Innateness and culture in the evolution of language

Simon Kirby*†, Mike Dowman‡, and Thomas L. Griffiths§

*School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, 40 George Square, Edinburgh EH8 9LL, United Kingdom; ‡Department of General Systems Sciences, Graduate School of Arts and Sciences, University of Tokyo, 3-8-1 Komaba, Tokyo 153-8902, Japan; and §Department of Psychology and Program in Cognitive Science, University of California, Berkeley, CA 94720

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved February 6, 2007 (received for review September 19, 2006)

Human language arises from biological evolution, individual learning, and cultural transmission, but the interaction of these three processes has not been widely studied. We set out a formal framework for analyzing cultural transmission, which allows us to investigate how innate learning biases are related to universal properties of language. We show that cultural transmission can magnify weak biases into strong linguistic universals, undermining one of the arguments for strong innate constraints on language learning. As a consequence, the strength of innate biases can be shielded from natural selection, allowing these genes to drift. Furthermore, even when there is no natural selection, cultural transmission can produce apparent adaptations. Cultural transmission thus provides an alternative to traditional nativist and adaptationist explanations for the properties of human languages.

cultural transmission | iterated learning | Bayesian learning | nativism

One of the key challenges for cognitive science is to explain the structure of human language. Although languages vary, they share many universal structural properties (1, 2). Where do these universals come from? A great deal of research has proceeded under the assumption that this is essentially a biological question (3): that languages have the structure they do because of our innate faculty for acquiring (4) and processing (5) language. Linguistic universals thus become evidence for strong innate constraints on language acquisition: if all languages share some feature, then that feature is assumed to arise from a constraint imposed by our language faculty. Naturally, this leads to an attempt to understand language in the light of biological evolution: if language structure has implications for our biological fitness and that structure is determined by our innate endowment, then natural selection seems like the most relevant explanatory mechanism (6). If this reasoning is sound, we can read-off properties of the human faculty of language (and learn about its evolution) by uncovering the universal structural generalizations underlying languages.

In this paper, we argue that there are serious problems with this orthodox evolutionary/biolinguistic approach. It treats language as arising from two adaptive systems, individual learning and biological evolution, but in doing so misses a third: cultural transmission (refs. 7–9, Fig. 1). The surprising consequences of taking all three adaptive systems into account are that strong universals need not arise from strong innate biases, that adaptation does not necessarily imply natural selection, and that cultural transmission may reduce the selection pressure on innate learning mechanisms. Our conclusions call into question the existence of strongly constraining biological predispositions for language, and the prominence of adaptationist explanations for the structural properties of languages.

The traditional evolutionary approach to language is missing an essential piece: a characterization of the mechanism linking our biological predispositions and the languages that are actually spoken in human societies (Fig. 2). Identifying the relationship between genes and languages is crucial, as it determines how we infer innate predispositions by looking at languages, and ultimately whether we need to take this linking mechanism into account when considering the biological evolution of the human language faculty. We can break this linking mechanism into two parts: the process by which innate biases influence the language learned by each individual, and the process by which cultural transmission affects the languages represented in a population. We will consider these two parts in turn.

To understand the link between biological predispositions and language structure, we need an account of the effect of innate biases on the language learned by each individual in a population. One such account assumes that learners apply the principles of Bayesian inference (10). This approach is widely used as a standard for rational inference in statistics (11), decision theory (12), and machine learning (13), and Bayesian methods are used in computational linguistics (14), psycholinguistics (15), and evolutionary linguistics (16). Formally, learners are faced with the problem of how to use the data provided by the linguistic behavior of others to select among a set of candidate hypotheses concerning the language they are exposed to. Letting h denote a particular hypothesis and d the data, we can express the prior biases of learners in a probability distribution, $P(h)$, indicating their degrees of belief concerning the different hypotheses before seeing d. Bayesian inference is a procedure for updating these degrees of belief in light of the evidence provided by the data. The "posterior" probability, $P(h|d)$, of a hypothesis h after seeing data d, is obtained via Bayes' rule,

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h'} (d|h')P(h')}. \quad [1]$$

In this approach, the degree to which a learner should believe in a particular hypothesis (i.e., a language) is a direct combination of their innate biases, as expressed in the prior, $P(h)$, and the extent to which the data are consistent with that hypothesis, given by $P(d|h)$. The learner can then choose to adopt a particular language based on these degrees of belief. For example, learners might select the language that has highest posterior probability, sample from their posterior distribution, or do anything in between.

