NATURAL SELECTION FOR COMMUNICATION FAVOURS THE CULTURAL EVOLUTION OF LINGUISTIC STRUCTURE

KENNY SMITH

Division of Psychology, Northumbria University, Northumberland Road, Newcastle-upon-Tyne, NE1 8ST, UK
kenny.smith@northumbria.ac.uk

SIMON KIRBY

Language Evolution and Computation Research Unit, University of Edinburgh, 40 George Square, Edinburgh, EH8 9LL, UK

There are two possible sources of structure in language: biological evolution of the language faculty, or cultural evolution of language itself. Two recent models (Griffiths & Kalish, 2005; Kirby, Dowman, & Griffiths, 2007) make alternative claims about the relationship between innate bias and linguistic structure: either linguistic structure is largely determined by cultural factors (Kirby et al., 2007), with strength of innate bias being relatively unimportant, or the nature and strength of innate machinery is key (Griffiths & Kalish, 2005). These two competing possibilities rest on different assumptions about the learning process. We extend these models here to include a treatment of biological evolution, and show that natural selection for communication favours those conditions where the structure of language is primarily determined by cultural transmission.

1. Introduction

Language is a consequence of two systems of transmission: biological and cultural. The human capacity for language uncontroversially has some grounding in specifically human biology — no other species uses a similar system in the wild. Language is also, again uncontroversially, socially learned — we learn the language of our speech community.

To what extent is the detailed structure of language determined by biology or culture, and how have cultural and biological evolution acted to shape language? The position here is less clear. The standard account attributes the structure of language to the biological evolution of an innate language faculty (Pinker & Bloom, 1990). An alternative account, grounded in the computational modelling of cultural transmission, allows a significant role for cultural evolution (e.g. Kirby & Hurford, 2002; Kirby, Smith, & Brighton, 2004): under this account, the structure of language is explained primarily as a consequence of the adaptation of language to the cultural transmission medium (e.g. partial, noisy, or frequency-skewed data:

Kirby, 2001).

Two recent studies have sought to explicitly address the link between language structure, biological predispositions, and constraints on cultural transmission (Griffiths & Kalish, 2005; Kirby et al., 2007). Both assume that learners apply the principles of Bayesian inference to language acquisition: a learner's confidence that a particular grammar h accounts for the linguistic data d that they have encountered is given by

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

and allows a contribution both from a prior (presumably innate) belief in each grammar, $P(h)$, and the probability that that grammar could have generated the observed data, $P(d|h)$. Based on their confidence in the various grammars, $P(h|d)$, the learner then selects a grammar and produces utterances which will form the basis, through social learning, of language acquisition in others.

Within this framework, Griffiths and Kalish (2005) show that cultural transmission factors (such as noise or the transmission bottleneck imposed by partial data) have *no* effect on the distribution of languages delivered by cultural evolution: the outcome of cultural evolution is solely determined by the prior biases of learners, given by $P(h)$.[a] Kirby et al. (2007) demonstrate that this result is a consequence of the assumption that learners select a grammar proportionally to $P(h|d)$ — if learners instead select the grammar which maximises $P(h|d)$, then cultural transmission factors play an important role in determining the distribution of languages delivered by cultural evolution: for example, different transmission bottlenecks lead to different distributions. Furthermore, for maximising learners, the strength of the prior bias of learners is irrelevant over a wide range of the parameter space.[b]

These models suggest two candidate components of the innate language faculty: firstly, the prior bias, $P(h)$, and secondly, the strategy for selecting a grammar based on $P(h|d)$ — sampling proportional to $P(h|d)$, or selecting the grammar which maximises $P(h|d)$. We can therefore straightforwardly extend models of this sort to ask how we might expect the evolution of the language faculty to unfold: does biological evolution favour sampling or maximising learners, strong or weak priors?

Specifically, we are interested in asking which selection strategies and priors are evolutionarily stable (Maynard Smith & Price, 1973; Smith, 2004): which strategies and priors are such that a population adopting that strategy or prior will

[a] Griffiths and Kalish (2005) point out that the prior need not necessarily take the form of a *language specific* innate bias in the traditional sense.

[b] For a treatment of both sampling and maximising learners, see Griffiths and Kalish (2007), who provide similar results to those of Griffiths and Kalish (2005) and Kirby et al. (2007).

not be invaded by some other strategy or prior under the influence of natural selection? This breaks down into two sub-questions: (1) what language will a population consisting entirely of individuals with a particular strategy and prior have?; (2) what level of communicative accuracy will some individual inserted into such a population have? The first question is answered by the work of Griffiths and Kalish (2005) and Kirby et al. (2007), which shows the relationship between prior, selection strategy, cultural transmission factors and distribution of languages in a population. Answering the second requires some additional machinery, described in Section 3.

