PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2010) **277**, 429–436
doi:10.1098/rspb.2009.1513
Published online 7 October 2009

Words as alleles: connecting language evolution with Bayesian learners to models of genetic drift

Florencia Reali* and Thomas L. Griffiths

Department of Psychology, 3210 Tolman Hall, MC 1650, University of California at Berkeley, Berkeley, CA 94720-1650, USA

Scientists studying how languages change over time often make an analogy between biological and cultural evolution, with words or grammars behaving like traits subject to natural selection. Recent work has exploited this analogy by using models of biological evolution to explain the properties of languages and other cultural artefacts. However, the mechanisms of biological and cultural evolution are very different: biological traits are passed between generations by genes, while languages and concepts are transmitted through learning. Here we show that these different mechanisms can have the same results, demonstrating that the transmission of frequency distributions over variants of linguistic forms by Bayesian learners is equivalent to the Wright–Fisher model of genetic drift. This simple learning mechanism thus provides a justification for the use of models of genetic drift in studying language evolution. In addition to providing an explicit connection between biological and cultural evolution, this allows us to define a 'neutral' model that indicates how languages can change in the absence of selection at the level of linguistic variants. We demonstrate that this neutral model can account for three phenomena: the s-shaped curve of language change, the distribution of word frequencies, and the relationship between word frequencies and extinction rates.

Keywords: language evolution; genetic drift; Bayesian inference; neutral models

1. INTRODUCTION

Natural languages, like species, evolve over time. The mechanisms of language evolution are quite different from those underlying biological evolution, with learning being the primary mechanism by which languages are transmitted between people. However, accounts of language evolution often appeal to forces that have analogues in biological evolution, such as selection or directed mutation. Recent computational work has emphasized the role of selective forces by focusing on the consequences of a language for the 'fitness' of its speakers in terms of communication success (Cavalli-Sforza & Feldman 1983; Hurford 1989; Oliphant 1994; Komarova & Nowak 2001). Other studies have emphasized the effects of differential learnability of competing linguistic variants, with selection or directed mutation operating at the level of sounds, words or grammatical structures (Batali 1998; Pearl & Weinberg 2007; Christiansen & Chater 2008). These functional explanations provide an intuitive and appealing account of language evolution. However, it is possible that the changes we see in languages over time could be explained without appealing to such factors, resulting from processes analogous to genetic drift.

Evaluating the role of selective forces in language evolution requires developing *neutral models* for language evolution, characterizing how languages can be expected to change simply as a consequence of being passed from one learner to another in the absence of selection or directed mutation. Neutral models have come to play a significant role in the modern theory of biological evolution, where they account for variation seen at the molecular level and provide a tool for testing for the presence of selection (Kimura 1983). The work mentioned in the previous paragraph illustrates that there are at least two levels at which evolutionary forces can operate in language evolution: at the level of entire languages (through the fitness of speakers or directed mutation when languages are passed from one speaker to another), and at the level of individual linguistic variants (with particular sounds, words or grammatical structures being favoured over others by learners). In this paper, we define a model that is neutral at the level of linguistic variants, indicating how languages can change in the absence of selection for particular variants.

Defining a model of language evolution that is neutral at the level of linguistic variants requires an account of learning that is explicit about the inductive biases of learners—those factors that make some variants easier to learn than others—so that it is clear that these biases do not favour particular variants. We model learning as statistical inference, with learners using Bayes' rule to combine the clues provided by a set of utterances with inductive biases expressed through a prior distribution over languages. We define a neutral model by using a prior that assigns equal probability to different variants of a linguistic form. While it is neutral at the level of variants, this approach allows for the possibility that learners have more general expectations about the structure of a language—such as the amount of probabilistic variation in the language, and the tendency for new

* Author for correspondence (florencia.reali@gmail.com).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2009.1513 or via http://rspb.royalsocietypublishing.org.

Received 21 August 2009
Accepted 21 September 2009

variants to arise—that can result in forces analogous to directed mutation at the level of entire languages. This allows us to explore the consequences of these expectations in the context of language evolution, determining which phenomena can be explained as a result of high-level inductive biases about the structure of languages without appealing to selective forces at the level of linguistic variants.

The basic learning problem we consider is estimation of the frequencies of a set of linguistic variants. Learning a language involves keeping track of the frequencies of variants of a linguistic form at various levels of representation, including phonology, morphology and syntax. The assumption that learners estimate the probabilities of different variants is shared by many researchers in the area of computational linguistics and sentence processing (e.g. Bod *et al.* 2003) and is supported by a growing body of experimental work in language acquisition (see Saffran 2003 for a review). From this perspective, a learner needs to estimate a probability distribution over variants. We assume priors on such distributions that differ in the amount of variation expected in a language but remain neutral between variants. We translate this Bayesian model of individual learning into a model of language evolution by considering what happens when learners learn from frequencies generated by other learners. The resulting process of 'iterated learning' (figure 1*a*) can be studied to examine the dynamics of language evolution and the properties of the languages that it produces (Kirby 2001; Griffiths & Kalish 2007; Kirby *et al.* 2007).