Bayesian inference provides a framework in which we can experiment with different assumptions about the effects of innate predispositions on language learning. However, learning is only part of the mechanism linking genes and the languages spoken in human societies. To determine the expected distribution of languages given a particular bias we also need to model the other part of this mechanism: the cultural transmission of language. The linguistic behavior a learner is exposed to as input is itself the output of learning by other individuals. Similarly, the language the learner acquires will ultimately produce data for a later generation of learners. The expected distribution of languages for a given prior bias is therefore a population-level

Author contributions: S.K., M.D., and T.L.G. designed research; S.K., M.D., and T.L.G. performed research; and S.K., M.D., and T.L.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS direct submission.

†To whom correspondence should be addressed. E-mail: simon@ling.ed.ac.uk.

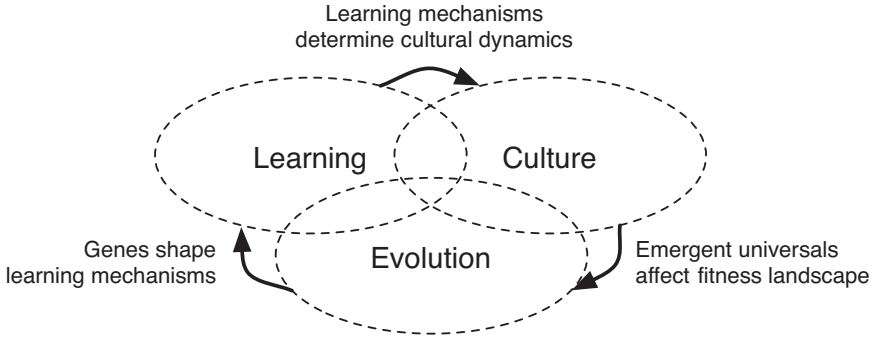

Fig. 1. The structure of language arises from the interactions between three complex adaptive systems. As individuals, we acquire language using learning mechanisms that are part of our biological endowment (characterized in this paper in terms of prior bias). This learning machinery acts as the mechanism by which language is transmitted culturally through a population of individuals over time. Ultimately, this process of cultural transmission leads to a set of language universals (which can be expressed as a distribution over types of languages). The relationship between learning machinery and consequent universals is nontrivial but can be uncovered using the framework developed here. Finally, the structure of languages that emerge from this process will affect the fitness of individuals using those languages, which in turn will lead to the biological evolution of language learners, closing the loop of interactions.

phenomenon that emerges out of the dynamics of cultural transmission, a process we call iterated learning (17–22).

Simplifying, we treat the population as consisting of a chain of individuals, one per generation, each learning from the output of the previous generation and producing utterances that are provided as input to the subsequent generation. If we focus just on the languages acquired by the sequence of learners, we can analyze iterated learning as a Markov process: the probability that a learner acquires a particular language depends only on the language acquired by the preceding learner (22–25). When these probabilities are calculated for all languages, they form a transition matrix, representing the probability of transitioning from any one language to any other. The transition probabilities are determined by the learning algorithm used by the learners, and the way in which the data they are exposed to are selected. Formally, the probability that the learner n chooses hypothesis i given that learner $n-1$ chose hypothesis j is

$$P(h_n = i|h_{n-1} = j) = \sum_d P_L(h_n = i|d)P_P(d|h_{n-1} = j), \qquad [2]$$

where $P_L(h|d)$ is the probability that a learner will select hypothesis h after observing data d, and $P_P(d|h)$ is the probability of producing the data d under hypothesis h. It is well known that the stationary distribution over states in the Markov chain is proportional to the first eigenvector of the transition matrix, providing the Markov chain is ergodic. (That is, so long as each state is reachable from every other state in a number of steps that has no fixed period.) Normalizing the first eigenvector so that it totals one thus reveals the probability of a learner speaking any particular language once iterated learning has converged on a stationary distribution; essentially, the expected distribution of languages emerging from cultural evolution.

