Journal of Theoretical Biology 259 (2009) 552–561

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Genetic biasing through cultural transmission: Do simple Bayesian models of language evolution generalise?

Dan Dediu *

Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6500 AH Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 5 December 2008
Received in revised form
22 March 2009
Accepted 3 April 2009
Available online 14 April 2009

Keywords:
Language evolution
Computer model
Bayesian agents

ABSTRACT

The recent Bayesian approaches to language evolution and change seem to suggest that genetic biases can impact on the characteristics of language, but, at the same time, that its cultural transmission can partially free it from these same genetic constraints. One of the current debates centres on the striking differences between *sampling* and *a posteriori maximising* Bayesian learners, with the first converging on the prior bias while the latter allows a certain freedom to language evolution. The present paper shows that this difference disappears if populations more complex than a single teacher and a single learner are considered, with the resulting behaviours more similar to the *sampler*. This suggests that generalisations based on the language produced by Bayesian agents in such homogeneous single agent chains are not warranted. It is not clear which of the assumptions in such models are responsible, but these findings seem to support the rising concerns on the validity of the "acquisitionist" assumption, whereby the locus of language change and evolution is taken to be the first language acquirers (children) as opposed to the competent language users (the adults).

1. Introduction

The role of genetic biases in language evolution and change is a very important topic (Kirby et al., 2007; Dediu and Ladd, 2007; Dediu, 2008b) because, on one hand, it is generally accepted that human language has some sort of genetic foundations but, on the other, it is not clear how specific or strong such biases must be.

More specifically, languages across the world show striking similarities as well as an amazing range of variation. Among these similarities are the definitional properties of language—Hockett (1963, 's) *design features*—which include, for example, *duality of patterning* (combinations of meaningless phonemes produce meaningful morphemes), *discretness* of the basic linguistic units (e.g., sounds) and *arbitrariness* of mapping between signal and meaning. Also included among these similarities are *language universals* (Greenberg, 1966), which are not part of the definition of language but seem to be true of all (called *absolute universals*) or most (*statistical universals* or *tendencies*) attested languages (e.g., the *word order universals* which constrain the relations between the ordering of various constituents like object, subject and verb).

However, languages also differ in myriad ways, ranging from the number of consonants (from 6 in *Rotokas* to 122 in *!Xóõ*; Maddieson, 2008a) and vowels (from 2 in *Yimas* to 14 in *German*; Maddieson, 2008c), the usage or not of voice pitch (fundamental frequency) to convey lexical or grammatical distinctions (slightly more than half the world's languages are *tone* languages, like Chinese and Yoruba; Yip, 2002; Maddieson, 2008b), the conceptualisation of space (Levinson, 2003) or the canonical order of subject and verb (Dryer, 2008).

One of the fundamental questions generated by this concerns the interplay between linguistic diversity and its fundamental constraints: what forces can produce such a bewildering variety of languages which can all be acquired by children using essentially the same "hardware" (albeit with large inter-individual variation in the actual implementation)? As hinted above, part of the answer must concern *genetics*, part of it must concern *individual learning* and yet another essential part must concern the process of *cultural transmission* across generations.

At one extreme of the spectrum of proposed explanations, it is suggested that genetic mechanisms are extremely specific and strong, coding for some constrained parameter values of a so-called "Universal Grammar" (UG) while, at the other, these influences are conceptualised as non-specific and very weak (for an overview see, e.g., Kirby et al. (2007) or Christiansen and Chater (2008) and associated comments).

In the first account, these genetic factors act at the scale of the individual by forcing the language to fit inside pre-determined constraints during language acquisition (Lightfoot, 1999), so that both the similarities across languages and the range of possible variation are a *direct* result of innate mechanisms. However, in the second account, the genetic biases act through the cultural

* Tel.: +31 643101537.
E-mail address: Dan.Dediu@mpi.nl

0022-5193/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2009.04.004

transmission of language across many generations (Dediu, 2008b; Dediu and Ladd, 2007; Kirby et al., 2007), so that languages are shaped by genetic and communicative constraints simultaneously, resulting in different solutions to similar problems, as well as contingent variation. As a consequence, in the UG account, the nature and strength of the genetic biases can be more or less directly inferred from the distribution of typological features, while in the second account the relationship between biases and distributional properties of languages is more indirect and complex (Kirby et al., 2007;Dediu and Ladd, 2007).

In this second category falls Robert Ladd's and my recent work relating tone languages and the derived haplogroups of *ASPM* and *Microcephalin*, where we propose that a very small genetic bias at the individual level can impact on the trajectory of language change through cultural transmission, influencing the distribution of tone and non-tone languages across the world (Dediu and Ladd, 2007; Ladd et al., 2008). Our proposal, thus, could allow a better understanding of the complex interplay between weak genetic biases and the process of cultural transmission in shaping language diversity and provide empirical arguments in favour of this account.

A popular methodology for studying the influence of cultural transmission on language evolution and change is represented by the *iterated learning model* or *ILM* (Kirby and Hurford, 2002), whereby naive agents learn their language using primary data produced by the previous generation. This methodology has been applied mainly to simulated agents, increasing our understanding of, for example, the emergence of compositionality (Kirby and Hurford, 2002), but also recently to real human participants (Kirby et al., 2008) and even bird song (Feher et al., 2008). A recent development of the ILM concerns the treatment of the agents as *Bayesian learners* (Kirby et al., 2007; Griffiths and Kalish, 2007; Smith and Kirby, 2008), which holds the promise to allow a more principled approach to modelling language change and evolution.

Such a Bayesian agent has a *prior distribution* over the possible languages, $P(h)$, which is updated according to the *observed linguistic data*, d, resulting in a *posterior distribution*, better accounting for these data:

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

where h is a hypothesis (language), $P(d|h)$ is the probability of the agent producing the observed linguistic data, d, assuming the hypothesis h, and $P(d) = \sum_h P(d|h)P(h)$. In this paradigm, the prior $P(h)$ is equated to the *learning bias* and is thought to reflect, at least partially, some genetic factors relevant for language (Kirby et al., 2007;Griffiths and Kalish, 2007;Smith and Kirby, 2008).

$P(.|d)$ represents the distribution of the posterior probabilities of all the possible languages, but in this approach a single "winning" hypothesis, h_w is chosen and taken to be the agent's knowledge of language (Kirby et al., 2007;Griffiths and Kalish, 2007). However, this assumption is controversial, given that it can be argued that h_w changes depending on time and context and, more importantly, that language learning and innovation continues throughout life (Croft, 2000). Picking this unique h_w is not trivial and seems to profoundly affect the outcomes of language evolution. Griffiths and Kalish (2007) proposed two such *learning algorithms*, namely the *sampling learner* (henceforth *SAM*), where h_w is randomly chosen according to its posterior probability, $P(h_w|d)$, and the *maximum a posteriori* (or *MAP*) learner, where h_w has the maximum posterior probability, $h_w = arg\,max_h P(h|d)$. Kirby et al. (2007) extended this scheme to a continuous spectrum of intermediate learning algorithms, whereby h_w is chosen with probability $(P(d|h)P(h))^r$, so that for $r = 1$ the learner is a *SAM*, while for $r \to \infty$ the learner is a *MAP*.

Language is transmitted across discrete generations, with the agents in the current generation, t, acting as *language models* for those in the next generation, $t + 1$, by producing data$d$ using their h_w. The agents in the first generation, having no teachers, produce utterances dictated only by their innate biases. In all subsequent generations, the *learners* use these produced data, d, to arrive at their own posterior and chose their own h_w, and they themselves will become language models in generation $t + 1$. Even if, theoretically, there can be many teachers with different priors and learning algorithms producing learner-dependent aggregate data for many different learners, the cases usually treated in the literature assume either a chain composed of a single teacher and a single learner (Kirby et al., 2007; Griffiths and Kalish, 2007), a pool of homogeneous teachers and learners with each learner using the data generated by a single teacher (Griffiths and Kalish, 2007), or homogeneous pools of teachers tested against invasion by a different type of learner (Smith and Kirby, 2008).

