Title

A Bayesian View of Language Evolution by Iterated Learning

Permalink

https://escholarship.org/uc/item/0vb7c896

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 27(27)

ISSN

1069-7977

Authors

Griffiths, Thomas L.
Kalish, Michael L.

Publication Date

2005

Peer reviewed

A Bayesian View of Language Evolution by Iterated Learning

Thomas L. Griffiths (tom_griffiths@brown.edu)
Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912

Michael L. Kalish (kalish@louisiana.edu)
Institute of Cognitive Science, University of Louisiana at Lafayette, Lafayette, LA 70504

Abstract

Models of language evolution have demonstrated how aspects of human language, such as compositionality, can arise in populations of interacting agents. This paper analyzes how languages change as the result of a particular form of interaction: agents learning from one another. We show that, when the learners are rational Bayesian agents, this process of iterated learning converges to the prior distribution over languages assumed by those learners. The rate of convergence is set by the amount of information conveyed by the data seen by each generation; the less informative the data, the faster the process converges to the prior.

Human languages form a subset of all logically possible communication schemes, with universal properties shared by all languages (Comrie, 1981; Greenberg, 1963; Hawkins, 1988). A traditional explanation for these linguistic universals is that they are the consequence of constraints on the set of learnable languages imposed by an innate, language-specific, genetic endowment (e.g., Chomsky, 1965). Recent research has explored an alternative explanation: that universals emerge from evolutionary processes produced by the transmission of languages across generations (e.g., Kirby, 2001; Nowak, Plotkin, & Jansen, 2000). Languages change as each generation learns from that which preceded it. This process of iterated learning implicitly selects for languages that are more learnable. This suggests a tantalizing hypothesis: that iterated learning might be sufficient to explain the emergence of linguistic universals (Briscoe, 2002).

Kirby (2001) introduced a framework for exploring this hypothesis, called the *iterated learning model* (ILM). In the ILM, each generation consists of one or more learners. Each learner sees some data, forms a hypothesis about the process that produced that data, and then produces the data which will be supplied to the next generation of learners, as shown in Figure 1 (a). The languages that succeed in being transmitted across generations are those that pass through the "information bottleneck" imposed by iterated learning. If particular properties of languages make it easier to pass through that bottleneck, then many generations of iterated learning might allow those properties to become universal.

The ILM can be used to explore how different assumptions about language learning influence language evolution. A variety of learning algorithms have been examined using the ILM, including a heuristic grammar inducer (Kirby, 2001), associative networks (Smith, Kirby, & Brighton, 2003), and minimum description length (Brighton, 2002). Iterated learning with these algorithms produces languages that possess one of the most compelling properties of human languages: *compositionality.* In a compositional language, the meaning of an utterance is a function of the meaning of its parts. The intuitive explanation for these results is that the regular structure of compositional languages means that they can be learned from less data, and are thus more likely to pass through the information bottleneck.

These instances of compositionality emerging from iterated learning raise an important question: what languages will survive many generations of iterated learning? While the circumstances under which compositionality will emerge from iterated learning with specific learning algorithms have been investigated (Brighton, 2002; Smith et al., 2003), there are no general results for arbitrary properties of languages or broad classes of learning algorithms. In this paper, we analyze iterated learning for the case where the learners are rational Bayesian agents. A variety of learning algorithms can be formulated in terms of Bayesian inference, and Bayesian methods underlie many approaches in computational linguistics (Manning & Schütze, 1999). The assumption that the learners are Bayesian agents makes it possible to derive analytic results indicating which languages will be favored by iterated learning. In particular, we prove the surprising result that the probability distribution over languages resulting from iterated Bayesian learning converges to the prior probability distribution assumed by the learners. This implies that the asymptotic probability that a language is used does not depend at all upon the properties of the language, being determined entirely by the assumptions of the learner.

