with unpredictable or inconsistent variation lose their unpredictability and become more regular over time. This process of *regularization* has come to play a prominent role in discussions of the role of innate constraints on language acquisition in linguistics and cognitive science (e.g., Bickerton, 1981; Pinker, 1994).

An example of regularization appears in the creolization of Pidgin languages and certain forms of learning of sign languages (e.g. Bickerton, 1981; Siegel, 2007, see Hudson Kam & Newport, 2005, for a review). Pidgin languages typically emerge when speakers of mutually intelligible languages come together and need to communicate. Speakers then create a new communication system based on the superstrate language, that is, the language that predominates in the region. *Creoles* are more stable forms of language that originate as pidgin and which are learned by the children of a community as their native language. Pidgin or early Creole languages contain variability that is not typical of natively acquired languages (Birdsong, 1999, Johnson, Shenkman, Newport, & Medin, 1996). For example, speakers are inconsistent in their use of morphological markers or word order. Importantly, unlike the kind of variation present in native speech, which is typically predictable and shared by all speakers, the variation in pidgin languages is largely unpredictable – as is typical of second language productions. For example, in early Hawaiian Creole the particular word order used by individual speakers was influenced by the word order used in their native language. In later stages of Creole, however, language forms typically lose this unpredictability and become more regular. Further evidence that learners exposed to incon-

* Corresponding author. Tel.: +1 (510) 642 7134; fax: +1 (510) 642 5293.
*E-mail address:* tom_griffiths@berkeley.edu (T.L. Griffiths).

0010-0277/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2009.02.012

sistent use of grammatical forms tend to regularize their input comes from learning during acquisition of sign language from inconsistent teachers (Singleton & Newport, 2004) and from the emergence of regular systems in the creation of new sign languages (Senghas & Coppola, 2001; see Hudson Kam & Newport, 2005, for a review).

Another example of regularization occurs in situations of language contact where unpredictable variability is introduced to a language. Consider the case of word order changes observed in the transition from Old to Modern English. Scandinavian influence is thought to have introduced the verb-object order to English, resulting in a temporarily mixed system composed of verb-object and object-verb word orders in Old English (Kroch & Taylor, 1997). Over time, however, verb-object gradually replaced object-verb word order, resulting in the regular system observed in Modern English (for a review see Pearl & Weinberg, 2007).

The tendency of learners to regularize inconsistent language forms has often been taken as evidence for innate language-specific constraints on language acquisition (e.g., Bickerton, 1981, 1999). For example, according to the Language Bioprogram Hypothesis (Bickerton, 1981), when children are exposed to reduced communication systems such as pidgin languages, they introduce universal properties of natural languages by drawing on their innate knowledge of natural language structure. Recent studies, however, have seriously challenged some of the fundamental tenets of the Language Bioprogram Hypothesis. The emergence of creole appears to be less abrupt than previously assumed and seems to depend on stabilized forms of pidgin spoke by adults as a second language (Siegel, 2007). This points toward the need to understand how the inductive biases of individual learners – those factors that constrain their inferences from limited linguistic data – contribute to the regularization of unpredictable variability.[1] Identifying this relationship can provide insight into why languages take the forms they do, and how words and grammars evolve over time. In this paper we begin to explore this question for the case of estimating the frequencies of linguistic variants.

Learning a language with any kind of probabilistic variation requires learning a probability distribution from observed frequencies. Over the last couple of decades, a number of studies have accumulated showing that learners are able to extract a variety of statistics from a wide range of linguistic input (see Gomez & Gerken, 2000, 2003, for reviews). Recent work has explored how the frequencies of linguistic forms are learned. In this context, regularization corresponds to collapsing inconsistent variation towards a more deterministic rule. In one study, Hudson Kam and Newport (2005) trained participants on artificial languages in which determiners occurred with nouns with varying probabilities. They found that children regularize the unpredictability in the input, producing consistent patterns that were not the same as the training stimuli. They also found that adult participants produced utterances with probabilities proportional to their frequency in training, a response referred as *probability matching*.

Subsequent studies by Hudson Kam and Newport (in press) showed that, for adult participants, regularization depends on the form and complexity of the inconsistency in the input. For example, when two variant forms in the artificial grammar were used in free alternation, that is, the determiner being either present or absent in a sentence, the most frequent form was not regularized. However, when many different determiners were used and one form was much more frequently and consistently used than the others, adults *did* regularize that most consistent form. In a different study, Wonnacott and Newport (2005) used a similar artificial language to show that when adults learners were tested on words different from those in the training stimuli, adults regularized. Taken together, these results suggest that the level of complexity in the probabilistic input might influence whether learners adopt a regularization strategy rather than probability matching.

Another recent study on word learning provides further insights into the learning biases operating during learning from inconsistent input. Vouloumanos (2008) examined how adults track the statistics of multiple-referent relations during word learning. Participants were trained on novel object-word pairs. Objects were associated with multiple words, which in turn were paired with multiple objects with varying probabilities. They were then presented with two objects while one of the words was playing, and asked to select the object that went best with the word. The results indicated that participants tended to select responses in proportion to their frequencies, suggesting that people might probability match rather than regularize in learning multiple-referent relations.

The studies outlined in the previous paragraphs suggest that language learners regularize under some circumstances, and probability match under others. However, identifying the inductive biases influencing frequency estimation can be challenging. Without a formal model that translates the inductive biases of learners into explicit predictions about behavior, it can be hard to determine what evidence a particular empirical result provides about those biases. For example, rather than a simple dichotomy between probability matching and regularization, we might imagine that biases towards regularization vary continuously in their strength, with different expectations, task demands, and processing limitations determining the strength of the bias in a given context. A formal model of the effects of inductive biases on frequency estimation would provide a way to make this distinction, and its predictions could be used to design experiments that test whether a given task results in probability matching or just a weaker bias towards regularization.

In this paper, we use a Bayesian model to make explicit the inductive biases that operate during frequency estimation of language forms. This model allows us to characterize the consequences of cultural transmission by *iterated learning* (Kirby, 2001) – the process by which one learner's

[1] While claims about innate constraints on language learning are clearly making statements about the inductive biases of learners, our use of the term should not be interpreted as only reflecting such constraints. Inductive biases relevant to language acquisition could come from a variety of domain-general factors, including innate constraints, a point that we return to in the General Discussion.

linguistic competence is acquired from observations of another learner's productions. The predictions of the model can be explored in the laboratory using an experimental method based on iterated learning (Griffiths, Christian, & Kalish, 2008; Kalish, Griffiths, & Lewandowsky, 2007, 2008). This provides a way to identify the conditions on the inductive biases of individual learners under which cultural transmission results in regularization. We applied this method to a variant on the word-object mapping task studied by Vouloumanos (2008). Our results show that while studying the responses of a single generation of participants does not reveal a bias towards regularization, this bias becomes extremely clear after a few generations. The results have implications for understanding both language evolution and language learning, revealing how weak biases can have a large effect on the languages spoken by a community, and how simulating language evolution in the laboratory can help to make these biases apparent.