It has been shown that for chains of identical agents, iterated learning with *SAM* ($r = 1$) agents is equivalent to a Gibbs sampler and always converges to the prior, while for *MAP* ($r \to \infty$) agents it is equivalent to an expectation-maximisation algorithm, and the behaviour is more complex but still largely influenced by the prior (Kirby et al., 2007; Griffiths and Kalish, 2007). Crucially, however, while for *SAM* chains the resulting distribution of languages can be directly and transparently used to infer the genetic biases (because they are identical), for *MAP* chains this inference is at best partial (concerning, e.g., the order of the hypotheses in the prior; Kirby et al., 2007). Smith and Kirby (2008) showed that *SAM* is not an evolutionarily stable strategy, being open to invasion by *MAP* learners, and suggested that "maximising is always preferred over sampling" (p. 289) in an evolutionary context. Moreover, Kirby et al. (2007) showed that while *SAM* chains are "uninteresting" by always converging to the prior, chains of *MAP* agents (or generally, learning algorithms with $r > 1$) can display complex dynamics, whereby the strength of the bias can be totally obscured by the cultural transmission process. In infinite homogeneous populations where each learner picks a teacher at random, Griffiths and Kalish (2007) showed that the results for single agent chains hold when translated in terms of population frequency of the languages h.

Together, these findings have been interpreting as suggesting that *MAP* (or $r > 1$), as opposed to *SAM*, agents could develop cultural systems largely free from genetic constraints, in the sense that very weak biases can produce very strong cultural universals, that the actual strength of the bias is largely irrelevant and that evolution prefers such systems (Kirby et al., 2007; Smith and Kirby, 2008). However, it is unclear how robust these results are to violations of the various assumptions (discussed in Griffiths and Kalish, 2007, p. 472) and how warranted is the transfer of these conclusions to natural human language. One of the goals of this paper is to explore the effects of altering some of the assumptions implicit in this work, especially those concerning the nature and structure of the populations of agents.

A related question concerns the nature of the impact of genetic biases on the trajectory of language change across generations and its detectability. This is motivated by our recent finding of a correlation between the geographic distribution of *linguistic tone* and the *derived haplogroups* of two brain growth and development-related human genes, *ASPM* and *Microcephalin* (Dediu and Ladd, 2007; Ladd et al., 2008). This correlation does not seem to be explained by the standard factors of shared ancestry and contact and, thus, we suggested that it represents the first case of a causal influence of genetic structure on the properties of language. It is not yet clear how such a bias works at the individual level and how it is amplified by the cultural transmission of language, but

we give some theoretical suggestions in Dediu and Ladd (2007) and especially in Ladd et al. (2008).

I have previously (Dediu, 2008b) investigated three possible operationalisations of such a genetic bias in a heterogeneous, spatially structured population of agents. The biases implemented are non-Bayesian and I found that only one of them, namely the *rate of learning* bias, can influence language in a manner similar to that suggested by the empirical data on tone, *ASPM* and *Microcephalin*. Given the importance of Bayesian models for language change, another goal of the present paper is to investigate the capacity of Bayesian agents to produce results compatible to those suggested by the currently available data (this was in part suggested by one anonymous reviewer of Dediu (2008b), who wondered what would happen if Bayesian learners were allowed to bias language transmission in such a complex population).

2. The agents, their language and genetic bias

In this paper, I will use the framework that I introduced and described in detail elsewhere (Dediu, 2008b). The genome of an agent consists of two independent genes, G_1 and G_2, each having two alleles (one denoted *), which are selectively neutral. The language is described by two linguistic features, F_1 and F_2, each with two possible values, one also denoted by *. G_1 can influence the language of the agents by coding specific linguistic biases which may affect F_1 only; therefore, G_2 and F_2 evolve independently and without biasing, acting as controls. An agent's internal representation of the language is given by the probabilities that each feature has value *, p_1 and p_2, as well as the joint probability that both have this value simultaneously, $p_{1\cdot2}$. Language production involves the generation of utterances containing F_i^* with probability p_i, and $F_1^* \wedge F_2^*$ with probability $p_{1\cdot2}$.

During language learning, the agent is presented with a sample of utterances from which the frequencies f_i and $f_{1\cdot2}$ of the utterances containing F_i^* and $F_1^* \wedge F_2^*$ are computed. These frequencies represent the observed data, d, which are used to update the agent's internal linguistic representation, p_i and $p_{1\cdot2}$, depending on the current *update rule*. There are three non-Bayesian and two Bayesian update rules.

2.1. The non-Bayesian agents

The difference between the observed frequency, f_i^t (with $i \in \{1, 2, 1\cdot2\}$), and the agent's internal probability, p_i^t, at time t is used to update the latter:

$$p_i^{t+1} = \begin{cases} p_i^t + \Delta_i^t \cdot r_i^+ & \text{if } p_i^t \leqslant f_i^t \\ p_i^t - \Delta_i^t \cdot r_i^- & \text{otherwise} \end{cases}$$

where $\Delta_i^t = |p_i^t - f_i^t|$, and the parameters $0 \leqslant r_i^+, r_i^- \leqslant 1$ are the *learning rates* adjusting the weight of the evidence in favour of or against F_i^*. Depending on the parameters r_i^+ and r_i^- and the initial internal probabilities p_i^0, the three models are defined as:

- $\mathbf{M_0}$ (*No genetic bias*): $r_1^+ = r_1^- = r_2^+ = r_2^- = r_{1\cdot2}^+ = r_{1\cdot2}^- = 1$ and $p_1^0 = p_2^0 = p_{1\cdot2}^0 = \frac{1}{2}$. Thus, there is no influence of the genes on the agent's language, allowing language to evolve in a purely cultural manner.
- $\mathbf{M_1}$ (*Genes bias the initial expectation*): The allele G_1^* determines $p_1^0 = 1$, while the other allele determines $p_1^0 = 0$; the other parameters are as for $\mathbf{M_0}$. In this case, the genes bias language acquisition by coding for *different initial starting points*, in the sense that the G_1^* allele very strongly "predisposes" the agent to expect a language of type F_1^*, while the other allele very strongly "predisposes" against such a language.
- $\mathbf{M_2}$ (*Genes bias the rate of learning*): The G_1^* allele encodes an asymmetric rate of learning bias, β, given by $r_1^- = \beta$; the other parameters are as for $\mathbf{M_0}$. For this, genes bias language acquisition by coding for *preferential rates of learning*, in the sense that the G_1^* allele makes the agent evaluate evidence favouring F_1^* as stronger than equivalent evidence against it; the *strength* of the bias is measured by β between 0.0 (extremely strong tendency towards F_1^*) to 1.0 (neutral).

2.2. The Bayesian agents

As described above, an utterance has the form $u = v_1 v_2$, where v_i is the value of the linguistic feature F_i. Therefore, there are four possible utterances (where 1 means that the * value is present and 0 that it is absent): 00, 01, 10 and 11. Let $\mathbf{p} = (p_{00}, p_{01}, p_{10}, p_{11})$ be the vector of probabilities of the four utterances and $\mathbf{n} = (n_{00}, n_{01}, n_{10}, n_{11})$ the vector of their frequencies of occurrence in a particular set of utterances.

Thus, an agent's internal representation of language as given by $\{p_1, p_2, p_{1\cdot2}\}$ is entirely equivalent to the vector $\mathbf{p}$ because

$$p_{00} = 1 - p_1 - p_2 + p_{1\cdot2}; \quad p_{01} = p_2 - p_{1\cdot2}; \quad p_{10} = p_1 - p_{1\cdot2}$$
$$p_{11} = p_{1\cdot2}$$
$$p_1 = p_{10} + p_{11}; \quad p_2 = p_{01} + p_{11}$$

and it produces utterances following a *multinomial distribution*, $\mathbf{n} \sim Multinom(\mathbf{p})$ with the probability mass function:

$$f(\mathbf{n}; \mathbf{p}) = \frac{(n_{00} + n_{01} + n_{10} + n_{11})!}{n_{00}! n_{01}! n_{10}! n_{11}!} p_{00}^{n_{00}} p_{01}^{n_{01}} p_{10}^{n_{10}} p_{11}^{n_{11}}$$

Assuming, as detailed in Griffiths and Kalish (2007), that an agent has full access to its own learning mechanism and that the best model for production it can have is its own (of course, all of these assumptions are open to debate), the conjugate Dirichlet prior distribution was used for the language, $\mathbf{p} \sim Dirichlet(\alpha)$, having the probability density function:

$$f(\mathbf{p}; \alpha) = \frac{1}{B(\alpha)} p_{00}^{\alpha_{00}-1} p_{01}^{\alpha_{01}-1} p_{10}^{\alpha_{10}-1} p_{11}^{\alpha_{11}-1}$$

where $\alpha = (\alpha_{00}, \alpha_{01}, \alpha_{10}, \alpha_{11}) > 0$ are parameters and $B(\alpha)$ is the beta function (Press, 2003). Therefore, given $\mathbf{p} \sim Dirichlet(\alpha)$ and $\mathbf{n} \sim Multinom(\mathbf{p})$, we have that after the application of Bayes' rule, $\mathbf{p}|\mathbf{n} \sim Dirichlet(\alpha + \mathbf{n})$.

