Language evolution is shaped by the structure of the world: An iterated learning analysis

Amy Perfors (amy.perfors@adelaide.edu.au)
School of Psychology, University of Adelaide

Daniel Navarro (daniel.navarro@adelaide.edu.au)
School of Psychology, University of Adelaide

Abstract

Human languages vary in many ways, but also show striking cross-linguistic universals. Why do these universals exist? Recent theoretical results demonstrate that Bayesian learners transmitting language to each other through iterated learning will converge on a distribution of languages that depends only on their prior biases about language and the quantity of data transmitted at each point; the structure of the world being communicated about plays no role (Griffiths & Kalish, 2005, 2007). We revisit these findings and show that when certain assumptions about the independence of languages and the world are abandoned, learners will converge to languages that depend on the structure of the world as well as their prior biases. These theoretical results are supported with a series of experiments showing that when human learners acquire language through iterated learning, the ultimate structure of those languages is shaped by the structure of the meanings to be communicated.
Keywords: language evolution; iterated learning; Bayesian modeling; linguistic structure

Introduction

Human languages have rich structure on many levels, from phonology to semantics to grammar. Where does this structure come from? Most researchers agree that linguistic structure is shaped by the structure of our minds – that our brains contain prior biases that favor the acquisition or retention of some linguistic systems over others. As such, debate generally centers around the nature and origin of these biases. Some suggest that the human language faculty is genetically specified, with natural selection operating on genes for language (e.g., Pinker & Bloom, 1990; Nowak, Komarova, & Niyogi, 2001; Komarova & Nowak, 2001) or else selecting for other capabilities (e.g., Hauser, Chomsky, & Fitch, 2002). Others have suggested that humans easily learn language not because of a language-specific genetically encoded mechanism, but because language evolved to be learnable and useable by human brains (e.g. Zuidema, 2002; Brighton, Smith, & Kirby, 2005; Christiansen & Chater, 2008). While these accounts disagree in many particulars, they agree that the structure of language arises from the structure of the brain.

In this paper we argue that language evolution is shaped by the structure of the world in addition to pre-existing cognitive biases. Because language involves communicating about the world, the structure of that world (i.e., the things to be communicated) can interact with people's prior biases to shape the languages that develop. We offer theoretical and experimental support of this proposition. On the theoretical side, we take as our starting point recent work within the "iterated learning" framework (in which new learners receive their data from previous learners). Previous research has shown that when learners are individually Bayesian, an iterated learning chain converges in the limit to the prior distribution over all possible languages (Griffiths & Kalish, 2005, 2007). However, the proof of this assumes *a priori* that a language carries no assumptions about the frequencies of events in the world. As we will show, when this assumption is relaxed, the iterated learning process converges to a distribution that depends on the distribution of meaningful events in the world as well as the prior biases of the learner. We experimentally test these theoretical results in a lab-based iterated learning experiment (as in, e.g., Kirby, Cornish, & Smith, 2008) and find that participants converge on different languages depending on the structure of the space of meanings they are shown.

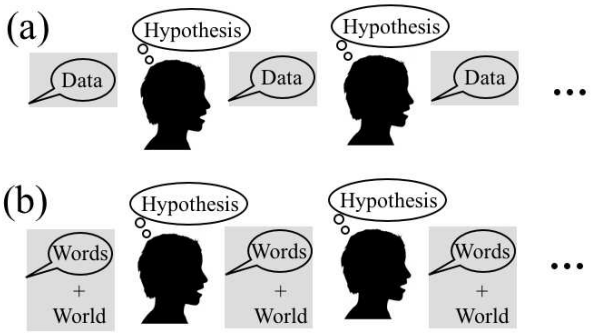

Figure 1: (a) Schematic illustration of the typical iterated learning paradigm, which assumes that learner n acquires language on the basis of the language data produced by learner $n-1$. (b) A different view of iterated learning recognizes that because individuals produce language in order to communicate about the world, the data available to learners includes meanings in the world as well as the linguistic data produced by the learner before them.

