Iterated learning in populations of Bayesian agents

Kenny Smith (kenny.smith@northumbria.ac.uk)
Cognition and Communication Research Centre, Division of Psychology, Northumbria University
Northumberland Building,Northumberland Road,Newcastle-upon-Tyne, NE1 8ST,UK

Abstract

Previous analytic results (Griffiths & Kalish, 2007) show that repeated learning and transmission of languages in populations of Bayesian learners results in distributions of languages which directly reflect the biases of learners. This result potentially has profound implications for our understanding of the link between the human language learning apparatus and the distribution of languages in the world. It is shown here that a variation on these models (such that learners learn from the linguistic behaviour of multiple individuals, rather than a single individual) changes this transparent relationship between learning bias and typology. This suggests that inferring learning bias from typology (or population behaviour from laboratory diffusion chains) is potentially unsafe.

Keywords: language learning; iterated learning; Bayesian learning; cultural evolution; language universals

Introduction

What is the relationship between the biases of language learners and the observed distribution of languages in the world? Under the standard generative account (e.g. Chomsky, 1965), a direct mapping is assumed between the mental apparatus of language learners and language structure. In the strongest possible form (e.g. Baker, 2001), the claim is that we can read off the structure of the language faculty from the typological distribution of languages in the world.

A second account which posits a similarly close match between the biases of language learners and the structure of language arises from considerations of cultural evolution (Christiansen & Chater, 2008). Rather than language structure being strongly constrained by a highly restrictive domain-specific learning apparatus, the idea is that languages have adapted over repeated episodes of learning and production in response to much weaker (and possibly domain-general) constraints arising from the biases of language learners. This process is sometimes called *iterated learning*: the outcome of learning at one generation provides the input to learning at the next. While typologically unattested languages might be both possible and even learnable, the languages we see in the world will typically be selected from the restricted set of *highly* learnable languages: languages which are hard to learn will tend to change, and those which are easy to learn will be preserved, eventually yielding languages which are uniformly well-fitted to the biases of language learners. We have previously termed this evolutionary pressure *cultural selection for learnability* (Brighton, Kirby, & Smith, 2005).

Are learner biases the only factor shaping the distribution of languages in the world? It has been argued (see e.g. Kirby, 2002; Zuidema, 2003; Brighton, Kirby, & Smith, 2005; Kirby, Dowman, & Griffiths, 2007) that, at a minimum, language must be seen as a compromise between two factors: the biases of learners, and other constraints acting on languages during their transmission. The classic example of this second constraint is the mismatch between the infinite expressivity of languages and the finite set of data from which such languages must be learned. This *transmission bottleneck* favours languages which can be recreated from a subset via generalisation. Recursive compositionality is one such generalisation (e.g. Kirby, 2002; Brighton, 2002), and therefore represents an adaptation by language in response to pressure arising from transmission factors external to the human mind. While this evolutionary process requires certain learner biases (e.g. ability to generalise), it does not arise as a consequence of these learning biases alone, but is modulated by the transmission bottleneck (Brighton, Smith, & Kirby, 2005). This suggests that the biases of language learners can't simply be read off from typological distributions.

However, this transmission-mediated view of the relationship between learning biases and typology has recently been thrown into doubt by some modelling work in the Bayesian framework. As discussed below, Griffiths and Kalish (2007) show that iterated learning in populations of Bayesian learners produces outcomes which are solely determined by the biases of language learners: in other words, in the linguistic case, the relationship between learning bias and language typology might be a transparent one after all.

It is shown here that a variant of Griffiths & Kalish's model (where each learner selects a single grammar after observing data produced by *multiple* individuals, rather than a single individual) leads to a blurring of the relationship between prior biases of learners and outcomes of cultural evolution: populations of Bayesian agents converge on distributions of languages which are dependent on both the biases of language learners and transmission factors (such as the diversity of models a learner is exposed to).