With these, the genetic bias of a Bayesian agent is given by the parameters $\alpha^{\mathbf{0}} = (\alpha_{00}^0, \alpha_{01}^0, \alpha_{10}^0, \alpha_{11}^0)$ of the distribution of $\mathbf{p^0} = (p_{00}^0, p_{01}^0, p_{10}^0, p_{11}^0) \sim Dirichlet(\alpha^{\mathbf{0}})$ with which the agent is born (to simplify the notation, in the following, the ^{0} superscript will be understood).

These initial probabilities, $\mathbf{p^0}$, are assigned deterministically with the following constraints:

(i) F_1 and F_2 are genetically independent and
(ii) F_2 is fully genetically unbiased.

Condition (i) means that $p_{1\cdot2} = p_1 p_2$, equivalent to $p_{11}^2 + p_{11}(p_{10} + p_{01} - 1) + p_{10} p_{01} = 0$, which is satisfied if $(p_{01} - 1)^2 + (p_{10} - 1)^2 \geqslant 1$, giving two solutions $p_{11}^{\pm} = ((1 - p_{10} - p_{01}) \pm \sqrt{\Delta})/2$.

Condition (ii) can be understood as requiring $\mathbf{E}[p_2] = \frac{1}{2}$ and $\mathbf{Var}[p_2]$ very large. The first implies $\mathbf{E}[p_{01}] + \mathbf{E}[p_{11}] = \frac{1}{2}$, but because $\mathbf{E}[p_{01}] = \alpha_{01}/\alpha_{sum}$ and $\mathbf{E}[p_{11}] = \alpha_{11}/\alpha_{sum}$, we have $\alpha_{01} + \alpha_{11} = \alpha_{sum}/2$. Because $\mathbf{Var}[p_2] = ((\alpha_{01} + \alpha_{11})(\alpha_{sum} - \alpha_{01} - \alpha_{11}))/(\alpha_{sum}^2 (\alpha_{sum} + 1)) = 1/(4(\alpha_{sum} + 1))$, the second condition requires α_{sum} small. Moreover, by requiring that $\alpha > 1$, we have that $\alpha_{sum} > 4$ and $\mathbf{Var}[p_2] < \frac{1}{20}$.

$\mathbf{E}[p_1] = \mu$ and $\mathbf{Var}[p_1] = \sigma$ define the genetic bias of F_1 and, using the previous results, we have $(\alpha_{10} + \alpha_{11})/\alpha_{sum} = \mu$ and $(\mu(1-\mu))/(\alpha_{sum}+1) = \sigma$. As $\alpha_{sum} > 4$, $\mu(1-\mu) > 5\sigma$ and given $\mu \in (0,1)$ and $max_\mu(\mu(1-\mu)) = \frac{1}{4}$ for $\mu = \frac{1}{2}$, we have that $\sigma \in (0, \frac{1}{20})$. Therefore, the genetic bias is described by the two parameters $\mu \in (0,1)$ (the *location*) and $\sigma \in (0, \frac{1}{20})$ (the *strength*).

Taking μ and σ as given, we have $\alpha_{sum} = ((\mu(1-\mu)/\sigma) - 1$, $\alpha_{10} + \alpha_{11} = \mu\alpha_{sum}$, $\alpha_{01} + \alpha_{11} = \alpha_{sum}/2$ and, by requiring that initially $\mathbf{p} = \mathbf{E}[\mathbf{p}]$, we have that $(\alpha_{sum} - \alpha_{01})^2 + (\alpha_{sum} - \alpha_{10})^2 \geqslant \alpha_{sum}^2$, in which case $\alpha_{11}^{\pm} = ((\alpha_{sum} - \alpha_{01} - \alpha_{10}) \pm \sqrt{(\alpha_{sum} - \alpha_{01} - \alpha_{10})^2 - 4\alpha_{01}\alpha_{01}})/2$, with solution $\alpha_{11} = (\alpha_{sum}\mu)/2$.

Furthermore, by requiring $\alpha_{11} > 1$, we have $\sigma < ((\mu^2(1-\mu))/(2+\mu))$. For function $f(\mu) = (\mu^2(1-\mu))/(2+\mu)$ the equation $0 = \partial f/\partial\mu = (2\mu(1-\mu) - \mu^2)/(2+\mu) - (\mu^2(1-\mu))/(2+\mu)^2$ has two real roots $\mu_\pm = (-5 \pm \sqrt{57})/4$, of which only $\mu_+ \approx 0.6375 \in (0,1)$ and $f(\mu_+) = \max_{x\in(0,1)} f(x) \approx 0.0558$.

Therefore, the two parameters describing the genetic bias, its location μ and its strength σ are *not* independent, with $\sigma < \min\{(\mu(1-\mu))/5, (\mu^2(1-\mu))/(2+\mu)\}$, $\mu \in (0,1)$ and $\sigma \in (0, \frac{1}{20})$. Given the symmetry of the model with regard to the *direction* of the bias (i.e., towards or against F_1^*), only the absolute deviation from unbiasedness, $|\mu - \frac{1}{2}|$ matters: therefore, only biases *towards* F_1^* will be considered (i.e., $\mu > \frac{1}{2}$) and μ will denote the deviation from unbiasedness, $|\mu - \frac{1}{2}|$.

To sum up, the Bayesian update rules are the *sampler*, **SAM**, and the *maximum a posteriori*, **MAP**, and their genetic bias can be described using just two non-independent parameters: the *bias location*, μ (given relative to the unbiasedness $\frac{1}{2}$), and the *bias strength*, σ. In the following, we will denote a *sampler* with bias (μ, σ) as $\mathbf{SAM}_\sigma^\mu$, a *maximum a posteriori* with the same bias as $\mathbf{MAP}_\sigma^\mu$, or, when the exact update rule is irrelevant, simply as $\mathbf{B}_\sigma^\mu$. Computationally, sampling from a Dirichlet distribution, as required by **SAM** agents, was implemented using the `GNU Scientific Library` (Galassi et al., 2006).

2.3. The parameter values

Obviously, most of the parameters used in the model are continuous, but due to computational constraints this continuous range was discretised. For the learning biases of the various models, the discretisation is as follows:

- for the $\mathbf{M_0}$, $\mathbf{M_1}$ agents: there are no parameters;
- for the $\mathbf{M_2}$ agents: the bias strength β took the values {0.0001 (extremely strong), 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99, 0.999, 0.99925, 0.9995, 0.99975, 0.9999, 0.99999 (extremely weak)} for the single agent chain and pair chain conditions, but only {0.10, 0.50, 0.80, 0.85, 0.90, 0.95, 0.99} for complex populations, as detailed in Dediu (2008b);
- for the Bayesian agents, **SAM** and **MAP**: the bias location, μ, and bias strength, σ, are given in Table 1 (the extreme values have been approximated due to computational rounding errors).

For all cases and for each parameter combination tested, I executed 20 independent runs, each lasting for 10,000 generations. Preliminary runs, testing the robustness of the model, were done as described in Dediu (2008b). All the statistical analyses use `R` (R Development Core Team, 2007) and I applied Holm (1979) multiple testing correction where appropriate, in which case I report adjusted p-values.

In the following sections, I will study in detail three cases: the case of *homogeneous chains composed of single agents*, which represents the standard model used in the literature and serves as a test for the current implementation, the case of *heterogeneous chains composed of pairs of agents* violates some assumptions of the standard model, while the case of the *heterogeneous, spatially structured populations with overallping generations* violates most of those assumptions.