Iterated learning

The iterated learning modeling (ILM) framework is widely used in language evolution research (e.g., Kirby & Hurford, 2002; Griffiths & Kalish, 2007; Kirby et al., 2008; Smith, 2009; Reali & Griffiths, 2009). It views the process of language evolution in terms of a chain of learners (or generations), shown schematically in Figure 1(a). The first learner in the chain sees some linguistic data (e.g., utterances), forms a hypothesis about what sort of language would have generated that data, and then produces their own data, which serves as input to the next learner in the chain. Over time, the languages that emerge from this process become non-arbitrary: Griffiths and Kalish (2005, 2007) (henceforth, GK) demon-

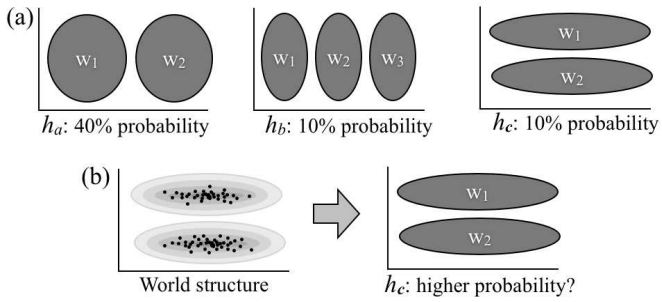

Figure 2: (a) Intuitive illustration of the results of Griffiths & Kalish (2005, 2007) (GK). Given a 2-dimensional semantic space, a learner with a prior bias to favor one dimension of that space (the *x*-axis) and languages with fewer words might have a prior distribution over languages that puts more probability on h_a and less on h_b or h_c. GK demonstrate that the languages that evolve will converge to this prior distribution. (b) If the natural categories in the world have a different structure, we might intuitively expect that languages that capture that structure, like h_c, should be more likely to evolve.

strate that when the learners are Bayesian, we should expect an iterated learning chain to converge to the prior distribution over all possible languages. That is, the probability of any given language emerging does not depend on the structure of the world or independent properties of the language – only the assumptions of the learner. The existence of a linguistic bottleneck (in which only a small amount of information is transmitted at each link in the chain) can speed the rate of convergence or create a pressure for certain kinds of linguistic structure like compositionality, but this result implies that neither the structure of the meaning space nor the nature of the initial language should have an affect on the language that eventually evolves.

GK's result can be conceptualized in intuitive terms as follows. Suppose learners must acquire languages that describe a two dimensional semantic space of some sort. For illustrative purposes, suppose further that the learners have a prior bias to prefer languages with fewer words and to pay more attention to one of the dimensions, as occurs in human category learning and development (e.g., Landau, Smith, & Jones, 1988). This prior bias might impose a distribution over hypotheses h about possible languages, like the illustrative one shown in Figure 2(a): languages like h_a with a few words that classify according to the preferred dimension (the *x*-axis in this case) have higher prior probability than languages like h_b, which have many words, or h_c, whose words classify according to the dis-preferred dimension. GK suggest that languages evolving to describe this space will converge to the prior distribution: 40% of the time h_a will emerge, 10% of the time h_b will emerge, and so forth. Although this prior and these precise numbers are imaginary, the picture provides a schematic illustration of what GK's results mean.

It also, however, highlights an apparent oddity within these results. Suppose that the world possesses structure in the form of natural categories of some sort, and these natural categories happen to group items according to the non-preferred dimension, as shown in Figure 2(b): the items observed by the learner – generated by the world – correspond to the black dots, which fall naturally into two clusters. We might intuitively expect that a language like h_c would be a better fit to this world (and hence be more likely to evolve) than a language like h_a, even though h_a has higher prior probability. The results of GK appear to suggest otherwise. Is our intuition simply wrong, or is there a mismatch between the GK derivation and the problem of language evolution within a structured world? In the next section, we argue for the latter.

