Language Evolution Can Be Shaped by the Structure of the World

Amy Perfors, Daniel J. Navarro

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 relationship between language 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; Linguistic structure

1. Introduction

Human languages have rich structure on many levels, from phonology to semantics to grammar. Where does this structure come from? Many 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 natural selection operates on genes for language (Komarova & Nowak, 2001; Nowak, Komarova, & Niyogi, 2001; Pinker & Bloom, 1990) or on capabilities then used by the language faculty (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 usable by human brains (Brighton, Smith, & Kirby, 2005; Christiansen & Chater, 2008; Zuidema, 2002). Although these accounts disagree in many particulars, they agree that the structure of language arises from the structure of the brain.

In this study we argue that language evolution can be 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 rests on certain premises: either languages carry no assumptions about the distribution of meanings/events in the world (Griffiths & Kalish, 2005), or if they do, the frequency with which those events are spoken about is entirely dependent on the language, rather than also the distribution of events in the world (Griffiths & Kalish, 2007). Here, we demonstrate that if either premise is abandoned, the iterated learning process converges to a distribution that depends on the distribution of events in the world as well as the prior biases of the learner. These theoretical results are experimentally tested in a laboratory-based iterated learning experiment, which shows that participants converge on different languages depending on the structure of the space of meanings they are shown.

1.1. Iterated learning

The iterated learning modeling framework is widely used in language evolution research (Griffiths & Kalish, 2007; Kirby et al., 2008; Kirby & Hurford, 2002; Reali & Griffiths, 2009a; Smith, 2009). It views the process of language evolution in terms of a chain of Bayesian learners (or generations), which is illustrated schematically in Fig. 1. 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 his or her own data by sampling from his or her posterior distribution; this data then serves as input to the next learner in the chain. Over time, the languages that emerge from this process become non-arbitrary: Two recent articles (both referred to, henceforth, as GK) demonstrate that when the learners are Bayesian, we should expect an iterated learning chain to converge to the prior distribution over all possible languages (Griffiths & Kalish, 2005, 2007). 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, but GK's result implies that neither the structure of the meaning space
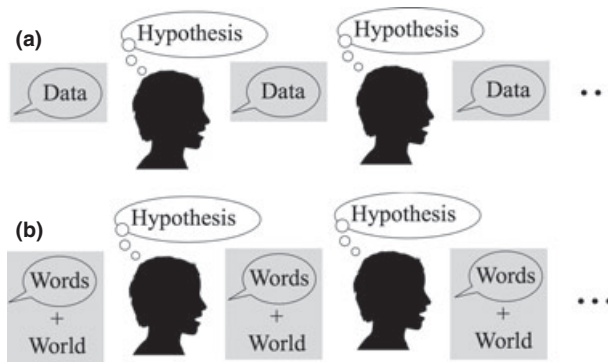
Fig. 1. (a) Schematic illustration of the typical iterated learning paradigm, which assumes that learner *n* acquires language on the basis of the language data produced by learner *n* − 1. (b) A different view of iterated learning recognizes that because individuals produce language to communicate about the world, the data available to learners includes meanings in the world as well as the language produced by the learner before them.

nor the nature of the initial language should have an effect on the language that eventually evolves.

GK's result can be conceptualized 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 (Landau, Smith, & Jones, 1988). This prior bias might impose a distribution over hypotheses *h* about possible languages, like the illustrative one shown in Fig. 2a: 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 less 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, so forth. Although this prior and these precise numbers are imaginary, the picture in Fig. 2a 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 Fig. 2b: 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
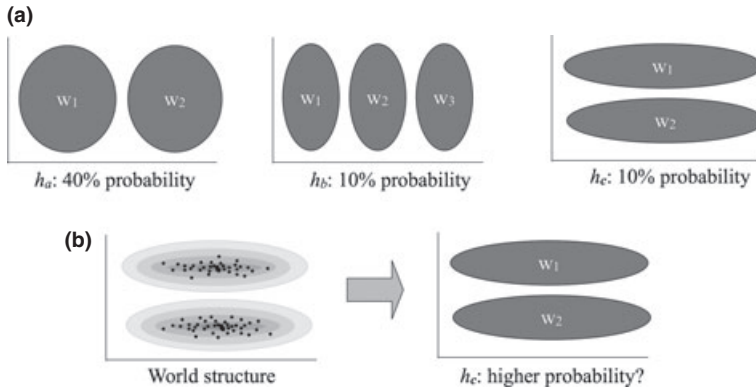
Fig. 2.   (a) Intuitive illustration of the results of Griffiths and Kalish (2005, 2007) (GK). Given a two-dimensional semantic space, a learner with a prior bias to favor one dimension of that space (the *x*-axis) and languages with fewer words might have a prior distribution over languages that puts more probability on h_a and less on h_b or h_c. GK demonstrates that the languages that evolve will converge to this prior distribution. (b) However, 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.

