INNATENESS AND CULTURE IN THE EVOLUTION OF LANGUAGE

MIKE DOWMAN, SIMON KIRBY
Language Evolution and Computation Research Unit, Linguistics and English Language
The University of Edinburgh, Edinburgh, EH8 9LL, UK

THOMAS L. GRIFFITHS
Cognitive and Linguistic Sciences, Brown University
Providence, RI 02912, USA

Is the range of languages we observe today explainable in terms of which languages can be learned easily and which cannot? If so, the key to understanding language is to understand innate learning biases, and the process of biological evolution through which they have evolved. Using mathematical and computer modelling, we show how a very small bias towards regularity can be accentuated by the process of cultural transmission in which language is passed from generation to generation, resulting in languages that are overwhelmingly regular. Cultural evolution therefore plays as big a role as prior bias in determining the form of emergent languages, showing that language can only be explained in terms of the interaction of biological evolution, individual development, and cultural transmission.

1. Introduction

Why is language the way it is and not some other way? Answering this *why* question is one of the key goals of modern linguistics. We can reframe this question as one about *language universals* in the broadest sense. Universals are constraints on variation, and include fundamental structural properties of language, such as compositionality, recursion, and semi-regularity. Language has arisen from the interactions of three complex adaptive systems: individual development, biological evolution, and cultural transmission, so a satisfactory approach to language should take into account each of these three systems, and how they interact.

Within generative linguistics, language structure is explained in terms of innately determined properties of the acquisition mechanism, and the constraints it places on developmental pathways (Chomsky, 1965). Whilst the generative approach is often contrasted with *linguistic functionalism*, which focuses on language use rather than acquisition, functionalism aims to explain aspects of language structure in terms of processing capacity, which is also an innately determined property of individual language users, though not necessarily one that is specific to language (see Kirby, 1999 for discussion). These approaches shift the burden of explaining linguistic structure to one of explaining how our innate learning biases arose. Therefore explanations of linguistic structure are shifted to the domain of biological evolution, in which our innate prior biases are shaped. These in turn affect individual development, and therefore ultimately the universal properties of human language.

2. Cultural Transmission and the Bottleneck Effect

The arguments made so far have neglected the problem of explaining through what mechanism the properties of individuals give rise to properties of languages (Kirby, 1999). Recent work has addressed this issue using *iterated learning models*, computer programs comprising simple models of language-learning individuals that are then placed in a simulated population so that the dynamics arising from their interactions can be studied (e.g. Batali 1998; Kirby, 2001). The simulated individuals can learn from one-another, and so language can be passed from one generation of speakers to the next, a process which gives rise to the cultural evolution of language. Fig. 1 shows how iterated learning forms a bridge between individual bias and language structure.

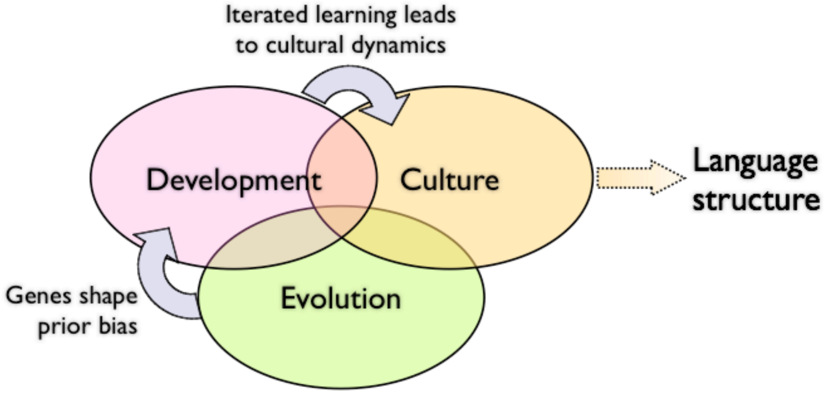

Figure 1: The universal structure of human language arises out of the process of iterated learning, an adaptive system operating on a cultural time-scale, driven by individual biases that are ultimately shaped by biological evolution.

One of the most basic and pervasive properties of human languages is *regularity*. That is, if one meaning is expressed using a particular rule or construction, then it is likely that another meaning will also make use of the same pattern. We might therefore expect regularity to arise from some fundamental aspect of our language faculty – in other words our prior bias might be expected to reflect this central universal strongly. However, in a wide variety of models using a diverse set of learning algorithms and assumptions about signal and meaning spaces (e.g. Batali, 1998; Kirby, 2001), the overwhelming conclusion is that strong biases are not necessary to explain the emergence of pervasive regularity. It seems that regularity emerges whenever the number of training samples that the learners are exposed to is small. If there is too little data stable languages do not emerge, while if there is too much training data, the emergent languages are not regular, but instead express meanings in an ad hoc way. Kirby, Smith & Brighton (2004) explain this behaviour in terms of adaptation to the *bottleneck* – the limited amount of linguistic examples from which each speaker must learn the language. A regular rule can persist into the next generation so long as the learner sees only one example of it, but an irregular expression can only persist into the next generation if the learner is exposed to exactly that expression.