Theoretical result

We formalize the iterated learning framework in much the same way as Griffiths and Kalish (2005). A learner sees m meanings or events, denoted $x = \{x^{(1)} \ldots x^{(m)}\}$. These meanings are paired with m corresponding utterances denoted $y = \{y^{(1)} \ldots y^{(m)}\}$. The first learner in the chain is shown some initial data consisting of meaning-utterance pairs (x_0, y_0). Then, when shown new events x_1, the learner produces utterances y_1, so that (x_1, y_1) are the input to the next learner. In general, learner $n+1$ sees data (x_n, y_n) and generates y_{n+1} given events x_{n+1}, so that the next learner receives input (x_{n+1}, y_{n+1}). The goal of each learner is to estimate the mapping between meanings and utterances, which corresponds to learning the language they are exposed to. It is assumed that each learner has the same countable hypothesis space $\mathcal{H}$ of possible languages, such that each $h \in \mathcal{H}$ corresponds to one language. For any learner, acquisition involves a learning step and a production step.

In the **learning step**, learner $n+1$ sees (x_n, y_n) and computes a posterior distribution over possible languages h_{n+1}. Bayes' rule implies that we can express this posterior distribution as follows:

$$P(h_{n+1}|x_n, y_n) = \frac{P(y_n|x_n, h_{n+1})P(h_{n+1}|x_n)}{\sum_{h \in \mathcal{H}} P(y_n|x_n, h)P(h|x_n)} \qquad (1)$$

In their derivation GK assume that each language h makes no assumption about which events x are more likely than any other; given that assumption, they note that $P(h|x) = P(h)$, and proceed with a version of Equation 1 based on that modification. Alternatively, however, it might be that the language carries with it certain assumptions about what events are possible or probable in the world, in which case the GK assumption is untenable.[1] In other words, simply observing meaningful events x may bias the learner to prefer some languages over others. If this is the case, then $P(h|x)$ does not equal $P(h)$, and the learning step is described by Equation 1.

To see what this shift does to the iterated learning chain, we now turn to the **production step**. In this step, the learner

[1] More formally, GK assume that each language h specifies $P(y|h,x)$, the *conditional* distribution over utterances y given the events x. Our formulation corresponds to assuming that each language maps onto a *joint* (subjective) probability distribution over events and utterances, $P(x,y|h)$. We can factorize the joint distribution $P(x,y|h) = P(y|x,h)P(x|h)$. Moreover, since $P(h|x) \propto P(x|h)P(h)$, in our set up $P(h|x) \neq P(h)$.

encounters new meanings x_{n+1}, generated from the (objective) distribution $Q(x)$ of meanings in the world. Given these meanings, the learner generates the new utterances y_{n+1} by sampling them from $P(y_{n+1}|x_{n+1},h_{n+1})$, where h_{n+1} is the learner's language (assumed to be sampled from the posterior distribution in Equation 1).

Since all people in the chain follow the same learning and production steps, we can calculate $P(h_{n+1}|h_n)$, the probability that learner $n+1$ acquires language h_{n+1} given that the previous learner used the language h_n, in the following way:

$$P(h_{n+1}|h_n) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P(h_{n+1}|x,y)P(y|x,h_n)Q(x). \quad (2)$$

Thus we have a sequence of random variables $h_1, h_2, h_3, \ldots$ describing the languages acquired by each person in the chain. This is generated by a Markov chain whose transition probabilities are given by $P(h_{n+1}|h_n)$. Assuming the chain is ergodic, then its stationary distribution $\pi(h)$ satisfies

$$\pi(h_{n+1}) = \sum_{h_n\in\mathcal{H}} P(h_{n+1}|h_n)\pi(h_n) \quad (3)$$

for all h_{n+1}. Put another way, the probability distribution over languages h_n approaches $\pi(h_n)$ as $n \to \infty$.

In the set up used by GK, the stationary distribution $\pi(h)$ corresponds to the prior $P(h)$. However, under our formalization this is no longer the case. To find the stationary distribution in this situation, we make the following "representativeness" assumption: that the posterior probability of a hypothesis given an actual dataset x is close to its expected posterior probability given the generating distribution $Q(x)$. In other words, we assume that $P(h|x) \approx E_{Q(x')}[P(h|x')] = \sum_{x'} P(h|x')Q(x')$, for some $x \sim Q(x)$. The math demonstrates that if this assumption holds, then the stationary distribution is approximately $\pi(h) = \sum_x P(h|x)Q(x)$. That is, the chain converges to the *expected posterior distribution* over languages given meaningful events in the world. This is because for $\pi(h) = \sum_x P(h|x)Q(x)$ to be the stationary distribution it must be true that:

$$\begin{aligned}
\pi(h_{n+1}) &= \sum_{h_n} P(h_{n+1}|h_n)\pi(h_n) \\
&= \sum_x \sum_y \sum_{h_n} P(h_{n+1}|x,y)P(y|x,h_n)Q(x)\pi(h_n) \\
&= \sum_x \sum_y \sum_{h_n} P(h_{n+1}|x,y)P(y|x,h_n)Q(x)\sum_{x'} P(h_n|x')Q(x') \\
&\approx \sum_x \sum_y \sum_{h_n} P(h_{n+1}|x,y)P(y|x,h_n)Q(x)P(h_n|x) \\
&= \sum_x \sum_y P(h_{n+1}|x,y)Q(x)\sum_{h_n} P(y|x,h_n)P(h_n|x) \\
&= \sum_x \sum_y P(h_{n+1}|x,y)Q(x)P(y|x) \\
&= \sum_x Q(x)\sum_y P(h_{n+1}|x,y)P(y|x) \\
&= \sum_x Q(x)P(h_{n+1}|x) \\
&= \pi(h_{n+1})
\end{aligned}$$

The assumption these results depend on is relatively weak: all it requires is that the events or meanings each learner sees be a representative sample from the true generating distribution $Q(x)$. In the limit where no learner sees any data, the stationary distribution converges to the prior, since $P(h|x) = P(h)$ in that situation. But as the amount of data increases, the languages that evolve will depend on the posterior distribution $P(h|x)$ and the distribution of meanings in the world $Q(x)$. Since the posterior depends on both prior and likelihood ($P(h|x) \propto P(h)P(x|h)$), this means that the languages that evolve will be sampled from a distribution depending on which ones are favored *a priori* as well as which ones best capture the meanings in the world. The additional $Q(x)$ term means that the distribution of those meanings matters as well. These results suggest that languages like h_c might be more likely to evolve in a world like the one in Figure 2(b) than the prior distribution over languages might suggest.

In the next section we report experimental results supporting these theoretical findings.

Experiment

Method

We adopt the standard iterated learning paradigm, in which participants form chains in which the output of the nth participant is the input of participant $n+1$ and the input for the first participant is random. In a training phase, participants see a number of meaning-word pairs and are asked to learn them. In a test phase, they are shown meanings and asked to produce the corresponding word; these are the pairings for the next participant and correspond to the "language" that exists at that point in the chain. Our question is whether the languages that evolve over the course of a chain depend on the distribution of meanings $Q(x)$.

In our experiments, the "meanings" consisted of 36 possible squares differing in size and color, as shown in Figure 3(a). In the CONTROL condition, the stimuli continuously varied along two dimensions: color and size.[2] In this condition there is no obvious or privileged way of categorizing the stimuli. In the SIZE condition, the stimuli were more discontinuous along the size dimension while in the COLOR condition they were discontinuous along the color dimension.[3]

These conditions, then, correspond to worlds with different event distributions $Q(x)$, and each favors languages that partition the stimuli in different ways, as shown in Figure 3(b). In the SIZE condition one would expect the words to categorize by size, in particular, to correspond to the distinction between smaller (w_1) and larger (w_2) items. Conversely, one would expect the words in the COLOR condition to evolve

[2]Color varied from 0% brightness (black) to 100% brightness (white) in increments of 20%, and size from smallest (10x10) to largest (60x60) in increments of 10.

[3]In particular, stimuli 2 and 5 from the CONTROL condition became 3 and 4, with the new 2 and 5 intermediate in value. Thus in the SIZE condition the final sizes were 10x10, 15x15, 20x20, 50x50, 55x55, and 60x60, and in the COLOR condition the final colors were 0%, 10%, 20%, 80%, 90%, and 100% brightness.