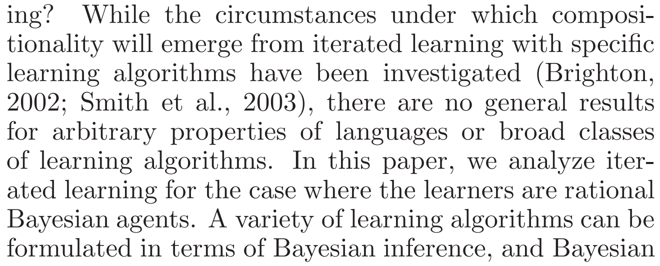

Figure 1: (a) Iterated learning. (b) Dependencies among variables in iterated iterated Bayesian learning.

Iterated Bayesian learning

Following most of the work applying iterated learning to language evolution, we will assume that our learners are faced with a *function learning* task: given a set of m inputs, $\mathbf{x} = \{x_1, \ldots, x_m\}$, and m corresponding outputs, $\mathbf{y} = \{y_1, \ldots, y_m\}$, the learner has to estimate the probability distribution over y for each x. In a language learning setting, $\mathbf{x}$ is usually taken to be a set of "meanings" or events in the world, and $\mathbf{y}$ is taken to be the set of utterances associated with those events. We will use $\mathcal{X}$ and $\mathcal{Y}$ to denote the set of values that $\mathbf{x}$ and $\mathbf{y}$ can take on.

Iterated learning begins with some initial data, $(\mathbf{x}_0, \mathbf{y}_0)$, presented to the first learner, who then generates outputs $\mathbf{y}_1$ in response to some new inputs $\mathbf{x}_1$. The second learner sees $(\mathbf{x}_1, \mathbf{y}_1)$, and generates $\mathbf{y}_2$ in response to $\mathbf{x}_2$. This process continues for each successive generation, with learner $n+1$ seeing $(\mathbf{x}_n, \mathbf{y}_n)$ and generating $\mathbf{y}_{n+1}$ in response to $\mathbf{x}_{n+1}$. The result of this process depends upon the algorithm used by the learners.

We will assume that our learners are Bayesian agents, supplied with a finite discrete[1] hypothesis space $\mathcal{H}$ and a prior probability distribution $p(h)$ for each hypothesis $h \in \mathcal{H}$. In a function learning task, each hypothesis h corresponds to a conditional probability distribution $p(\mathbf{y}|\mathbf{x}, h)$, specifing the distribution over all sets of outputs for any set of inputs. In the *learning* step of the process illustrated in Figure 1 (a), learner $n+1$ sees $(\mathbf{x}_n, \mathbf{y}_n)$, and computes a posterior distribution over h_{n+1} using Bayes' rule

$$p(h_{n+1}|\mathbf{x}_n, \mathbf{y}_n) = \frac{p(\mathbf{y}_n|\mathbf{x}_n, h_{n+1})p(h_{n+1})}{p(\mathbf{y}_n|\mathbf{x}_n)}$$

where

$$p(\mathbf{y}_n|\mathbf{x}_n) = \sum_{h \in \mathcal{H}} p(\mathbf{y}_n|\mathbf{x}_n, h)p(h).$$

We will assume that the learners consider each y_i independent given x_i and h, so $p(\mathbf{y}|\mathbf{x}, h) = \prod_i p(y_i|x_i, h)$. In the *production* step, learner $n+1$ sees $\mathbf{x}_{n+1}$, generated from a distribution $q(\mathbf{x}_n)$ that is independent of all other variables. The change in notation is a reminder that unlike all of the other distributions we have mentioned, $q(\mathbf{x})$ expresses the objective probability of an event in the world, rather than a subjective probability assessed by the learner. Given $\mathbf{x}_{n+1}$, the learner samples a hypothesis, h_{n+1}, from $p(h_{n+1}|\mathbf{x}_n, \mathbf{y}_n)$, and generates $\mathbf{y}_{n+1}$ from the distribution $p(\mathbf{y}_{n+1}|\mathbf{x}_{n+1}, h_{n+1})$.