3. Iterated Bayesian Learning

While a wide range of learning algorithms, with a correspondingly wide range of biases, have resulted in the emergence of regular linguistic structure, it is very hard to determine exactly what the biases inherent in the various models of learning actually are. Because of this, it is difficult to be sure of the generality of the result. To address this issue, we have begun to explore a Bayesian version of iterated learning, which has the advantage that we can make the biases of learners completely explicit, and manipulate them freely.

In the Bayesian framework, constraints on both acquisition and processing can be characterised as a probability distribution over possible languages. In this view, the language learner is faced with the task of forming a hypothesis about the language of her speech community on the basis of data that she is exposed to. She aims to assess the *posterior* probability, $P(h|d)$, of a hypothesis (i.e. a language) h, based upon the observed data d. Bayes' rule indicates that this should be done by combining two quantities: the *likelihood*, $P(d|h)$, being the probability of the observed data d given hypothesis h, and the *prior probability*, $P(h)$, indicating the strength of the learner's *a priori* bias towards the hypothesis h. According to Bayes rule the posterior probability of a hypothesis is proportional to the product of the associated likelihood and prior probability, as shown in Eq. (1). In other words, the learner must take into account how well each possible hypothesis predicts the data seen *and* how likely each hypothesis is *a priori*. Conceived in this way, the influence of the learner's language acquisition device (Chomsky, 1965) and language-processing machinery is characterised as the prior probability distribution over hypotheses.

$$P(h \mid d) \propto P(d \mid h)P(h) \tag{1}$$

Griffiths & Kalish (2005) investigated iterated learning under the assumption that the learners first calculate the posterior distribution over languages, and then *sample* from this distribution (i.e., they choose a language with a probability equal to its posterior probability). By viewing the process of iterated learning as a Markov chain (c.f. Nowak et al., 2001), they were able to prove that such a process will result in a distribution of languages that exactly mirrors the prior bias.[a] This is a startling result, and one that renders the results from previous simulations mysterious, as the process of cultural evolution makes no independent contribution to the emergent languages. In particular, the number of training samples – the bottleneck size – has *no effect whatsoever* on the probability of each type of language emerging. Why then is the bottleneck size the crucial factor in all the simulation models of iterated learning?

The answer to this puzzle turns out to hinge on our assumptions about what a rational learning agent should do when faced with a choice between languages. It might seem more rational for learners to pick the language with the maximum *a posteriori* probability (which in Bayesian learning theory is called the MAP hypothesis), rather than sampling from this distribution, as in Griffiths and Kalish's approach. This would maximise the chance of picking the same language as that spoken by the previous agent. This small difference between the sampling and MAP learner turns out to have huge implications for the dynamics of iterated learning.

To demonstrate why this is the case, we can work through a simple idealised model which nevertheless reflects the general case of iterated Bayesian learning. The first step in constructing such a model is to decide on the form of the space of logically possible languages. For this example, we aim to explain the origins of *regularity* in language. Regularity here can be seen as an umbrella concept that covers a variety of aspects of language. We will treat a language as a deterministic function from discrete *meanings* to discrete *classes*. Depending on how the model is interpreted these could be thought of as classes across a morphological paradigm, or an indication of the form of a compositional encoding of a particular meaning. The idea is that a language in this model is completely *regular* if all its meanings belong to the same class, and is completely *irregular* if all the meanings belong to different classes.

Learners are exposed to a set of m utterance-meaning pairs, each of which consists of a meaning and the class used to express it. They then use Bayes rule to find the most likely language to have generated this set of meaning-class pairs. Finally, this language is used to generate a new set of meaning-class pairs. This is done by sampling m meanings at random and using the hypothesized language to generate classes for each. In addition, we model noise in the system by randomly picking a different class to the correct one for each meaning with probability ϵ. The dynamics of this system can be completely characterized for a particular value of m (the bottleneck) and ϵ (the noise factor), by looking at the

[a] Strictly speaking, this is only the case if the prior bias results in a dynamical system that is *ergodic*. Simplifying, this essentially means that every language in the space must be at least potentially "reachable" by cultural evolution.

probability that a language l_1 spoken by one agent will give rise to a learner choosing a language l_2, for all pairs of languages (l_1,l_2), as shown in Eq. (2). $\boldsymbol{x} = \{x_1,\ldots,x_m\}$ is the set of meanings chosen at random, and $\boldsymbol{y} = \{y_1,\ldots,y_m\}$ is the corresponding set of classes output by the speaker, and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the possible sets of meanings and output classes respectively.

