Iterated Learning

Bas Cornelissen

August 9, 2017

Could it be that structure in language emerges because it is transmitted from one generation to the next? Is cultural *transmission* the force shaping language? Early models of iterated learning suggested precisely that. Bayesian models improved the early work by separating the biases of the learners from the effects of transmission. But they also indicated that cultural evolution only allows the prior biases to surface, a result that sparked a small controversy. The 'convergence to the prior' was shown to break down in more complicated populations, again creating room for the shaping force of cultural evolution. This chapter introduces the iterated learning tradition and ends with a list of desiderata for models of cultural language evolution. The list serves as a guide to the remainder of this thesis.

Early iterated learning models

In the early years of this century, James Hurford, Simon Kirby, Kenny Smith and others, developed the idea that cultural transmission, in the form of *iterated learning* (IL), could be the source of structure in language. Early work in this tradition tried to isolate a "minimal set of assumptions and hypotheses with which linguistic structure can be explained" (Brighton 2002). The result was a simple model of cultural transmission between generations consisting of a single agent each. In the model, language alternates between an internal representation (I-language in Chomskyan parlance; Chomsky 1986, pp. 19–24) or an external representation in the form of actual utterances (E-language), as figure 1 illustrates. The first agent (the parent) is presented with several objects for which it produces some utterances. Those utterances form the primary linguistic data from which the second agent (the child) has to learn a language.[1] The child goes on to become the parent of the next generation, forms expressions for several (other) objects, which are observed by the next agent, and so on.

Every generation learns a language by observing the language of the previous generation, who themselves learned it from the generation before them. The target of learning is therefore the outcome of the same learning process and this gives rise an evolutionary dynamics on the cultural level: the fact that a language has to be learned over again

1 The utterances alone are not enough, unless you assume the child can mind-read. Instead meaning-signal pairs are often communicated.

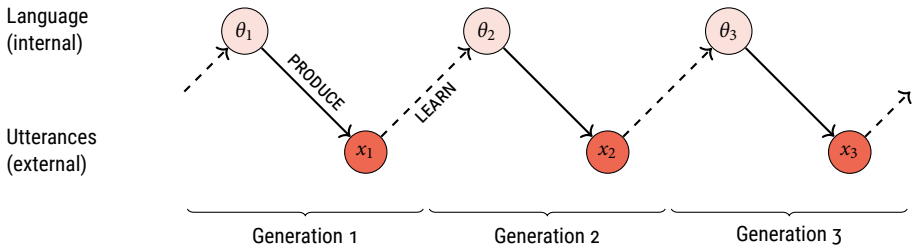

FIGURE 1 In the iterated learning model, the language produced by the previous generation serves as the primary linguistic data for the next.

Adapted from Kirby (2001).

shapes the language itself to become better learnable, hence better transmissible. And key to better transmission, many studies suggested, was the acquisition of some form of systematicity. Paraphrasing Hurford (2000), language appeared to be structured, because cultural transmission favours systematicity.

THE EMERGENCE OF COMPOSITIONALITY, I This conclusion was primarily based on computer simulations of the emergence of compositionality which I briefly want to discuss. Suppose, following Brighton (2002), that agents are positioned in an environment with various objects. The objects have F possible features, each taking V values, and thus correspond to points in a F-dimensional meaning space $\mathcal{M}$. The features might be color and shape, taking values triangular, rectangular or circular and orange, blue and black respectively. A language associates meanings $m \in \mathcal{M}$ to signals s in a space $\mathcal{S}$ of signals, typically strings over some alphabet. Certain languages are compositional, meaning that the signals can be decomposed in subsignals that each bear one aspect of the meaning. Compositional languages should be distinguished from *holistic* languages where meanings correspond to a signals without there being any underlying regularity.