The paper is organized as follows. First, we describe the Bayesian model for frequency estimation, and consider how the expectations of learners influence the outcome of iterated learning. We then explore the predictions of the model by conducting three experiments. In the final section, we discuss the implications of these results: that iterated learning can reveal weak biases towards regularization, and that these biases can have strong effects on the structure of languages over time.

## 2. A Bayesian model of frequency estimation

Our goal in studying the estimation of linguistic frequency distributions is to understand how the inductive biases of learners influence their behavior. To satisfy this goal, we need a formalism for describing learning that makes these inductive biases explicit. In this section, we outline how the frequency estimation problem can be solved using methods from Bayesian statistics. This allows us to identify how a rational learner with particular expectations about the nature of the frequency distributions in a language should behave, providing a basis for exploring the effects of these inductive biases on the evolution of frequency distributions and a method for inferring such biases from human behavior. Our focus will be on learning the relative frequencies of word-object associations. However, the models we develop apply to all problems that require learning probability distributions.

Assume that a learner is exposed to $N$ occurrences of a referent (e.g., an object), which is paired with multiple competing linguistic variants with certain probability. We will use the example of estimating the relative frequency of two competing words, but our analysis generalizes naturally to larger numbers of variants, and to variants of different kinds. We will use $x_1$ to denote the frequency of word 1 ($w_1$) and $x_2 = N - x_1$ to denote the frequency of word 2 ($w_2$), and $\theta_1$ and $\theta_2$ to denote the corresponding estimates of the probabilities of these words. The learner is faced with the problem of inferring $\theta_1$ and $\theta_2$ from $x_1$ and $x_2$.

This estimation problem can be solved by applying Bayesian inference. The hypotheses being considered by the learner are all possible values of $\theta_1$ (since $\theta_2$ follows directly from this). The inductive biases of the learner are expressed in a *prior* probability distribution $p(\theta_1)$ over this set of hypotheses, indicating which hypotheses are considered more probable before seeing any data. The degrees of belief that the learner should assign to these hypotheses after seeing $x_1$ are the posterior probabilities $p(\theta_1|x_1)$ given by Bayes' rule

$$p(\theta_1|x_1) = \frac{P(x_1|\theta_1)p(\theta_1)}{\int P(x_1|\theta_1)p(\theta_1)\,d\theta_1} \quad (1)$$

where $P(x_1|\theta_1)$ is the *likelihood*, giving the probability of observing each value of $x_1$ for each value of $\theta_1$.

For the case of two competing words, the likelihood $P(x_1|\theta_1)$ is defined by the Bernoulli distribution, with the probability of $N$ object-word pairings containing $x_1$ instances of $w_1$ is

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

where we assume that $N$ is known to the learner. This likelihood is equivalent to the probability of a sequence of coin flips containing $x_1$ heads being generated by a coin which produces heads with probability $\theta_1$.

Specifying the prior distribution $p(\theta_1)$ specifies the inductive biases of the learners, as it determines the conclusions that a learner will draw when given a particular value for $x_1$. We will assume that the frequency of $w_1$ and $w_2$ have a prior probability distribution given by a Beta distribution with parameters $\frac{\alpha}{2}$. This flexible prior corresponds to the distribution

$$p(\theta_1) = \text{Beta}\left(\frac{\alpha}{2},\frac{\alpha}{2}\right) = \frac{\theta_1^{\frac{\alpha}{2}-1}(1-\theta_1)^{\frac{\alpha}{2}-1}}{B(\frac{\alpha}{2},\frac{\alpha}{2})} \quad (3)$$

where $B(\cdot,\cdot)$ is the beta function (Boas, 1983).

The Beta distribution can take on different shapes depending on the values of $\alpha$. As shown in Fig. 1, when $\alpha/2 = 1$ the density function is simply a uniform distribution. When $\alpha/2 < 1$, the density function is U-shaped and when $\alpha/2 > 1$, it is a bell-shaped unimodal distribution. Thus, despite the apparent complexity of the formula, the Beta distribution captures prior biases that are intuitive from a psychological perspective. For example, when $\alpha/2 < 1$ the prior bias is such that the learner tends to assign high probability to one of two competing variants, consistent with regularization strategies. When $\alpha/2 > 1$, the learner tends to weight both competing variants equally, disfavoring regularization.

Substituting the likelihood from Eq. (2) and the prior from Eq. (3) into Eq. (1) gives the posterior distribution $p(\theta_1|x_1)$. In this case, the posterior is also a Beta distribution, with parameters $x_1 + \frac{\alpha}{2}$ and $N - x_1 + \frac{\alpha}{2}$, due to the fact that the Bernoulli likelihood and Beta prior form a conjugate pair. The mean of this distribution is $\frac{x_1+\frac{\alpha}{2}}{N+\alpha}$, so estimates of $\theta_1$ produced by a Bayesian learner will tend to be close to the empirical probability of $w_1$ in the data, $\frac{x_1}{N}$, for a wide range of values of $\alpha$ provided $N$ is relatively large. Thus, even learners who have quite different inductive biases can be expected to produce similar estimates of $\theta_1$, making