We show that this simple model of language evolution with Bayesian learners has two surprising consequences. First, it is equivalent to a classic neutral model of allele transmission that is well known in population genetics, the Wright–Fisher model (Fisher 1930; Wright 1931). This equivalence involves treating the linguistic variants as different alleles of a gene, and establishes a mathematical connection between biological and cultural evolution. Second, the model reproduces several basic regularities in the structure and evolution of languages—the s-shaped curve of language change, the power-law distribution of word frequencies and the inverse power-law relationship between word frequencies and extinction rates—suggesting that these regularities can be explained without needing to appeal to forces analogous to selection or directed mutation at the level of linguistic variants.

The plan of the paper is as follows. First, we introduce the Bayesian model of individual learning in more detail, explaining how we define priors corresponding to different expectations about the properties of languages. Next, we show how this model can be used to make predictions about language evolution via iterated learning, and outline the connections to the Wright–Fisher model and the implications of these connections. Finally, we present a series of analytical results and simulations illustrating that the model reproduces the three basic regularities seen in the structure and evolution of human languages mentioned above.

2. MODELING INDIVIDUAL LEARNING

We model individual learning of frequency distributions by assuming that our learners use Bayesian inference, a rational procedure for belief updating that explicitly

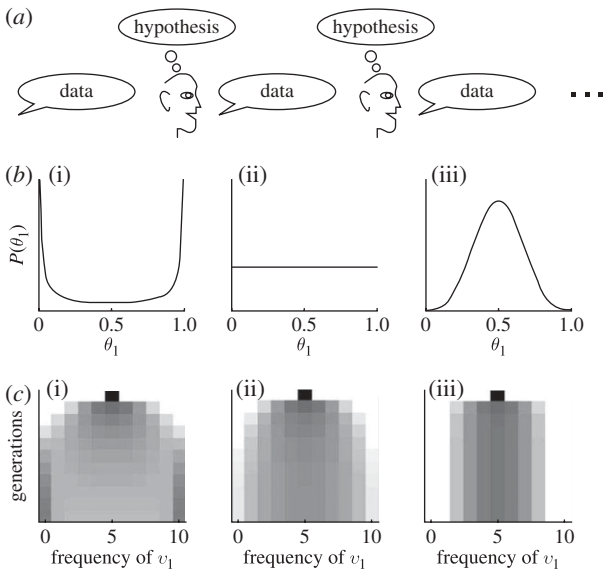

Figure 1. (*a*) Iterated learning: each learner sees data—i.e. utterances—produced by the previous learner, forms a hypothesis about the distribution from which the data were produced, and uses this hypothesis to produce the data that will be supplied to the next learner. (*b*) Prior distribution of θ_1 for the case of two competing variants ($K = 2$), for values of (i) $\alpha/2 = 0.1$, (ii) $\alpha/2 = 1$, (iii) $\alpha/2 = 5$. When $\alpha/2 = 1$ the density function is simply a uniform distribution. When $\alpha/2 < 1$ the prior is such that most of the probability mass is in the extremes of the distribution, favouring the 'regularization' of languages towards deterministic rules. When $\alpha/2 > 1$, the learner tends to weight both variants equally, expecting languages to display probabilistic variation. (*c*) The effects of these expectations on the evolution of frequencies for values of $\alpha/2$ indicated at the top of each column. Each panel shows changes in the probability distribution of one of the two variants (v_1) (horizontal axis) over five iterations (vertical axis). The frequency of v_1 was initialized at $x_1 = 5$ from a total frequency of $N = 10$. White cells have zero probability, darker grey indicates higher probability.

represents the expectations of learners (Robert 1997). Assume that a learner is exposed to N occurrences of a linguistic form, such as a sound, word or grammatical construction, partitioned over K different variants. Let the vectors $\mathbf{x} = [x_1, x_2, \ldots, x_K]$ and $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_K]$ denote the observed frequencies and the estimated probabilities of the K variants, respectively. The learner's expectations are expressed in a prior probability distribution, $p(\boldsymbol{\theta})$. After seeing the data $\mathbf{x}$, the learner assigns posterior probabilities $p(\boldsymbol{\theta}|\mathbf{x})$ specified by Bayes' rule,

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}, \tag{2.1}$$

where $p(\mathbf{x}|\boldsymbol{\theta})$ is the *likelihood*, indicating the probability of observing the frequencies $\mathbf{x}$ from the distribution $\boldsymbol{\theta}$, being the probability of obtaining the frequencies in $\mathbf{x}$ via N draws from a multinomial distribution with parameters $\boldsymbol{\theta}$. Each draw is a statistically independent event. The learner estimates the parameter $\boldsymbol{\theta}$ from a sample of N tokens produced by a speaker before producing any utterances himself. The posterior combines the learner's expectations—represented by the prior—with the

evidence about the underlying distribution provided by the observed frequencies.

We select the prior distribution to be neutral between linguistic variants, with no variant being favoured *a priori* over the others. This assumption differs from other computational models that emphasize the selection or directed mutation at the level of linguistic variants, as discussed above. However, being neutral between variants is not enough to specify a prior distribution: learners can also differ in their expectations about the amount of probabilistic variation in a language. For example, learners facing unpredictable variation may either reproduce this variability accurately or collapse it towards more deterministic rules—a process referred to as *regularization* (Hudson & Newport 2005; Reali & Griffiths 2009). A way to capture these expectations, while maintaining neutrality between variants, is to assume that the prior is a K-dimensional Dirichlet distribution, a multivariate generalization of the Beta distribution (Bernardo & Smith 1994). This is a standard prior used in Bayesian statistics (Bernardo & Smith 1994). In the context of language, Dirichlet priors have been recently used in models of iterated learning (Kirby *et al.* 2007) and language acquisition (Goldwater *et al.* 2009; Reali & Griffiths 2009).