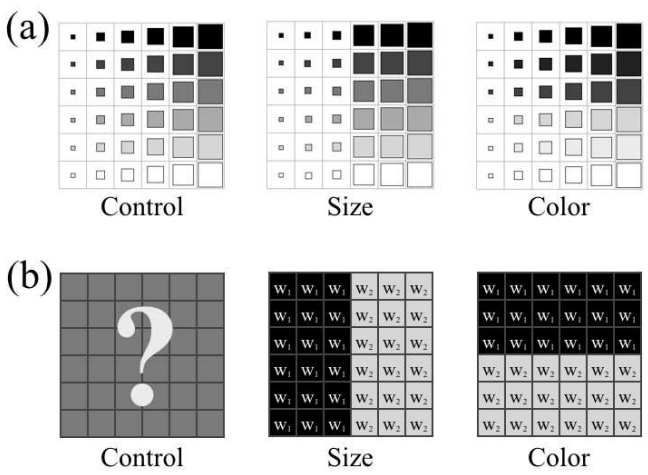

Figure 3: (a) Space of stimuli seen in each of the three conditions of the experiment. Stimuli in the CONTROL condition varied continuously along the dimensions of size and color; in the SIZE condition they varied discontinuously according to size, and in the COLOR condition they varied discontinuously along the color dimension. These different spaces thus impose different event distributions $Q(x)$. (b) Schematic illustration of the predictions about what the evolved language should look like in each condition. In the SIZE condition, the words should evolve to categorize the stimuli according to size, with one word (w_1) applying to the smaller objects and the other (w_2) applying to the larger ones; in the COLOR condition the words should split the space into the dark (w_1) and light (w_2) objects. Predictions for the CONTROL condition are more uncertain, since there are no natural boundaries within this space.

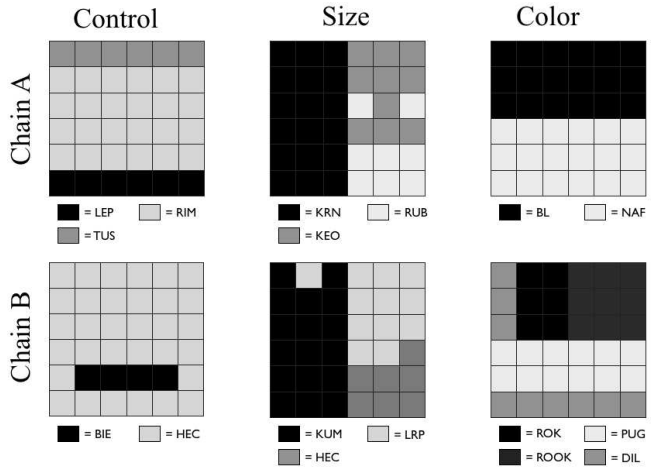

Figure 4: Final languages (at the 20th participant in the chain) in each of the two chains in each of the conditions. It is evident that the structure of the stimulus space has a considerable impact on the structure of the resulting language; both languages in the SIZE condition evolved words that categorized more according to size, both languages in the COLOR condition evolved words that categorized more according to color, and both languages in the CONTROL condition were not strongly driven by either dimension.

to distinguish between darker (w_1) and lighter (w_2) stimuli. Because the CONTROL condition contains stimuli that vary continuously along both dimensions, it is more unclear what the resulting language should look like. If participants have a prior bias to favor one dimension more than another, one might expect the resulting language to have six words, one for each value along the most important dimension; if they do not have any strong prior bias, one might expect languages to vary idiosyncratically, or to evolve towards having one word for all stimuli. Which of these happens is somewhat irrelevant for our purposes; the main goal of running the CONTROL condition was to provide a comparison for the other conditions, and to make apparent any prior biases that might exist.

Our main question was whether the structure of the resulting language would be different in the SIZE and COLOR conditions. We tested this by running two chains of 20 participants in each of the conditions using a methodology based on Kirby et al. (2008). For each participant, stimuli were pseudo-randomly divided into two sets of equal size: the SEEN and UNSEEN sets.[4] Each participant acquired the language in a single session consisting of three rounds, each containing a training and a testing phase, with an optional break in between rounds. In the training phases, participants were shown two randomized exposures to the SEEN set (36 trials in total) in which each stimulus was shown on a computer screen with the corresponding word printed below it. In the testing phases, participants were shown the stimuli and asked to type the corresponding word; they were never given feedback. The testing phases in the first two rounds contained a random half of the SEEN set and a random half of the UNSEEN set (18 trials total). The final round of testing contained the entire stimulus set (i.e., all 36 stimuli).