GK derivation and the problem of language evolution within a structured world? In the next section, we argue for the latter.

2. Theoretical result

Our primary result formalizes the learning framework similarly to Griffiths and Kalish (2005), although later we will also address the slightly different formalization of Griffiths and Kalish (2007). For now, we consider a Bayesian learner who 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} based on 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 he or she is 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:

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

In their derivation, Griffiths and Kalish (2005) 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 Eq. 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 Eq. 1.

To see what this shift does to the iterated learning chain, we now turn to the *production step*. In this step, the learner 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 (sampled from the posterior distribution in Eq. 1).

As 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:

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

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 constitutes 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

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

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

In the setup used by Griffiths and Kalish (2005), 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 $x \sim Q(x)$. 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 $Q(x)$.[2] In the limit where no learner sees any data, the stationary distribution converges to the prior, as $P(h|x) = P(h)$ 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 $Q(x)$. As 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 Fig. 2b than the prior distribution over languages might suggest.

The derivation offered by Griffiths and Kalish (2007) removes the premise made by Griffiths and Kalish (2005) that languages h and events x are independent. However, their new proof of convergence to the prior depends on a different premise: When h and x are not independent, then the distribution of events x depends entirely upon the language rather than the world.[3] Indeed, the 2007 derivation makes no reference to an external generating distribution, like $Q(x)$, at all.

Our results taken together with both GK derivations lead to the following conclusion: Language evolution should by affected by the structure of the world as long as (a) different languages make different predictions about the distribution of events, and (b)

the actual distribution of which events are spoken about *does* depend, at least in part, on the nature of the external world. This second premise, put another way, simply states that language is shaped by the structure of the world as long as which things people talk about (and how often they talk about them) is affected by the distribution of those things in the external world, not just the language the people speak.

One key difference between our result and both GK results is that we predict that the amount of data that is transmitted from one speaker to the next (i.e., the information bottleneck) actually changes the stationary distribution of the chain. This contrasts with the prediction made by the GK derivations that although smaller bottlenecks lead to faster convergence, the language will always be a sample from the prior distribution over languages regardless of bottleneck size.[4] Our results, however, predict that the nature of the emergent languages will change as a function of the size of the bottleneck. When it is large or non-existent (i.e., there is a great deal of data available to each learner), the stationary distribution will converge to the expected posterior distribution over languages given meaningful events in the world; when the bottleneck is small, it will converge to the prior. This provides a natural test of our prediction: When the bottleneck is small, do the resulting languages resemble the ones that emerge when it is not?

In the next section we report experimental results supporting our theoretical findings. When there is a great deal of data available to the learner at each generation, the languages that emerge from an iterated learning chain with humans are shaped by the structure of the world. However, when there is little data, they look substantially different and appear unaffected by the structure of the world. This is consistent with the predictions of our theoretical analysis but not that of GK.

3. 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 $Q(x)$, and whether the predicted dependence on bottleneck size is observed.

3.1. Participants

A total of 225 participants from the University of Adelaide participated in what they were told was an "alien language" study in which they were asked to learn a language consisting of labels for visual stimuli (squares of different sizes). Participants consisted of undergraduates from all departments as well as some recruited from a paid participant

pool drawing from the surrounding community. The experiment was conducted at a computer terminal and the instructions did not reveal that it was an iterated learning experiment; participants did not know that their answers were to be given to the next participant.