More formally, we assume that the prior $p(\boldsymbol{\theta})$ is a symmetric K-dimensional Dirichlet distribution with parameters α/K, giving

$$p(\boldsymbol{\theta}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \theta_k^{\alpha/K-1}, \tag{2.2}$$

where $\Gamma(\cdot)$ is the generalized factorial function. By using a distribution that is symmetric we maintain neutrality between different variants. When $K = 2$, the prior reduces to a Beta distribution—denoted as Beta($\alpha/2$, $\alpha/2$). The use of the same parameter, α/K, for all variants ensures that the prior does not favour one variant over the others, with the mean of the prior distribution being the uniform distribution over variants for all values of α and K. However, the value of α/K determines the expectations that learners have about probabilistic variation. When $\alpha/K < 1$ the learner tends to assign high probability to one of the K competing variants. This situation reflects a tendency to regularize languages, with probabilistic variation being reduced towards more deterministic rules. When $\alpha/K > 1$, the learner tends to weight all competing variants equally, producing distributions closer to the uniform distribution over all variants (see figure 1*b* for examples). Thus, despite the apparent complexity of the formula, the Dirichlet prior captures a wide range of biases that are intuitive from a psychological perspective.

Some intuitions for the consequences of using different priors can be obtained by considering how they affect the predictions that learners would make about probability distributions. Under the model defined above, the probability that a learner assigns to the next observation being variant k after seeing x_k instances of that variant from a total of N is $(x_k + \alpha/K)/(N + \alpha)$ (see the electronic supplementary material for details). This formula captures two aspects of the learners' behaviour. First, the probability that the learner assigns to a variant is approximately proportional to its frequency x_k. This means that individual variants get strengthened by use. Second, the parameter α/K acts like a number of additional observations of each variant. The largest effect of these additional observations will be when there are no actual observations, with $x_k = 0$. In this case, a learner expecting a more deterministic language (with α/K small) will assign a very small probability to the unobserved variant, while a learner expecting probabilistic variation (with α/K large) will assign it a much higher probability. The prior thus expresses the willingness of learners to consider unobserved variants part of the language.

This model can be extended to cover learning a distribution over an unbounded set, such as the vocabulary of a language. In this case, word production can be viewed intuitively as a *cache* model: each word in the language is either retrieved from a cache or generated anew. Using an infinite-dimensional analogue of the Dirichlet prior (see the electronic supplementary material for details), the probability of a variant that occurred with frequency x_k is $x_k/(N + \alpha)$, while the probability of a completely new variant is $\alpha/(N + \alpha)$. The parameter α thus controls the tendency to produce new variants, as before. There is also a two-parameter generalization of the infinite-dimensional Dirichlet model, which gives a variant that occurred with frequency x_k probability $(x_k - \delta)/(N + \alpha)$, while the probability of a completely new variant is $(\delta K_+ + \alpha)/(N + \alpha)$, where $\delta \in (0,1)$ is a second parameter allowing K_+, the number of variants for which $x_k > 0$, to influence the probability of producing a new variant (see the electronic supplementary material for details).

3. MODELS OF CULTURAL AND BIOLOGICAL TRANSMISSION

(a) *Cultural transmission through iterated learning*

We can now consider the predictions that the model of individual learning defined in the previous section makes about language evolution. As shown in figure 1*a*, we translate this model of frequency estimation into a model of language evolution by assuming that each learner not only observes the frequencies of a set of variants and estimates their probabilities, but then produces a sample of variants from the estimated distribution, with this sample providing the frequencies given to the next learner. This process of iterated learning provides a simple way to examine the consequences of this form of cultural transmission in isolation from other factors that might be involved in language evolution, such as the pressure to create a shared communication system (Komarova & Nowak 2001, 2003; Steels 2003) or the effects of learning from multiple teachers (Niyogi 2006). This emphasis on a single factor is consistent with our goal of providing a simple null hypothesis against which other models can be compared: by identifying which properties of human languages can be produced by iterated learning alone, we can begin to understand when explanations that appeal to other factors are necessary.

More formally, we assume that each learner selects a hypothesis $\boldsymbol{\theta}$ based on the observed data, and then generates the data presented to the next learner—in this case a frequency vector $\mathbf{x}$—by sampling from the distribution $p(\mathbf{x}|\boldsymbol{\theta})$ associated with that hypothesis. We take $\boldsymbol{\theta}$ to be

the distribution used in making predictions about the next variant, with the probability of v_k being $\theta_k = (x_k + \alpha/K)/(N + \alpha)$ (we motivate this choice and discuss the consequences of using other estimators in the electronic supplementary material). This results in a stochastic process defined on estimates $\boldsymbol{\theta}$ and frequency vectors $\mathbf{x}$. Since the frequencies generated by a learner depend only on the frequencies generated by the previous learner, this stochastic process is a Markov chain. It is possible to analyse the dynamics of the process by computing a transition matrix indicating the probability of moving from one frequency value to another across generations, and to characterize its asymptotic consequences by identifying the stationary distribution to which the Markov chain converges as the number of generations increases (for other analyses of this kind, see Griffiths & Kalish 2007; Kirby *et al.* 2007).