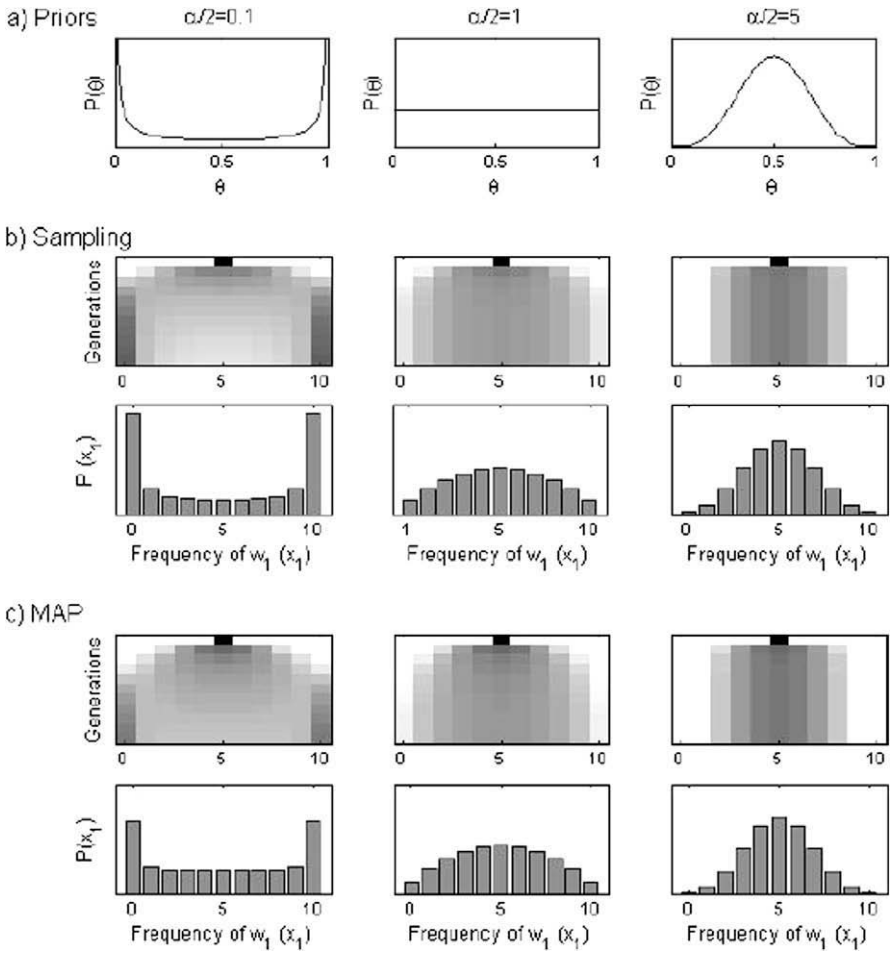


**Fig. 1.** The effects of inductive biases on the evolution of frequencies. (a) Prior distributions on $\theta_1$ for $\frac{\alpha}{2} = 0.1$ (left), $\frac{\alpha}{2} = 1$ (center), $\frac{\alpha}{2} = 5$ (right). Iterated learning by (b) sampling or (c) MAP estimation. Upper panels in (b) and (c) show the changes in the probability distribution on the frequency of $w_1$ (horizontal axis) over several iterations of learning (denoted "Generations" on the vertical axis). The frequency of $w_1$ was initialized at 5 from a total frequency of 10. White cells have zero probability, darker grey indicates higher probability. Lower panels in (b) and (c) show the final probability distribution for each frequency value ($P(x_1)$) after 10 generations.

it difficult to draw inferences about their inductive biases from these estimates.

## 3. Language evolution by iterated learning

Having considered how a single Bayesian learner should solve the frequency estimation problem, we can now explore what happens when a sequence of Bayesian learners each learn from data generated by the previous learner. In learning object-word relations, this corresponds to observing a set of objects being named, making an inference about the relative probabilities of the names, and then producing names for a set of objects which are observed by the next learner. More formally, we assume that each learner is provided with a value of $x_1$ produced by the previous learner, forms an estimate of $\theta_1$ based on this value, and then generates a value of $x_1$ by sampling from $P(x_1|\theta_1)$, with the result being provided to the next learner. The key question is how the biases of the learners influence the outcome of language evolution via this process of iterated learning.

Griffiths and Kalish (2007) analyzed the consequences of iterated learning when learners are Bayesian agents. The first step in this analysis is recognizing that iterated learning defines a Markov chain, with the hypothesis selected by each learner depending only on the hypothesis selected by the previous learner. This means that it is possible to analyze the dynamics of this process by computing a *transition matrix*, indicating the probability of moving from one value of $\theta_1$ to another or one value of $x_1$ to another across iterations, and the asymptotic consequences by identifying the *stationary distribution* to which the Markov chain converges as the number of iterations increases.

Further analysis of this Markov chain requires stating how the posterior distribution is actually translated into an estimate of $\theta_1$. Griffiths and Kalish (2007) identified two such estimation procedures: sampling a hypothesis from the posterior distribution, and choosing the hypothe-

sis with the highest posterior probability. They demonstrated that when learners sample from the posterior, the stationary distribution of the Markov chain on hypotheses is the prior distribution. That is, as the number of iterations increases, the probability of selecting a particular hypothesis converges to the prior probability of that hypothesis. In more intuitive terms, this means that each generation learners will reproduce the frequency distribution in their input in such a way that frequencies gradually move – over the course of generations – towards the frequency distribution specified by the prior. In the case of frequency estimation, this means that we should expect that iterated learning with learners whose priors favor regularization (ie. with $\frac{\alpha}{2} < 1$) will ultimately produce strongly regularized languages.

It is typically more difficult to analyze the case where learners choose the hypothesis with highest posterior probability, known as the maximum *a posteriori* (MAP) hypothesis. The MAP estimate is the same as the posterior predictive distribution, which corresponds to the probability of a variant $i$ given the frequencies in the language. This estimate of $\theta$ is also the mean of the posterior distribution and it corresponds to the case of just estimating the probability with the empirical frequencies, deterministically. Interestingly, in the case of frequency estimation, the Markov chain defined by iterated learning using the MAP estimate is equivalent to a model that has been used in population genetics, the Wright-Fisher model of genetic drift with mutation (Ewens, 2004). This means that the simple learning mechanism based on Bayesian inference outlined in the previous section provides a justification for the use of genetic drift models to study language change. A proof of the equivalence between iterated learning and the Wright-Fisher model is provided in Reali and Griffiths (2008). In addition to providing an explicit connection between biological and cultural evolution, this equivalence makes it possible to use mathematical results from population genetics to identify an approximate stationary distribution on $\theta_1$, which is a Beta distribution with parameters $\frac{\alpha}{1+\frac{\alpha}{N}}$, where $N$ is the total frequency. Unlike the case of sampling, frequencies do not converge to the prior distribution. However, the shape of the stationary distribution depends on the value of priors' parameter $\alpha$. For example, it can be shown that for all values of $\alpha < \frac{N}{N-1}$, the stationary distribution is U-shaped.

The transition matrices associated with these two forms of estimation can also be computed. We will focus on the transition matrices for the values of $x_1$, as these values are easily observed in behavioral data. For the case of sampling, the probability that learner $t$ generates a particular value of $x_1$ given the value generated by learner $t-1$ is given by

$$P(x_1^{(t)}|x_1^{(t-1)}) = \int P(x_1^{(t)}|\theta_1)p(\theta_1|x_1^{(t-1)})d\theta_1 = \binom{N}{x_1^{(t)}} \frac{B\left(x_1^{(t-1)}+x_1^{(t)}+\frac{\alpha}{2}, 2N-x_1^{(t-1)}-x_1^{(t)}+\frac{\alpha}{2}\right)}{B\left(x_1^{(t-1)}+\frac{\alpha}{2}, N-x_1^{(t-1)}+\frac{\alpha}{2}\right)} \quad (4)$$

where $P(x_1^{(t)}|\theta_1)$ is the likelihood from Eq. (2), $p(\theta_1|x_1^{(t-1)})$ is computed by applying Bayes' rule as in Eq. (1), and the final result follows from the fact that the integral is of a standard form used to characterize the beta function (Boas, 1983). For the MAP case, the value of $\theta_1$ produced as an estimate is deterministically related to $x_1^{(t-1)}$, so $P(x_1^{(t)}|x_1^{(t-1)})$ is given by Eq. (2) with $\hat{\theta}_1 = \frac{x_1+\frac{\alpha}{2}}{N+\alpha}$ substituted for $\theta_1$. These transition matrices can be used to compute the probability distribution $P(x_1^{(t)}|x_1^{(0)})$ as a function of the initial frequency of $w_1$, $x_1^{(0)}$, and the number of iterations of learning, $t$. The predictions of the sampling and MAP models are shown in Fig. 1. Consistent with the analysis given above, the figure shows that when the prior distribution is bell-shaped, frequencies of linguistic variants converge over time to a distribution where the probability mass is concentrated around the mean. When the prior is U-shaped, the frequencies converge to a distribution where the probability mass is concentrated in the extremes of the distribution. Under these conditions, the most likely situation is that one variant becomes the vast majority in the population, while the other one becomes very infrequent, regardless of initial conditions. This situation can be interpreted as a regularization process.