3.2. Stimuli

The "meanings" in our experiments consisted of 36 possible squares differing in size and color, as shown in Fig. 3a. In the CONTROL condition, color varied continuously from 0% brightness (black) to 100% brightness (white) in increments of 20%, and size from smallest (10×10) to largest (60×60) in increments of 10 pixels. 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: The sizes were 10×10, 15×15, 20×20, 50×50, 55×55, and 60×60. In the COLOR condition they were discontinuous along the color dimension: The colors were 0%, 10%, 20%, 80%, 90%, and 100% brightness.
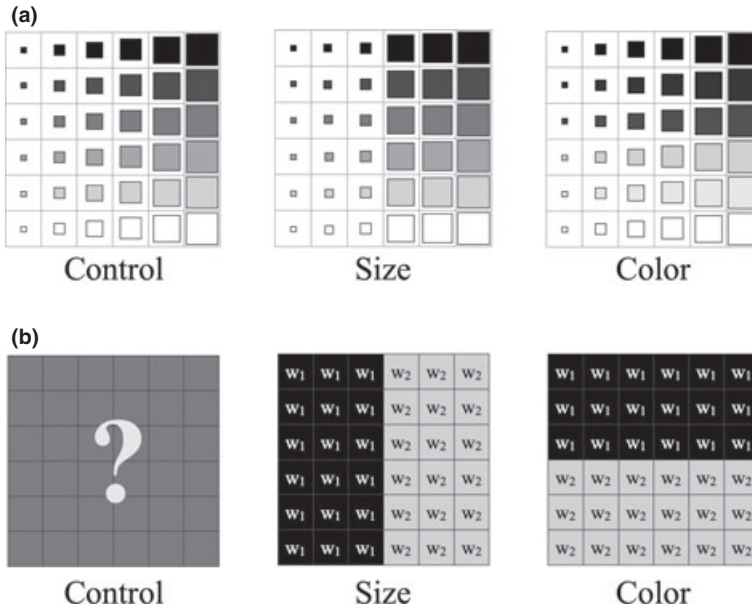

Fig. 3. (a) Space of stimuli seen in each of the three conditions of the experiment. Stimuli in the CONTROL condition varied continuously along the dimensions of size and color; in the SIZE condition they varied discontinuously according to size, and in the COLOR condition they varied discontinuously 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, as there are no natural boundaries within this space.

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 Fig. 3b. 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 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 toward 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.

3.3. Procedure

Our main question was whether the structure of the resulting language would be different in the SIZE and COLOR conditions. We also wanted to test the prediction about bottleneck size made by our theoretical analysis. This was accomplished by additionally running each of the three conditions through either a small or large bottleneck, resulting in a 2 × 3 design. There were three chains of participants in each of the six conditions; Because convergence is faster when the bottleneck is small, the SMALL bottleneck group had five participants per chain, whereas the LARGE bottleneck group had 20.

Following Kirby et al. (2008), in all conditions, stimuli were pseudo-randomly divided into a *seen* and an *unseen* set for each participant. The two bottleneck conditions differed only according to how many items were in the *seen* set: Participants in the LARGE bottleneck group were presented with half of the entire dataset (i.e., 18 meanings), whereas those in the SMALL bottleneck group saw only four.

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 in which each stimulus was shown on a computer screen with the corresponding word printed below it. Participants could see the next item by pressing the NEXT button. 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 an equal-sized number of items randomly sampled from the *unseen* set. The final round of testing in all conditions contained the entire stimulus set (i.e., all 36 stimuli).

For the first participants in each chain, the *seen* set consisted of a random subset of a full 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. With each new person in the chain, different items were randomly allocated into the *seen* and *unseen* sets. We performed no filtering on the data, not even to remove typos or ensure that the words followed a CVC pattern.

4. Results

One prediction made by both our analysis and GK's is that convergence should be more rapid when there is less data available to the learner, as in the SMALL bottleneck group. This was borne out in our experiment, as shown in Fig. 4. We measured convergence, as in Kirby et al. (2008), by calculating the Levenshtein distance (Levenshtein, 1966) between adjacent pairs of participants in each chain at generation g. It is given by:

$$E(g) = \frac{1}{36} \sum_{m=1}^{36} \text{LD}(y_g^m, y_{g-1}^m), \tag{4}$$

where y_g^m is the word associated with meaning m by the participant at generation g, $\text{LD}(y_g^m, y_{g-1}^m)$ is the normalized Levenshtein distance between words y_g and y_{g-1}, and 36 is the number of total meanings in our experiment. Levenshtein distance is the number of edits required to turn one string into another (for instance, the distance between *bik* and *bok* is 1), and therefore it captures the transmission error from generation to generation.