2. The model of learning and cultural transmission

We adopt Kirby et al.'s (2007) model of language and language learning. A language consists of a system for expressing m meanings, where each meaning can be expressed using one of k means of expression, called *classes* (e.g. a particular morphosyntactic paradigm). We will assume two types of prior bias. For *unbiased* learners, all grammars have the same prior probability: $P(h) = 1/k^m$. *Biased* learners have a preference for languages which use a consistent means of expression, such that each meaning is expressed using the same class. Following Kirby et al. (2007), this prior is given by the expression

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

where $\Gamma(x) = (x-1)!$, n_j is the number of meanings expressed using class j and α determines the strength of the preference for consistency: low α gives a strong preference for consistent languages, higher α leads to a weaker preference for such languages.

The probability of a particular data set d (consisting of b meaning-form pairs) being produced by an individual with grammar h is:

$$P(d|h) = \frac{1}{m} \prod_{\langle xy \rangle \in d} P(y|x,h)$$

where all meanings are equiprobable, x is a meaning, y is the signal class associated with that meaning in the data, and $P(y|x,h)$ gives the probability of y being produced to convey x given grammar h and noise ϵ:

$$P(y|x,h) = \begin{cases} 1-\epsilon & \text{if } y \text{ is the class corresponding to } x \text{ in } h \\ \frac{\epsilon}{k-1} & \text{otherwise} \end{cases}$$

Bayes' rule can then be applied to give a confidence distribution over hypotheses given a particular set of utterances. This confidence distributions is

used by a learner to select a grammar, according to one of two strategies. Sampling learners simply select a grammar proportional to its posterior probability: $P_L(h|d) = P(h|d)$. Maximising learners select the grammar with the highest posterior probability:

$$P_L(h|d) = \begin{cases} 1 & \text{if } P(h|d) > P(h'|d) \text{ for all } h' \neq h \\ 0 & \text{otherwise} \end{cases}$$

A model of cultural transmission follows straightforwardly from this model of learning: the probability of a learner at generation n arriving at grammar h_n given exposure to data produced by grammar h_{n-1} is simply

$$P(h_n = i|h_{n-1} = j) = \sum_d P_L(h_n = i|d)P(d|h_{n-1} = j)$$

The matrix of all such transition probabilities is known as the Q matrix (Nowak, Komarova, & Niyogi, 2001): entry Q_{ij} gives the transition probability from grammar j to grammar i. As discussed in Griffiths and Kalish (2005) and Kirby et al. (2007), the stable outcome of cultural evolution (the *stationary distribution* of languages) can be calculated given this Q matrix, and is proportional to its first eigenvector. We will denote the probability of grammar i in the stationary distribution as Q^*_i.

Table 1 gives some example prior probabilities and stationary distributions, for various strengths of prior and both selection strategies.[c] As shown in Table 1, strength of prior determines the outcome of cultural evolution for sampling learners, but is unimportant for maximising learners as long as some bias exists.

Table 1. $P(h)$ for three grammars given various types of bias (unbiased, weak bias [$\alpha = 40$], strong bias [$\alpha = 1$], denoted by u, bw and bs respectively), and the frequency of those grammars in the stationary distribution for sampling and maximising learners. Grammars are given as strings of characters, with the first character giving the class used to express the first meaning and so on.

h	$P(h)$			Q^*, sampler			Q^*, maximiser		
	u	bw	bs	u	bw	bs	u	bw	bs
aaa	0.0370	0.0389	0.1	0.0370	0.0389	0.1	0.0370	0.2499	0.2499
aab	0.0370	0.0370	0.0333	0.0370	0.0370	0.0333	0.0370	0.0135	0.0135
abc	0.0370	0.0361	0.0167	0.0370	0.0361	0.0167	0.0370	0.0014	0.0014

3. Evaluating evolutionary stability

In order to calculate which selection strategies and priors are evolutionarily stable we need to define a measure which determines reproductive success. We make

[c] All results here are for $m = 3, k = 3, b = 3, \epsilon = 0.1$. Qualitatively similar results are obtainable for a wide range of the parameter space.