The analyses presented in the last two sections support two conclusions. First, since the estimates of $\theta_1$ produced by an individual learner will be only weakly affected by their prior, it can be hard to identify inductive biases by studying individual learners. Second, iterated learning can magnify these weak biases, resulting in rapid convergence to a regular language when learners have priors supporting regularization. The strength of learning biases is determined by the parameters of the prior and is reflected by the speed of convergence. This means that weak biases will produce gradual changes in the distribution of frequencies that may not be obvious in a single generation. These conclusions motivate the three experiments presented in the remainder of the paper. Experiment 1 demonstrates the difficulty of inferring the biases of learners by studying a single generation. Experiment 2 uses an iterated version of the same task to reveal that human learners favor regular languages, and to explore the consequences of this bias for language evolution by iterated learning. The results reveal biases towards regularization that were not obvious in Experiment 1. Experiment 3 is a control study that uses the same iterated learning experimental design but where learners are trained on a non-linguistic task. After observing a sequence of coin flips, participants are asked to predict the outcome of a new sequence. In contrast with Experiment 2, learners show a tendency toward variability that becomes evident in a few generations. This suggests that the regularization bias observed in the object-name matching task is not an artifact of the experimental method.

## 4. Experiment 1: a single generation

The design of Experiment 1 was inspired by Vouloumanos (2008, Experiment 1). The experiment had a training phase where participants were exposed to novel word-object associations and a test phase assessing their knowledge of these associations. However, the design differs from Vouloumanos (2008) in that each word was associ-

ated with just one object, and the test trials consisted of a forced choice between words instead of objects.

### 4.1. Method

#### 4.1.1. Participants

Thirty undergraduates from the University of California, Berkeley, participated in exchange for course credit.

#### 4.1.2. Materials

The materials used in Experiment 1 were the same used in Vouloumanos (2008). The auditory stimuli consisted of twelve words recorded by a native English female speaker. All words consisted of consonant-vowel-consonant syllables with consonants p, t, s, n, k, d, g, b, m, l and vowels æ, i, a, e, ʌ and u. Place of articulation was controlled both between and within words. Word pairs assigned to a common referent (object) were controlled so that they differed in the place of articulation, the vowels and letters they contained. The visual stimuli consisted of six out of the twelve three dimensional objects used in Vouloumanos (2008). The objects differed in color and shape and were animated to move horizontally as a cohesive unit. They were presented in short videos shown on a computer screen.

#### 4.1.3. Design and procedure

The experiment consisted of a training phase followed by a test phase. Participants were instructed that they would learn a novel language. No further information regarding the nature of the study was given in the instructions. During the training block participants were exposed to novel word-object associations. Participants were exposed to 60 training trials in total. Each of the six objects were presented a total of ten times, each time paired with one of two words ($w_1$ and $w_2$) with varying probabilities. The frequency with which each object occurred with $w_1$ and $w_2$ obeyed one of six different conditions. Conditions 0, 1, 2, 3, 4 and 5, corresponded to $w_1$ frequencies of 0, 1, 2, 3, 4 and 5, and $w_2$ frequencies of 10, 9, 8, 7, 6 and 5 respectively. For example, an object assigned to Condition 4 was presented 4 times with $w_1$ and 6 times with $w_2$ in the training phase. A unique pair of $w_1$ and $w_2$ was presented with a unique object. Therefore, the *overall* frequency of a word was determined by the frequency with which it appeared with its referent. Each of the six objects were randomly assigned to one of the six frequency conditions for every participant, so that there was only one object per frequency condition per participant.[2]

The word pairs ($w_1$ and $w_2$) used to refer to each object were also randomized for every participant. On each trial, the object was presented for 3000 ms separated by 3000 ms, and the word was played concurrently with the visual stimuli. In addition to the auditory stimuli, the word was visually presented below the moving object.

The test block consisted of a forced choice selection task. Participants saw one object in the center of the screen and the two words associated with it were visually presented below the object image (bottom left and bottom right). Each object was presented with the two words that co-occurred with it during the training block. Participants were instructed to select one of the two words pressing a key. The position of the word in the screen (left or right) was randomized across trials and participants. The six objects were presented 10 times each to match the number of presentations used in the training block, producing a total of 60 test forced-choice trails. The order of training and test trials was randomized for every participant.

### 4.2. Results and discussion

There was a significant effect of $w_1$ frequency in the training stimuli on mean production of $w_1 (F(5, 29) = 13.32, p < .0001)$. In response to relative frequency values of 0, 1, 2, 3, 4, and 5 in the input, the mean number of $w_1$ in participants' productions were 0.3, 0.9, 1.6, 3.6, 4.7, and 5, respectively. Fig. 2 compares the mean frequencies of $w_1$ produced by participants to the frequencies of $w_1$ in the training stimuli.

As shown in Fig. 2a, the mean frequency of $w_1$ in the productions was close to the corresponding frequencies in the training phase. However, this pattern of performance does not necessarily indicate that participants are probability matching rather than regularizing. The results displayed in Fig. 2a are the group means and they could have resulted from averaging across individuals who each are using only one of the two competing words to name each object. To rule out this possibility, we examined the consistency of production among individual participants. We found that only 6 out of 30 participants regularized all of their productions. The responses of all participants in all conditions are shown in Fig. 2b.

The results of this experiment seem to suggest that people probability match when learning the probabilities with which words can be used to describe objects. These results are consistent with the conclusions of Vouloumanos (2008). However, the formal analyses presented above suggested that it may be difficult to detect a weak bias towards regularization in a single generation of learners, even though such a bias might still have a significant effect on language evolution. Experiment 2 was designed to investigate the possibility that people have biases towards regularization that only emerge over several generations of iterated learning.

## 5. Experiment 2: iterated learning

### 5.1. Method

#### 5.1.1. Participants

Fifty undergraduates from the University of California, Berkeley, participated in exchange for course credit. The participants formed five generations of learners in ten "families". The responses of each generation of learners

[2] As pointed out by one of the anonymous reviewers, the fact that the various frequency conditions vary within participants could affect participants' hypotheses regarding predictability at the language level relative to the word level. Future directions of the present work may include the implementation of between-subject design to test this interesting possibility.