It is straightforward to calculate the transition matrix on frequencies for particular values of α, K and N, allowing us to examine how frequencies evolve over generations. Figure 1*c* shows the changes in the frequency of one variant (of two) over five generations, for three different values of α/K. However, providing a general analysis of the dynamics and asymptotic consequences of this process—identifying how languages will change and what languages will be produced by iterated learning—is more challenging. In the remainder of this section, we show that an equivalence between this model of cultural transmission and a model of biological transmission of alleles allows us to use standard results from population genetics to investigate the consequences of iterated learning.

(b) *Biological transmission and the Wright–Fisher model*

The Wright–Fisher model (Fisher 1930; Wright 1931) describes the behaviour of alleles evolving under random mating in the absence of selection. In the simplest form of the model, we have just two alleles which have frequencies x_1 and x_2. If the next generation is produced by a process equivalent to selecting an allele at random from the previous generation and duplicating it, then the values of x_1 in the next generation will come from a binomial distribution in which N draws are taken where the probability of v_1 is $\theta_1 = x_1/N$.[1] Introducing mutation into the model modifies the value of θ_1 (and thus implicitly $\theta_2 = 1 - \theta_1$). In the presence of symmetric mutation, allele types can mutate to any other allele type with equal probability u. Then x_1 in the next generation follows a binomial distribution with N draws and $\theta_1 = ((1 - u)x_1 + u(N - x_1))/N$. This generalizes naturally to K variants, with the frequencies $\mathbf{x}$ being drawn from a multinomial distribution where the probabilities $\boldsymbol{\theta}$ are given by $\theta_k = (x_k\,(1 - u) + (N - x_k)\,u/(K - 1))/N$.

The Wright–Fisher model with symmetric mutation is a neutral model, characterizing how the proportions of alleles change over time as a result of genetic drift and mutation. This model can be analysed by observing that it defines a stochastic process on allele frequencies $\mathbf{x}$ and their corresponding probabilities $\boldsymbol{\theta}$. This stochastic process can be reduced to a Markov chain on $\mathbf{x}$ or $\boldsymbol{\theta}$, with the transition matrix and stationary distribution being used to characterize the dynamics and asymptotic consequences of this process, respectively. Standard results from population genetics show that the stationary distribution of the vector $\boldsymbol{\theta}$ in the K variant case with mutation is approximated by a Dirichlet distribution with parameters $2Nu/(K - 1)$ (see Ewens 2004 for details).

The Wright–Fisher model also extends beyond the case where there are only finitely many allelic variants. A significant amount of research in population genetics has been directed at analysing the Wright–Fisher model with infinitely many alleles (see Ewens 2004). As in the symmetric version of the finite model, every allele can mutate to any other allele with equal probability. Since the number of possible alleles is infinite, the probability of mutation to an existing allele is negligible. Thus, all mutants are considered to represent new allelic types not currently or previously seen in the population. An analytical expression for the stationary distribution over frequencies can also be obtained in this case, known as the Ewens sampling formula (Ewens 1972; see the electronic supplementary material for details).

(c) *Equivalence between cultural and biological transmission*

The Markov chain produced by iterated learning is equivalent to that produced by the Wright–Fisher model of genetic drift, provided we equate linguistic variants with alleles. The basic structure of the biological and linguistic models of the evolution of frequency distributions are the same: in both cases, a value of $\boldsymbol{\theta}$ is computed from the frequencies $\mathbf{x}$, and the next value of $\mathbf{x}$ is obtained by sampling from a multinomial distribution with parameter $\boldsymbol{\theta}$. We can show that these two processes are equivalent by demonstrating that the values taken for $\boldsymbol{\theta}$ are the same in the two processes. With K variants, equivalence to iterated learning by Bayesian learners can be shown by taking the mutation rate to be $u = (\alpha/(K - 1))/(\alpha + N))K$. In this case, the frequency estimate for allele k is $\theta_k = (x_k + \alpha/K)/(N + \alpha)$, identical to the estimate of $\boldsymbol{\theta}$ derived in the iterated learning model. Note that the equivalence holds in general and not for a restricted set of parameters. This is because for each value of mutation rate u, there exists values of α and K satisfying the condition for equivalence. It is easy to see that the reciprocal is true also. We can thus use results from population genetics to characterize the dynamics and stationary distribution of the Markov chain defined by iterated learning, indicating what kind of languages will emerge over time.

Other recent work has pointed out connections between the Wright–Fisher model in population genetics and theories of language change (Baxter *et al.* 2006). The model presented here differs from this previous work in a number of ways. First, our learning model provides a way to explicitly relate language change to individual inductive biases. This allows us to investigate the consequences of iterated learning by manipulating the parameters of the prior. For example, in the case of K variants, the stationary distribution of each probability θ_k is approximated by a Dirichlet distribution with parameters $2\alpha/K(1 + \alpha/N)$. The stationary distribution can now be interpreted as follows: languages in which a single variant dominates are favoured when the prior

parameters meet the condition $2\alpha/K(1 + \alpha/N) < 1$, while languages in which all variants are weighted equally will prevail when $2\alpha/K(1 + \alpha/N) > 1$.[2]

Another way in which our approach differs from previous related work is that our mathematical formulation allows us to generalize the biological–linguistic equivalence to the case of an unbounded number of variants.[3] Following an argument similar to that for the finite case, iterated learning with Bayesian learners considering distributions over an unbounded vocabulary can be shown to be equivalent to the Wright–Fisher model for infinitely many alleles (see the electronic supplementary material for a detailed proof). The stationary distribution for this process is thus the Ewens sampling formula. The two-parameter generalization discussed above has not previously been proposed in population genetics, and its stationary distribution is unknown. Consequently, we assess the predictions of this model through computer simulation, although a related analytic result is presented in the electronic supplementary material.