$$P(l_2 \mid l_1) = \sum_{\boldsymbol{x} \in \boldsymbol{X}} \sum_{\boldsymbol{y} \in \boldsymbol{Y}} P(l_2 \; is \, MAP \mid \boldsymbol{x}, \boldsymbol{y}) P(\boldsymbol{y} \mid \boldsymbol{x}, l_1) P(\boldsymbol{x}) \tag{2}$$

The probability that a language is the MAP language will normally be either 1 or 0, but where several languages are tied with equally high posterior probability, this value will be equal to 1 divided by the number of such languages. Eq. (3) shows how the classes are produced for a given language. We can think of $P(l_2|l_1)$ as a matrix of transitions from language to language (what Nowak et al. 2001 call the *Q*-matrix), defining a Markov chain over languages. The first eigenvector of this transition matrix gives the *stationary distribution* for the Markov chain, indicating the distribution over languages that will emerge out of iterated learning (provided the underlying Markov chain is ergodic). We now have everything in place to determine what universal properties will emerge for a given bottleneck, noise-term and prior bias.

$$P(\boldsymbol{y} \mid \boldsymbol{x}, l) = \begin{cases} 1 - \varepsilon & \textit{if } \boldsymbol{y} \textit{ is the class corresponding to } \boldsymbol{x} \textit{ in } l \\ \dfrac{\varepsilon}{|\boldsymbol{y}| - 1} & \textit{otherwise} \end{cases} \tag{3}$$

What is a reasonable prior bias for the language model we have described? One possibility is to have a completely uninformative prior, with every language being equally likely. Unsurprisingly, this results in no clear preference for one language over another in the final result. We can infer from this directly that the prior does indeed matter. Since language exhibits regularity, human language learners cannot be "blank slates". Instead, we make the minimal assumption that learners will expect future events to be similar to previous events. If we have *n* meanings and *k* classes, this assumption is embodied in the prior specified by Eq. (4), where n_j is the number of meanings assigned to class *j*, α is a parameter of the prior, and Γ(*x*) is the generalised factorial function, with Γ(*x*) = (*x*-1)! when *x* is an integer. The specific form of the prior can be justified both from the perspective of minimum description length (Rissanen, 1978), and from the perspective of Bayesian statistics, where *P*(*l*) appears as a special case of the Dirichlet-multinomial distribution (Johnson & Kotz, 1972). α controls the strength of the prior, or alternatively how much of a bias towards regularity is built into the model. Low values of α create a strong regularity bias, and high values a much weaker one.

$$p(l) = \frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k \Gamma(n + k\alpha)} \prod_{j=1}^{k} \Gamma(n_k + \alpha) \tag{4}$$

The key question which we wish to address is: how strong does this bias towards regularity need to be? As we mentioned above, Griffiths and Kalish (2005) have shown that, for a sampling learner, the expected distribution of languages exactly reflects the prior. So, since languages are overwhelming regular, this suggests the prior bias must be very strong. With the MAP learner, however, this turns out not to be the case as long as there is a bottleneck on linguistic transmission.

For example, we looked at a simple model of languages with four possible meanings ($n = 4$), and four possible classes for each meaning ($k = 4$). From the perspective of regularity, there are five different types of language in this space: all meanings in the same class; three meanings in one class and one in another; two meanings in one class and the other two in a second class; two meanings in one class and the other two in two different classes; and all four meanings in different classes. For shorthand, we label these types: *aaaa, aaab, aabb, aabc,* and *abcd* respectively. The first row of Table 1 shows the prior probability of these five types of languages under the prior $P(l)$ described above, with $\alpha = 10$, so that there is only a weak preference for regularity.

Table 1. The predicted distribution of language types for different bottleneck sizes, in terms of the probability of a particular language of each type. The prior distribution is shown for comparison. As the bottleneck on linguistic transmission tightens, the preference for regularity is increasingly over-represented in the distribution of languages.

Languag e	*aaaa*	*aaab*	*aabb*	*aabc*	*abcd*
Prior	0.00579	0.00446	0.00409	0.00371	0.00338
m = 10	0.145	0.00743	0.000449	0.000324	0.0000335
m = 6	0.175	0.00566	0.000150	0.000158	0.0000112
m = 3	0.209	0.00329	0	0.0000370	0

Given this prior, the expected distributions of languages by type for different bottleneck sizes and an error-term (ϵ) of 0.05 are shown in the remainder of Table 1. What is immediately obvious from these results is that the prior is *not* a good predictor of the emergent properties of the languages in the model. The *a priori* preference for regularity is being hugely over-represented in the languages that evolve culturally. In fact the strength of the bias often has no effect on the resulting stationary distribution – it is the ordering it imposes on languages that is most important. Furthermore, as with the simulation models, the tightness of the bottleneck on linguistic transmission is a determining factor in how regular the languages are.