Summary of iterated learning results for Bayesian learners

Bayesian learners select a hypothesis h according to its posterior probability in light of some data d:

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_h P(d|h)P(h)} \qquad (1)$$

$P(d|h)$ gives the likelihood of data d being produced under hypothesis h, and $P(h)$ gives the prior probability of each hypothesis. For models of iterated learning of language, the set of hypotheses are interpreted as the set of possible grammars, data are sets of utterances from which learners must induce a language, and the prior probability distribution over grammars arises from the bias (domain-specific or domain-general, innate or learned) of learners.

Previous analytic and numerical results in this framework (primarily Griffiths & Kalish, 2007; Kirby et al., 2007) show that the relationship between the prior biases of learners and the outcomes of cultural evolution depends critically on how learners select a grammar given the posterior probability distribution over possible grammars.

When learners select a grammar with probability proportional to its posterior probability (known as *sampling* from the posterior), the stable outcome of cultural evolution (the stationary distribution) is simply the prior distribution (Griffiths & Kalish, 2007). This is true regardless of the initial distribution over grammars or transmission factors such as the amount of data learners receive or the amount of noise on transmission: iterated learning in populations of samplers results in convergence on the prior. As discussed above, this suggests a transparent relationship between the prior bias of learners and the observed distribution of languages in the world: the typological distribution exactly reflects the biases of learners.

On the other hand, when learners select the hypothesis with the maximum a posteriori probability (MAP selection), the relationship between prior bias and the stationary distribution is more complex (Griffiths & Kalish, 2007; Kirby et al., 2007). The distribution of languages produced by cultural evolution will reflect the ordering of hypotheses in the prior, but differences in prior probability are magnified, such that the a priori most likely hypothesis is overrepresented in the stationary distribution. Furthermore, different priors can lead to the same stationary distribution, and changing transmission factors (amount of data, noise, etc) can result in convergence to a different stationary distribution. In MAP populations, the relationship between learner biases and typological distributions is therefore somewhat opaque.

These models suggest that the sampling / MAP opposition is a critical one for understanding the relationship between learner biases and the distribution of languages in the world. While the true nature of the human hypothesis selection strategy is ultimately an empirical question, it is worth probing the assumptions behind the formal results presented above. Griffiths & Kalish's sampling result holds in two cases:

1. Populations are treated as long thin chains, with a single individual per generation, and transmission occurring between adjacent generations in the chain (in the classic iterated learning model configuration). In this case, the temporal distribution of grammars over multiple generations converges to the prior: while any grammar may be in use in the chain at a particular generation, on average the usage of the various grammars reflects their prior probability.

2. Populations are infinitely large, organised into discrete generations, and each individual learns from a single model at the previous generation.[1] In this case the distribution of interest is the proportion of individuals in the population using each grammar at a particular generation, which converges to the prior over time.

This equivalence between chains and populations is an interesting and potentially important one, since it suggests that we can obtain useful insights into the behaviour of real-world populations by studying long thin diffusion chains (either formally or in the laboratory: Griffiths, Kalish, & Lewandowsky, 2008).

A two-grammar model

Given the potential implications of these results, it would be interesting to know whether the equivalence between chains and populations holds in situations where each learner learns from more than a single model, potentially including models from the same generation. A simple two-grammar model can be used to explore (at present, numerically) this issue.

Model details

We assume that populations are infinitely large, and are organised into discrete generation. Learners observe a set of b utterances, produced by (one or more) models selected from the immediately preceding generation of the population, and subsequently select a grammar with probability proportional to its posterior probability in light of that data (i.e. they sample from the posterior). Note that, importantly, learners are required to select a single grammar, despite potentially being provided with data produced by multiple grammars, an issue we return to in the discussion.

There are two grammars, h_0 and h_1, and two utterances, d_0 and d_1. Individuals produce single utterances as follows (where ε gives the probability of noise on production):

$$\begin{aligned} P(d_x|h_x) &= 1-\varepsilon \\ P(d_{y\neq x}|h_x) &= \varepsilon \end{aligned} \tag{2}$$

Given that the population is infinitely large and there are only two grammars, we simply track p_t, which is the proportion of individuals at generation t who select hypothesis h_0 after learning ($1-p_t$ gives the proportion selecting h_1).