To illustrate the behavior of this model, we will assume that language is a noisy mapping between meanings and signals and that, in each generation, learners are exposed to a random subset of the pairs defined by this mapping for the previous generation's language. The size of this subset imposes an informational "bottleneck" on cultural transmission, and is a crucial parameter in our model. The other important parameter is, of course, the prior bias. For this example, we will assume that learners have a prior expectation of predictability. That is, languages which employ a systematic scheme for expressing different meanings will be assigned a higher prior probability than those that treat each meaning separately and idiosyncratically.

To simplify, we represent languages as a pairing of meanings to classes rather than signals. These classes correspond to different possible strategies for expressing a meaning. By abstracting away from an explicit representation of signals, we have a straightforward way of interpreting our bias for predictable systematicity: a systematic language will be one in which all of the meanings belong to the same class, whereas a completely idiosyncratic language will have no two meanings in the same class. To give a concrete example, in the case of morphology, we

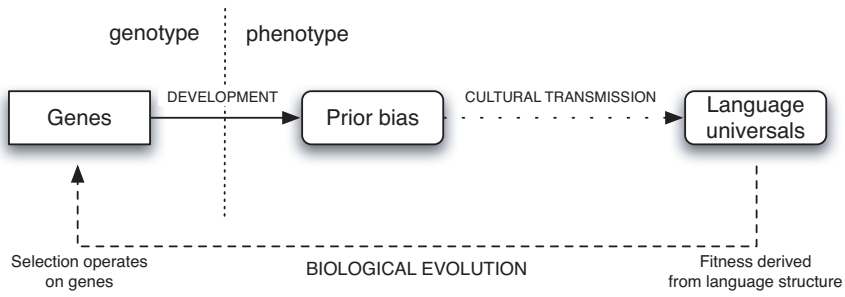

Fig. 2. The link between biological predispositions and language structure. Genes (in combination with the nonlinguistic environment) give rise to mechanisms for learning and processing language. These determine our innate predispositions with respect to language (our prior linguistic bias). Bias is a property of an individual, but the (universal) structure of human language emerges from the interaction of many individuals over time. Therefore, cultural transmission bridges the link between bias and universals. Although genes code for bias, biological fitness will in part be governed by the extended phenotype (i.e., language structure). To understand language evolution, we must understand this linking mechanism.

can consider different ways of making past tense forms of verbs in a language as corresponding to distinct classes. A completely regular language would use the same past-tense form for every verb; that is, the same class would be assigned to every meaning. A language with a great deal of irregularity, on the other hand, would have a less predictable pairing of meanings and classes. Similarly, we can envisage a higher-level interpretation of our scheme by applying it to the syntax of a language as a whole. Languages with compositional syntax assign signals to meanings in a predictable and systematic manner; in other words, they use the same encoding strategy for every meaning. An evolutionarily early form of protolanguage that has been hypothesized (26) has no such systematic syntax, but instead treats every meaning holistically. In such a protolanguage, the signal for every meaning must be learned individually, and no generalizations are possible. Recasting this in terms of meanings and classes, a compositional language is simply one which treats each meaning as belonging to the same class, whereas a nonstructured protolanguage assigns each meaning a distinct class.

We use a scheme for assigning prior probabilities to languages that allows us to vary the strength of the prior; in other words, how skewed the expectation of the learner is toward systematic languages, in which the assignment of classes to meanings is relatively predictable (see *Methods* for details of the prior). Our central question is: how does this parameter of the bias (our model of innateness) relate to the stationary distribution (the types of language that emerge)? Using the Bayesian model outlined above, and the initial assumption that learners always choose the language with the highest posterior probability, we find striking evidence that the prior bias is not a good predictor of the resulting distribution of languages (Fig. 3). In particular, for a range of parameters, the strength of the bias has no effect whatsoever on the languages that emerge. As long as the relative ranking of languages is preserved, even a tiny innate preference for systematicity can have a large effect, due to the process of cultural evolution. Equally, it is not simply the case that the language with the highest prior probability is the only one represented in the stationary distribution. Rather, it is the number of training examples, the cultural bottleneck, that determines how systematic languages become.