Consider the following language with alphabet $\{t, r, c,\ o, b, k\}$:

$$
\begin{array}{lll}
(\triangle, \color{#e8604c}{\bullet}) \mapsto to, & (\triangle, \color{#6ab0d4}{\bullet}) \mapsto tb & (\triangle, \bullet) \mapsto tk \\
(\square, \color{#e8604c}{\bullet}) \mapsto so, & (\square, \color{#6ab0d4}{\bullet}) \mapsto sb & (\square, \bullet) \mapsto sk \\
(\bigcirc, \color{#e8604c}{\bullet}) \mapsto co, & (\bigcirc, \color{#6ab0d4}{\bullet}) \mapsto cb & (\bigcirc, \bullet) \mapsto ck
\end{array}
$$

This language is clearly compositional, since the first subsignal indicates the shape, (*t*riangle, *r*ectangle, *c*ircle) and the second subsignal the color (*o*range, *b*lue, blac*k*). In fact, that description is much more efficient:

$$
\begin{array}{lll}
\triangle \mapsto t, & \square \mapsto s, & \bigcirc \mapsto c \\
\color{#e8604c}{\bullet} \mapsto o, & \color{#6ab0d4}{\bullet} \mapsto b & \bullet \mapsto k
\end{array}
$$

Rather than listing the signals corresponding to each of the $V^F = 3^3$ meanings (the worst case scenario for a holistic language), a compositional languages can be *compressed* to $F \cdot V = 2 \cdot 3$ rules listing to which *sub*signal every feature maps. That also means that one can faithfully reconstruct a compositional language from $F \cdot V$ signals, whereas it would need to observe *all* signals to reconstruct a holistic language (in the worst case). A compositional language is, in short, more *compressible* and as a result better *transmissible*.

TRANSMISSION BOTTLENECKS AND GENERALISATION In reality, children do not observe their entire language (e.g. all English sentences), but only a subset of it. They face a *transmission bottleneck*[2] better known as the *poverty of the stimulus*. If there is no such bottleneck all languages can be transmitted in their entirety, and faithfully so. The language can consequently not be changed by transmission and the initial language marks a *steady state*, maintained throughout all future generations. In the presence of a bottleneck, however, the learner is forced to *generalize* the observed data to a full language, in which case systematic errors can slowly accumulate.

The exact generalisation mechanism can take many different forms, such as (heuristic) grammar induction (Kirby 2001; Zuidema 2003), training a neural network (Kirby and Hurford 2002; Smith 2002) or constructing a finite state transducer (Brighton 2002). All these mechanisms try to discern some structure (e.g. compositionality) in the language. Sometimes, that allows the child to produce signals for unobserved meanings. But in other cases, the child is forced to invent a new signal. Incidentally, the new signal introduces a structure previously absent in the language. The next generation is then more likely to infer a language that reproduces that structure. As time passes, the differences between successive generations shrink and the language becomes more and more stable: the transmission bottleneck forced the language to become better transmissible. This is how the poverty of the stimulus solves the poverty of the stimulus (cf. Zuidema 2003).

A variety of different models confirmed this account. To name a symbolic and connectionist example, Kirby (2001) showed that a bottleneck caused the emergence of a stable, compositional language in agents representing language with definite-clause grammars. In another study, Kirby and Hurford (2002) found that the number of training instances passed between neural-network agents acted as a bottleneck, with a medium-sized training set leading to structured meaning-signal mappings. The fact that these, and many other, different models gave rise to similar behaviour is in itself striking. But it also makes it difficult to decipher what exactly is going on.

The shape-color example gives a hint, since there is a language much more compressible than a compositional one: the degenerate language that expresses *every* meaning with the same signal. The fact that none of the early studies seem to have produced degenerate languages, suggests that a bias against those must have been present (Cornish 2011). Or, conversely, that the learning algorithms implicitly pressured towards compositional languages. This opens up the possibility that "cultural evolution does no more than transparently map properties of the biology of an individual to properties of language" (Kirby 2017). Kirby points out that there are reasons to doubt this conclusion: The size of the bottleneck and the structure of the domain for example influence the simulations. Nonetheless, it became clear that in order to make claims about the shaping force of cultural evolution, one needs to know 1) what the *implicit biases* in the model are, 2) what the biases of the agents are and 3) how those interact with the cultural process.

2 In fact, various different bottlenecks have been put forward; see Cornish (2011, ch. 4) for an overview and a discussion of the empirical findings regarding the presence of such a bottleneck.