If learners at generation $t+1$ learn from a *single model* selected at random from generation t, the proportion of individuals using h_0 at time $t+1$ will be

$$p_{t+1} = \sum_d P(h_0|d).\left(p_t.\prod_{x\in d} P(x|h_0) + (1-p_t).\prod_{x\in d} P(x|h_1)\right) \tag{3}$$

where the sum is over all possible data sets and the products are over the individual utterances in each data set. In other words, each learner learns from an h_0 model with probability p_t and an h_1 model with probability $1-p_t$, and subsequently selects h_0 with probability determined by the data produced by that model.

Alternatively, if a learner learns from *multiple models*, each utterance in their data set may be produced by a different individual, possibly using a different grammar. We will assume

[1] It is worth noting that a number of non-Bayesian, population-biology inspired models of language evolution similarly focus on situations where learners learn from a single model (e.g. Nowak, Komarova, & Niyogi, 2001).

that the model for each item of data is independently selected from the population at the preceding generation, which gives the following expression for the proportion of individuals selecting h_0 at generation $t+1$:

$$p_{t+1} = \sum_{d} P(h_0|d) . \prod_{x \in d} \left(p_t . P(x|h_0) + (1-p_t) . P(x|h_1) \right) . \quad (4)$$

Again, the sum is over all possible sets of data, and the product is over the items in each data set, where each utterance is produced by an individual using either grammar h_0 or h_1 (according to the proportions of those two grammars in the population).

Results

The main result (see Figure 1a) is that, when learners learn from multiple models, the proportion of individuals using each grammar (after cultural evolution has run its course) is no longer the same as the prior distribution. Rather, one language predominates, with the winning language being determined by the starting proportions of the two grammars and their prior probability.[2]

Figure 1b shows this sensitivity to initial proportions of the two languages in a little more detail. There is a critical value of the initial proportion of h_0 (at around 0.4465 for this combination of parameters): for initial proportions below this, h_1 eventually dominates, otherwise h_0 dominates. This sensitivity to initial conditions is not found in the single model treatments discussed above.

When learners learn from multiple models, the insensitivity to transmission factors such as amount of data (b) normally seen in populations of samplers also disappears. This is illustrated in Table 1. Notice that the effect of increased amounts of data (higher b) runs in the opposite direction to that seen in MAP populations: whereas in the chains of MAP learners described in Kirby et al. (2007) *less* data gives greater exaggeration of the prior, here *more* data gives greater exaggeration of the prior preference for h_0. Note also that, as b increases, the impact of the strength of prior preference for h_0 on the final proportion of h_0 in the population diminishes — in essence, when $b \geq 3$, the population converges on h_0 regardless of strength of prior bias in favour of that grammar. This is reminiscent of the MAP phenomenon of insensitivity to strength of prior bias, but is modulated by b.

Why does b, the amount of data learners receive, lead to this departure from the known sampler results? In a mixed population, increasing b increases the diversity of learner's sample of the population's linguistic behaviour (unlike in the case where learners learn from a single model, when they simply receive an increasingly accurate reflection of the grammar of that model). Consequently, if one grammar predominates in the population, this is likely to be reflected in the data learners see.