It is clear from Fig. 4 that in all conditions, the chains ultimately converged to a low error rate. Convergence was slower in the LARGE bottleneck group, as predicted.
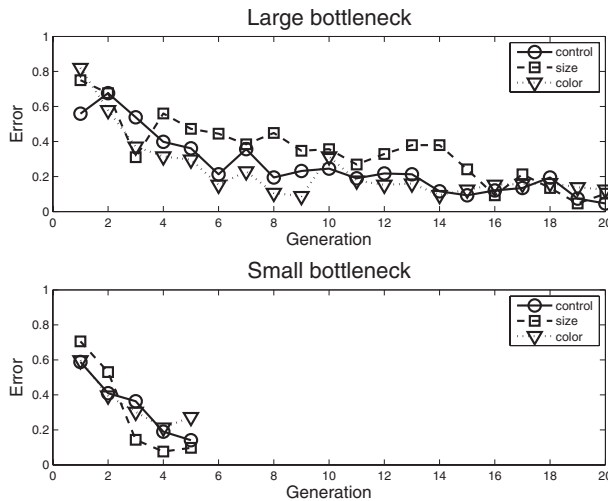

Fig. 4. Transmission error by generation in each condition. As predicted, languages evolve to be increasingly learnable (decreasing transmission error in later generations), and convergence is faster when the bottleneck is smaller. Note that the slightly lower convergence in the COLOR condition with the SMALL bottleneck is entirely carried by the final participant in one of the chains (Chain A).

Our main prediction was that the structure of the world would affect the languages when the bottleneck was LARGE but not SMALL, as in the former case the data would exist in sufficiently high quantity that the stationary distribution would be closer to the expected posterior distribution, which includes the distribution of events in the world $Q(x)$. The results, shown in Fig. 5, are consistent with this prediction. In the LARGE bottleneck groups it is evident that there was a substantial effect of the structure of the meaning space on the structure of the resulting languages; all chains in the SIZE condition evolved words whose primary categorization divided the stimuli by size, and all chains in the COLOR condition evolved words which categorized according to color (although this effect was stronger for some chains than others). By contrast, languages in the SMALL bottleneck groups tended to evolve toward having only one word, and there are no observable effects of world structure.

These differences can be quantified using the adjusted Rand Index or adjR (Hubert & Arabie, 1985). This measure captures the similarity between clusterings; an adjR of 1 indicates that the clusters are identical, whereas 0 is the score one would expect when
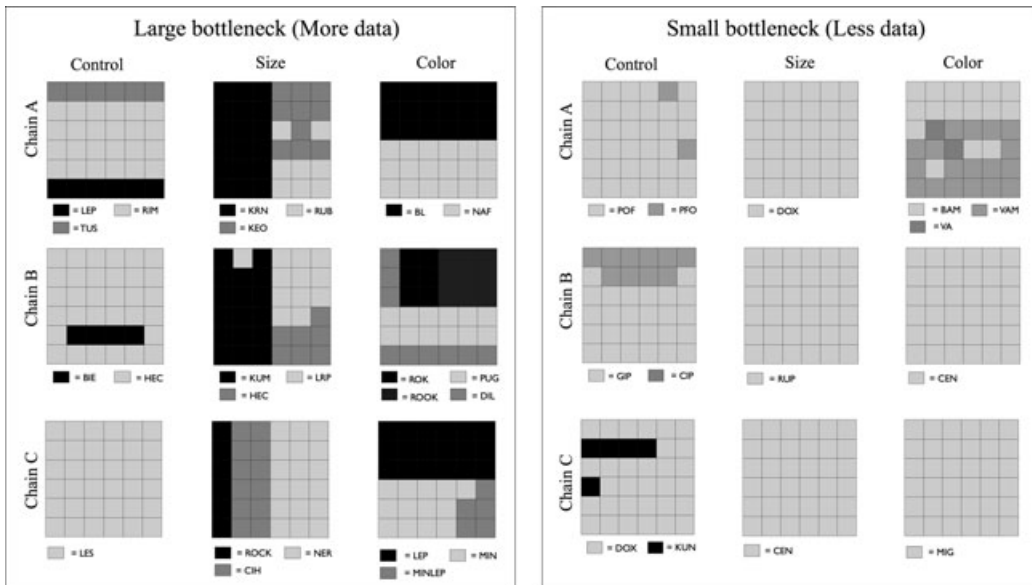