Of course, languages typically are not completely regular. Irregularity is a common feature of morphological paradigms. This irregularity is not randomly distributed throughout a language, but rather appears to correlate with frequency – for example, the ten most common verbs in English are all irregular in the past tense. Previous iterated learning simulations have suggested that this can be explained in terms of adaptation to cultural transmission. Put simply, frequent verbs can afford to be irregular, since they will have ample opportunity to be transmitted faithfully through the bottleneck (Kirby, 2001).

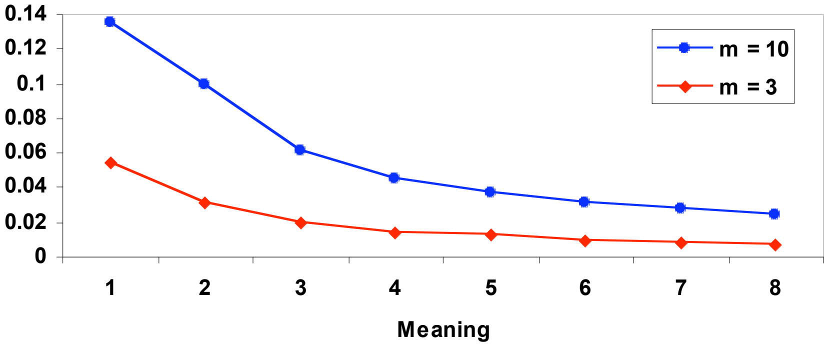

Figure 2. Irregularity correlates with frequency. These are results from simulations using 8 meanings and 4 classes, with $\epsilon = 0.05$ and $\alpha = 1$. They show the proportion of languages in which each meaning was expressed using an irregular construction. The frequency of each meaning was inversely proportional to its rank (counting left to right). A meaning was counted as regular if its class was in the majority.

To test this intuition, we made a simple modification to our model. Rather than picking meanings at random from a uniform distribution, they were skewed so that some were more common than others. The results in Fig. 2 show how often each meaning was irregular (i.e. in a minority class) for a language model with 8 meanings and 4 classes. The frequency of each meaning decreases from left to right in this graph, demonstrating that the model results in a realistic frequency/regularity interaction.

4. Conclusion

Language involves three adaptive systems: biological evolution, individual development, and cultural transmission. An adequate account of the origins of linguistic structure must crucially focus on the interactions between these systems. Our contribution has been to demonstrate that the innovation of cultural transmission radically alters the relationship between our innate learning biases and our linguistic behaviour.

The implications of this for the study of language evolution are, firstly, that our innately given bias cannot be directly inferred from our phenotype (i.e. language). More specifically, weak innate biases can nevertheless lead to strong universals wherever there is a bottleneck on the cultural transmission of language. This leads naturally to an explanation of frequency-related patterns of regularity and irregularity in language assuming only a weak expectation of predictability on behalf of learners. Finally, this result demonstrates that we must be cautious in assuming that adaptive structure necessitates an explanation in terms of the selective evolution of innate traits that are specifically linguistic. Regularity is an adaptive feature of language but we have shown that the mechanism for adaptation need not be biological.

In this paper we have not looked at the final interaction in Fig. 1 – between culture and evolution. Much work remains to be done. However, we note that our result shows iterated learning can shield the strength of innate biases from the view of natural selection. The implications of this are beginning to be worked out, but it is clear that an account of the biological evolution of the human language faculty cannot be complete if it fails to take into account the interactions between innateness, culture and linguistic structure.

References

Batali, J. (1998). Computational simulations of the emergence of grammar. In J. R. Hurford, M. Studdert-Kennedy and C. Knight (Eds.), *Approaches to the Evolution of Language: Social and Cognitive Bases*. Cambridge: Cambridge University Press.

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

Griffiths, T. and Kalish, M. (2005). A Bayesian view of language evolution by iterated learning. In B. G. Bara, L. Barsalou and M. Bucciarelli (Eds.), *Proceedings of the XXVII Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Johnson, N. L. and Kotz S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions*. New York, NY: John Wiley & Sons.

Kirby, S. (1999). *Function, Selection and Innateness: the Emergence of Language Universals*. Oxford: Oxford University Press.

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(2): 102--110.

Kirby, S., Smith, K. and Brighton, H. (2004) From UG to Universals: Linguistic Adaptation through Iterated Learning. *Studies in Language,* 28(5): 587-607.

Nowak, M. A., Komarova, N. L., and Niyogi, P. (2001). Evolution of universal grammar. *Science*, 291: 114–118.

Rissanen, J. (1978). Modelling by shortest data description. *Automatica*, 14: 465–471.