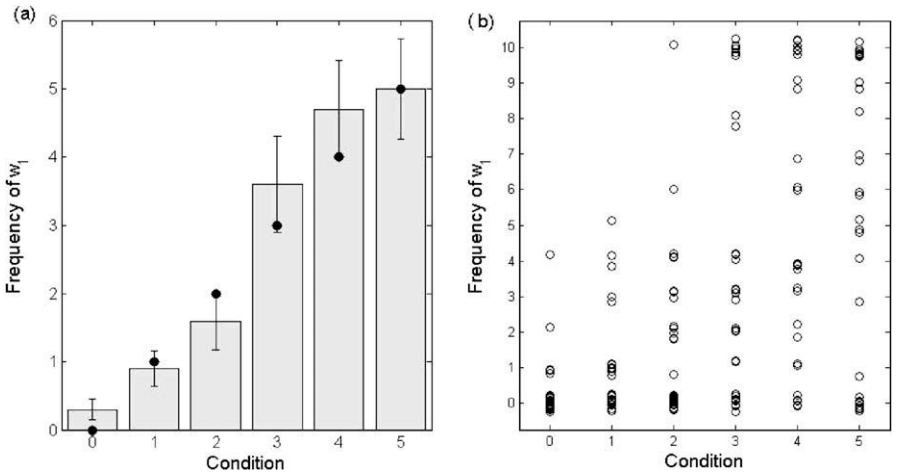


**Fig. 2.** Results of Experiment 1. (a) Mean frequency of $w_1$ selected by participants. Black dots correspond to the frequency of $w_1$ in the training stimuli and error bars indicate one standard error. (b) Individual data points for all 30 participants across all conditions.

during the test phase were presented to the next generation of learners as the training stimuli.

### 5.1.2. Materials

The materials used in Experiment 2 were the same as in Experiment 1.

### 5.1.3. Procedure

For the ten learners who formed the first generation of any family, the methods and procedure of the experiment were identical to Experiment 1. In subsequent generations, the method and procedure were the same, except that the frequency conditions in the training phase were determined by the productions of the previous participant within a family. That is, intergenerational transfer was implemented by letting the frequencies of $w_1$ (and $w_2$) produced by a single participant during the test phase be the frequencies of the training items for the participant in the next generation of that family. Participants were not made aware that their test responses would serve as training for later participants and intergenerational transfer was conducted without personal contact between participants. The actual words and objects used in each condition were assigned randomly for each participant.

## 5.2. Results and discussion

The results of Experiment 2 are shown in Fig. 3. The top row shows participants' productions for each of the ten families. The data is broken down across the six different initial conditions of relative frequency of $w_1$. Across all conditions, the frequencies of $w_1$ moved rapidly towards 0 and 10, reflecting a bias towards regularization. In fact, by the fourth generation, all productions were completely regular.

The sampling and MAP models introduced above were both fit to these data by maximum-likelihood estimation of the parameter $\alpha/2$. The predictions of these models are shown in the middle and bottom rows of Fig. 3. As can be seen from the figure, the models do a good job of capturing the dynamics of iterated learning. For the sampling model, the value of $\alpha/2$ that best fit the data was 0.026, giving a log-likelihood of $-266$, equivalent to a probability of 0.41 of correctly predicting the next value of $x_1$ from the previous one. For the MAP model, the value of $\alpha/2$ that best fit the data was 0.045, with a log-likelihood of $-357$, equivalent to a probability of 0.3 of correctly predicting the next value of $x_1$. These results suggest that the human data are better characterized in terms of learners sampling from their posterior distributions than by MAP estimation.[3] Two aspects of the data are nicely captured by the model. First, as shown in the middle and bottom panels in Fig. 3, the values of $x_1$ selected by learners in early iterations are close to the initial frequency of $w_1$. Thus, the model predicts responses that are consistent with probability matching when a single generation is considered. Second, since the value of $\alpha/2$ that best fit the data is smaller than 1, the best fitting model is one where the prior distribution is U-shaped (see Fig. 1, left panels). This means that the distribution over frequencies should converge to an equilibrium where one variant becomes the vast majority in the population, while the other one becomes very infrequent. Thus, the model predicts regularization of inconsistent language forms as a consequence of learners' inductive biases.

Taken together, Experiments 1 and 2 demonstrate that a bias toward regularization exists that is not obvious in a single generation. More generally, these results show that iterated learning provides a way to test whether weak biases operate during individual learning, as well as the

[3] A possibility not explored in the present study is that participants may use a weaker form of MAP of the kind studied in Kirby, Dowman, and Griffiths (2007), which is defined by raising the posterior distribution to some power. We thank an anonymous reviewer for suggesting this possibility. Future directions of the present work may include the analysis of weaker forms of MAP estimation and its comparison with the sampling model.

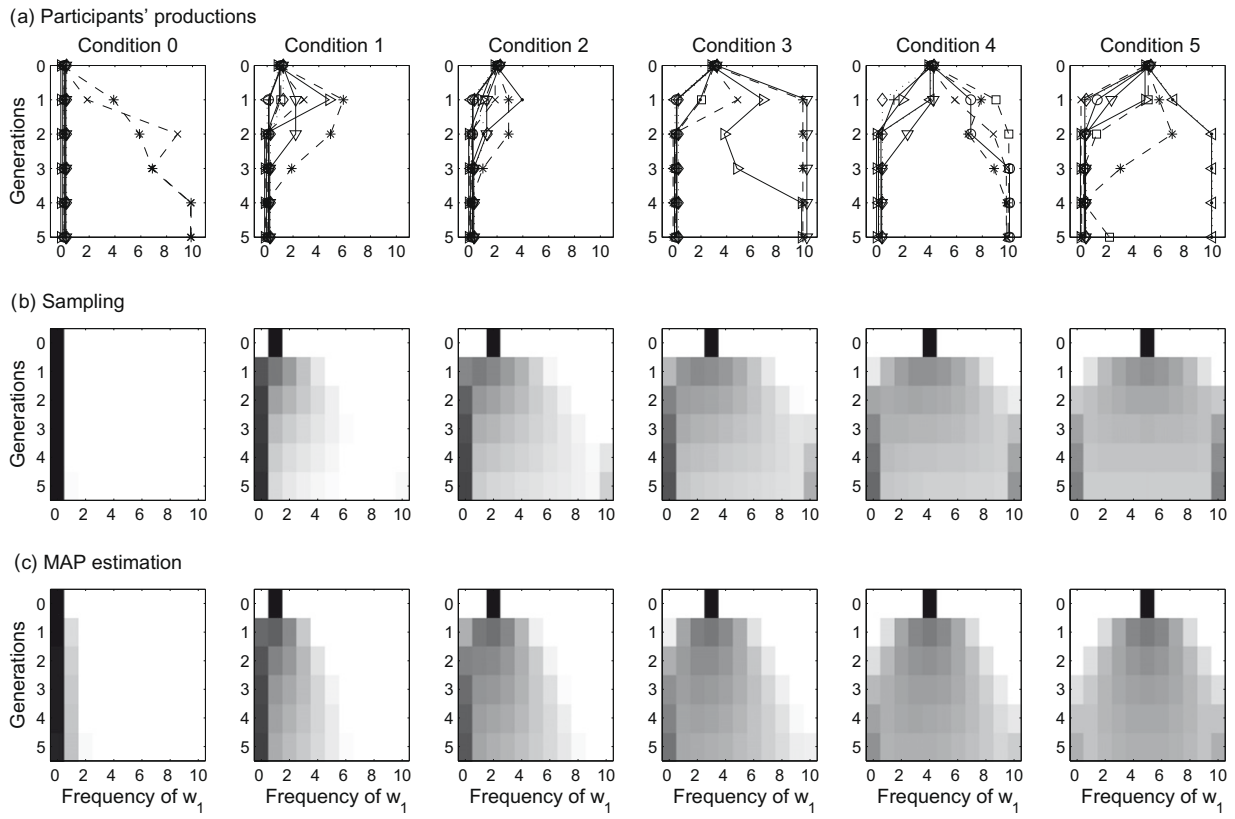


**Fig. 3.** Results of Experiment 2. (a) Frequency of $w_1$ produced by participants (horizontal axis) per generation (vertical axis). Each panel corresponds to increasing values of the frequency of $w_1$ in the input to the first learner (right to left 0, 1 2, 3, 4, 5), and each line to one "family" of participants. Iterated learning with Bayesian agents using (b) sampling and (c) MAP estimation produce predictions in correspondence with these results. White cells have zero probability, darker grey indicates higher probability. The sampling model provides a better account of the participants' responses.