Fig. 5. Languages from the final participant in each of the chains in the LARGE and SMALL bottleneck groups. Each square is laid out as in Fig. 2, where the *x* axis is size and the *y* axis is color. Items labeled with the same word are depicted with the same shade; words that are very similar but not identical are depicted with similar shades. Specific words are detailed below each square. It is evident that in the LARGE bottleneck group, as predicted, the stimulus space has a considerable impact on the structure of the resulting language. All languages in the SIZE condition evolved words that categorized more according to size, all languages in the COLOR condition categorized more according to color, and all languages in the CONTROL condition were not strongly driven by either dimension. Also as predicted, in the SMALL bottleneck group all languages evolved similarly; there was not enough data at each generation for the structure of the meaning space to affect how the languages evolved.

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 in the LARGE bottleneck group corresponds 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 Fig. 3b. The results are shown in Fig. 6. When the bottleneck is LARGE, the languages in the COLOR condition have a much higher adjR when compared with the canonical color clustering, and languages in the SIZE condition have a much higher adjR when compared with the canonical size clustering. However, in the SMALL bottleneck groups, none of the languages is consistent with either canonical clustering.

To determine if these results are significant, for each condition and bottleneck group we calculated a *structural precision score*, which captures the degree to which the evolved languages in each condition differentially match either the canonical color or canonical size clusterings. This score is given by the absolute value of the difference between the adjR value of that language for the canonical color clustering and its adjR value for the canonical size clustering.[5] If the structural precision score is close to 0, it indicates that the evolved language fits with both the canonical color and canonical size clusterings equally well—in other words, that it has not evolved specifically to match either the size or color cluster. A higher structural precision score indicates that the language matches one of the two clusterings better than the other—in other words, that it has evolved to fit one "world" more than the other. If the structure of the world did have a significant role on the structure of the language, we would therefore expect that the structural precision score would be significantly higher in the COLOR and SIZE conditions (where the adjR values capturing the consistency of the language to the canonical size and canonical color should be markedly different from each other) than in the CONTROL condition (where those adjR values should be similar). Moreover, we should expect to
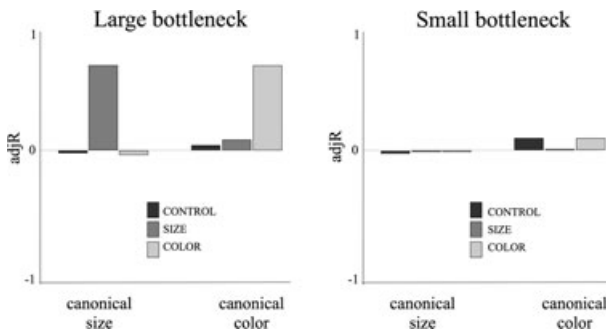

Fig. 6. Average adjR values for the final languages in each condition, compared with the canonical clusterings according to size and color: An adjR of 0 indicates that they match no more than would be expected by chance, whereas an adjR of 1 indicates perfect alignment. In the LARGE BOTTLENECK condition, the languages in the SIZE and COLOR conditions match their respective canonical clusterings far above chance; all other languages do not match either clustering. Error bars reflect standard error.

see this only in the SMALL bottleneck group, where the effect of world structure should be discernible.

As expected, the mean values of the structural precision score are larger in the SIZE and COLOR conditions than in the CONTROL conditions when the bottleneck is large (SIZE: 0.685, COLOR: 0.670, and CONTROL: 0.055) but not when the bottleneck is small (SIZE: 0.0, COLOR: 0.117, and CONTROL: 0.109). A two-way ANOVA on bottleneck × condition reveals a significant main effect of bottleneck type ($F(1,12) = 15.988$, $p = .002$) and a significant interaction ($F(2,12) = 5.333$, $p = .022$). The nature of the interaction is evident in planned comparisons, which show that in the LARGE bottleneck group there is a significant effect of condition ($F(2,6) = 6.155$, $p = .035$), but in the SMALL bottleneck group there is not ($F(2,6) = 0.528$, $p = .615$). Together, these analyses indicate that, as predicted, the structure of the world has a significant effect on the structure of the language when the bottleneck is LARGE but not when it is SMALL.