the following assumptions: (1) a population consists of several subpopulations; (2) each subpopulation has converged on a single grammar through social learning, with the probability of each grammar being used by a subpopulation given by that grammar's probability in the stationary distribution; (3) natural selection favours learners who arrive at the same grammar as their peers in a particular subpopulation, where peers are other learners exposed to the language of the subpopulation. Given these assumptions, the communicative accuracy between two individuals A and B is given by:

$$ca(A, B) = \sum_{h} \sum_{h'} Q^{A}_{hh'}.Q^{B}_{hh'}.Q^{*}_{h'}$$

where the superscripts on Q indicates that learners A and B may have different selection strategies and priors. The *relative communicative accuracy* of a single learner A with respect to a large and homogeneous population of individuals of type B is therefore given by $ca(A, B)/ca(B, B)$. Where this quantity is greater than 1 the combination of selection strategy and prior (the *learning behaviour*) of individual A offers some reproductive advantage relative to the population learning behaviour, and may (through natural selection acting on genetic transmission) come to dominate the population. Where relative communicative accuracy is less than 1 learning behaviour A will tend to be selected against, and when relative communicative accuracy is 1 both learning behaviours are equivalent and genetic drift will ensue.

Table 2. Relative communicative accuracy of each strategy played off against all alternatives. s denotes sampling, m maximising, bias types are as for Table 1. Cases in which the minority learning behaviour can potentially invade the population via drift are boxed. Cases where the minority learning behaviour will be positively selected for are boxed and shaded. Values are given to two decimal places unless rounding would obscure a selection gradient.

		Minority behaviour					
		⟨s,u⟩	⟨s,bw⟩	⟨s,bs⟩	⟨m,u⟩	⟨m,bw⟩	⟨m,bs⟩
Majority behaviour	⟨s,u⟩	—	0.99998	0.98	1.12	1.12	1.12
	⟨s,bw⟩	0.9997	—	0.99	1.12	1.14	1.14
	⟨s,bs⟩	0.81	0.82	—	0.92	1.39	1.39
	⟨m,u⟩	0.88	0.88	0.86	—	1.00	1.00
	⟨m,bw⟩	0.38	0.38	0.60	0.45	—	1.00
	⟨m,bw⟩	0.38	0.38	0.60	0.45	1.00	—

Table 2 gives the relative communicative accuracies of 6 learning behaviours when played against each other: two selection strategies and three types of prior bias. Several results are apparent. Firstly, none of the sampling behaviours are

evolutionarily stable: all are prone to invasion by biased maximisers, and all but the strongly biased samplers are subject to invasion by unbiased maximisers.

Secondly, abstracting away from strength of prior, maximising is an ESS: samplers entering a maximising population have low relative communicative accuracy. In other words, natural selection prefers maximisers, at least under the fitness function described above. Maximisers boost the probability that the most likely grammar will be learned, and are consequently more likely to arrive at the same grammar as some other learner exposed to the same data-generating source.

Thirdly, strength of prior is relatively unimportant. In sampling populations (where the stationary distribution is determined by strength of prior), it is best to have the same strength of prior as the rest of the population. If your prior is stronger than the norm, you will be less likely to learn the less common languages from the stationary distribution, if it is weaker you be more likely to misconverge on those minority languages, which are themselves less likely to occur due to the stronger bias of the population.

The situation regarding the evolution of priors in maximising populations is slightly more complex. Strong and weak biases for maximisers turn out to be equivalent: for the parameter settings used here (and a wide range of other parameter settings) $\alpha = 1$ and $\alpha = 40$ generate equivalent Q matrices (and hence equivalent stationary distributions, as shown by Kirby et al., 2007). Strong and weak biases in maximising populations are therefore equivalent in terms of communicative accuracy, and can invade each other by drift.

In unbiased maximising populations, all levels of bias are interchangeable: all languages are equally probable, and the preference of biased learners for consistent languages is counterbalanced by their difficulty in acquiring the equally probable inconsistent languages. Unbiased maximising populations can therefore be invaded by drift by biased maximisers. However, unbiased maximisers cannot in turn invade biased maximising populations: in such populations, as can be seen in Table 1, the distribution of languages is skewed in favour of consistent languages, and it therefore pays to be biased to acquire these languages. While unbiased maximisation is therefore not an ESS, biased (strong or weak) maximisation forms a compound ESS.