A Markov chain on hypotheses

The stochastic process that iterated Bayesian learning defines on $\mathbf{x}$, $\mathbf{y}$, and h has the dependency structure shown in Figure 1 (b). It is straightforward to analyze the properties of this stochastic process. In particular, if we sum over the data $(\mathbf{x}_n, \mathbf{y}_n)$ we obtain the distribution

$$p(h_{n+1}|h_n) = \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{y} \in \mathcal{Y}} p(h_{n+1}|\mathbf{x}, \mathbf{y})p(\mathbf{y}|\mathbf{x}, h_n)q(\mathbf{x}), \quad (1)$$

defining a sequence of random variables in which h_{n+1} is independent of all previous hypotheses given h_n. This is a Markov chain, with state space $\mathcal{H}$ and transition matrix $T(h_n, h_{n+1}) = p(h_{n+1}|h_n)$.[2]

Markov chains are a common form of stochastic process with well-understood properties (see Norris, 1997). In particular, identifying a process as a Markov chain immediately provides insight into its asymptotic behavior. If a Markov chain with transition matrix $T(h_n, h_{n+1})$ is *ergodic*, then it will converge to a *stationary distribution* $\pi(h)$ satisfying the equation

$$\pi(h_{n+1}) = \sum_{h_n \in \mathcal{H}} T(h_n, h_{n+1})\pi(h_n) \quad (2)$$

for all h_{n+1}. Intuitively, convergence to the stationary distribution means that regardless of the initial state h_1 (or, in our setting, the data on which that hypothesis is based), the probability distribution over h_n approaches $\pi(h_n)$ as $n \to \infty$. Equation 2 indicates why $\pi(h_n)$ is the "stationary" distribution: if h_n follows this distribution, then so will h_{n+1}, and likewise for every h_{n+k} for $k > 0$.

The conditions for ergodicity are given in Norris (1997). The most important condition in our setting is *irreducibility*: for every pair of hypotheses h and h', there must be some k such that the probability of going from h to h' in k steps is greater than zero. If this condition is violated, it is possible to enter sets of states from which there is no departure, preventing the chain from visiting all of the states which have some probability under the stationary distribution.

The stationary distribution, $\pi(h)$, for the Markov chain with transition matrix $T(h_n, h_{n+1}) = p(h_{n+1}|h_n)$ satisfies the equation

$$\pi(h_{n+1}) = \sum_{h_n \in \mathcal{H}} p(h_{n+1}|h_n)\pi(h_n).$$

If we take $\pi(h) = p(h)$, we obtain

$$\begin{aligned}
p(h_{n+1}) &= \sum_{h_n \in \mathcal{H}} \left[\sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{y} \in \mathcal{Y}} p(h_{n+1}|\mathbf{x}, \mathbf{y})p(\mathbf{y}|\mathbf{x}, h_n)q(\mathbf{x}) \right] p(h_n) \\
&= \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{y} \in \mathcal{Y}} p(h_{n+1}|\mathbf{x}, \mathbf{y}) \left[\sum_{h_n \in \mathcal{H}} p(\mathbf{y}|\mathbf{x}, h_n)p(h_n) \right] q(\mathbf{x}) \\
&= \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\mathbf{y} \in \mathcal{Y}} \frac{p(\mathbf{y}|\mathbf{x}, h_{n+1})p(h_{n+1})}{p(\mathbf{y}|\mathbf{x})} p(\mathbf{y}|\mathbf{x})q(\mathbf{x}) \\
&= p(h_{n+1}) \sum_{\mathbf{x} \in \mathcal{X}} \left[\sum_{\mathbf{y} \in \mathcal{Y}} p(\mathbf{y}|\mathbf{x}, h_{n+1}) \right] q(\mathbf{x}).
\end{aligned}$$

The two sums on the last line evaluate to 1, showing that $p(h)$ is the stationary distribution of this chain. Consequently, provided the underlying Markov chain is ergodic, the distribution over hypotheses entertained by Bayesian learners engaged in iterated learning will converge to the prior over hypotheses held by the learners.

[1] This assumption is not necessary for our results to hold. Similar results can be obtained with continuous hypothesis spaces, but the proofs are more involved.

[2] The process can also be reduced to a Markov chain on data, $(\mathbf{x}_n, \mathbf{y}_n)$, by summing out the hypotheses, h_n. Similar results hold for this chain, but we omit them due to space.