4. IMPLICATIONS FOR THE PROPERTIES OF LANGUAGES

Determining how languages change purely as a result of iterated learning allows us to explore the forces that drive language evolution. Our model is neutral with respect to both selection and directed mutation at the level of linguistic variants: no fitness-related factor is associated with any of the competing linguistic variants, and we assume symmetric mutation between them. Thus, our analysis of iterated learning identifies a 'null hypothesis' that is useful for evaluating claims about the importance of selective pressures at this level in accounting for statistical regularities found in the form and evolution of languages, playing a role similar to that of neutral models of genetic drift such as the Wright–Fisher model in biology (Kimura 1983). The model also allows for a kind of directed mutation at the level of entire languages, with expectations about the amount of probabilistic variation in a language shaping the structure of that language over time. These expectations play a role analogous to setting the mutation rate in the Wright–Fisher model. The equivalence between these models implies that the outcome of the transmission of linguistic variants will be the same as that expected under the neutral version of the Wright–Fisher model. In this section, we show that this model can account for three basic regularities in the form and evolution of languages.

(a) *S-shaped curves in language change*

When old linguistic variants are replaced by new ones, an s-shaped curve is typically observed in plots of frequency against time (Yang 2001; Pearl & Weinberg 2007). This phenomenon has been documented in numerous studies of language change (e.g. Weinreich *et al.* 1968; Bailey, 1973; Kroch 1989). An example is the way in which modern verb–object (VO) word order gradually replaced object–verb (OV) in English (Pearl & Weinberg 2007; Clark *et al.* 2008). Speakers' preferences shifted from OV phrases such as *you God's commandment keep will* in Old English to modern VO phrases such as *you will keep God's commandment* (Clark *et al.* 2008). Existing

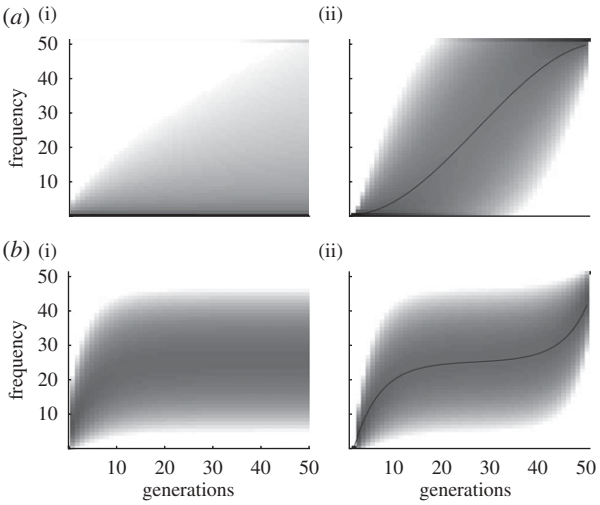

Figure 2. Changes in the probability (vertical axis) of a new variant (v_1) over 50 iterations of learning (horizontal axis) as a function of the value of α. Total frequency of v_1 and v_2 was $N = 50$, but the same effects are observed with larger values of N. (*a*) Changes in the probability of v_1 using $\alpha = 0.05$ corresponding to a prior that favours regularization. (i) probability changes when conditioning on initial frequency only ($x_1 = 0$), (ii) changes in the probability of v_1 when conditioning on both initial frequency ($x_1 = 0$) and final frequency ($x_1 = 50$), corresponding to the situation when the new variant (v_1) eventually takes over the language. Under these conditions, s-shaped curves are observed, consistent with historical linguistic data. (*b*) Changes in the probability of v_1 when $\alpha = 10$ is used, corresponding to a prior that favours probabilistic variation. (ii) Case conditioning on initial frequency only ($x_1 = 0$), (i) case of conditioning on both initial ($x_1 = 0$) and final ($x_1 = 50$) frequencies, illustrating that the appearance of the s-shaped curve depends on the expectations of the learners. White cells have zero probability, darker grey indicates higher probability.

models of this phenomenon have typically assumed that the data used by the learner are filtered in some way (Pearl & Weinberg 2007) or explore the effect of the different distributions with which sentences are presented to the learner (Niyogi & Berwick 1997). Other models have assumed that the emerging linguistic variant has functional advantages over the replaced one (Yang 2001; Christiansen & Chater 2008; Clark *et al.* 2008). By contrast, our neutral model assumes that the learner uses the entire input and that variants carry no functional advantages. Consider the specific case of two competing variants v_1 and v_2, such as the VO and OV word orders.[4] When learners have a prior near or below the threshold for favouring regularization ($\alpha < N/(N - 1)$), the model predicts that v_1 frequency should gradually converge to extreme values, meaning that v_1 can emerge even when its initial frequency is zero (figure 2). Similar to the case of biological genetic drift, the probability of a variant appearing and going to fixation is very small for large values of N. However, historical documentation of the s-shaped curve comes from cases where a change is observed, which corresponds to conditioning on fixation taking place. We therefore restrict our analyses to such cases. When we condition on v_1 eventually taking over the language, the trajectory of its frequencies follows an s-shaped curve, provided learners have priors favouring

regularization. This suggests that the dynamics of language change partly reflect the expectations that human learners have about the form of languages.