How does this model relate to real language? If we return to the morphological example given above, we can see that there is variation in systematicity within and across languages. For example, the verbal paradigm of English is partially regular (e.g., walk–walked) and partially idiosyncratic (e.g., go–went). The regular pattern is by far the most dominant if we look across verbs, but interestingly, the irregular verbs tend to be highly frequent (17, 27). This pattern is seen in many languages and has the hallmarks of an adaptation. Regularity is adaptive for infrequently expressed meanings because it maximizes the chance of being understood by another individual with different learning experience to you. It is less relevant for frequently expressed meanings because there is a greater chance that two individuals will have previously been exposed to the same form. In fact, irregularity might be preferred for these meanings if, for example, it enables the use of a shorter and therefore more economical form.

To examine whether the relationship between frequency and regularity needs to be explained as an adaptation, we can use the model to compute the distribution of regulars and irregulars when some meanings are expressed more frequently than others. When the frequency of meanings is skewed in this way, we find precisely the attested frequency/irregularity interaction (Fig. 4). Note that this relationship is not coded anywhere in the innate predispositions of the individuals in the population, nor is there any selective pressure favoring optimal communication. The apparent adaptation thus arises purely from the process of cultural transmission, providing an alternative to the adaptationist explanation for the prevalence of this relationship across languages.

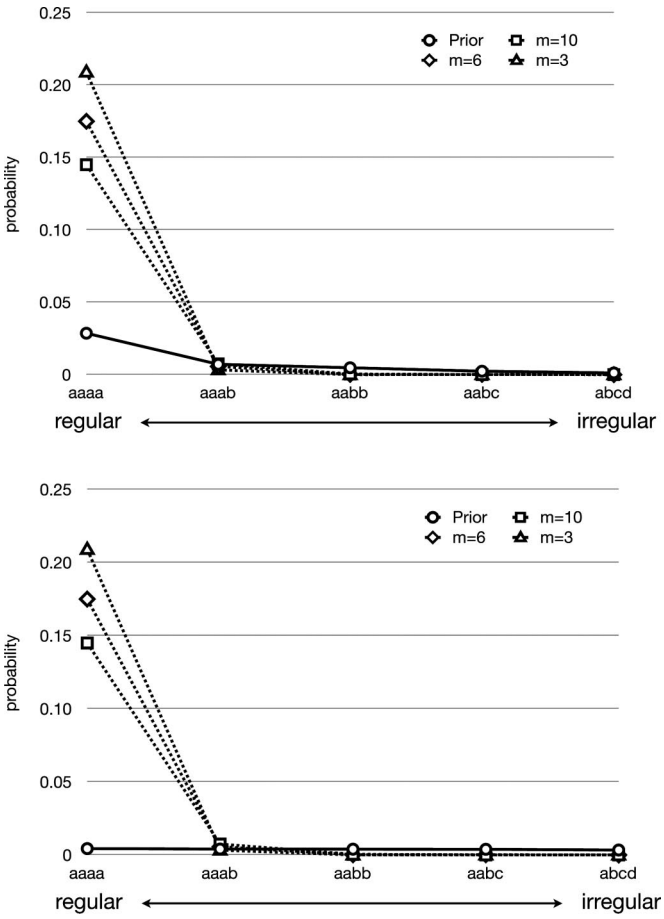

Fig. 3. Results of iterated learning. Cultural transmission amplifies innate bias. Even with a very weak bias in favor of regularity (i.e., a consistent mapping from meanings to classes), regular languages predominate in the emergent distribution of languages. These graphs show the probability of five particular languages, each with a different degree of regularity, on the same plot as the learners' prior expectation of those languages. (Each language has four meanings and four classes, represented here by letters.) As the number of training examples is reduced (i.e., the bottleneck becomes tighter), regularity is increasingly favored. The strength of the bias (how skewed it is in favor of regularity) has no effect on the results.