An example: evolving compositionality

The results in the previous section imply that the asymptotic probability with which a language is spoken depends only upon its prior probability, and is not affected by any of the properties of the language. This result is counter-intuitive, particularly in the light of previous results indicating that iterated learning seems to favor particular properties of languages, such as compositionality. To gain a deeper understanding of our results, we examined the consequences of iterated Bayesian learning in a simplified version of the scenario used by Kirby (2001), Smith et al. (2003), and Brighton (2002) for exploring the evolution of compositionality.

In our scenario, meanings and utterances each vary along two binary dimensions. This yields a total of four meanings and four utterances, each corresponding to the set $\{00, 01, 10, 11\}$. In a *holistic* language, the mapping between meanings and utterances is arbitrary, and a single word is chosen to represent each meaning without any constraints. There are $4^4 = 256$ such languages. In a *compositional* language, the mapping between meanings and utterances depends upon their parts: the two dimensions of meanings are mapped onto the two dimensions of utterances (for simplicity, we assume that the order is preserved), and the only uncertainty is in which values map to one another. There are $2^2 = 4$ such languages.

The hypothesis space $\mathcal{H}$ thus contains 260 hypotheses, each a mapping between meanings and utterances. For each $h \in \mathcal{H}$, we defined the probability distribution over outputs y given the input x to be

$$p(y|x,h) = \begin{cases} 1-\epsilon & x \text{ maps to } y \text{ in } h \\ \frac{\epsilon}{3} & \text{otherwise} \end{cases} \tag{3}$$

where ϵ is the "error rate" of production. The prior probability of each hypothesis was

$$p(h) = \begin{cases} \frac{\alpha}{4} & h \text{ is compositional} \\ \frac{1-\alpha}{256} & h \text{ is holistic} \end{cases} . \tag{4}$$

This is a *hierarchical prior*, allocating a probability of α to the set of compositional languages and $1-\alpha$ to the set of holistic languages, and then spreading this probability uniformly over the hypotheses within those sets.

Since every language is simply a mapping from meanings to utterances, our hypothesis space includes four holistic languages that each give the same mapping as one of the four compositional languages. These languages make the same predictions about inputs and outputs, as determined by Equation 3, and thus cannot be discriminated by any data. The advantage of the compositional languages over their holistic counterparts results from the prior defined in Equation 4. If compositional and holistic languages are equally probable a priori ($\alpha = 0.5$), then the relatively small number of compositional languages means that any particular compositional language is more probable than any particular holistic language. Consequently, it would be very unlikely to see a holistic language that just happened to produce a compositional mapping. As α becomes smaller, it becomes less likely that one would see a compositional language at all, and a holistic language that just happened to produce a compositional mapping becomes more plausible. However, the number of holistic languages grows much faster than the number of compositional languages as the space of meanings and utterances becomes larger, so the value of α needed to overwhelm the advantage of compositional languages rapidly becomes extremely small.

The matrix of transition probabilities $p(h_{n+1}|h_n)$ can be obtained by summing over all $(\mathbf{x}, \mathbf{y})$ pairs, as shown in Equation 1. Since there are $(2^2 2^2)^m$ such pairs, this is intractable for large m. Consequently, matrices for $m > 4$ were computed approximately using a Monte Carlo method with 1000 samples for each hypothesis. We computed transition matrices for $\alpha \in \{0.01, 0.5\}$, $\epsilon \in \{0.01, 0.05\}$, and $m \in \{1, 2, \ldots, 10\}$. The first column of Figure 2 shows a portion of some of these transition matrices.

The effect of priors

The second column of Figure 2 shows 1000 iterations sampled from four Markov chains (initialized by choosing h_1 at random), while the third and fourth columns (labelled "Chain" and "Prior") show the distribution over hypotheses from a single sample of 10000 iterations from those chains and the prior $p(h)$ respectively. The results in (a)-(c) all use $\alpha = 0.5$ and $\epsilon = 0.05$, giving equal prior probability to compositional and holistic languages and allowing a reasonable amount of error. The number of datapoints seen by the learners varies, with $m = 1$ in (a), $m = 3$ in (b), and $m = 10$ in (c). While the Markov chain develops a greater tendency to remain in the same state as m increases, indicated by the strong diagonal in the transition matrices and the length of the streaks in the samples, it maintains the same distribution over hypotheses, as shown in the "Chain" column. This distribution matches the prior over hypotheses, consistent with our mathematical analysis.