Table 1
The bias parameters for the Bayesian learners ($\mathbf{B}_\mu^\sigma$; **B** stands for both **SAM** and **MAP**).

μ	σ					
	0.00	0.01	0.02	0.03	0.04	0.05
0.50	$\mathbf{B}_{0.001}^{0.45}$					
0.40	$\mathbf{B}_{0.002}^{0.40}$					
0.30	$\mathbf{B}_{0.002}^{0.30}$	$\mathbf{B}_{0.010}^{0.30}$	$\mathbf{B}_{0.014}^{0.30}$			
0.20	$\mathbf{B}_{0.002}^{0.20}$	$\mathbf{B}_{0.010}^{0.20}$	$\mathbf{B}_{0.020}^{0.20}$	$\mathbf{B}_{0.027}^{0.20}$		
0.10	$\mathbf{B}_{0.002}^{0.10}$	$\mathbf{B}_{0.010}^{0.10}$	$\mathbf{B}_{0.020}^{0.10}$	$\mathbf{B}_{0.030}^{0.10}$	$\mathbf{B}_{0.039}^{0.10}$	
0.00	$\mathbf{B}_{0.002}^{0.00}$	$\mathbf{B}_{0.010}^{0.00}$	$\mathbf{B}_{0.020}^{0.00}$	$\mathbf{B}_{0.030}^{0.00}$	$\mathbf{B}_{0.040}^{0.00}$	$\mathbf{B}_{0.049}^{0.00}$

3. Chains of single agents

This represents the standard model in the literature (Kirby et al., 2007; Kirby and Hurford, 2002; Griffiths and Kalish, 2007) and feature *discrete, non-overlapping generations*, the population in each generation being reduced to *a single agent* learning from the agent in the previous generation. Moreover, these chains of single agents are *homogeneous* in the sense that the agents in all generations are "born the same", all having identical genetic biases and learning algorithms. Therefore, there are five such types of chains, composed, respectively, of $\mathbf{M_0}$, $\mathbf{M_1}$, $\mathbf{M_2}$, **SAM** and **MAP** agents, respectively, and the following describes their behaviour.

For $\mathbf{M_0}$, the probability p_1 goes very quickly to fixation at either 0 or 1 in on average 1516.5 generations, with no preference for 0 or 1 (7 vs 13 cases, $\chi^2(1) = 1.8$, $p = 0.179$) and no difference in speed when converging to either ($t(17.16) = 1.56$, $p = 0.137$). This confirms that $\mathbf{M_0}$ is indeed unbiased and language evolves through cultural drift towards fixation at a uniform language featuring F_1^* or not. As expected, p_2 behaves in the same manner: it converges to 0 or 1 in on average 1379.5 generations, with no difference in speed ($t(8.26) = 0.35$, $p = 0.731$) and no preferences for either ($\chi^2(1) = 3.2$, $p = 0.073$).

For $\mathbf{M_1}$, p_1 is constantly 1, suggesting that this bias invariably and immediately forces the languages to converge towards the biased value, F_1^*, while p_2 behaves as for $\mathbf{M_0}$, as expected.

For $\mathbf{M_2}$ and all biases β, p_2 behaves as for $\mathbf{M_0}$, as expected. However, p_1 invariably converges to 1 in on average 1395.3 generations, for all biases $\beta < 0.99$, with a speed independent of β (one-way ANOVA $F(10, 209) = 1.31$, $p = 0.226$). However, when $\beta \geqslant 0.999$, as the bias weakens (as for $\mathbf{M_2}$, weaker biases mean higher βs), more and more runs fail to converge to 1 and, when the bias is extremely weak ($\beta = 0.999999$), p_1 behaves very similar to $\mathbf{M_0}$, suggesting that, indeed, very weak $\mathbf{M_2}$ biases converge to no bias at all. For this range, $0.999 \leqslant \beta < 1.0$, the number of runs converging to 1 (notated $n_{\to 1}$) or to 0 ($n_{\to 0}$) does depend on β (Pearson's $r_{n_{\to 1}} = -0.90$, $p = 0.005$ and $r_{n_{\to 0}} = 0.93$, $p = 0.002$), as does the speed of convergence to 1 (Spearman's $\rho = -0.24$, $p = 0.017$). Taken together, these suggest that $\mathbf{M_2}$ behaves as intuitively expected, with the biases stronger than the threshold bias $\beta \approx 0.999$ invariably producing the biased language, while weaker biases gradually converge towards unbiasedness.

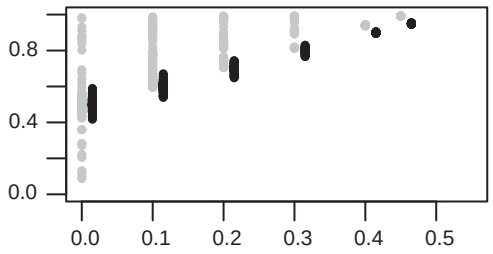

Fig. 1. The average of the internal representation of language, $mean(p_1)$ (vertical axis), function of the bias location, μ (horizontal axis), for chains of single **MAP** (grey) and **SAM** (black; displaced to the right by 0.015 for display purposes) agents (please note that the actual value of μ is $\mu+0.5$).

For **SAM**, the language maintains the location of the bias, μ, across generations, and the variation around this value is controlled by the bias strength, σ. The regression

$$mean(p_1)\sim 0.50+0.998\mu \quad (1)$$

has an adjusted $R^2=0.978$ and all $p<2\times10^{-16}$. These confirm the claim in the literature (Kirby et al., 2007; Griffiths and Kalish, 2007) that chains of samplers converge on their prior.

For **MAP**, however, the languages produced by very strong biases ($\sigma=0.002$) cluster tightly around μ, but for more relaxed biases ($\sigma>0.002$) they tend to spread around higher locations, depending on σ:

$$mean(p_1)\sim 0.430+1.397\mu+4.974\sigma \quad (2)$$

with adjusted $R^2=0.705$ and all $p<10^{-12}$. Therefore, in this case, the actual biasing of language depends on both the bias location, μ, and its strength, σ, and supports the findings in (Kirby et al., 2007; Griffiths and Kalish, 2007) that **MAP** learners can amplify weak biases, but the mechanism is complex and strong biases are stable.

Fig. 1 represents the average internal representation of language to which the agents arrive, $mean(p_1)$, function of the bias location, μ, for **MAP** and **SAM** single-agent chains. It can be clearly seen that sampler chains tightly track their prior (genetic bias), while MAP learner chains tend to be more scattered around higher values, showing that they have a more complex and unpredictable dynamics around an—on average—amplified bias.

4. Chains of pairs of agents

This represents a slight modification of the standard, single-agent homogeneous chain model, in the sense that now there are *two agents* (*one pair*) per generation and they are not necessarily "born the same", meaning that the chain is potentially *inhomogeneous*. More exactly, if we let the two agents in generation t be denoted A_1^t and A_2^t, then we can have that $A_1^t\neq A_2^t$. However, the two *lineages* are identical, namely for any two generations, t_1,t_2, we have that $A_1^{t_1}=A_1^{t_2}$ and $A_2^{t_1}=A_2^{t_2}$.

The language models for generation $t+1$, A_1^t and A_2^t, produce n utterances each, u_i^1 and u_i^2, respectively, $i=\overline{1,n}$, which are used by the current learners, A_1^{t+1} and A_2^{t+1}, to learn a language from exactly the same *mixed primary data*. This alteration is small but still a step towards increased realism, and any disagreements with the standard model are potentially important for the latter's generalisability.

In the following I will present the behaviour of the resulting shared language and individual language representations function of the pair's composition. I tested all possible pairs between all possible agent types and parameter values, leading to 931 cases.

4.1. The shared language

The pairs of agents managed to always converge on a shared language, with only six cases showing significant differences between the agents' internal representations and their common language: $\mathbf{M_1}$–$\mathbf{B}^{0.00}_{0.002}$ and $\mathbf{B}^{0.10}_{0.002}$–$\mathbf{B}^{0.40}_{0.002}$, where **B** stands for any type of Bayesian learner. These failures to agree are due to the widely different starting points of the Bayesian agents, μ, and their very strong biases, σ, suggesting that, in general, even agents with strong and very incongruous genetic biases can still learn the language of their community.