(b) *Emergence of power-law distributions*

One of the most striking statistical properties of human languages is that word frequencies follow a power-law distribution (Zipf 1949). More formally, the probability that a word has frequency x_k, is $p(x_k) \propto x_k^{-\gamma}$. Thus, when log $p(x_k)$ is plotted as a function of log x_k a linear relationship is observed with slope $-\gamma$. What are the forces that underlie the emergence of power-law distributions in language? One proposal is that they are a consequence of ambiguity constraints in communication systems (Ferrer i Cancho 2005). According to another view, power laws are evidence that large and redundant vocabularies may have evolved as a result of sexual selection (Miller 2000). Recently, some patterns of verb use have been used to critically explore how power-law distributions could have emerged as a consequence of iterated learning, providing some challenges to this approach (Briscoe 2008).

We conducted simulations to show that languages with this property are naturally produced by the infinite version of the neutral model. Simulations consisted of the implementation of the stochastic process described above: For each simulated learner, the posterior predictive distribution is used as the estimate $\boldsymbol{\theta}$ of the frequency of each word, and this estimate is used to produce words that serve as input to the next learner (see the electronic supplementary material for details). In this context, a 'variant' represents a word type and each occurrence of a variant corresponds to a token. Thus, for each generation, the size of the vocabulary is given by the number of different variant types that co-exist in the population. Word frequencies were initialized so that all tokens correspond to a single type, and the sampling process was repeated until frequencies stabilized. As shown in figure 3, word frequencies converge to power-law distributions that are consistent with those observed in natural languages. Comparison of the distribution over frequencies produced by the two-parameter model with data from a corpus of child-directed speech (Bernstein-Ratner 1984) shows that the model produces a power law with exponent $\gamma = 1.74$, providing a close match with that estimated from a corpus ($\gamma = 1.7$).

(c) *Frequency effects in lexical replacement rates*

As languages evolve, new words replace old ones. Recent work shows that frequency of use predicts the rates at which verbs change from irregular to regular forms (Lieberman *et al.* 2007) as well as word replacement rates in Indo-European languages (Pagel *et al.* 2007). Frequently used words are replaced much more slowly than less frequent ones, as revealed by an inverse power-law relationship between frequency of use and replacement rate (Lieberman *et al.* 2007; Pagel *et al.* 2007). For example, words such as *one*, *who* or *night* are at the slow end of the distribution, while words such as *dirty* or *guts* are replaced at a much faster rate (Lieberman *et al.* 2007). It has been proposed that some form of selection may be responsible for the slow rate of replacement

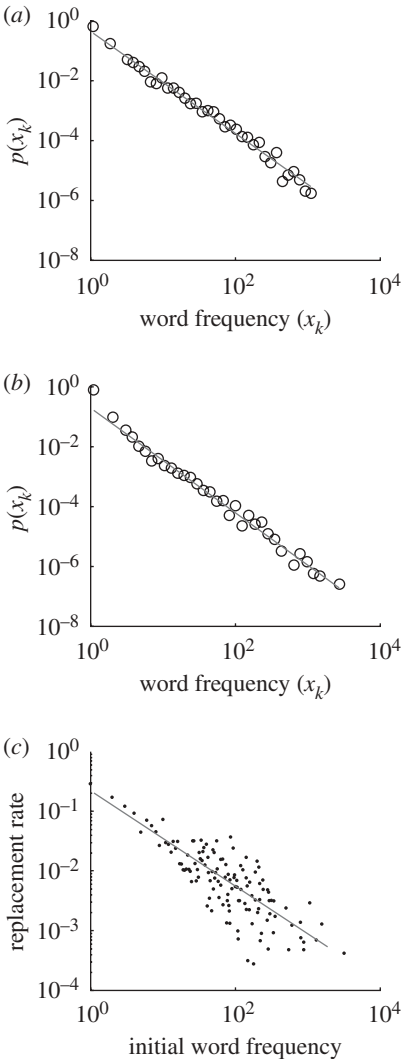

Figure 3. (*a*) Power-law distribution on word frequencies obtained from corpus data, consisting of $N = 33\,399$ word tokens. Horizontal axis corresponds to word frequency (x_k) in log scale, and the vertical axis corresponds to the probability $p(x_k)$ that a certain word-type would fall within the bin at that frequency level. A power-law distribution is indicated by a linear relationship with slope $\gamma = 1.70$ (Bernstein-Ratner corpus). (*b*) Iterated learning using a two-parameter Poisson–Dirichlet distribution as a prior on distributions over infinitely many variants also produces a power-law relationship with $\gamma = 1.74$. Simulations were implemented by sampling over a population of arbitrarily assigned 33 399 numerical word tokens to match the size of the corpus. Frequencies were initialized by setting all word tokens to the same unique type. The frequency distribution stabilized after 10 000 iterations of learning, and the result shown here reflect the distribution produced by a single learner after 20 000 iterations. We ran the simulations across a range of values of δ (from 0.1 to 1, with steps of 0.1), and the value of α was set to 10 (see the electronic supplementary material, for details). Simulations with $\delta = 0.3$ produced the closest match to the corpus data, and this is the case shown in the figure model. (*c*) Initial lexical frequency x_k plotted against the replacement rate, estimated as $r = 1/t$, where t is the number of iterations before absorption (i.e. $x_k = 0$). For each frequency value, time of absorption was directly measured over 5000 iterations after frequencies reached a steady state. The resulting linear relationship on a log–log plot reflects an underlying power law with $\gamma = 0.8$ (the correlation between log frequency and log replacement rate is $r = -0.81$, $p < 0.00001$).