The results shown in Figure 2 (d) illustrate how changing the prior changes the stationary distribution. Keeping ϵ and m at the same values as in (b), α is set to 0.01, giving extremely low prior probability to the set of compositional languages. Consequently, each holistic language has a slightly higher probability than any compositional language. The distribution over hypotheses produced by the Markov chain is markedly different from that in (b), and matches the prior. If compositional languages have low prior probability, then they do not emerge as a result of iterated learning.

Convergence rates

The rate at which a discrete Markov chain converges to its stationary distribution is determined by the second eigenvalue of the transition matrix, λ_2, with smaller values of λ_2 resulting in faster convergence (Norris, 1997). Figure 3 (a) shows how λ_2 is affected by α, ϵ, and m. As α brings $p(h)$ away from uniformity, it increases the probability that learners n and $n+1$ will share the same hypothesis. Thus, increasing α increases λ_2. Changing ϵ and m decreases the rate of convergence as the data

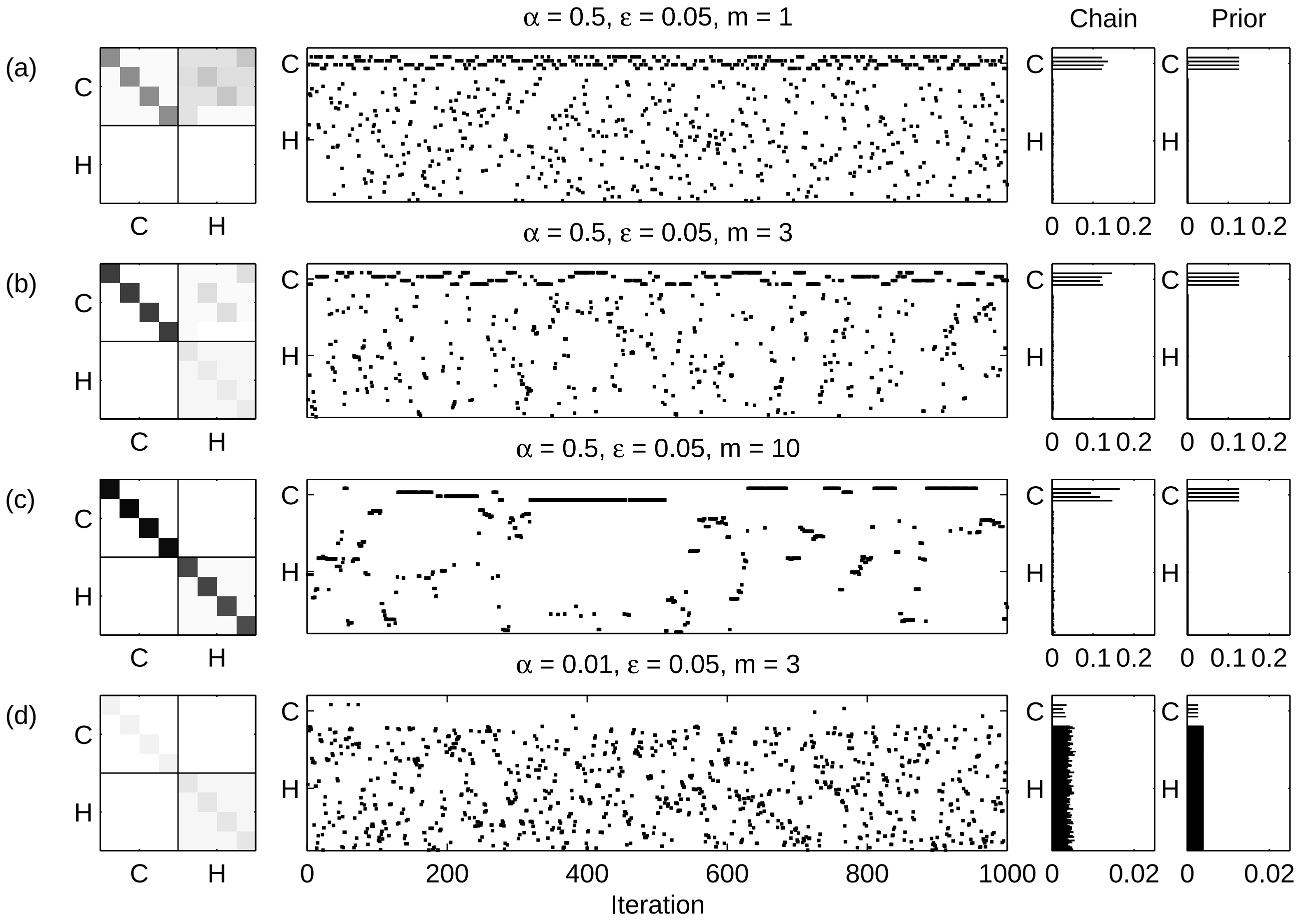

Figure 2: Markov chains on hypotheses for the evolution of compositionality. Different rows correspond to different parameter values. For each set of parameters, the first column shows a portion of the transition matrix, with four compositional languages (C) and four holistic languages (H). Columns are h_n, rows are h_{n+1}, and darker grey indicates a higher value of $p(h_{n+1}|h_n)$. The second column shows a sample of 1000 iterations from this matrix, the third shows the relative frequency of hypotheses across 10000 iterations, and the fourth shows the prior, $p(h)$.