The following subsections will analyse the effects model **X** has on model **Y** when paired together, by comparing the internal representations of language, p_1^Y, p_2^Y and $p_{1\cdot2}^Y$, of the **Y** agent in the single and paired conditions. Moreover, unless specified, these comparisons will focus on the biased feature, F_1.

4.2. The effects of $\mathbf{M_0}$

$\mathbf{M_0}$ does not have any major effects when paired with itself. However, it does alter $\mathbf{M_1}$ by making it very similar to $\mathbf{M_0}$ ($t(37.89)=-0.66$, $p=0.51$). By contrast, $\mathbf{M_2}$ is mostly unaffected, except for the extremely strong bias $\beta=0.0001$, which seems to be slightly buffered by this pairing, slowing down its convergence towards 1 (mean 3223.8 generations).

Interestingly, both Bayesian models are profoundly affected by their pairing with $\mathbf{M_0}$, by becoming *very similar* to each other and by having their biases *attenuated*, as shown by the regression of their internal representation of language on their bias:

$$\mathbf{SAM}: mean(p_1)\sim 0.506+0.569\mu-0.338\sigma$$
$$\mathbf{MAP}: mean(p_1)\sim 0.507+0.573\mu-0.357\sigma$$

with adjusted $R^2=0.950$ and 0.951, respectively, and all $p<10^{-5}$. The correlation between the $mean(p_1)$ of **SAM** and the $mean(p_1)$ of **MAP** is very strong and positive when they are paired with $\mathbf{M_0}$ (Pearson's $r=0.958$, $p<2.2\times10^{-16}$), but is negative and small when in single chains ($r=-0.157$, $p=0.0015$). These show that even if they are very different when alone, the two Bayesian types of agents become indistinguishable when paired with the non-biased model, $\mathbf{M_0}$, throwing some doubt on the generalisability of their different behaviour in the standard model.

4.3. The effects of $\mathbf{M_1}$

$\mathbf{M_1}$ does not have any effects on itself, and $\mathbf{M_2}$ is only slightly affected by pairing it with $\mathbf{M_1}$, but still behaving in a similar manner to its single condition.

However, both Bayesian agents change radically in the sense that they become *very similar* to each other and have their bias *amplified*, as shown by the regression of their internal representation of language on their bias

$$\mathbf{SAM}: mean(p_1)\sim 0.680+0.638\mu+1.639\sigma$$
$$\mathbf{MAP}: mean(p_1)\sim 0.684+0.652\mu+1.680\sigma$$

adjusted $R^2=0.946$ and 0.942, respectively, and all $p<10^{-16}$. As for the paring with $\mathbf{M_0}$ discussed above, the correlation between the $mean(p_1)$ of **SAM** and the $mean(p_1)$ of **MAP** is very strong and positive when they are paired with $\mathbf{M_1}$ ($r=0.981$, $p<2.2\times10^{-16}$).

4.4. The effects of $\mathbf{M_2}$

Pairing $\mathbf{M_2}$ with itself has only the effect of slightly lowering the convergence speed. Interestingly, the bias strength β does not make any difference when paired with Bayesian agents (one-way ANOVAs are *ns*), and there are no differences between **SAM** and

MAP ($t(3197.56) = 0.346$, $p = 0.729$). Moreover, paring $\mathbf{M_2}$ with a Bayesian agent is similar to pairing that agent with $\mathbf{M_0}$ (all *t*-tests are *ns*). This suggests that the Bayesian and $\mathbf{M_2}$ types of genetic bias are somehow orthogonal, referring to profoundly different manners of being biased and that, from the point of view of a Bayesian agent, all $\mathbf{M_2}$s are just a sort of $\mathbf{M_0}$, while for $\mathbf{M_2}$, both types of Bayesian agents are the same thing.

4.5. The effects of the Bayesian agents

Pairing two identical Bayesian agents (either both **MAP** or both **SAM** and both having the same bias location, μ, and strength, σ) does not impact significantly on their behaviour. More specifically, the pairs **MAP**–**MAP** behave effectively identical to single **MAP** chains of the same bias, and the pairs **SAM**–**SAM** behave like single **SAM** chains of the same bias, except that they tend to "be more on target" (they have smaller standard deviations around the mean). However, pairs of identically biased **SAM**–**MAP** agents behave very much like the corresponding (same bias) single **SAM** chains (all *t*-tests are *ns*), but different from single same bias **MAP** chains (14 of 20 runs have *t*-tests significant at α-level 0.05).

Fig. 2 shows these results as the mean of the internal representation of language of the agents, $mean(p_1)$ across runs, as a function of the (common) bias location μ. Comparing this with the behaviour of single agent chains represented in Fig. 1, it can be seen that, indeed, pairs of identically biased **MAP** agents (light grey in Fig. 2) behave like single **MAP** agents (grey in Fig. 1), pairs of identically biased **SAM** agents (dark grey in Fig. 2) behave similar to single **SAM** agents (black in Fig. 1), but pairs of identically biased **MAP** and **SAM** agents (black in Fig. 2) are very similar to pairs of identical **SAM** agents and, thus, to single **SAM** chains.

In the general case, when pairing any two Bayesian agents, each having any of all the possible biases, an interesting behaviour emerges: pairs of **MAP**–**MAP** agents differ from both pairs of **SAM**–**MAP** agents and pairs of **SAM**–**SAM** agents, but pairs of **SAM**–**MAP** agents behave in the same manner as pairs of **SAM**–**SAM** agents (one-way ANOVA $F(2, 12597) = 241.03$, $p < 2.2 \times 10^{-16}$, post-hoc adjusted pairwise comparisons, respectively, $p = 0.000, 0.000$ and 0.913). Briefly put in a symbolic form (**SAM**–**SAM** $\approx$ **SAM**–**MAP**) $\neq$ **MAP**–**MAP**.

Unfortunately, directly comparing the behaviour of chains of pairs of agents to the behaviour of chains of single agents is hindered by the different dimensionalities of the results (single bias vs two biases), but from visually inspecting the language function of the bias location (μ_1, μ_2 for pairs of agents and μ for single agents), it can be observed that the behaviour of **MAP**–**MAP** pairs is similar to the single **MAP**s, while the pairs **SAM**–**MAP** and **SAM**–**SAM** resemble the single **SAM**s. Taken together, these suggest that when **SAM** and **MAP** agents are mixed, the winning strategy is **SAM**.

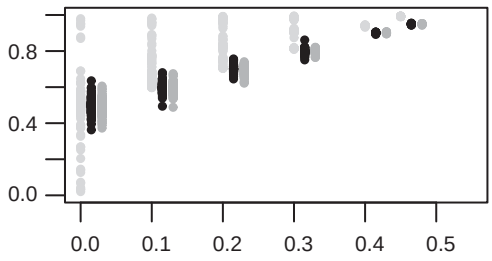

Fig. 2. The average of the internal representation of language, $mean(p_1)$ (vertical axis), function of the bias location, μ (horizontal axis), for chains of pairs of Bayesian agents with identical biases: **MAP**–**MAP** (light grey, leftmost dots), **SAM**–**MAP** (black; middle dots, displaced by 0.015) and **SAM**–**SAM** (dark grey, rightmost dots, displaced by 0.030). Compare with Fig. 1.

This can also be seen from the linear regression of the shared language, $mean(f_1)$, on the biases of **MAP**–**MAP**, **SAM**–**MAP** and **SAM**–**SAM** pairs:

$$mean(f_1) \sim 0.408 + 0.668\mu_1 + 0.730\mu_2 + 4.726\sigma_1 + 0.510\sigma_2$$
$$mean(f_1) \sim 0.488 + 0.478\mu_1 + 0.550\mu_2 + 2.120\sigma_1 - 1.494\sigma_2$$
$$mean(f_1) \sim 0.489 + 0.481\mu_1 + 0.549\mu_2 + 2.066\sigma_1 - 1.551\sigma_2$$

with adjusted R^2 of 0.811, 0.889 and 0.887, respectively, and all $p < 2.2 \times 10^{-5}$. These suggest that while **MAP**–**MAP** pairs always amplify their individual biases ($mean(f_1) > max(\mu_1, \mu_2)$ in 98.73% cases, strongly reminiscent the behaviour of single **MAP** chains), the pairs involving at least a **SAM** behave in a different manner, by tending to settle on an average bias location. In a symbolic notation (**SAM**–**SAM** $\approx$ **SAM**–**MAP** $\approx$ **SAM**) $\neq$ (**MAP**–**MAP** $\approx$ **MAP**). However, it is clear that in this general case, the bias(es) cannot be directly and transparently inferred from the resulting distribution of languages.