consequences of these biases in shaping the form of languages over time. One possibility is that the priors operating during frequency estimation may vary continuously as a function of task demands, rather than a simple dichotomy between probability matching and regularization.

## 6. Experiment 3: revealing a different kind of prior

A possible objection to the conclusions drawn from Experiment 2 is that the bias toward regularization could be an artifact of the iterated learning experimental paradigm. We are interested in ruling out this possibility regardless of the specific mechanism potentially involved. To do that, we designed a non-linguistic task, in which priors are expected to favor competing variants equally, that is, the inductive biases are not expected to favor regularization. In Experiment 3, participants were exposed to a sequence of coin flips and then asked to predict the outcome of another sequence of coin flips during the test phase. Since participants presumably have experience with fair coins, they are not expected to have a bias towards heads or tails. Rather, priors operating during this task should weight both outcomes equally. Thus, Experiment 3 is a control study designed to test whether a bias that does not favor regularization would be revealed by iterated learning, even under conditions when participants are exposed to a highly uneven number of tails and heads in a sequence of coin flips in the initial generations. Showing that the iterated learning task can produce this bias will illustrate that our previous results on regularization are not merely a consequence of the task, but genuinely reflect the consequences of the learning process.

### 6.1. Method

#### 6.1.1. Participants

Fifty participants took part in the experiment in exchange for course credit or financial compensation of $10/h. Participants were undergraduates from the University of California, Berkeley, or other members of the university community. As in Experiment 2, participants formed five generations of learners in ten families.

#### 6.1.2. Materials

Three different coins were used to produce the sequences of coin flips: a two headed quarter, a two tailed quarter and a regular unbiased quarter. In addition, we used a deck of cards and an unbiased die to familiarize participants with the idea of predicting random processes.

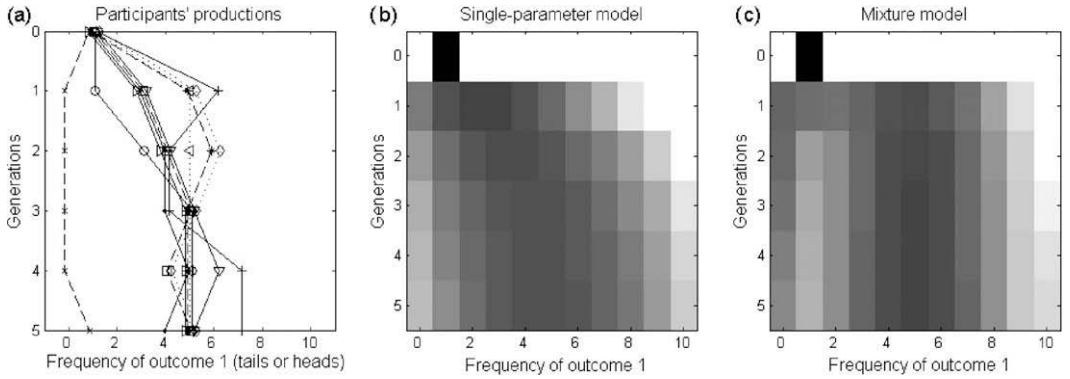


**Fig. 4.** Results of Experiment 3. (a) Frequency of outcome 1 produced by participants (horizontal axis) per generation (vertical axis). Each line corresponds to one "family" of participants. Iterated learning with Bayesian agents using the sampling model with (b) the single-parameter model and (c) the mixture model. White cells have zero probability, darker grey indicates higher probability. The mixture model provides a better account of the participants' responses.

#### 6.1.3. Procedure

The experiment had three consecutive parts, each of them consisting of a training and a production phase. The first two parts consisted of filler trials that served to familiarize participants with the task. In the first part participants were exposed to a sequence of ten random die rolls during the training phase and were asked to predict the outcome of ten die rolls during the production phase. Similarly, in the second part, participants watched a sequence of ten random card draws during the training phase and were asked to predict the card color in ten card draws during the production phase. Participants were not allowed to see the actual outcomes of rolling the die or drawing a card during the production phases.

The third part of the experiment consisted of the experimental trials. During the training phase, participants were exposed to a sequence of ten coin flips. The experimenter used a two headed quarter, a two tailed quarter and a regular quarter to manipulate the relative frequencies of heads and tails in the sequence. The number of heads and tails was manipulated as follows: The experimenter used an regular quarter until the number of tails or heads reached the desired frequency and then secretly switched the coin to either the two headed or two tailed quarter for the remaining of the flips. The ten participants who formed the first generation of any family were exposed to frequencies suggesting a bias in the coin: five participants saw a sequence that contained nine tails and one head, while the other five saw a sequence that contained nine heads and one tail. Since we were using one type of coin for each participant, we could only reproduce one of the frequency conditions used in Experiment 2. We chose nine to one because it provides the strongest test of whether there is a bell-shaped prior, subject to the constraint that participants are exposed to both tails and heads at least once during the training block. During the production (test) phase, participants were asked to predict the outcome of a sequence of ten coin flips with no feedback. The relative frequencies of heads and tails in a sequence predicted by a participant were used as the input in the training phase for the next participant within a family.

### 6.2. Results and discussion

The results of Experiment 3 are shown in Fig. 4. The left panel (Fig. 4a) shows participants' predictions in each of the ten families. We will use the term *outcome 1* to refer to the least frequent outcome during training – heads or tails – in the initial generation. In nine of the ten families, the relative frequency of outcome 1 converged over time towards the center of the distribution. In contrast with the results of Experiment 2, the data show that heads and tails were weighted equally by participants, consistent with a bias that does not favor regularization.

We fit the model introduced above to these data by maximum-likelihood estimation of the parameter $\alpha/2$. In this case, $\theta_1$ and $\theta_2$ correspond to the probability estimates of the two possible outcomes, heads or tails, and $x_1$ and $x_2$ correspond to their relative frequencies. Since the sampling version of the model fit the data better in Experiment 2, we used this version of the model to fit the data in Experiment 3.