received by each learner become more informative. Decreasing ϵ increases the fidelity with which information is transferred between generations, increasing the correspondence between the hypotheses of successive learners and thus increasing λ_2. Increasing m increases the amount of information language production provides for language learning, and thus the probability that learner $n+1$ will acquire the language of learner n, increasing λ_2. With large m, it is likely that a single hypothesis will be maintained across several generations. The effect of large m is demonstrated in Figure 2 (c), where some compositional hypotheses are dominant over hundreds of iterations.

The information bottleneck

In applications of the ILM, it is common to explain the emergence of languages with particular properties in terms of the "information bottleneck" imposed by transmission of a language across generations. This bottleneck provides a selection pressure for languages which are more learnable. For example, Kirby (2001) explained the emergence of compositionality in terms of the relative ease of transmitting a compositional language: languages that contain more generalizations are more compressible, and can thus be learned from smaller amounts of data. In support of this claim, Smith et al. (2003) showed that tightening the information bottleneck, by reducing the amount of data a learner saw, increased the advantage in stability for compositional languages over holistic languages with their learning algorithms.

Figure 3 (b) shows how the relative stability of compositional and holistic languages changes as a function of α, ϵ, and m. The relative stability was assessed by computing the ratio of the mean probability that a particular compositional language would appear as both h_n and h_{n+1} to the mean probability that a particular holistic language would appear as both h_n and h_{n+1}. The figure shows that this stability ratio is strongly affected by α: if the prior probability of a compositional language is high, it is more likely that a learner will acquire that