5. Complex populations

Here I extend my previous model of complex populations (described in detail in Dediu, 2008b) to the Bayesian case. The world is composed of a square grid of 10×10 regions, each region being able to support a population. The optimal population size is fixed for each region, and the actual population size fluctuates around it through birth, death and migration. The time is discretised in years, generations are desynchronised (overlapping) and each agent has a limited non-deterministic lifespan. Language acquisition takes place during the critical period, with most utterances learned from the agent's mother, followed by members of the agent's own group and with some influence from the members of neighboring groups. After reaching sexual maturity, there is mating and reproduction. All demographic and linguistic processes depend on space (through Moore neighbourhoods), and the genes are exposed to drift, there being no natural selection. Following the procedure described in Dediu (2008b), I tested the model against various settings of these parameters and I found it to be robust.

The biasing allele, G_1^*, has initial frequency v in the population, and this together with the biasing mechanism ($\mathbf{M_0}$, $\mathbf{M_1}$, $\mathbf{M_2}$, **SAM** or **MAP**) and the appropriate bias parameters (β, or μ and σ) identifies such a complex population model (see Dediu, 2008b for more details on the procedure). For all cases and for each parameter combination tested, 20 independent runs were executed, each lasting for 10,000 simulation years (Dediu, 2008b). Preliminary runs, testing the robustness of the model, were done as described in Dediu (2008b).

The following sections will focus on two goals:

- the description of the emerging language function of the composition of the population and its comparison with the simpler pair and single agent chains and
- the analysis of the resulting spatial pattern of biased language and the detectability of this bias.

5.1. The language

In this case, the language is represented by the global (world-wide) frequency of the starred value F_i^*, denoted f_i, and the genetics by the global frequency of starred allele, G_i^*, denoted g_i,

for $i \in \{1,2,1\cdot 2\}$. As expected, f_2, g_1 and g_2 do not depend on the model or its parameters, fluctuating around $\frac{1}{2}$, with f_2 being globally very stable (see Dediu, 2008b for details). However, f_1—the frequency of the biased linguistic value—does depend on the parameter values and the following will study this dependency. Due to computational constraints, the initial population frequency of the biasing allele G_1^*, v, takes the values {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}.

For $\mathbf{M_0}$ (meaning that the population is entirely composed of unbiased agents), as expected, f_1 is constant across generations for each individual run, and does not depend on v, the initial frequency of the biasing allele. The $mean(f_1) \approx \frac{1}{2}$, and the standard deviation, $sd(f_1) \approx 0.004$, showing that, indeed, $\mathbf{M_0}$ does not bias the language.

When $\mathbf{M_1}$ agents form part of the population, f_1 increases with their increasing frequency, v, in a non-linear fashion:

$$mean(f_1) \sim 1.992v - 0.990v^2$$

with adjusted $R^2 = 0.999$ and all $p < 2.2 \times 10^{-16}$. Moreover, the variation, $sd(f_1)$, declines with increasing population frequency, v: Pearson's $r = -0.67$, $p < 2.2 \times 10^{-16}$.

As opposed to all other models ($\mathbf{M_0}$, $\mathbf{M_1}$, **MAP** or **SAM**), where f_1 tends to be relatively constant across time (except for very slight random fluctuations), for $\mathbf{M_2}$ it follows an asymptotic growth curve, modelled here as a linear rise possibly followed by a plateau at the maximum possible value of 1.0 (see Fig. 3).

Given that the rise always starts at $\frac{1}{2}$ and the plateau at 1 is not always reached during the time allowed for the run, a valid measure of the dynamics of f_1 is represented by the actual *maximum* value $M_1 = max(f_1)$ and the *moment* it was first realised, $Mx_1 = min(arg\,max(f_1))$. From these, the *speed* of the rise in the frequency of the biased linguistic value, F_1^*, can be estimated by the angle of the rising part, $\alpha = a\tan(M_1 - \frac{1}{2}, Mx_1)$ (measured in degrees, °). The speed of rise α and, consequently, the maximum reached in the allowed time, M_1, are higher for stronger biases and higher initial frequencies of the biasing allele:

$$\alpha(^\circ) \sim 43.85^\circ + 22.92^\circ v - 48.43^\circ \beta$$

with adjusted $R^2 = 0.839$ and all $p < 2.2 \times 10^{-16}$.

When the biased agents in the population are Bayesian learners (**MAP** or **SAM**), the language is also more or less constant across time for each run (as for the other non-Bayesian agents except $\mathbf{M_2}$). Moreover, it is indistinguishable between the two models ($t(7993.66) = 1.004$, $p = 0.32$) and depends on all three parameters in a complex and non-transparent manner:

$$mean(f_1) \sim 0.40 + 0.31v + 0.73\mu - 0.64\sigma - 0.16v^2 + 11.35\sigma^2$$

with adjusted $R^2 = 0.905$, all $p < 10^{-6}$.

However, in order to allow the comparison with the languages produced by the single agent chains, $mean(f_1)$ was regressed

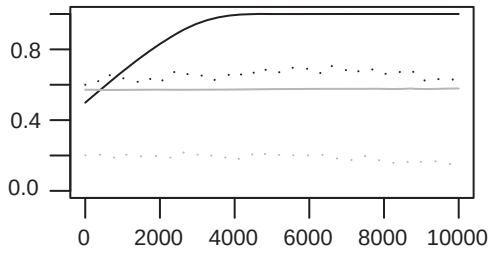

Fig. 3. The trajectory through time (horizontal axis, in simulation years) of the frequency of the biased linguistic value, f_1 (solid lines) and the global frequency of the biasing allele, g_1 (dotted lines) for representative runs of $\mathbf{M_2}$ ($\beta = 0.5$, $v = 0.6$, black) and **MAP** ($\mu = 0.2$, $\sigma = 0.027$, $v = 0.2$, grey). g_1 changes by random drift around its initial value, v, irrespective of the model, while f_1 behaves dramatically different for the two models.

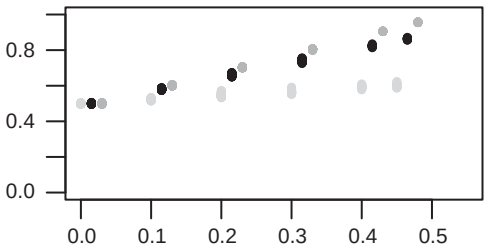

Fig. 4. The average of the internal representation of language, $mean(p_1)$ (vertical axis), function of the bias location, μ (horizontal axis), for complex populations of **MAP** agents for three initial frequencies of the biasing allele, v: 0.1 (light grey, leftmost dots), 0.5 (black; middle dots, displaced by 0.015) and 0.9 (dark grey, rightmost dots, displaced by 0.030). The same happens also for **SAM** agents. It can be seen by comparing with Figs. 1 and 2 that this is more similar to single **SAM** chains.

linearly just on μ and σ:

$$mean(f_1) \sim 0.502 + 0.734\mu$$

with adjusted $R^2 = 0.789$ and all $p < 2.2 \times 10^{-16}$; it can be seen that this is much more similar to single **SAM** chains than to single **MAP** chains (compared with Eqs. (1) and (2)).

Focusing on the extreme case of *homogeneous Bayesian populations* ($v = 1.0$), **SAM** and **MAP** are again identical ($t(797.57) = 0.35$, $p = 0.72$) and very similar to the single **SAM** (compared with Eq. (1)):

$$mean(f_1) \sim 0.50 + 1.024\mu$$

with adjusted $R^2 = 0.999$ and all $p < 2.2 \times 10^{-16}$. Therefore, it is again clear that populations of samplers and MAP learners more complex than the homogeneous single agent chains do not show any differences in the language produced and behave more similar to single **SAM** chains than to single **MAP** chains, calling into question the generality of the results in Kirby et al. (2007) (see also Fig. 4).