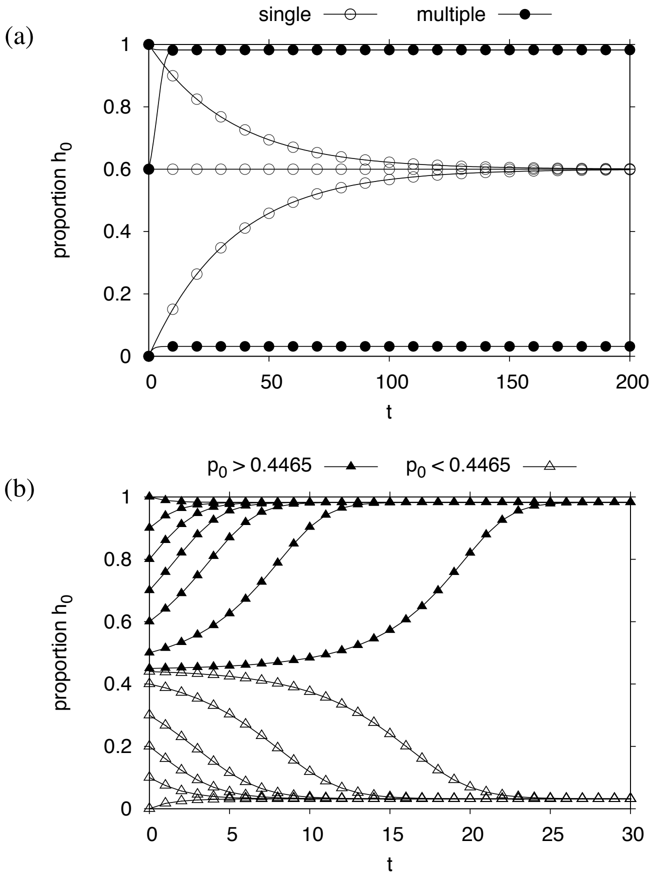

Figure 1: $P(h_0) = 0.6$, $b = 3$, $\varepsilon = 0.05$. (a) When individuals learn from a single model, the population converges to the prior. When learners learn from (potentially) multiple models, populations converge to one of two stable states, depending on initial conditions. (b) When learners learn from multiple models, the eventual distribution is sensitive to the starting proportions of the two grammars.

Table 1: Stable proportions of h_0 for various values of $p(h_0)$ and b. Populations initialised with equal proportions of h_0 and h_1, $\varepsilon = 0.05$.

	b				
$P(h_0)$	1	2	3	4	5
0.51	0.51	0.548	0.978	0.989	0.997
0.6	0.6	0.822	0.983	0.992	0.998
0.7	0.7	0.92	0.986	0.994	0.998
0.8	0.8	0.961	0.99	0.996	0.999
0.9	0.9	0.983	0.993	0.998	0.999

How do Bayesian learners respond to mixed samples? The grammar which matches with the majority of the data has higher posterior probability and is therefore likely to be selected. Importantly, under a wide range of conditions, the grammar matching the more common data is *disproportion-*

[2] See Niyogi (2006) for a number of more general analytic results providing the dynamics of transmission in populations associated with various non-Bayesian learning algorithms.

ately preferred. Given a data set consisting of i d_0 items and j d_1 items, $i \geq j$, the ratio of likelihoods $P(d|h_0)/P(d|h_1)$ is $(\frac{1-\varepsilon}{\varepsilon})^{i-j}$. This quantity is generally greater than the corresponding ratio of data items (i/j) for low noise rates. In other words, learners exposed to a mixed sample and required to select a single grammar are disproportionately likely to pick the more frequently represented grammar, making Bayesian learning in this context a type of conformist frequency-dependent learning (Boyd & Richerson, 1985). The well-know consequence of conformist learning is the rich-get-richer behaviour seen here, with the mismatch in frequencies of the two grammars increasing generation on generation.

Conformity bias is not, however, the whole story. Increasing b has a second effect: as well as increasing the representativeness of the sample of the population's linguistic behaviour, it also increases the fidelity of transmission of the majority grammar in a sample of a fixed diversity (holding i/j constant, increasing b increases the quantity $i-j$). Both these effects lead to an increase in the dominant grammar's share of the population. The impact of the two effects can be probed by implementing a minor extension to the model outlined above, where learners learn from a specified number of models (c), with b/c data items from each parent. Table 2 shows the eventual proportion of h_0 in converged populations for various c and b (for convenience we only consider cases where b/c yields integer values). As can be seen from the table, increasing c or b independently increases the dominance of the winning grammar. Importantly, *any* diversity of models ($c \geq 2$) results in a single grammar winning out. For $b = 2$, the grammar favoured by the prior wins out in situations where learners received perfectly mixed input, and for $b > 2$ the conformity effect outlined above also comes into play.