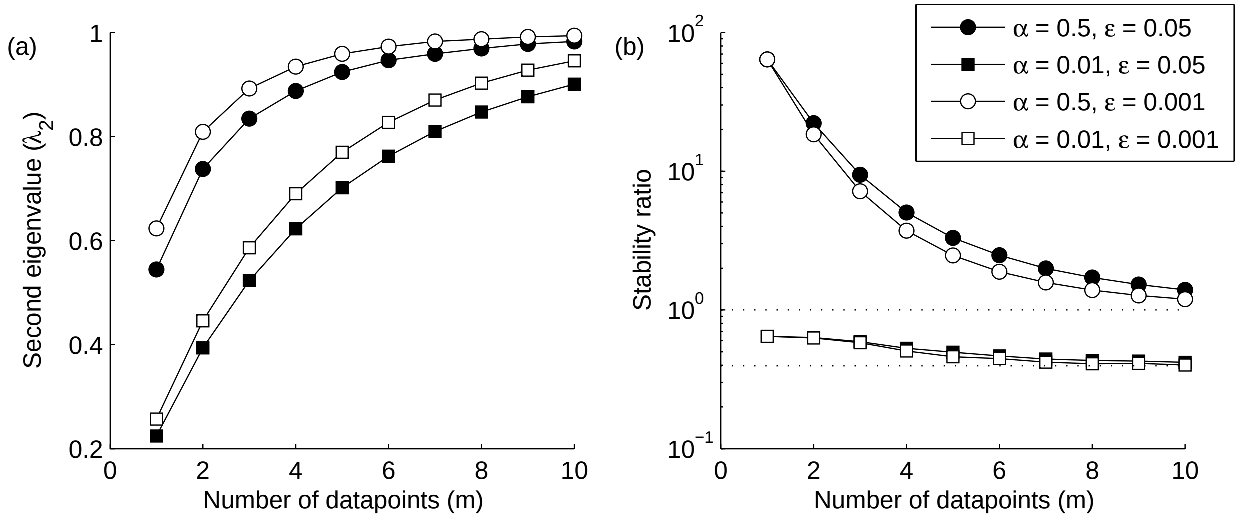

Figure 3: Quantities derived from Markov chains on hypotheses as a function of number of datapoints, m, prior on composite languages, α, and error rate, ϵ. (a) Second eigenvalue of transition matrix, λ_2. (b) Stability ratio. The dotted line shows the stability ratio as $m \to \infty$. Lower values of m constitute a tighter "information bottleneck".

language, and consequently that language is more stable. The magnitude of this effect is modulated by the number of datapoints, m, with α having the greatest effect when m is small. As m increases, the data begin to overcome the influence of the prior.[3]

The results shown in Figure 3(b) are evocative of those of Smith et al. (2003): tightening the information bottleneck produces a greater advantage for compositional languages when each of those languages has higher prior probability than any holistic language. The explanation is the same: a tight bottleneck favors more learnable languages. For a Bayesian, an important aspect of learnability is consistency with the prior.

Discussion

Our results provide simple conditions for determining when a particular property of languages will emerge through iterated Bayesian learning: languages will appear in proportion to their prior probability, provided the Markov chain defined by the learners is ergodic. In closing, we will consider connections between these results and previous work on iterated learning, how they bear upon the issue of linguistic universals, and their implications for understanding the diversity and dynamics of human languages.

Previous work on iterated learning

We view our results as broadly consistent with previous work on iterated learning, but suggesting a different approach to understanding its consequences. With Bayesian learners, iterated learning results in convergence to the prior distribution over hypotheses. Consequently, the iterated learning process is only the engine by which languages with particular properties emerge: the real object of analysis should be the assumptions behind the algorithms used by the learners. This conclusion is quite different from that of previous work, in which the emphasis is placed upon the learning process itself as the source of linguistic universals.

Interpreting previous results in terms of our framework is not straightforward, as the learning algorithms used in most previous analyses do not have a clean Bayesian interpretation. It is also not clear whether these algorithms satisfy the requirements for the underlying Markov chain to be ergodic: in many cases, the criterion for ending simulations was reaching a steady state, which is not something that should happen with an irreducible Markov chain. To the extent that our results are applicable, they suggest that the algorithms that Kirby (2001), Brighton (2002), and Smith et al. (2003) used to demonstrate the emergence of compositionality through iterated learning implicitly define a prior distribution over hypotheses that favors compositional languages. Making these connections explicit is an important direction for future work.

Prior probabilities and linguistic universals

By tying the probability of a language emerging through iterated learning to its prior probability, our analysis locates the stability of languages firmly in the algorithm applied by the learners. The structure of languages has no effect on their stability, except insofar as it determines

[3] A small influence of the prior remains even at asymptote due to the presence of holistic hypotheses that are equivalent to compositional hypotheses. These hypotheses cannot be separated by any amount of data, so the stability ratio approaches $\frac{64}{62+1/\alpha}$ as $m \to \infty$. If $\mathcal{H}$ did not include hypotheses that make equivalent predictions about the data, the stability ratio would approach 1 as $m \to \infty$. Consequently, the decrease in the stability ratio as a function of m for the cases where $\alpha = 0.01$ is due to the specific structure of $\mathcal{H}$, rather than being a general trend.

prior probability. From this perspective, linguistic universals simply manifest the prior probability distribution over languages entertained by the learner. Explaining linguistic universals thus requires explaining why particular properties of language have high prior probability.