The first participants in each chain were shown a language consisting of 36 consonant-vowel-consonant (CVC) words randomly assigned to each of the possible 36 possible stimuli. For subsequent participants, the language consisted of the meaning-word pairs given by the previous participant in their final round of testing. We performed no filtering at any stage.

Results

The final languages in the two chains in each condition are shown in Figure 4. It is evident that there was a substantial effect of condition on the structure of the resulting languages; both chains in the SIZE condition evolved words whose primary categorization divided the stimuli by size, and both chains in the COLOR condition evolved words which categorized according to color (although this effect was stronger for Chain A than Chain B).

The difference between conditions can be quantified using the adjusted Rand Index (*adjR*) of Hubert and Arabie (1985). This measure captures the similarity between clusterings; an *adjR* of 1 indicates that the clusters are identical, while 0 is the score one would expect when comparing two random clusterings; scores below 0 indicate that the clusters match less than one would expect by chance. Here, each of the resulting languages corresponds to one "clustering" of the stimuli; for instance, the language in Chain A of the COLOR condition cor-

[4]Stimuli were randomly assigned except for the constraints that there had to be at least 4 stimuli from each quadrant and 1 stimulus from each row and column in the SEEN set.

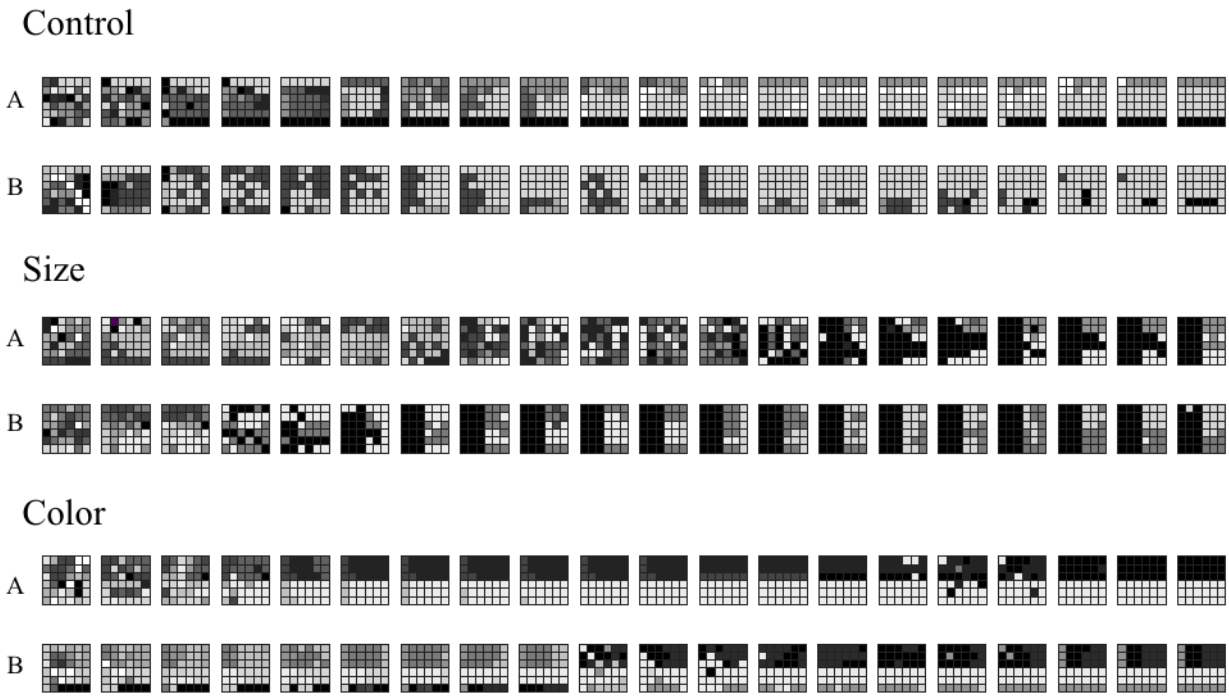

Figure 5: All participants in all of the chains in the iterated learning experiment. Languages in different conditions evolved in different ways, reflecting the different structure of the meaning space across conditions. Different shades indicate different words.