Table 2: Stable proportion of h_0 for various c and b. $P(h_0) = 0.6$, $\varepsilon = 0.05$, both grammars initially equally frequent.

	b						
c	1	2	3	4	6	8	12
1	0.6	0.6	0.6	0.6	0.6	0.6	0.6
2	-	0.822	-	0.964	0.993	0.999	1
3	-	-	0.983	-	0.999	-	1
4	-	-	-	0.992	-	1	1

A more complex model

While the results for the two-grammar model are potentially interesting, one might reasonably worry that they are reliant on some feature of the simplest possible two-grammar model. Of particular interest are the models in the literature which allow multiple grammars with equal prior probability. With this in mind, the grammar model from Kirby et al. (2007) is adopted here: similar results can be obtained for the 260-grammar model of Griffiths and Kalish (2007).

The model

In this more complex model, a language consists of a system for expressing m meanings, where each meaning can be expressed using one of k means of expression, called *signal classes*. In a perfectly regular (or systematic) language the same signal class will be used to express each meaning — for example, the same compositional rules will be used to construct an utterance for each meaning. Following Kirby et al. (2007), we assume that learners have a preference for languages which use a consistent means of expression, such that each meaning is expressed using the same signal class. This prior is given by the expression

$$P(h) = \frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k \Gamma(m+k\alpha)} \prod_{j=1}^{k} \Gamma(n_j + \alpha) \tag{5}$$

where $\Gamma(x) = (x-1)!$ when x is an integer,[3] n_j is the number of meanings expressed using class j and $\alpha \geq 1$ determines the strength of the preference for regularity: low α gives a strong preference for regular languages, higher α leads to a weaker preference for such languages.

The probability of a particular meaning-form pair $\langle x,y \rangle$ (consisting of a meaning x and a signal class y) being produced by an individual with grammar h is:

$$P(\langle x,y \rangle | h) = \frac{1}{m} \cdot \begin{cases} 1-\varepsilon & \text{if } y \text{ is the class for } x \text{ in } h \\ \frac{\varepsilon}{k-1} & \text{otherwise} \end{cases} \tag{6}$$

where ε gives the noise probability on production and all meanings are equiprobable (hence the scaling by $1/m$).

We can then plug this production model into the two population learning models outlined above. $p_{i,t}$ gives the proportion of individuals at generation t who select h_i (again, learners are required to select a single grammar). If learners at generation $t+1$ learn from a *single model*:

$$p_{i,t+1} = \sum_d \sum_j P(h_i|d).p_{j,t}.\prod_{x \in d} P(x|h_j) \tag{7}$$

where the sums are over all possible data sets and all possible model grammars, and the products are over the b items in each data set. If a learner learns from *multiple models*:

$$p_{i,t+1} = \sum_d P(h_i|d).\prod_{x \in d} \left(\sum_j p_{j,t}.P(x|h_j) \right) . \tag{8}$$

Results

The main features of the two-grammar model are preserved in the more complex model: sensitivity to initial conditions, a dependency on b, and an interaction between strength of prior and b.

Figure 2 shows the final stable distribution over all grammars for strong and weak prior preferences in favour of regularity, for various values of b. For $b = 1$ the standard sampling

[3] We will only consider the case where α takes integer values.

result for learning from a single cultural parent is retrieved. For high b, the majority of the population converges on one of the a priori more likely grammars (with the identity of the winning grammar depending on the initial frequencies). Indeed, for $b = 10$ the strength of the prior preference in favour of regularity makes little difference to the final distribution.

Finally, there appears to be a critical value of b required for the population to converge on a single majority grammar. For b below this critical value, the would-be dominant grammar suffers from a lack of transmission fidelity: learners tend to receive data sets which underspecify the target language, and the posterior probabilities of the various languages are therefore heavily constrained by the prior. Note, however, that the stable distribution is not identical to the prior: the differences in prior probability are smoothed out somewhat. Above the critical value of b (which is around $b = 2m$, but is somewhat sensitive to α), transmission fidelity becomes sufficiently high to allow one grammar to dominate through the processes discussed for the two-grammar model. This constraint on b is analogous to the coherence threshold described in Nowak et al. (2001).