Iterated learning with Bayesian agents

In 2005, Thomas Griffiths and Michael Kalish reinterpreted the iterated learning model in a population of Bayesian agents. One reason for doing so is that it connects the iterated learning model to a rich Bayesian modelling tradition in cognitive science (see e.g. Perfors et al. 2011; Goodman and Tenenbaum 2016; Griffiths, Kemp, and Tenenbaum 2008) and the formal models of human behaviour that have been proposed there. The Bayesian model of Griffiths and Kalish also solved the issues arising from implicit biases, since it *explicitly* encodes the biases of the learners. Moreover, the authors managed to characterise the long-term behaviour of the model — *convergence to the prior* — which sparked a small controversy. In the years that followed, the Bayesian paradigm appears to have surfaced as the primary approach to modelling iterated learning (Kirby, Griffiths, and Smith 2014; Kirby 2017). For that reason, and for its role in the next chapter, I want to go through the model in detail.

Recall from figure 1 that in iterated learning, a language alternates between a 'latent' internal representation θ and an 'overt' external representation x. Agents use a *production* and *language algorithm* (PA and LA) to move between these representations.[3] The idea put forward by Griffiths and Kalish (2007) is to model these production and language algorithms with probability distributions. An agent using language θ_t has a distribution $p_{\text{PA}}(x_t \mid \theta_t)$ over productions describing how to select a utterance. Conversely, it has a distribution $p_{\text{LA}}(\theta_t \mid x_{t-1})$ from which the agent picks a language after observing data x_{t-1} produced by the previous agent. Note that, as figure 1 illustrates, these are the only dependencies. Productions are conditionally independent from previous productions and the same goes for languages. This seems reasonable as an agent cannot use the previous production when making a new one (only its representation thereof) and clearly an agent cannot use the *unobservable* language of the previous agent directly. In short, iterated learning becomes a stochastic process on the random variables x_t and θ_t, which are conditionally independent from previous x_t's and θ_t's respectively.

What makes these agents 'Bayesian' is that their language algorithm reuses the the production algorithm and the prior beliefs of the agents using Bayes' rule. When confronted with data x_t, the agents infer the *posterior* distribution

$$p(\theta_t \mid x_{t-1}) \propto p_{\text{PA}}(x_{t-1} \mid \theta_t) \cdot p(\theta_t), \tag{1}$$

which captures how likely every language θ_t is in light of the observed the data. The posterior distribution balances two factors. First — and this is where the production algorithm is reused — how probable the agent itself regards the observed data to be, if it were to use language θ_t. This is the *likelihood* term $p(x_t \mid \theta_t)$. And second, how likely the language is in the first place: the *prior* $p(\theta_t)$.

Interestingly, *before* Griffiths and Kalish published their Bayesian interpretation, Kirby, Smith, and Brighton (2004) also noted that the language acquisition can be seen as Bayesian inference. The prior, they state, corresponds to Universal Grammar or the Language Acquisition Device: "everything the learner brings to the task *independent of the data*" (italics in original). However, Griffiths and Kalish (2007) stress that the prior "should not be interpreted as reflecting innate constraints *specific* to language acquisition" (my italics). The prior is, in other words, not necessarily domain specific,

3 The language algorithm is usually called a *learning* algorithm. Since that terminology causes some confusion in chapter **??**, I use the term *language algorithm*.

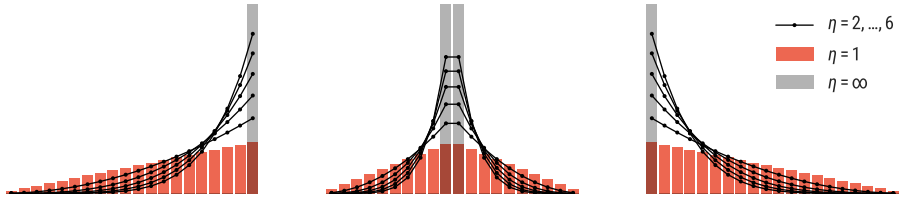

FIGURE 2 Exponentiating a distribution moves the probability mass towards the mode. Illustrated for three different distributions.