5. Discussion

Our work indicates that if there is no a priori assumption that a learner's hypotheses about languages are independent of the external world the learner inhabits, and the external world *does* actually affect which events or meanings are spoken of, 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 external structure of the meaning space. Here, we consider the implications and limitations of our findings.

These 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.[6] This divergence arises because GK either assume that languages make no assumptions about the distribution of events in the world (2005) or that, if they do, the distribution of events spoken about is shaped entirely by the language and not at all by the external world (2007). Note that one cannot simply interpret GK's results in such a way that their sense of a prior itself includes these additional external environmental influences. Interpreting the prior in such a way that it includes the structure of the world strips the GK result of much of its theoretical relevance: If one were to do this, it would be impossible to distinguish between the effects of the cognitive biases of the learner and the effects of the environment in which the learner operates.

As the difference between our results and GK's stems from different underlying premises about the relationship between the structure of the world and the nature of language, a natural question is which set of premises are correct. This is an open question, although we suggest that in at least some circumstances ours are plausible. First, the premise that the distribution of events x that are talked about is shaped at least in part by the external structure of the world—and not just by language—seems eminently reasonable: We talk about cats not just because our language has the word "cat" but also because cats actually do exist, and we have direct experience of them. Second, the premise that languages make

assumptions about the distribution of events in the world also appears sensible, at least sometimes. 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—for instance, expecting salient or frequent characteristics, like number or gender, to be marked grammatically, as has been suggested earlier (e.g., Bybee, 2000; Du Bois, 1987; Evans, 2003). Further exploring in what circumstances this premise is correct is a subject for future research.

Our experimental results further support our theoretical predictions by showing that when people are exposed to different worlds with different distributions of events, the language that emerges reflects that event structure—even when the people, as in most iterated learning experiments, do not realize that they are in the middle of "evolving" a language at all. It is important to note here that the experiments do *not* indicate, nor did we ever predict, that people's prior biases would play no role in shaping the structure of the language; in fact, it is quite likely that cognitive factors like memory are a major reason that languages in all conditions evolved to have fairly few words. Rather, the implication of the experiments is that world structure plays an *additional* role on top of the effect of whatever prior biases exist. This was evident in the fact that the difference between the evolved languages in SIZE, COLOR, and CONTROL conditions can only be explained by the differences in the distribution of events $Q(x)$ in those conditions, as the participants were sampled from the same underlying population and thus presumably all shared similar prior biases.

There has been a great deal of experimental work supporting the finding that iterated learning experiments reveal human learners' inductive biases (Griffiths, Christian, & Kalish, 2008; Kalish, Griffiths, & Lewandowsky, 2007; Kirby et al., 2008; Reali & Griffiths, 2009a; 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 in certain circumstances they are not the *only* factor. Second, our results do not invalidate the broad research program of using iterated learning experiments to illuminate people's inductive biases. Many experiments are constructed to either not have events be generated by the external world (i.e., have no $Q(x)$ distribution) or to have all participants see the same distribution of events (i.e., they do not manipulate $Q(x)$ between conditions). We do predict that significant changes in the distribution $Q(x)$ should result in different stationary distributions of the chains, although how large that effect may be in different kinds of experiments is an open empirical question.

Our findings may also resolve an apparent contradiction in the literature. While some theoretical results suggest that language evolution should converge to the prior, there is also theoretical work showing that the structure of the meaning space can affect the nature of the evolving language (Brighton & Kirby, 2001; Kirby, 2001; Maurits, Perfors, & Navarro, 2010; Smith, Jones, Yoshida, & Colunga, 2003). In addition, many empirical cross-linguistic results are consistent with the idea that aspects of world structure can affect linguistic structure. Communicative complexity has been argued to increase with

social complexity, both in humans and non-humans (Freeberg, Dunbar, & Ord, 2012; McComb & Semple, 2005), whereas phonological and morphological complexity may be negatively related to community size or isolation (Lupyan & Dale, 2010; Trudgill, 2009). Others have suggested that environmental factors like frequency of use may also affect language structure (Bybee, 2000; Evans, 2003). Although the link between some of these factors and the distribution $Q(x)$ is not always exact, our work may help explain how these findings are possible in light of the GK results.