The predictions of the model are shown in Fig. 4b. As can be seen from the figure, the model does a good job of capturing the dynamics of iterated learning. The value of $\alpha/2$ that best fit the data was 4.38, giving a log-likelihood of $-89$, equivalent to a probability of 0.16 of correctly predicting the next value of $x_1$ from the previous one.[4] Crucially, the value of $\alpha/2$ is larger than 1, corresponding to a bell-shaped prior distribution where the probability mass is concentrated around the mean (see Fig. 1), which corresponds to a prior that does not favor regularization.

As shown in Fig. 4a, all participants tended to predict extreme frequency values in one of the families, differing significantly from the trajectory observed in the other nine families. The reason for this might be that, within that family, the participant in the first generation predicted ten tails

[4] The probability of correctly predicting the next frequency value is considerably smaller than the probabilities found in the case of word-object mapping. A possible reason for this difference is that there is more intrinsic variability in the coin flipping task, that is, trials are less predictable due to the kind of prior involved.

after being exposed to nine tails and one head. Thus, in subsequent generations, participants were exposed to sequences of "all" tails, and, in turn, predicted extreme frequency values during the production phase. It is conceivable that participants considered the coin to be biased in some way, in which case, their expectations would be consistent with a different kind of prior. To explore this possibility we used a Bayesian model in which the prior was a mixture model, a linear combination of two Beta distributions with different parameters (see the Appendix for details). The prior is then given by the following expression,

$$p(\theta_1) = \pi p_1(\theta|\alpha_1) + (1-\pi)p_2(\theta|\alpha_2) \quad (5)$$

where $\pi$ reflects the weight assigned to the first component, and $\alpha_1$ and $\alpha_2$ characterize the shape of the Beta distribution associated with the two components.

We fit the mixture model to the data by maximum-likelihood estimation of the parameters $\pi$, $\alpha_1/2$, and $\alpha_2/2$. As in the single-parameter case, the probability estimate $\theta$ was derived using sampling. The predictions of the mixture model are shown in Fig. 4c. The values that best fit the data were $\pi = 0.05$, $\alpha_1/2 = 0.31$ and $\alpha_2/2 = 525$, corresponding to log-likelihood of $-79$, equivalent to a probability of 0.2 of correctly predicting the next value of $x_1$ from the previous one. Crucially, the value of one of the parameters ($\alpha_1/2$) is smaller than 1, corresponding to a U-shaped prior distribution that favors regularization, while the value of the second parameter ($\alpha_2/2$) is larger than 1, corresponding to a bell-shaped prior distribution. Thus, the mixture model is capable of accounting for complex inductive biases which assign high prior probability to multiple distinct outcomes. As illustrated in Fig. 4, the mixture model does a better job than the single parameter model of capturing the dynamics of iterated learning, and the improvement in fit obtained by adding the extra parameter is statistically significant ($\chi^2(1) = 20$, $p < .001$).[5]

In sum, Experiment 3 shows that, when learners estimate frequencies in a sequence of coin flips, a prior that does not favor regularization is revealed by iterated learning. This bias clearly differs from the one observed in the linguistic task. This means that iterated learning alone is not sufficient to produce regularization, suggesting that the results of Experiment 2 truly reflect the consequences of the learning process. Moreover, our analysis using the mixture model demonstrates that this method provides a way to identify complex priors, such as those in which multiple outcomes are consistent with people's expectations.

## 7. General discussion

We tested the predictions of our Bayesian model of the evolution of frequency distributions in three experiments. Experiment 1 revealed that when participants were exposed to inconsistent word-meaning mappings, the frequencies determined by their responses were close to the frequencies present in training stimuli. This is consistent with the predictions of our Bayesian model. Moreover, the results are in accord with the pattern of responses reported by Vouloumanos (2008), which showed that participants were sensitive to fine-grained patterns of word-meaning mappings. The results of Experiment 2, however, revealed a trend toward regularization that was not obvious in a single generation. The distribution over competing words converged toward an equilibrium where one of the variants becomes the vast majority in the population. The dynamics of convergence again matched the predictions of our Bayesian model. Experiment 3 shows that the iterated learning task can produce a different kind of bias, suggesting that the results on regularization illustrated in Experiment 2 are not merely a consequence of the task, but reflect the learner's expectations. This pattern of results indicates that weak regularization biases may have a strong effect on how languages evolve over time, and that iterated learning provides an effective method for revealing the inductive biases of human learners.

Our results are consistent with recent studies of creolization that challenge the traditional view that creoles were created in one generation from a rudimentary pidgin as input (Siegel, 2007). This view is compatible with our findings showing that weak biases may only become evident after a number of generations. An important question that remains to be answered, however, is how these inductive biases – represented in our model as a prior distribution – should be interpreted from a psychological viewpoint. One possible interpretation is that the model's prior distribution corresponds to innate constraints specific to language learning. Alternatively, the prior could be interpreted as learning biases affecting the formation of linguistic representations deriving from a number of domain-general innate constraints on learning such as information-processing constraints, resource limitations or the inductive bias associated with some kind of general-purpose learning algorithm. Another interesting possibility is that the biases reflected by the model's prior distribution result from limited cognitive resources, such as working memory limitations that may vary with development or operate differently as a function of task difficulty. The studies mentioned in the introduction (e.g., Hudson Kam & Newport, 2005, in press) showed that the tendency to regularize inconsistent input seems to be related with participant age and task complexity. More precisely, they found that children consistently engage in regularization strategies when exposed to unpredictable variation, while adult's preference toward regularization or probability matching seems to be a function of the task complexity. One interesting possibility is the bias toward regularization might be inherently stronger for children than for adults due to resource-based limitations, as suggested by Hudson Kam and colleagues (Hudson Kam & Chang, submitted for publication). Moreover, if regularization priors resulted from memory constraints, these biases would more evident as a function of the task difficulty. Learning

[5] We ran the mixture model on the data of Experiment 2 and found that the best fitting values of $\alpha_1$ and $\alpha_2$ both matched the value of $\alpha$ that fitted the data best in the case of the single parameter model, and that the more complex model showed no statistically significant improvement in fit.

biases would be perceived as weak when tasks demand less memory resources, and strong otherwise.

In line with recent work on iterated learning (Kalish et al., 2007; Griffiths et al., 2008; Kirby, Cornish, & Smith, 2008) and recent diffusion chain studies in animals (Whiten & Mesoudi, 2008), the experiments presented in this paper suggest that simulating language evolution on the lab provides a way to reveal biases – in this case regularization biases – that might otherwise be hard to detect. Moreover, the results are compatible with recent work in language evolution (Kirby et al., 2007) showing that weak priors may have strong effects on language change over time. Our experiments illustrate how individual learning biases may help explain regular properties found in natural languages, with regular languages emerging quickly for the simple cases we studied. Taken together, our mathematical and empirical analyses suggest that a full understanding of the constraints on language acquisition might require the combination of multiple approaches, including theoretical investigation of language evolution and the simulation of this process in the laboratory.

## Acknowledgements

We thank Athena Vouloumanos for providing the materials used in the experiments, and Carla Hudson Kam and Fei Xu for suggestions. We also thank Aaron Beppu, Matt Cammann, Jason Martin, Vlad Shut and Linsey Smith for assistance in running the experiments. This work was supported by grants BCS-0631518 and BCS-0704034 from the National Science Foundation.