Discussion

The two models described above represent a first attempt to explore the impact of population structure on the outcomes of iterated learning in populations of Bayesian agents. While much remains to be done, they show that the analytic result provided by Griffiths and Kalish (2007) can break down under some model configurations. Before considering the implications of this finding, it is worth considering some of the model's more serious limitations.

Learners are required to select a single grammar based on exposure to a potentially diverse sample (or equivalently, learners use each grammar probabilistically, with probabilities determined by their posterior probability). It may be that there are more sophisticated treatments of the hypothesis selection task for which the Griffiths and Kalish (2007) result can be retrieved. One obvious possibility is to consider cases where learners have a structured model, such that they appreciate that their data potentially comes from multiple individuals who may (or may not) use different grammars and who may (or may not) exhibit consistent usage. Hierarchical Bayesian approaches can be straightforwardly used to model this type of learner, and can therefore be used to explore the evolutionary consequences of learning from multiple models in a more satisfying (and cognitively plausible) fashion than that described here.

The population model used here also offers only a minimal increase in complexity over its predecessors. Although it adds the possibility of learners learning from multiple (equally-weighted) models, it entirely ignores horizontal (within-generation) transmission. Populations also lack any interesting internal structure. Transmission in real-world populations takes place over complex social networks, with implications for language structure (see e.g. Kerswill & Williams, 2000),

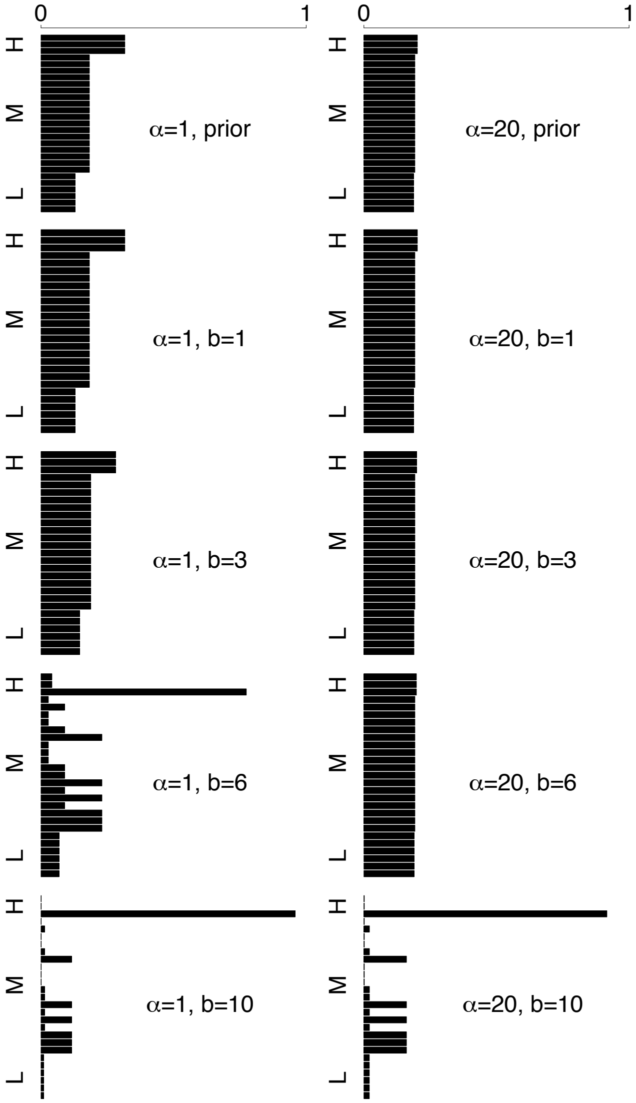

Figure 2: $m = 3, k = 3, \varepsilon = 0.05$. Stable proportions of all 27 grammars, grouped by prior probability (H are the highly regular grammars, L are the low regularity grammars, M are the grammars of intermediate regularity). The top row gives prior probability distributions for two values of α. The remaining rows give the stable proportions for various values of b. In all cases the population is initialised with one regular grammar having a slightly boosted frequency and all other grammars being equally frequent. All proportions have undergone a square root transformation to show the variability among the less frequent grammars.

a phenomenon little explored in the modelling literature.