If we assume that strong prior biases have some (arbitrarily small) cost, then only weak bias and unbiased maximisation would be evolutionarily stable. Given that unbiased maximising populations are more prone to invasion by biased maximisers than the reverse, there will be some high value of α, which we will call α^*, for which: (1) the prior is sufficiently weak that its costs relative to the unbiased strategy are low enough to allow the $\langle$m,$\alpha^*\rangle$ behaviour to invade $\langle$m,u$\rangle$ populations by drift; (2) the prior remains sufficiently strong that the $\langle$m,$\alpha^*\rangle$ population is resistant to invasion by $\langle$m,u$\rangle$, due to the selection asymmetry discussed above.

Under such a scenario, $\langle$m,$\alpha^*\rangle$ becomes the sole ESS: evolution will favour maximisation and the weakest possible (but not flat) prior. The actual value of

α^* will depend on the cost function used. For example, if we assume that higher values of α are associated with decreasing costs, but high α (say $\alpha = 100$, which yields a Q matrix identical to that for $\alpha = 40$ under the parameters used here) has a cost very close to that associated with a flat prior, then $\langle \text{m}, \alpha = 100 \rangle$ becomes the sole ESS: it benefits from both low costs and a skewed stationary distribution. While a more principled cost function is desirable, the insensitivity of the stationary distribution to α for maximising learners and the factorial in the expression for $P(h)$ means we have been unable to explore sufficiently large values of α under more complex treatments of cost.

4. Discussion and conclusions

The main result from this analysis of evolutionary stability is that maximising is always preferred over sampling: combining this with the findings of Griffiths and Kalish (2005) and Kirby et al. (2007), we can conclude that evolution prefers precisely those circumstances in which strength of prior bias has least effect and cultural evolution (driven by transmission factors such as the bottleneck and utterance frequency) has the greatest scope to shape the linguistic system.

The second result to highlight is that the strength of the prior is relatively unimportant from the perspective of biological evolution. In the (disfavoured) sampling strategies, it is best to have the same bias as the rest of the population. In maximising populations some bias is better than no bias, but strength of that bias is unimportant. Furthermore, if we assume that strong biases have some cost, then evolution will prefer the weakest bias possible.

This latter result runs counter to the phenomenon known as the Baldwin effect (Baldwin, 1896; Briscoe, 2000), whereby initially learned traits tend to become nativised. We note that this model is not designed to elicit the Baldwin effect — nativisation of a particular language is not allowed by our definition of prior bias, and the Baldwin effect requires that learning be costly, whereas in our model it is costless. However, we also note in passing that *not* learning is also likely to have some cost in social coordination tasks such as language: in situations where the stationary distribution admits multiple languages, committing to one language through a strong and specific prior is likely to prove costly in the event of being born into a subpopulation which does not speak that language.

The model described above deals with a limited range of learning behaviours. Strength of prior, given by α, is a continuous parameter and amenable to a more fine-grained analysis. Similarly, the dichotomy between sampling and maximising can be recast into a continuum by a means suggested in Kirby et al. (2007): if $P_L(h|d)$ is proportional to $[P(d|h)P(h)]^r$, then a range of strategies lie between sampling (given by $r = 1$) and maximising (infinitely large r). Preliminary analysis of this much larger space yields results broadly similar to those presented here: higher values of r are preferred, and α exhibits large-scale neutrality in populations with any maximising tendency (Smith & Kirby, in preparation). The general

picture remains that natural selection for communication favours those conditions where cultural transmission factors plays a significant role in shaping language, and strength of innate predispositions is relatively unimportant.

Acknowledgements

Kenny Smith is funded by a British Academy Postdoctoral Research Fellowship. The initial stages of this research took place at the Masterclass on Language Evolution, organised by P. Vogt and B. de Boer and funded by NWO.

References

Baldwin, J. M. (1896). A new factor in evolution. *American Naturalist*, *30*, 441–451.

Briscoe, E. J. (2000). Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language*, *76*, 245–296.

Griffiths, T. L., & Kalish, M. L. (2005). A Bayesian view of language evolution by iterated learning. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the cognitive science society* (pp. 827–832). Mahwah, NJ: Erlbaum.

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

Kirby, S. (2001). Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*, 102–110.

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

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 121–147). Springer Verlag.

Kirby, S., Smith, K., & Brighton, H. (2004). From UG to universals: linguistic adaptation through iterated learning. *Studies in Language*, *28*, 587–607.

Maynard Smith, J., & Price, G. R. (1973). The logic of animal conflict. *Nature*, *146*, 15–18.

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

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, *13*, 707–784.

Smith, K. (2004). The evolution of vocabulary. *Journal of Theoretical Biology*, *228*, 127–142.

Smith, K., & Kirby, S. (in preparation). *The evolution of lanaguge learning in Bayesian agents*.