of highly frequent words (Pagel *et al.* 2007). However, the neutral model is sufficient to account for the frequency effect. In the infinite case, mutation occurs only to new variants, thus, all variants are eventually lost from the population. A new cognate is represented by a new variant. Replacement happens when the variant corresponding to the old cognate becomes extinct. The case of verb regularization is modelled by assuming that irregular and regular verbs coexist as two variants among other words in the vocabulary. Extinction of the irregular verb happens when the regular form replaces completely the irregular one. Time of absorption is given by the number of iterations of learning before the frequency of a variant drops to zero. Analytic results can be obtained for this quantity in the one-parameter model (see the electronic supplementary material), and these results and our simulations indicate that replacement rate—calculated as the inverse of the absorption time—follows an inverse power-law relationship with frequency (figure 3*c*).

5. CONCLUSION

The idea that models of biological evolution can be used to shed light on processes of cultural evolution is not new (Komarova & Nowak 2003; Bentley *et al.* 2004; Fontanari & Perlovsky 2004). Likewise, models of genetic drift have been applied to cultural evolution in the past (Bentley *et al.* 2004; Baxter *et al.* 2006) and even offered as a possible account of the distribution of word frequencies (Fontanari & Perlovsky 2004; see the electronic supplementary material for further discussion of this related work). However, much of this previous work has used biological models as a metaphor to study the effects of random copying and selective processes when applied to the case of cultural transmission. In contrast, we have started with a simple learning mechanism—Bayesian inference—and shown that this mechanism provides a direct justification for the use of models of genetic drift as an account of how languages can change over time in the absence of selection at the level of linguistic variants. This is because our model provides an explicit connection between models of random copying and mutation and individual inductive biases. More precisely, the effect of iterated learning on language change is investigated by manipulating the parameters of the prior, allowing us to explore the consequences of learners having different expectations while maintaining neutrality between variants. While this analysis of one of the most basic aspects of language—the frequencies of different variants—emphasizes the connections between biological and cultural evolution, it also illustrates that the models developed in population genetics cover only one small part of the process of cultural evolution. We anticipate that developing neutral models that apply to the transmission of more richly structured aspects of language will require developing a deeper understanding of the mechanisms of cultural transmission—in this case, language learning.

We thank A. Bouchard-Côté, M. Ettlinger, P. Liang and J. McClelland for comments on the manuscript. This work was supported by grant numbers BCS-0631518 and BCS-070434 from the United States National Science Foundation.

ENDNOTES

[1]Our description corresponds to a 'haploid' version of the Wright–Fisher model, with just one allele per organism. In the more conventional diploid model, an additional factor of 2 appears in front of N, as N is taken to be the number of organisms rather than the number of alleles.

[2]Note that, under certain conditions, the shape of the stationary distribution does not correspond exactly to the shape of the prior. For example, as illustrated in figure 1 for the case of $K = 2$ and $N = 10$, a uniform prior given by $\alpha/2 = 1$ is associated with a bell-shaped stationary distribution.

[3]As suggested by an anonymous reviewer, it is possible that the model in Baxter *et al.* (2006) admits an extension to accommodate the infinite case, but this possibility has not been explored in previous work.

[4]In our example we assume that languages can only maintain one kind of word order (and that there are only two possibilities). This is simplifying assumption, as illustrated by languages such as German where word order in subordinate clauses can differ from word order in main clauses.

REFERENCES

Bailey, C. J. 1973 *Variation and linguistic theory.* Washington, DC: Center for Applied Linguistics.

Batali, J. 1998 Computational simulations of the emergence of grammar. In *Approaches to the evolution of language: social and cognitive bases* (eds. J. Hurford & M. Studdert-Kennedy), Cambridge, UK: Cambridge University Press.

Baxter, G. J., Blythe, R. A., Croft, W. & McKane, A. J. 2006 Utterance selection model of language change. *Phys. Rev. E* **73**, 046118. (doi:10.1103/PhysRevE.73.046118)

Bentley, R., Hahn, M. W. & Shennan, S. J. 2004 Random drift and culture change. *Proc. R. Soc. Lond. B* **271**, 1443–1450. (doi:10.1098/rspb.2004.2746)

Bernardo, J. M. & Smith, A. F. M. 1994 *Bayesian theory.* Chichester, UK: Wiley.

Bernstein-Ratner, N. 1984 Patterns of vowel modification in Motherese. *J. Child Lang.* **11**, 557–578.

Bod, R., Hay, J. & Jannedy, S. 2003 *Probabilistic linguistics.* Cambridge, MA: MIT Press.

Briscoe, T. 2008 Language learning, power laws, and sexual selection. *Mind Soc. Cogn. Stud. Econ. Soc. Sci.* **7**, 65–76.