5.2. Correlations between linguistics, genetics and geography

As briefly described in the *Introduction*, our recent discovery of a correlation between the geographical distributions of tone languages and the derived haplogroups of *ASPM* and *Microcephalin*, correlation which apparently cannot be entirely explained by contact and shared ancestry, has lead us to propose that a genetic bias *causes* language to change in a certain direction when transmitted across generations in a population containing enough such biased individuals. Methodologically, we used *Mantel (partial) correlations* (Mantel, 1967) (computed using the ZT software Bonnet and Van de Peer, 2002) between geographical, genetic, historical linguistic and typological distances to control for these factors, where historical linguistic distances encode the degree of relatedness due to sharing a common ancestor, while the typological distances reflect the degree of structural similarity between languages (for details, please see Dediu and Ladd, 2007). For a detailed discussion of the assumptions, earlier proposals and implications for language change, see Ladd et al. (2008), and for language evolution, see Dediu (2008a).

In Dediu (2008b) I implemented a computer model designed to test the conditions under which such a genetic biasing of language can be detected using the methodology we proposed in Dediu and Ladd (2007) using non-Bayesian learners and the type of complex populations described above. There, I found that while $\mathbf{M_1}$ agents cannot produce such a bias, $\mathbf{M_2}$ agents can and, moreover, the region of the parameter space allowing this bias to be detected is quite large. Here I extend this work for the two types of Bayesian learners and investigate their capacity to induce a detectable

genetic biasing of the language of the type we reported for tone and *ASPM* and *Microcephalin*.

The relationships between (structural) linguistic, genetic and geographic distances between populations are measured using (partial) Mantel correlations, where *GenGeo* is the correlation between genes and geography, *LingGeo* is the correlation between linguistics and geography, *GenLing* is the relationship between genetic and linguistic distances and *GenLingGeo* is the residual correlation between genes and languages after controlling for geography. Also, we measure the Pearson correlations between the population frequencies of the starred allele, G_j^*, and starred linguistic feature, F_i^*, denoted F_iG_j, for $i,j \in \{1,2\}$. In order to capture the dynamics of a series of such correlations through time (simulation years), ρ was defined in Dediu (2008b) to represent the *proportion of significant correlations* for a given α-level.

The first finding is that the behaviour of the correlations produced by the two Bayesian learners, **SAM** and **MAP**, in such complex populations is almost indistinguishable: randomisation ANOVAs (Edgington, 1987) are *ns* after multiple testing correction (Holm, 1979). Therefore, only **MAP** learners will be analysed in the following.

Concerning the separate correlations of linguistic and genetic distances with geography, *GenGeo* and *LingGeo* are both very high, confirming that the model is behaving as expected (Dediu, 2008b). Moreover, *GenGeo* depends very slightly only on v—the initial frequency of the biasing allele, G_1^*, in the population which—in an inverted "U" shape, being highest either for very rare or very common biasing agents. However, *LingGeo* is higher for lower v (Spearman's $\rho = -0.17$, $p < 2.2 \times 10^{-16}$), higher μ ($\rho = 0.26$, $p < 2.2 \times 10^{-16}$) and small σ ($\rho = -0.084$, $p = 3.33 \times 10^{-7}$), suggesting that the correlation between linguistics and geography is most detectable for rare but strongly biased agents.

The correlation between genetics and linguistic structure, *GenLing*, is moderately strong and depends in an inverted "U" shape on v ($\rho = -0.21$, $p < 2.2 \times 10^{-16}$), and gets stronger for higher μ ($\rho = 0.30$, $p < 2.2 \times 10^{-16}$) and lower σ ($\rho = -0.19$, $p < 2.2 \times 10^{-16}$), being most detectable for rare or common strongly biased agents. Moreover, controlling for geography by using *GenLingGeo* diminishes the values of the correlations but does not alter their pattern, suggesting that geography does not fully explain the similarity between genes and language structures.

In order to understand the nature and detectability of the genetic biases, we need to focus on the correlations between individual alleles and linguistic features. As expected, F_1F_2 (the correlation between the two linguistic features), F_1G_2, F_2G_1 and F_2G_2 are constantly very low, reflecting their mutual independence (built in the simulation). The correlation between the two genes, G_1G_2, is a bit higher, reflecting the slower peace of genetic drift. However, the correlation between the biasing allele and the biased feature, F_1G_1, is very low for small biases and high population frequencies, but extremely high for strong (small σ), marked (large μ) infrequent (low v) biases:

$$F_1G_1 \sim 0.257 - 0.266v + 0.663\mu - 1.903\sigma$$

with adjusted $R^2 = 0.411$ and all $p < 2 \times 10^{-16}$. This function has a maximum for $v = 0$, $\mu = \frac{1}{2}$ and $\sigma = 0$, suggesting that it is easiest to detect very strong infrequent Bayesian biases as opposed to moderately strong and relatively frequent $\mathbf{M_2}$ biases, as was found in Dediu (2008b).

It can be concluded that such non-homogeneous, spatially structured complex populations with overlapping generations confirm the results of chains of pairs of agents, of which probably the most important is that the two types of Bayesian learners behave in the same way and similar to single chains of samplers. Moreover, the biases generated by such Bayesian learners in a complex population can be detected using the methodology of Dediu and Ladd (2007), but probably only for very strong infrequent biases.

6. Discussion

Despite their low realism, homogeneous chains of single agents probably represent the standard model of language evolution, both theoretically, through mathematical and computational modelling (Kirby et al., 2007; Griffiths and Kalish, 2007) and experimentally (Kirby et al., 2008). This is primarily due to the fact that such homogeneous single agent chains are well understood thanks to a relatively long history of computational studies in the *iterated learning model* tradition, focused mainly on the emergence of compositionality (Kirby and Hurford, 2002) but also to the recent seminal approach of Griffiths and Kalish (2007). Moreover, they are also relatively easy to implement and study in the laboratory using real human subjects, as recently shown by Kirby et al. (2008). However, these models present a series of potential pitfalls and the following enumeration is not intended to be exhaustive:

- they abstract away from horizontal and oblique social interactions (by having a single agent in each generation), assuming that their impact is not dramatic;
- the assumption of "acquisitionism" (Honeybone, 2003)—the view of language change and evolution as due to children acquiring a different language because of the reinterpretation of data produced by adults (Kirby and Hurford, 2002)—a position prevalent in certain views of historical linguistics (Lightfoot, 1999), but largely contradicted by a number of empirical observations suggesting that actually it is the adults that have the active role in language change (Croft, 2000);
- they also assume that the (degenerate) population is homogeneous, which is not warranted by the data on inter-individual diversity, both at the genetic and at the cultural levels (e.g., Stromswold, 2001; Plomin et al., 2001).

Moreover, the Bayesian approach to these models carries with it a supplementary set of assumption, including the fact that human language learning is validly approximated by Bayes' rule, that a single fixed hypothesis is selected from the posterior after the learning has ceased, and that learners have extensive access to their own learning algorithm so that they can use it to measure the likelihood that the observed data were produced by a given hypothesis (for a larger list and discussion see Griffiths and Kalish, 2007).

While I do not want to imply that *all* of these assumptions are wrong, I want to suggest that most are empirical questions not currently well supported by the data available, and that the effects of their violation for the validity of theoretical results obtained in the homogeneous single agent chains tradition were not investigated in proportion to its importance. The present study has attempted to explore the effects of changing some of these assumptions.

First, with all assumptions in place, the present model succeeded in confirming the theoretical and computational results in Griffiths and Kalish (2007) and Kirby et al. (2007), namely that chains of samplers, **SAM**, converge to the "pure" genetic bias, while the language of chains of a posteriori maximisers, **MAP**, depends in a complex manner on their genetic bias.

Second, when enlarging the population to two agents and allowing the homogeneity assumption to be dropped, I found that,

in general, even learners with very different genetic biases do converge on a common language. However, Bayesian agents involved in heterogeneous pair chains behave in the same manner, with the strong differences between **SAM** and **MAP** learners in the single chain condition being totally lost. Moreover, chains composed of **SAM**–**MAP** pairs tend to behave like single chains of **SAM**.