responds to a clustering in which the 18 darkest stimuli are in one cluster and the 18 lightest stimuli are in another. We can compare each of the actual clusterings to the canonical color and size clusterings in Figure 3(b). The results are shown in Table 1. It is evident that the languages in the COLOR condition have a much higher *adjR* when compared to the canonical color clustering, and languages in the SIZE condition have a much higher *adjR* when compared to the canonical size clustering. These results are somewhat preliminary since they incorporate only two chains per condition; nevertheless, they are consistent, and this number of chains is not unusual for iterated learning studies.

	canonical size	canonical color
CONTROL	-0.0204	0.0618
SIZE	0.704	0.079
COLOR	0.065	0.696

Table 1: Average *adjR* values for the final languages in each condition (rows), compared to the canonical clusterings according to size and color (columns). The languages in the CONTROL condition match with both of the canonical sortings no more than they would by chance, but the languages in the other conditions match with their canonical clusterings far above chance.

Our mathematical derivation implies that an iterated learning chain will converge to a *distribution* over languages, not a single language. We therefore examine the languages at each step in the chain, shown in Figure 5. They support the theoretical result: after an initial period in which the number of words decreases dramatically, which is typical for iterated learning experiments, the chains in different conditions stabilize on languages that carve up the meaning space in ways appropriate to the structure of that space in that condition.

Discussion

Our work indicates that if there is no *a priori* assumption that a learner's hypotheses about languages are independent of the world they inhabit, then the languages evolved by Bayesian learners through iterated learning will converge to a distribution that depends on the posterior probability over languages as well as the structure of the meaning space. Here we consider some of the implications and limitations of our findings.

Our results differ significantly from previous results by Griffiths and Kalish (2005, 2007) that suggest that the stationary distribution of a chain of Bayesian iterated learners depends only on their prior. This divergence arises because GK assume that learners' distribution over languages is independent of the structure of the world,[5] whereas we make no such assumption. Which assumption is correct is an open question, although we suggest that in at least some circumstances – especially in the case of semantic categories – ours is plausible. Language learners only start acquiring words after having observed many objects and events in the world, and it seems reasonable for them to expect word meanings to map onto these objects and events in a sensible way. The mapping between grammar and world structure is less obvious, but one might expect that learners' grammatical expectations are affected by their observations of the world (e.g., expecting salient or frequent characteristics, like number or gender,

[5]One might be tempted to just redefine the prior $P(h)$ in GK's results to include the collection of items in the world. However, unless all learners have observed the exact same set of items, their formalism cannot not in fact be interpreted this way, since their proof assumes that all learners share the same prior. Nor is this consistent with how the GK results are usually discussed in the literature.

to be marked grammatically).

It is important to clarify one subtle point that may be confusing. The original Griffiths and Kalish (2007) did identify a dependence on the quantity of data transmitted each generation. However, this is a very *different* dependence than we identify here. When learners sample languages from their posterior, the only effect of increasing quantities of data is to decrease the rate of convergence to the prior; it does not change the actual stationary distribution. They also show that if learners maximize the posterior rather than sample from it, the stationary distribution is centered at the maximum of the posterior. However, this is still different from our results, because there is no role of the structure of meaning space $Q(x)$.

There has been a lot of experimental work supporting the finding that iterated learning experiments reveal human learners' inductive biases (e.g., Kalish, Griffiths, & Lewandowsky, 2007; Griffiths, Christian, & Kalish, 2008; Kirby et al., 2008; Reali & Griffiths, 2009; Smith & Wonnacott, 2010). How do we reconcile our results with this research? First, we do not deny that prior biases are a factor; our results simply suggest that they are not the *only* factor. Second, in all of these experiments, the world never has significant structure: the set of meanings x occur with approximately equal probability. The world structure $Q(x)$ is also never manipulated between conditions: all participants see the same distribution of events.[6] As a result, any effect of world structure may be easy to miss. Our work does not invalidate any of these results, since none of these experiments were made investigate the role of world structure. We do predict that in these experiments, significant changes in the distribution $Q(x)$ should result in different stationary distributions of the chains.