These results demonstrate that strong universals need not imply strong innate constraints on learning and that biological evolution is not the only potential explanation for adaptive structure in language. This raises an important question: under what circumstances do weak biases result in strong universals? To investigate this question, we examined the consequences of learners using a more general class of strategies for choosing a particular language given the posterior distribution and an approach that potentially allows the hypotheses and data to take arbitrary forms rather than the meaning-class mappings used in our previous analyses. If we assume that learners choose a particular hypothesis with probability $P_L(h|d)$ proportional to $[P_P(d|h)\ P(h)]^r$, we obtain a class of strategies that interpolates between two special cases: sampling from the posterior distribution when $r = 1$, and selection of the hypothesis with highest posterior probability when r approaches infinity. We can then examine the consequences that different values of r have on the stationary distribution of the resulting Markov chain.

In the special case where learners sample from the posterior (i.e., $r = 1$), the stationary distribution is simply the prior (22).

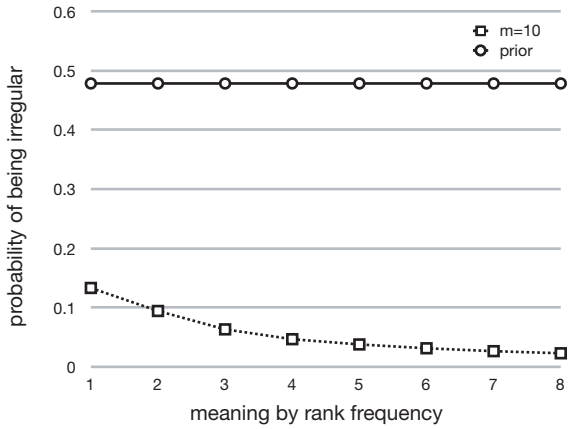

Fig. 4. The emergence of an adaptive irregularity/frequency interaction. Cultural transmission results in languages where the probability of a meaning being irregular (i.e., not being assigned the majority class) is correlated with its frequency; this is despite the fact that learners in this model have a prior expectation that all meanings are equally likely to be irregular. This result mirrors what is found in real languages and has the hallmarks of an adaptation. This graph shows the probability of each meaning not being in a majority class, and the frequency of each meaning is inversely proportional to its rank. It was derived through simulation over a million iterations because the more complex languages used in this simulation made calculation of the whole transition matrix infeasible.

Obtaining general results for the consequences of increasing r is complicated, but if we place some constraints on the structure of languages we can still determine the stationary distribution analytically. Here, we constrain our languages such that $P(d|h)$ is either constant or zero across all hypotheses h for all data d. This is not an overly restrictive constraint; for example, it is satisfied by the set of deterministic languages, with a unique signal for each meaning and an arbitrary distribution over meanings. With a set of languages that satisfies this constraint, the probability that a particular hypothesis h will be produced by iterated learning is proportional to $P(h)^r$ (see *Methods* for proof). The implications of this are clear: languages will be systematically overrepresented with respect to their prior probabilities for values of $r > 1$. That is, weak biases will produce strong universals if learners choose hypotheses in a fashion that disproportionately favors hypotheses with higher posterior probabilities.

Conclusion

Our analyses demonstrate that, by mediating between innate bias and resulting behavior, culture may profoundly influence the evolutionary process. We have shown that the strength of bias can be completely obscured by iterated learning. Genes may code for the strength of a learning bias, but fitness (and hence selection of those genes) is determined by the extended phenotype: in this case, the properties of languages that emerge in populations. Genes controlling strength of bias could therefore be shielded from selection, so culture may introduce neutrality to the fitness landscape of learners. This has potentially far reaching consequences. For example, if strong learning biases must be maintained against mutation pressure (28), the introduction of cultural transmission may lead to a weakening of these innate biases.

The implications of our results are not restricted to human language. They have relevance to any behavior that is passed between generations through learning. For example, some bird species produce songs that exhibit particular structural universals, but they have nevertheless been shown to be capable of learning artificially constructed songs that violate these universal constraints (29). This is exactly the sort of result we would predict if a weak learning bias is being amplified by cultural transmission through iterated learning.

Language is therefore the result of nontrivial interactions between three complex adaptive systems: learning, culture, and evolution. As such, it is an extremely unusual natural phenomenon. Taking the role of culture into account provides alternative explanations for phenomena that might otherwise require an explanation in terms of innate biases or biological evolution. Ultimately, if we are to understand why language has the universal structural properties that it does, we need to consider how learning impacts on cultural transmission, and how this affects the evolutionary trajectory of learners.