## Appendix A. The mixture model prior

The Bayesian model for frequency estimation can be modified to assume that the prior distribution is a linear combination of two Beta priors with different parameter values as given by Eq. (5). In this case, the posterior distribution is given by

$$p(\theta_1|x_1) = \frac{\pi P(x_1|\theta_1)p(\theta_1|\alpha_1) + (1-\pi)P(x_1|\theta_1)p(\theta_1|\alpha_2)}{\int \pi P(x_1|\theta_1)p(\theta_1|\alpha_1) + (1-\pi)P(x_1|\theta_1)p(\theta_1|\alpha_2)\,d\theta_1} \tag{6}$$

where we simply expand out the terms of the prior $p(\theta_1)$. Substituting this posterior $p(\theta_1|x_1)$ into Eq. (4) we obtain an expression for the transition probability $P(x_1^{(t)}|x_1^{(t-1)})$ for the mixture model,

$$\begin{aligned}
P(x_1^{(t)}|x_1^{(t-1)}) &= \int P(x_1^{(t)}|\theta_1)\frac{\pi P(x_1^{(t-1)}|\theta_1)p(\theta_1|\alpha_1) + (1-\pi)P(x_1^{(t-1)}|\theta_1)p(\theta_1|\alpha_2)}{\int \pi P(x_1^{(t-1)}|\theta_1)p(\theta_1|\alpha_1) + (1-\pi)P(x_1^{(t-1)}|\theta_1)p(\theta_1|\alpha_2)\,d\theta_1}\,d\theta_1 \\
&= \frac{\pi \int P(x_1^{(t)}|\theta_1)P(x_1^{(t-1)}|\theta_1)p(\theta_1|\alpha_1)\,d\theta_1 + (1-\pi)\int P(x_1^{(t)}|\theta_1)P(x_1^{(t-1)}|\theta_1)p(\theta_1|\alpha_2)\,d\theta_1}{\pi \int P(x_1^{(t-1)}|\theta_1)p(\theta_1|\alpha_1)\,d\theta_1 + (1-\pi)\int P(x_1^{(t-1)}|\theta_1)p(\theta_1|\alpha_2)\,d\theta_1} \\
&= \binom{N}{x_1^t}\left[\frac{\pi\frac{B\left(x_1^{(t)}+x_1^{(t-1)}+\alpha_1/2,2N-x_1^{(t)}-x_1^{(t-1)}+\alpha_1/2\right)}{B(\alpha_1/2,\alpha_1/2)}}{\pi\frac{B\left(x_1^{(t-1)}+\alpha_1/2,N-x_1^{(t-1)}\right)}{B(\alpha_1/2,\alpha_1/2)} + (1-\pi)\frac{B\left(x_1^{(t-1)}+\alpha_1/2,N-x_1^{(t-1)}\right)}{B(\alpha_2/2,\alpha_2/2)}} + \frac{(1-\pi)\frac{B\left(x_1^{(t)}+x_1^{(t-1)}+\alpha_2/2,2N-x_1^{t}-x_1^{(t-1)}+\alpha_2/2\right)}{B(\alpha_2/2,\alpha_2/2)}}{\pi\frac{B\left(x_1^{(t-1)}+\alpha_1/2,N-x_1^{(t-1)}\right)}{B(\alpha_1/2,\alpha_1/2)} + (1-\pi)\frac{B\left(x_1^{(t-1)}+\alpha_1/2,N-x_1^{(t-1)}\right)}{B(\alpha_2/2,\alpha_2/2)}}\right]
\end{aligned} \tag{7}$$

via a derivation similar to that used in Eq. (4), where $B(\cdot,\cdot)$ is the beta function (Boas, 1983). We used this expression to fit the values of $\pi$, $\alpha_1/2$, and $\alpha_2/2$ to the experimental data via maximum-likelihood estimation.

## References

Bickerton, D. (1981). *Roots of language*. Ann Arbor, MI: Karoma.
Birdsong, D. (1999). Introduction: Whys and why nots of the Critical Period Hypothesis for second language acquisition. In D. Birdsong (Ed.), *Second language acquisition and the Critical Period Hypothesis* (pp. 178–198). Mahwah, NJ: Lawrence Erlbaum Associates.
Boas, M. L. (1983). *Mathematical methods in the physical sciences* (second ed.). New York: Wiley.
DeGraff, M. (1999). Creolization, language change, and language acquisition: An epilogue. In M. DeGraff (Ed.), *Language creation and language change: Creolization, diachrony, and development* (pp. 473–543). Cambridge: MIT Press.
Ewens, W. (2004). *Mathematical population genetics*. New York: Springer-Verlag.
Gomez, R., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences, 4*, 178–186.
Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2008). Using category structures to test iterated learning as a method for revealing inductive biases. *Cognitive Science, 32*, 68–107.
Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science, 31*, 441–480.
Hudson Kam, C. L., Chang, A. (submitted for publication). Investigating the cause of language regularization in adults: Memory constraints or learning effects?
Hudson Kam, C. L., Newport, E. L. (in press). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*.
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*, 151–195.
Johnson, J., Shenkman, K., Newport, E., & Medin, D. (1996). Indeterminacy in the grammar of adult language learners. *Journal of Memory and Language, 35*, 335–352.
Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin and Review, 14*, 288–294.
Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation, 5*, 102–110.
Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences, 105*, 10681–10686.
Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences, 104*, 5241–5245.
Kroch, A., & Taylor, A. (1997). Verb movement in old and middle english: Dialect variation and language contact. In A. vanKemenade & N. Vincent (Eds.), *Parameters of morphosyntactic change*. Cambridge, England: Cambridge University Press.
Pearl, L., & Weinberg, A. (2007). Input filtering in syntactic acquisition: Answers from language change modeling. *Language Learning and Development, 3*, 43–72.

Pinker, S. (1994). *The language instinct: How the mind creates language*. Harper Collins.
Reali, F., Griffiths, T. (2008). Words as alleles: Equivalence of iterated learning and neutral models from population genetics [Manuscript under review].
Saffran, J. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science, 12*, 110–114.
Senghas, A., & Coppola, M. (2001). Children creating language: How Nicaraguan sign language acquired a spatial grammar. *Psychological Science, 12*, 323–328.
Siegel, J. (2007). Recent evidence against the Language Bioprogram Hypothesis. *Studies in Language, 31*, 51–88.
Singleton, J. L., & Newport, E. L. (2004). When learners surpass their model: The acquisition of american sign language from impoverished input. *Cognitive Psychology, 49*, 370–407.
Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, *107*, 729–742.
Whiten, A., & Mesoudi, A. (2008). Establishing an experimental science of culture: Animal social diffusion experiments. *Philosophical Transactions of the Royal Society B, 363*, 3477–3488.
Wonnacott, E., Newport, E. (2005). Novelty and regularization: The effect of novel instances on rule formation. In *BUCLD 29: proceedings of the 29th annual Boston University conference on language development*. Somerville, MA: Cascadilla Press.