Cavalli-Sforza, L. L. & Feldman, M. W. 1983 Paradox of the evolution of communication and of social interactivity. *Proc. Natl Acad. Sci. USA* **80**, 2017–2021. (doi:10.1073/pnas.80.7.2017)

Christiansen, M. H. & Chater, N. 2008 Language as shaped by the brain. *Behav. Brain Sci.* **31**, 489–558. (doi:10.1017/S0140525X08004998)

Clark, B., Goldrick, M. & Konopka, K. 2008 Language change as a source of word order correlations. In *Variation, selection, development: probing the evolutionary model of language change* (eds R. Eckardt & G. Jäger, and Veenstra), pp. 75–102. Berlin, Germany: Mouton de Gruyter.

Ewens, W. J. 1972 The sampling theory of selectively neutral alleles. *Theor. Pop Biol.* **3**, 87–112. (doi:10.1016/0040-5809(72)90035-4)

Ewens, W. J. 2004 *Mathematical population genetics.* New York, NY: Springer.

Ferrer i Cancho, R. 2005 Decoding least effort and scaling in signal frequency distributions. *Physica A* **345**, 275–284.

Fisher, R. A. 1930 *The genetical theory of natural selection.* Oxford, UK: Clarendon Press.

Fontanari, J.F. & Perlovsky, L. I. 2004 Solvable null model for the distribution of word frequencies. *Phys. Rev.* **E70**, 042901. (doi:10.1103/PhysRevE.70.042901)

Goldwater, S., Griffiths, T. L. & Johnson, M. 2009 A Bayesian framework for word segmentation: exploring the

effects of context. *Cognition* **112**, 21–54. (doi:10.1016/j.cognition.2009.03.008)
Griffiths, T. L. & Kalish, M. L. 2007 Language evolution by iterated learning with Bayesian agents. *Cogn. Sci.* **31**, 441–480.
Hudson Kam, C. & Newport, E. 2005 Regularizing unpredictable variation: the roles of adult and child learners in language formation and change. *Lang. Learn. Dev.* **1**, 151–195. (doi:10.1207/s15473341lld0102_3)
Hurford, J. R. 1989 Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua* **77**, 187–222. (doi:10.1016/0024-3841(89)90015-6)
Kimura, M. 1983 *The neutral theory of molecular evolution.* Cambridge, UK: Cambridge University Press.
Kirby, S. 2001 Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE J. Evol. Comput.* **5**, 102–110. (doi:10.1109/4235.918430)
Kirby, S., Dowman, M. & Griffiths, T. L. 2007 Innateness and culture in the evolution of language. *Proc. Natl Acad. Sci. USA* **104**, 5241–5245. (doi:10.1073/pnas.0608222104)
Komarova, N. L. & Nowak, M. A. 2001 Natural selection of the critical period for language acquisition. *Proc. R. Soc. Lond. B* **268**, 1189–1196. (doi:10.1098/rspb.2001.1629)
Komarova, N. L. & Nowak, M. A. 2003 Language dynamics in finite populations. *J. Theor. Biol.* **221**, 445–457. (doi:10.1006/jtbi.2003.3199)
Kroch, A. 1989 Reflexes of grammar in patterns of language change. *Lang. Var. Change* **1**, 199–244. (doi:10.1017/S0954394500000168)
Lieberman, E., Michel, J., Jackson, J., Tang, T. & Nowak, M. 2007 Quantifying the evolutionary dynamics of language. *Nature* **449**, 713–716. (doi:10.1038/nature06137)
Miller, G. 2000 *The mating mind: how sexual choice shaped the evolution of human nature*. London, UK: William Heinemann.
Niyogi, P. 2006 *The computational nature of language learning and evolution.* Cambridge, MA: MIT Press.
Niyogi, P. & Berwick, R.C. 1997 A dynamical systems model for language change. *Complex Syst.* **11**, 161–204.
Oliphant, M. 1994 The dilemma of Saussurean communication. *BioSystems* **37**, 31–38. (doi:10.1016/0303-2647(95)01543-4)
Pagel, M., Atkinson, Q. D. & Meade, A. 2007 Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* **449**, 717–720. (doi:10.1038/nature06176)
Pearl, L. & Weinberg, A. 2007 Input filtering in syntactic acquisition: answers from language change modeling. *Lang. Learn. Dev.* **3**, 43–72. (doi:10.1207/s15473341lld0301_2)
Reali, F. & Griffiths, T. L. 2009 The evolution of frequency distributions: relating regularization to inductive biases through iterated learning. *Cognition* **111**, 317–328. (doi:10.1016/j.cognition.2009.02.012)
Robert, C. P. 1997 *The Bayesian choice: a decision-theoretic motivation.* New York, NY: Springer.
Saffran, J. 2003 Statistical language learning: mechanisms and constraints. *Cur. Directions Psychol. Sci.* **12**, 110–114. (doi:10.1111/1467-8721.01243)
Steels, L. 2003 Evolving grounded communication for robots. *Trends. Cogn. Sci.* **7**, 308–312. (doi:10.1016/S1364-6613(03)00129-3)
Weinreich, U., Labov, W. & Herzog, M. 1968 In *Directions for historical linguistics* (eds W. Lehmannand & Y. Malkiel), pp. 97195. Austin, TX: University of Texas Press.
Wright, S. 1931 Evolution in Mendelian populations. *Genetics* **16**, 97–159.
Yang, C. 2001 Internal and external forces in language change. *Lang. Var. Change* **12**, 231–250. (doi:10.1017/S0954394500123014)
Zipf, G. K. 1949 *Human behavior and the principle of least-effort.* Cambridge, MA: Addison-Wesley.