Methods

Meaning-Class Mapping Model. In this model, we assume that a language consists of a mapping from a set of n meanings to a set of k classes. The data observed (and produced) by each learner consist of m pairs of meanings and classes. The probability of the set of meaning-class pairs d being produced given that a learner speaks the language corresponding to h is given by

$$P_P(d|h) = \prod_{\langle xy\rangle \in d} P(y|x, h)P(x), \qquad [3]$$

where x is a meaning and y is a class that is produced in response to that meaning. This equation assumes that the class produced in response to each meaning is independent of the other meanings for which that learner has produced classes. In the initial study (Fig. 3), $P(x)$ is equal for each x. Noise in the linguistic transmission process is modeled by incorporating a parameter ε that corresponds to the probability that a different class to the correct one will be chosen for each production. The probability of producing a particular class in response to a given meaning if a learner speaks language h is therefore

$$P(y|x, h) = \begin{cases} 1 - \varepsilon & \text{if } y \text{ is the class corresponding to } x \text{ in } h \\ \dfrac{\varepsilon}{k-1} & \text{otherwise.} \end{cases} \qquad [4]$$

The prior probability assigned to each language, h, is

$$P(h) = \frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k \Gamma(n + k\alpha)} \prod_{j=1}^{k} \Gamma(n_k + \alpha), \qquad [5]$$

where n_j is the number of meanings expressed using class j. $\Gamma(x)$ is the generalized factorial function, with $\Gamma(x) = (x\text{-}1)!$ when x is an integer. α is a parameter that controls the strength of the prior, with low values of α creating a strong prior bias in favor of regularity, and high values creating a relatively flat prior, in which the probability assigned to the most regular languages is only slightly greater than that assigned to the most irregular. This prior is a special case of the Dirichlet-multinomial distribution (30). Its use means that the Bayesian inference mechanism can be seen as a form of minimum description length (31). This is because the probability assigned to each language corresponds to the amount of information needed to encode it in a minimally redundant form if information theory (32) is used to relate probability to entropy. In the cases considered in this paper, there was a language with each possible mapping of meanings to classes, given the number of meanings and classes available.

Proof of Weak Biases Producing Strong Universals. We now allow h and d to correspond to any form of language, not just meaning-class mappings, so long as the Markov chain on h is ergodic. By definition, the stationary distribution π of a Markov chain satisfies the expression

$$\pi(h_{n+1}) = \sum_{h_n} P(h_{n+1}|h_n)\pi(h_n). \quad [6]$$

For the Markov chain defined by Eq. **2**, this becomes

$$\pi(h_{n+1}) = \sum_{h_n} \sum_{d} P_L(h_{n+1}|d)P_P(d|h_n)\pi(h_n). \quad [7]$$

Taking $P_L(h|d)$ to be the exponentiated posterior distribution, as described above, we obtain

$$\pi(h_{n+1}) = \sum_{h_n} \sum_{d} \frac{[P_P(d|h_{n+1})P(h_{n+1})]^r}{\sum_h [P_P(d|h)P(h)]^r} P_P(d|h_n)\pi(h_n). \quad [8]$$

In general, finding an analytic solution to this equation can be challenging. However, we can make the simplifying assumption that for each hypothesis, any data d have a probability $P_P(d|h)$ of either 0 or some constant value $f(d)$. Under this assumption, the stationary distribution reduces to

$$\pi(h_{n+1}) = \sum_{h_n} \sum_{d \subset h_{n+1}} \frac{P(h_{n+1})^r}{\sum_{h \supset d} P(h)^r} P_P(d|h_n)\pi(h_n), \quad [9]$$

where $d \subset h$ indicates that $P_P(d|h) = f(d)$. Exchanging the sums produces

$$\pi(h_{n+1}) = P(h_{n+1})^r \sum_{d \subset h_{n+1}} f(d) \frac{\sum_{h_n \supset d} \pi(h_n)}{\sum_{h \supset d} P(h)^r}, \quad [10]$$

which it is easy to check is satisfied by $\pi(h) = P(h)^r/\Sigma_{h'}P(h')^r$ because $\Sigma_{d \subset h} f(d) = 1$ for any h. Note that the noisy meaning-class mapping model used in our previous analyses does not fall within the set of languages to which this result applies unless $\varepsilon = 0$ and that this result does not predict the "bottleneck" effect discussed in the text because the posterior distribution is invariant to the amount of information provided by the data d. From this, we infer that some form of noise in the system is critical for the "bottleneck" effect to occur, although establishing the exact conditions under which this effect arises is an interesting problem for future research.