FIG03

but aggregates all factors that influence language acquisition, including learned biases. Therefore, "the prior is better seen as determining the amount of evidence that a learner would need to see in order to adopt a particular language". Nevertheless many later papers use the prior primarily to capture innate learning biases (e.g. Kirby, Griffiths, and Smith 2014; Kirby 2017).

So how does a Bayesian agent adopt a particular language? Kirby, Smith, and Brighton (2004) assume agents pick the language with the highest probability under the posterior, the *maximum a posteriori* (MAP) estimate $\arg\max_\theta p(\theta \mid x)$. Griffiths and Kalish (2005), however used a different strategy where agents *sample* a language from their posterior, i.e. they are probability matching. The two strategies can be seen as extreme cases of a more general strategy: sampling from a *exponentiated* (or 'exaggerated') version of the posterior (Kirby, Dowman, and Griffiths 2007):

$$p_\eta(\theta_t \mid x_{n-1}) \propto p(\theta_t \mid x_{n-1})^\eta, \qquad \eta \geq 1. \tag{2}$$

For $\eta = 1$ this is the same as the sampling strategy, but as η increases, more and more of the probability mass is moved towards the maximum of the distribution (the mode) until sampling becomes indistinguishable from the MAP strategy (see figure 2). The language *algorithm* thus takes the posterior distribution and applies the language *strategy* (sample or maximise) to adopt a language.

THE EMERGENCE OF COMPOSITIONALITY, II It might be helpful to go through a concrete example. Griffiths and Kalish (2005) introduced a 'binary' language, which figured in several later studies (Griffiths, Canini, et al. 2007; Burkett and Griffiths 2010; Kirby, Tamariz, et al. 2015). It is a special case of the shape-color example introduced earlier, with two colours and two shapes (so $F = V = 2$). The language was introduced to study the emergence of compositionality. If we simplify the encoding, it is easier to see what the compositional languages are. Write 0 for a triangle, 1 for a square, 0 for black and 1 for orange, such that (□, ●) for instance becomes 10 and (△, ●) becomes 01. Using alphabet $\{a, b\}$ there are 4 compositional languages given by the feature-subsignal mappings

$$\begin{aligned}
&(1) \qquad 0 \mapsto a, \quad 1 \mapsto a \\
&(2) \qquad 0 \mapsto a, \quad 1 \mapsto b \\
&(3) \qquad 0 \mapsto b, \quad 1 \mapsto a \\
&(4) \qquad 0 \mapsto b, \quad 1 \mapsto b
\end{aligned}$$

In this scenario there are 4 meanings (▲, ▲, ■, ■) and $4^4 = 256$ ways to map four meanings to four signals $\{aa, ab, ba, bb\}$. This gives 256 languages of which 4 compositional and 252 holistic.

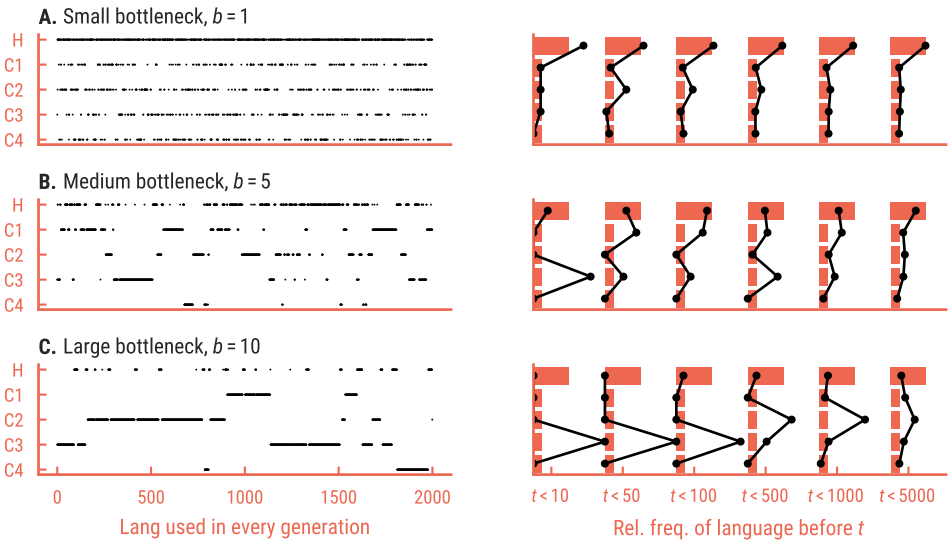

FIGURE 3 Emergence of compositionality in the Bayesian iterated learning model of Griffiths and Kalish (2007). On the left, the language used in every generation with H one of 252 holistic languages and C1–4 the compositional languages. On the right the relative frequency of every language up to a certain time t. These relative frequencies converge to the prior (orange). Larger bottlenecks (subfigures A–C) slow down convergence.