Third, in the case of complex, spatially structured populations with overlapping generations, the resulting language is largely influenced by the frequency, nature and strength of the genetic biases. The two Bayesian learners are behaving indistinguishably from each other and similar to single chains of samplers, **SAM**, even in the case of populations composed entirely of a single agent type.

Taken together, these results suggest that the stark differences between the two types of Bayesian learners when in single agent chains disappear for more complex settings, both behaving similar to Bayesian samplers. Therefore, the language resulting from Bayesian biased cultural transmission is very much influenced by the characteristics of the bias (the prior). However, the resulting distribution of languages *does not* allow the direct and transparent inference of these same genetic biases, supporting the view that the process of cultural transmission plays a very important mediating role (Kirby et al., 2007).

Moreover, I found that the methods we have previously used to find the correlation between the geographical distributions of tone, *ASPM* and *Microcephalin* (Dediu and Ladd, 2007), can indeed detect such Bayesian biases, especially when they are rare in the population and strong. However, it seems that the type of bias produced by the non-Bayesian $\mathbf{M_2}$ (Dediu, 2008b) better fits the intuitive notion of a genetic bias for tone (Dediu and Ladd, 2007; Ladd et al., 2008), but this fit is still far from perfect.

For now, the ultimate source(s) of this imperfection in modelling the notion of such a genetic bias is not known but it seems probable that it stems, at least partially, from the acquisitionist assumption. As extensively discussed in the language change literature (see Croft, 2000 for a good review), the theory that first language acquirers (children) determine language change by reanalysing the language they hear as being produced by a slightly different grammar (''acquisitionism'') fails to explain most of the empirical data available. Likewise, in the type of models discussed in this paper, after the acquisition period is over and the single winning grammar, h_w, is chosen, it becomes frozen in the adults. However, probably a more realistic alternative would be to allow adults not only to change their grammar through time and social context but also to be active innovators. This would certainly add extra complexity to our models but it would make them more plausible, as well. The kind of genetic biases discussed throughout this paper can act in the children as well as competent adult language users, but their action in the second condition has so far been neglected, representing a promising direction for future research.

In conclusion, it seems premature to draw any general conclusions concerning the process of language evolution and the specific interplay between nature (genetic biases) and nurture (cultural transmission) based on the models of first language acquisition in homogeneous chains of single Bayesian agents. However, these models do seem to suggest that when the structure of the language community is more complex, the relationship between the distribution of languages and the genetic bias(es) depends less on the particular type of language learner, and is complex and non-transparent. Therefore, they support the view that theories which explain the similarities and differences between languages as direct consequences of a genetically endowed ''Universal Grammar'' are probably wrong. The alternative view, that language variation and its constraints are due to a complex interaction between genetics, individual learning and cultural transmission across generations in populations is, in my view, better as describing the empirical data we have. I hope that further refining our models to include more complex communicative contexts and groups, and shifting the focus towards the competent language users as the agents of change will provide a better account of the complex phenomena involving the expression of genetic factors mediated by cultural transmission(Dediu and Ladd, 2007; Ladd et al., 2008).

Acknowledgements

The author thanks K. Smith, M. Dowman, S. Höfler, S. Kirby, A. Dima, M. Cysouw and an anonymous reviewer for discussions and comments.

References

Bonnet, E., Van de Peer, Y., 2002. Zt: a software tool for simple and partial mantel tests. J. Stat. Softw. 7, 1–12.
Christiansen, M.H., Chater, N., 2008. Language as shaped by the brain. Behav. Brain Sci. 31 (5), 489–508 (Discussion 509–58).
Croft, W., 2000. Explaining Language Change: An Evolutionary Approach. Pearson Eduction Limited, Harlow, England.
Dediu, D., 2008a. Causal correlations between genes and linguistic features—the mechanism of gradual language evolution. In: Smith, A.D.M., Smith, K., Ferrer i Cancho, R. (Eds.), The Evolution of Language: Proceedings of the 7th International Conference (EVOLANG7). World Scientific, Singapore, pp. 83–90.
Dediu, D., 2008b. The role of genetic biases in shaping language-genes correlations. J. Theor. Biol. 254, 400–407.
Dediu, D., Ladd, D.R., 2007. From the cover: linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and microcephalin. Proc. Natl. Acad. Sci. USA 104 (26), 10944–10949.
Dryer, M.S., 2008. Order of subject and verb. In: Haspelmath, M., Dryer, M.S., Gil David, B., Comrie (Eds.), The World Atlas of Language Structures Online. Max Planck Digital Library, Munich ⟨http://wals.info/feature/82⟩.
Edgington, E.S., 1987. Randomization Tests, second ed. Marcel Dekker, New York.
Feher, O., Mitra, P.P., Sasahara, K., Tchernichovski, O., 2008. Evolution of song culture in the zebra finch. In: Smith, A.D., Smith, K., Ferrer i Cancho, R. (Eds.), The Evolution of Language. World Scientific, Singapore, pp. 423–424.
Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Booth, M., Rossi, F., 2006. GNU Scientific Library Reference Manual, second ed. (v1.8). Network Theory Limited, UK.
Greenberg, J.H., 1966. Language Universals, Mouton.
Griffiths, T., Kalish, M., 2007. Language evolution by iterated learning with Bayesian agents. Cogn. Sci. 31 (3), 441–480.
Hockett, C.F., 1963. The problem of universals in language. In: Greenberg, J.H. (Ed.), Universals of Language. MIT Press, Cambridge, MA, pp. 1–29.
Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.
Honeybone, P., 2003. Review of: Holt, D.E. (Ed.), 2003 'optimality theory and language change. J. Linguist. 42, 726–731.
Kirby, S., Hurford, J., 2002. The emergence of linguistic structure: an overview of the iterated learning model. In: Cangelosi, A., Parisi, D. (Eds.), Simulating the Evolution of Language. Springer, London, pp. 121–148.
Kirby, S., Cornish, H., Smith, K., 2008. Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. Proc. Natl. Acad. Sci. USA 105 (31), 10681–10686.
Kirby, S., Dowman, M., Griffiths, T.L., 2007. Innateness and culture in the evolution of language. Proc. Natl. Acad. Sci. USA 104 (12), 5241–5245.
Ladd, D.R., Dediu, D., Kinsella, A.R., 2008. Languages and genes: reflections on biolinguistics and the nature–nurture question. Biolinguistics 2 (1), 114–126.
Levinson, S., 2003. Space in Language and Cognition: Explorations in Cognitive Diversity. Cambridge University Press, Cambridge.
Lightfoot, D., 1999. The Development of Language: Acquisition, Change and Evolution. Blackwell, Oxford.
Maddieson, I., 2008a. Consonant inventories. In: Haspelmath, M., Dryer, M.S., Gil David, B., Comrie (Eds.), The World Atlas of Language Structures Online. Max Planck Digital Library, Munich ⟨http://wals.info/feature/1⟩.
Maddieson, I., 2008b. Tone. In: Haspelmath, M., Dryer, M.S., Gil David, B., Comrie (Eds.), The World Atlas of Language Structures Online. Max Planck Digital Library, Munich ⟨http://wals.info/feature/13⟩.
Maddieson, I., 2008c. Vowel quality inventories. In: Haspelmath, M., Dryer, M.S., Gil David, B., Comrie (Eds.), The World Atlas of Language Structures Online. Max Planck Digital Library, Munich ⟨http://wals.info/feature/2⟩.
Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27 (2), 209–220.

Plomin, R., DeFries, J.C., McClearn, G.E., P.M., 2001. Behavioral Genetics, fourth ed. Worth Publishers, New York.
Press, S.J., 2003. Subjective and Objective Bayesian Statics. Wiley Series in Probability and Statistics, second ed. Wiley, New York.
R Development Core Team, 2007. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
Smith, K., Kirby, S., 2008. Cultural evolution: implications for understanding the human language faculty and its evolution. Phil. Trans. R. Soc. B 363, 3591–3603.
Stromswold, K., 2001. The heritability of language: a review and metaanalysis of twin, adoption and linkage studies. Language 77, 647–723.
Yip, M., 2002. Tone: Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge, UK.