We thank the members of the Language Evolution and Computation research unit in Edinburgh, M. Johnson, M. Kalish, S. Lewandowsky, and T. Lombrozo for many discussions of this work during its infancy. M.D. was supported by Economic and Social Research Council (ESRC) and Japan Society for the Promotion of Science Postdoctoral Fellowships (ESRC award PTA-026-27-0760), and T.L.G. was supported by National Science Foundation Grant BCS-0544708.

1. Croft W (1990) *Typology and Universals* (Cambridge Univ Press, Cambridge, UK).
2. Hawkins JA , ed (1988) *Explaining Language Universals* (Blackwell, Oxford).
3. Hauser M, Chomsky N, Fitch WT (2002) *Science* 298:1569–1579.
4. Chomsky N (1965) *Aspects of the Theory of Syntax* (MIT Press, Cambridge, MA).
5. Hawkins JA (1994) *A Performance Theory of Order and Constituency* (Cambridge Univ Press, Cambridge, UK).
6. Pinker S, Bloom P (1990) *Behav Brain Sci* 13:707–784.
7. Kirby S (1999) *Function, Selection and Innateness: The Emergence of Language Universals* (Oxford Univ Press, Oxford).
8. Christiansen MH (1994) PhD thesis (Univ of Edinburgh, Edinburgh).
9. Deacon TW (1997) *The Symbolic Species: The Co-evolution of Language and the Brain* (Norton, New York).
10. Bayes T (1763) *Philos Trans R Soc London* 53:370–418.
11. Bernardo JM, Smith AFM (1994) *Bayesian Theory* (Wiley, Chichester, UK).
12. Robert C (1995) *The Bayesian Choice* (Springer, New York).
13. MacKay D (2003) *Information Theory, Inference and Learning Algorithms* (Cambridge Univ Press, Cambridge, UK).
14. Manning C, Schütze H (1999) *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, MA).
15. Jurafsky D (1996) *Cognit Sci* 20:137–194.
16. Briscoe EJ (2002) in *Linguistic Evolution Through Language Acquisition*, ed Briscoe EJ (Cambridge Univ Press, Cambridge, UK), pp 255–300.
17. Kirby S (2001) *IEEE Trans Evol Comput* 5:102–110.
18. Kirby S, Hurford J (2002) in *Simulating the Evolution of Language*, eds Cangelosi A, Parisi D (Springer, London), pp 121–148.
19. Smith K, Kirby S, Brighton H (2003) *Artificial Life* 9:371–386.
20. Kirby S, Smith K, Brighton H (2004) *Studies Lang* 28:587–607.
21. Brighton H, Smith K, Kirby S (2005) *Phys Life Rev* 2:177–226.
22. Griffiths TL, Kalish ML (2007) *Cognit Sci*, in press.
23. Niyogi P, Berwick RC (1997) *Complex Syst* 11:161–204.
24. Nowak MA, Komarova NL, Niyogi P (2001) *Science* 291:114–118.
25. Nowak MA, Komarova NL, Niyogi P (2002) *Nature* 417:611–617.
26. Wray A (1998) *Language Commun* 18:47–67.
27. Francis N, Kucera H (1982) *Frequency Analysis of English Usage: Lexicon and Grammar* (Houghton Mifflin, New York).
28. Deacon TW (2003) in *Evolution and Learning: The Baldwin Effect Reconsidered*, eds Weber B, Depew D (MIT Press, Cambridge, MA).
29. Hultsch H (1991) *Anim Behav* 42:883–889.
30. Johnson NL, Kotz S (1972) *Distributions in Statistics: Continuous Multivariate Distributions* (Wiley, New York).
31. Rissanen J (1978) *Automatica* 14:465–471.
32. Shannon CE (1948) *Bell System Tech J* 27:379–423 and 623–656.