GKO1 WebPPL simulation with $\alpha = 0.5, \varepsilon = 0.001$ and samplers ($\eta = 1$).

Not all languages are equally likely. A hierarchical prior that puts a fraction α of the probability mass on the compositional languages:

$$p(\theta) = \begin{cases} \frac{\alpha}{4} & \text{if } \theta \text{ is compositional} \\ \frac{1-\alpha}{256} & \text{otherwise} \end{cases} \tag{3}$$

Once an language θ has been fixed, the agent is presented with new meaning m for which it then produces a signal s by sampling from the distribution

$$p(s \mid m, z) = \begin{cases} 1 - \varepsilon & \text{if } m \mapsto s \text{ in language } \theta \\ \varepsilon/3 & \text{otherwise} \end{cases} \tag{4}$$

This means the agent will pick the signal s corresponding to m under language θ most of the time, but has a small probability ε of making an error and uniformly picking one of the other signals. Together with a completely independent distribution $p(m)$, typically a uniform one, this specifies the production algorithm

$$p_{\text{PA}}(x \mid \theta) = p(s \mid m, \theta) \cdot p(m), \qquad x = (m, s) \tag{5}$$

If $\mathrm{x} = ((m_1, s_1), \ldots (m_b, s_b))$ is the list of the utterances produced by the previous agent, then the posterior distribution is

$$p(\theta \mid \mathrm{x}) \propto p(\theta) \cdot \prod_{i=1}^{b} p_{\text{PA}}(x_i \mid \theta), \tag{6}$$

and, as usual, the language algorithm takes the form $p_{\text{LA}}(\theta \mid \mathrm{x}) \propto p(\theta \mid \mathrm{x})^{\eta}$.

Figure 3 illustrates the resulting simulation in a population of samplers ($\eta = 1$). It shows which language was used in every generation (left): one of the 252 holistic languages (H) or a compositional language (C1–4). The compositional languages seem to be used much more frequently, which is confirmed by the plots on the right. There

we see the relative frequency of every language up to several points t in the simulation. These plots indicate that the relative frequencies converge to the prior, shown in orange. Since the compositional languages have a higher prior probability than each of the holistic languages, they are more frequent. The convergence rate towards the prior is much faster when the bottleneck is small ($b = 1$, subfigure A) than when it is large ($b = 10$, subfigure C). It is clear why this happens: the more data is transmitted, the greater the probability that the child can reconstruct the language. The result is that languages will be stable throughout multiple generations, as seen from the lines in figure 3C. Nevertheless, even with a large bottleneck the relative frequencies seem to converge to the prior, be it very slowly. We will discuss all these findings in more detail later. First, what discuss the observed 'convergence to the prior'.

Convergence to the prior

Let me briefly summarise what we have seen so far. Bayesian agents observe utterances x_{t-1} produced by the previous agent, and then use Bayes' rule to infer a language. This language is θ is drawn from $p_\eta(\theta_t \mid x_{t-1}) = p(\theta_t \mid= x)^\eta$, where η interpolates between a sampling- and MAP-strategy for $\eta = 1$ and $\eta = \infty$ respectively. All this results in a chain of the form

$$x_0 \longrightarrow \theta_1 \longrightarrow x_1 \longrightarrow \theta_2 \longrightarrow x_2 \cdots . \tag{7}$$

Griffiths and Kalish (2005) noted that several Markov chains can be discerned in eq. 7, of which the long-term behaviour is well-studied: They often converge to a so called stationary distribution. This characterised the long-term behaviour of the iterated learning model.