Our findings may also resolve an apparent contradiction in the literature. While many results have suggested that language evolution should converge to the prior, there is also work showing that the structure of the meaning space can also affect the nature of the evolving language (e.g., Kirby, 2001; Brighton & Kirby, 2001; Smith, Kirby, & Brighton, 2003; Maurits, Perfors, & Navarro, 2010). Our result offers an explanation for why such a dependence might exist.

This work is still preliminary. Additional experimental tests of our theoretical predictions include varying the frequency of meanings and initializing chains with languages that do not match the space of meanings (e.g., initializing participants who see the meaning space from the COLOR condition with a language conforming to the canonical size pattern). In addition, a great deal of theoretical work remains. Existing work investigates how GK's results are affected if the chain consists of more than one learner per generation (Smith, 2009; Burkett & Griffiths, 2010), or if learners are capable of "teaching" subsequent learners in the chain (Beppu & Griffiths, 2009). How would our results be affected under these circumstances? There are many remaining open questions in addition to these, but our results indicate that the world may matter more than we previously thought. Perhaps language has the structure it does not just because of our brains, but because of the world as well.

[6]Note that in some experiments, for instance Kalish et al. (2007) and Griffiths et al. (2008), the *mapping* between meanings x and utterances (or utterance equivalents) y is different, at least for the initial person in the chain. However, $Q(x)$ itself is never varied.

Acknowledgments

We thank Natalie May, Tin Yim Chuk, Jia Ong, and Kym McCormick for their help recruiting participants and running the experiment. DJN was supported by an Australian Research Fellowship (ARC grant DP0773794).

References

Beppu, A., & Griffiths, T. L. (2009). Iterated learning and the cultural rachet. In *Proc. 31st CogSci conf.* (p. 2089-2094).

Brighton, H., & Kirby, S. (2001). *Meaning space structure determines the stability of culturally evolved compositional language* (Tech. Rep.). Language Evolution and Computation Research Unit: University of Edinburgh.

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

Burkett, D., & Griffiths, T. L. (2010). Iterated learning of multiple languages from multiple teachers. In *Evolang* (Vol. 8).

Christiansen, M., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences, 31*, 489-558.

Griffiths, T. L., Christian, B., & Kalish, M. (2008). Using category structures to test iterated learning as a method for identifying inductive biases. *Cognitive Science, 31*(1), 68-107.

Griffiths, T. L., & Kalish, M. (2005). A Bayesian view of language evolution by iterated learning. In *Proc. 27th CogSci conf.*

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

Hauser, M., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science, 298*(5598), 1569–1579.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Jn of Classification*, 193–218.

Kalish, M., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psych. Bulletin and Review, 14*(2), 288-294.

Kirby, S. (2001). Spontaneous evolution of linguistic structure – an iterated learning model of the emergence of regularity and irregularity. *IEEE Trans. on Evolutionary Computation, 5*(2), 102-110.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences, 105*(31), 10681-10686.

Kirby, S., & Hurford, J. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi & D. Parisi (Eds.), (p. 121-148). London: Springer Verlag.

Komarova, N., & Nowak, M. (2001). Natural selection of the critical period for language acquisition. *Pr Roy Soc B, 268*, 1189-1196.

Landau, B., Smith, L., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*, 299–321.

Maurits, L., Perfors, A., & Navarro, D. J. (2010). Why are some word orders more common than others? A uniform information density account. In *NIPS* (Vol. 23, p. 1585-1593).

Nowak, M., Komarova, N., & Niyogi, P. (2001). Evolution of universal grammar. *Science, 291*, 114-118.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences, 13*(4), 707-784.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition, 111*, 317-328.

Smith, K. (2009). Iterated learning in populations of Bayesian agents. In *Proc. 31st CogSci conf.* (p. 697-702).

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

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition, 116*, 444-449.

Zuidema, W. (2002). How the poverty of the stimulus solves the poverty of the stimulus. In *NIPS* (Vol. 15).