The results presented here suggest that caution must (at least at present) be used when extrapolating from cultural evolution in convenient one-individual iterated learning chains to larger populations. While diffusion chain experiments provide a powerful tool for identifying the prior biases of learners, in real populations those prior biases are fed in to a population dynamic whose consequences are largely not understood. Exploring transmission in larger and more complex laboratory populations may prove necessary.

Secondly, the implication of this modelling work is that assuming a straightforward or one-to-one mapping between the biases of learners and the typological distributions of languages in the world may be unsafe. While the Griffiths and Kalish (2007) result offers some hope in this direction (and may indeed still hold given a more sophisticated treatment of the multiple-model case, of the sort discussed above), the results presented here shows that, under certain assumptions about the nature of the learning problem, changing the population dynamic can change the outcomes of cultural evolution: the typological distribution of languages in these models is emphatically not the prior.

Furthermore, in some situations the *ranking* of languages in the prior is not even preserved in the final distribution of languages. Previous work in the Bayesian framework suggests that cultural evolution returns some distribution which is the same shape as the prior distribution — either the prior itself, or (in MAP populations) some stretched version of the prior where the magnitudes of the differences are changed but the overall rank ordering of hypotheses in the stationary distribution is the same as in the prior. However, close inspection of Figure 2 shows a different picture: some languages with high or intermediate prior probability end up being less frequent than languages with the lowest prior probability. In other words, attempting to read off even the ranking of languages in the prior from this typological distribution would lead to error. In the model this is largely due to competition in a homogeneous population between the languages with high prior probability, combined with the coupling of each mid-ranking language to high-ranking languages with which they overlap. However, the point remains: the outcome of cultural transmission in populations where learners learn from multiple models is not necessarily simply determined by the prior, but also by transmission factors such as the amount of data learners receive and the diversity of the set of models they receive it from. The relationship between prior biases of learners and observed typological distributions is not transparent.

Conclusions

Some work on the outcomes of cultural transmission in populations of Bayesian learners suggests that cultural evolution will deliver up the prior distribution. This implies that we can gain insights into population behaviours by studying diffusion chains (highly amenable to laboratory investigation), and that we can read off the prior biases of learners from the typological distributions of languages in the world. The results presented here show that this modelling result is dependent on learners learning from a single model. When this idealisation is relaxed, the straightforward mapping from prior bias to typology breaks down.

Acknowledgements

Thanks to Dan Dediu, Mike Dowman, Tom Griffiths, Mike Kalish, Tim O'Donnell and Jelle Zuidema for helpful comments.

References

Baker, M. (2001). *The atoms of language: The mind's hidden rules of grammar*. New York, NY: Basic Books.

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. Chicago, IL: University of Chicago Press.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life*, *8*, 25–54.

Brighton, H., Kirby, S., & Smith, K. (2005). Cultural selection for learnability: Three principles underlying the view that language adapts to be learnable. In M. Tallerman (Ed.), *Language origins: Perspectives on evolution* (pp. 291–309). Oxford: Oxford University Press.

Brighton, H., Smith, K., & Kirby, S. (2005). Language as an evolutionary system. *Physics of Life Reviews*, *2*, 177–226.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Christiansen, M., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489–509.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*, 441–480.

Griffiths, T. L., Kalish, M. L., & Lewandowsky, S. (2008). Theoretical and experimental evidence for the impact of inductive biases on cultural evolution. *Philosophical Transactions of the Royal Society*, *363*, 3503–3514.

Kerswill, P., & Williams, A. (2000). Creating a new town koine: Children and language change in Milton Keynes. *Language in Society*, *29*, 65–115.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In E. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (pp. 173–203). Cambridge: Cambridge University Press.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences, USA*, *104*, 5241–5245.

Niyogi, P. (2006). *The computational nature of language learning and evolution*. Cambridge, MA: MIT Press.

Nowak, M. A., Komarova, N. L., & Niyogi, P. (2001). Evolution of universal grammar. *Science*, *291*, 114–117.

Zuidema, W. H. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15* (pp. 51–58). Cambridge, MA: MIT Press.