The statement that linguistic universals result from the priors of learners initially seems consistent with traditional explanations, with innate, language-specific, genetic endowment providing these priors. This need not be the case. The priors that a Bayesian agent brings to a learning task reflect their general cognitive capacities, and the expectations yielded by their experience with all other independent sources of evidence. Furthermore, priors can be motivated by a priori symmetry arguments – such as the belief that holistic and compositional languages should be equally likely – or information theoretic constraints – such as the relative difficulty of encoding languages (c.f. Brighton, 2002). Either of these latter considerations would be sufficient to explain the evolution of compositionality. Framing the explanation of linguistic universals in terms of accounting for the prior probability distribution entertained by language learners provides a well-defined formal setting in which to explore these issues.

Language change and regularization

If we take the idea that language evolves through iterated Bayesian learning seriously, it has several interesting implications for understanding the diversity and dynamics of language. First, if iterated learning is the only force influencing language change, then all languages used by human beings should be considered samples from the prior probability distribution over languages. Even though the assumption that no other selective pressures are at work in language evolution – such as the relative communicative utility of speaking one language rather than another – is probably false, this result provides a direct connection between the mind and world that is provocative. It provides a formal justification for the idea that examining the diversity of human languages has the potential to reveal interesting properties of the human mind.

A second implication of this view of language evolution is that language change should be viewed as a random walk through this prior probability distribution. Once a Markov chain has converged, it will move through its state space in a fashion determined by its dynamics, visiting each state with probability determined by the stationary distribution. If a language is well-determined by the evidence available to a learner (together with their prior probabilities), the rate at which this random walk moves between states should be very low, as in Figure 2 (c). However, examining the dynamics of language change can shed further light on the structure of the prior, and the transition probabilities.

Finally, the development of new languages can be understood in terms of convergence to this prior. Cases like creolization (e.g., Bickerton, 1981) are striking in terms of the sudden regularization introduced by a single generation of learners. Similarly, a Markov chain initialized in a state with very low probability under the stationary distribution will rapidly move towards states with higher probability. By studying how new languages develop, we have the opportunity to map out languages with low and high prior probability, and to estimate the rate at which the Markov chain converges.

Conclusion

We have presented a novel mathematical framework for exploring the consequences of iterated learning, based upon the assumption that learners are rational Bayesian agents. Making this assumption allows us to obtain precise results that characterize the circumstances under which iterated learning will produce languages with particular properties. These results have the potential to provide connections between the formal and functional approaches to explaining the existence of linguistic universals, showing that the results of iterated learning – a process that would seem to emphasize the functional properties of languages – do not depend upon the structure of the languages involved, except insofar as that structure determines their prior probability. We anticipate that this framework will prove useful in the further study of language evolution by iterated learning, as well as in other settings. Iterated learning provides a natural model of cultural evolution, suggesting that our framework could be applied to a range of cultural phenomena other than language. Our results demonstrate that iterated learning also provides a means of assessing the priors of learners, which could be exploited in a laboratory setting as a means of determining learning biases.

References

Bickerton, D. (1981). *Roots of language.* Karoma, Ann Arbor, MI.

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

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

Comrie, B. (1981). *Language universals and linguistic typology.* University of Chicago Press, Chicago.

Greenberg, J., editor (1963). *Universals of language.* MIT Press, Cambridge, MA.

Hawkins, J., editor (1988). *Explaining language universals.* Blackwell, Oxford.

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

Manning, C. and Schütze, H. (1999). *Foundations of statistical natural language processing.* MIT Press, Cambridge, MA.

Newmeyer, F. J. (1998). *Language form and language function.* MIT Press, Cambridge, MA.

Norris, J. R. (1997). *Markov Chains.* Cambridge University Press, Cambridge, UK.

Nowak, M. A., Plotkin, J. B., and Jansen, V. A. A. (2000). The evolution of syntactic communication. *Nature*, 404:495–498.

Smith, K., Kirby, S., and Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life*, 9:371–386.