Further tests of our theoretical predictions will include additional cross-linguistic work as well as experimental, laboratory-based research—for instance, 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 (Burkett & Griffiths, 2010; Smith, 2009), or if learners are capable of "teaching" subsequent learners in the chain (Beppu & Griffiths, 2009). How would our results be affected under these circumstances? Other work has built on the GK result to show that iterated learning with Bayesian learners is equivalent to the Wright–Fisher model of genetic drift (Reali & Griffiths, 2009b); does iterated learning correspond to a different evolutionary model when the stationary distribution it converges to is different?

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.

Acknowledgments

Notes

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, as $P(h|x) \propto P(x|h)P(h)$, in our set up $P(h|x) \neq P(h)$.

2. Note that this is actually a specific case of a more general result: languages may be shaped by world structure even in situations where the meanings are not a representative sample from $Q(x)$. We focus on the special case when representativeness holds because it does hold in a lot of situations, and the stationary distribution is more interpretable in this case. Appendix A presents the more general proof.

3. More formally, this derivation assumes that data d are generated from $P(d|h)$ and that h are sampled from $P(h|d)$, for *any* construal of d. Thus, d can include events x as well as utterances y. Note that if $d = (x, y)$, however, the events and utterances are entirely generated by $P(d|h) = P(x, y|h)$; that is, an external event-generating distribution like $Q(x)$ plays no role.

4. The GK results *do* predict that there will be an effect of bottleneck size if one assumes that learners are selecting the hypothesis with the maximum posterior probability (MAP estimation) rather than sampling directly from the posterior distribution (Griffiths & Kalish, 2007). However, that bottleneck effect is different than the one we predict; it suggests, unlike us, that the stationary distribution will still always be centered on the prior, implying that the bottleneck size does not influence which language is most likely. The difference is that under MAP estimation, the variance of that distribution depends on the bottleneck size.

5. The reason we calculate the statistics based on this score, rather than on something more intuitive like the adjR values themselves, is that the adjR values are not all independent of each other; for example, if a language is highly consistent with the canonical size clustering, then it will not be consistent with the canonical color clustering. The structural precision score, by combining both values into one measure, ensures that all the values entering into the statistical test are independent of each other.

6. Note that they are also distinct from (although related to) the results of Beppu and Griffiths (2009), which reveal a different way that the stationary distribution of iterated learning can change when learners are provided with data generated by the external world. In that work, which is not focused on language, the hypotheses to be learned are hypotheses about the nature of the external distribution x directly, rather than mappings between x and y (as occurs in language). Like us, Beppu and Griffiths (2009) find that the stationary distribution no longer converges to the prior, although the precise nature of the stationary distribution is not the same as we find here.

References

Beppu, A., & Griffiths, T. L. (2009). Iterated learning and the cultural rachet. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2089–2094). Austin, TX: Cognitive Science Society.

Brighton, H., & Kirby, S. (2001). *Meaning space structure determines the stability of culturally evolved compositional language (Tech. Rep.)*. Edinburgh, Scotland: 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 Proceedings of Evolang* (Vol. *8*), Utrecht, the Netherlands.

Bybee, J. (2000). Lexicalization of sound change and alternating environments. In M. Broe & J. Pierrehumbert (Eds.), *Papers in laboratory phonology v. acquisition and the lexicon* (pp. 250–268). Cambridge, UK: Cambridge University Press.

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

Du Bois, J. (1987). The discourse basis of ergativity. *Language*, *63*, 805–855.

Evans, N. (2003). Context, culture, and structuration in the languages of Australia. *Annual Reviews in Anthropology*, *32*, 13–40.

Freeberg, T., Dunbar, R., & Ord, T. (2012). Social complexity as a proximate and ultimate factor in communicative complexity. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*, 1785–1801.

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 B. G. Bara, L. W. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 827–832). Austin, TX: Cognitive Science Society.

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. *Journal of Classification*, *2*, 193–218.

Kalish, M., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic 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 Transactions 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 of the United States of America*, *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.), *Simulating the evolution of language* (pp. 121–148). London: Springer Verlag.

Komarova, N., & Nowak, M. (2001). Natural selection of the critical period for language acquisition. *Proceedings of the Royal Society of London B: Biological Sciences*, *268*, 1189–1196.

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

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *6*, 707–710.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE*, *5*(1), e8559.

Maurits, L., Perfors, A., & Navarro, D. J. (2010). Why are some word orders more common than others? A uniform information density account. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel & A. Culotta (Eds.), *Neural information processing systems* (Vol. *23*) (pp. 1585–1593). Cambridge, MA: MIT Press.

McComb, K., & Semple, S. (2005). Coevolution of vocal communication and sociality in primates. *Biology Letters*, *1*, 381–385.

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. *The Behavioral and Brain Sciences*, *13*(4), 707–784.

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

Reali, F., & Griffiths, T. L. (2009b). Words as alleles: Connecting language evolution with Bayesian learners to models of genetic drift. *Proceedings of the Royal Society of London B: Biological Sciences*, *277*(1680), 429–436.

Smith, K. (2009). Iterated learning in populations of Bayesian agents. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 697–702). Austin, TX: Cognitive Science Society.

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

Smith, L., Jones, S., Yoshida, H., & Colunga, E. (2003). Whose DAM account? Attentional learning explains Booth and Waxman. *Cognition*, *87*, 209–213.

Trudgill, P. (2009). Typology and sociolinguistics: Linguistic structure, social structure and explanatory comparative dialectology. *Folia Linguistica*, *31*, 349–360.

Zuidema, W. (2002). How the poverty of the stimulus solves the poverty of the stimulus. In S. Becker, S. Thrun & K. Obermayer (Eds.), *Neural information processing systems* (Vol. *15*) (pp. 43–50). Cambridge, MA: MIT Press.

Appendix A

The derivation presented in the theoretical results is actually a special case of a more general derivation (we thank Tom Griffiths for a discussion of this). This more general derivation diverges from the one in the main article at step 4 below:

$$
\begin{aligned}
\pi(h_{n+1}) &= \sum_{h_n} P(h_{n+1}|h_n)\pi(h_n) \\
&= \sum_{h_n}\sum_{x}\sum_{y} P(h_{n+1}|x,y)P(y|x,h_n)Q(x)\pi(h_n) \\
&= \sum_{h_n}\sum_{x}\sum_{y} P(h_{n+1}|x,y)P(y|x,h_n)Q(x)\sum_{x'} P(h_n|x')Q(x') \\
&= \sum_{h_n}\sum_{x}\sum_{y} \frac{P(h_{n+1}|x)P(y|x,h_{n+1})}{\sum_{h'} P(h'|x)P(y|h',x)} P(y|x,h_n)Q(x)\sum_{x'} P(h_n|x')Q(x') \\
&= \sum_{x} P(h_{n+1}|x)Q(x)\left[\sum_{y} P(y|h_{n+1},x)\frac{\sum_{h_n}\left(\sum_{x'} P(h_n|x')Q(x')\right)P(y|h_n,x)}{\sum_{h'} P(h'|x)P(y|h',x)}\right]
\end{aligned}
$$

When the bracketed term is equal to 1, then $\pi(h_{n+1}) = \sum_{x} P(h_{n+1}|x)Q(x)$ (or, equivalently $\pi(h_n) = \sum_{x} P(h_n|x)Q(x)$). One circumstance when this occurs is, as in the main article, when $P(h|x) \approx \sum_{x'} P(h|x')Q(x')$:

$$\pi(h_{n+1}) = \sum_x P(h_{n+1}|x)Q(x)\left[\sum_y P(y|h_{n+1},x)\frac{\sum_{h_n}\left(\sum_{x'} P(h_n|x')Q(x')\right)P(y|h_n,x)}{\sum_{h'} P(h'|x)P(y|h',x)}\right]$$

$$= \sum_x P(h_{n+1}|x)Q(x)\left[\sum_y P(y|h_{n+1},x)\frac{\sum_{h_n} P(h_n|x)P(y|h_n,x)}{\sum_{h'} P(h'|x)P(y|h',x)}\right]$$

$$= \sum_x P(h_{n+1}|x)Q(x)$$

There are presumably other circumstances under which the bracketed term is equal to 1, but they are much less straightforward and harder to offer an intuitive interpretation of. Because of this, and because the general point of this study is simply that language evolution *can* be shaped by the structure of the world—a point which is strengthened rather than weakened by this more general derivation—our interpretation in the main article focuses on the version presented there.