Appendix **??** introduces the relevant convergence results for Markov Chains; I only summarise them here. Consider a system with a set of possible states S. If the variables $x_0, x_1, x_2, \ldots$ indicate the state at every time step, they form a Markov chain if the probability of moving to another state only depends on the last state: $p(x_t \mid x_0, \ldots, x_{t-1}) = p(x_t \mid x_{t-1})$. If the number of states is finite, these *transition probabilities* can be collected in the transition matrix T. Suppose the initial distribution over states is given by vector π, then the next distribution is $p(x_1 = i) = (\mathrm{T}\pi)_i$ and after t steps, $p(x_t = i) = (\mathrm{T}^t\pi)_i$. These probabilities can converge to the so called *stationary distribution* $\pi^\star$ which must be an eigenvector of T since $\mathrm{T}\pi^\star = \pi^\star$. If the Markov chain is *ergodic* it is guaranteed to have a unique stationary distribution to which it converges: $p(x_t = i) \to \pi_i^\star$ as $t \to \infty$. Ergodicity, briefly, ensures that the chain keeps revisiting the entire state space and has a positive probability of reaching any other state from any given state in a finite number of steps. How often it visits every state is given by the stationary distribution, in the sense that the relative frequencies of visited states converges to the stationary distribution.

PROOF OF THE CONVERGENCE TO THE PRIOR Griffiths and Kalish (2005) noted that by marginalising out the productions x_t in eq. 7 one obtains the following Markov chain

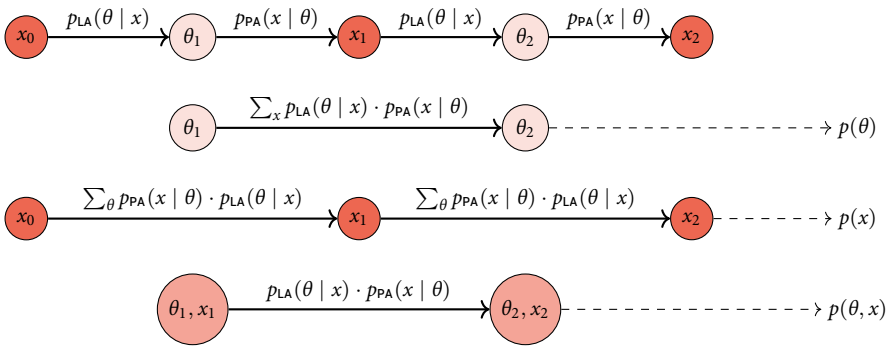

FIGURE 4 Different Markov chains hidden in the Bayesian iterated learning model, and to which stationary distribution they converge (right).

Figure adapted from Griffiths and Kalish (2007).

(see also figure 4):

$$p(\theta_t \mid \theta_{t-1}) = \sum_{x_{t-1}} p_{LA}(\theta_t \mid x_{t-1}) \cdot p_{PA}(x_{t-1} \mid \theta_{t-1}). \tag{8}$$

We hitherto assumed that the transition probabilities remain constant over time, that is, we looked at time-homogeneous chains. The Markov chain in eq. 8 is only homogeneous if all agents use the same production and language algorithms. In particular, they should all use the same prior. We will later discuss the validity of this assumption. If these assumptions hold and the chain is moreover ergodic, *then* the long-term behaviour of iterated learning is known: convergence to the stationary distribution, independent of the initial distribution.

The stationary distribution $\pi^\star$ of this distribution happens to be the prior $q(\theta) := p(\theta)$. To show this, one has to see that

$$q(\theta_{t+1}) = \sum_{\theta_t} p(\theta_{t+1} \mid \theta_t) \cdot q(\theta_t) \tag{9}$$

I have written q for the prior to highlight that we do not know whether $q(\theta_{t+1})$ is a marginal distribution of $p(\theta_{t+1}, \theta_t)$. In that case, the equality would hold trivially. Otherwise, the following derivation shows that eq. 9 holds:

$$\begin{aligned}
\sum_{\theta_t} q(\theta_t) \cdot p(\theta_{t+1} \mid \theta_t) &= \sum_{\theta_t} q(\theta_t) \cdot \sum_{x_t} p_{LA}(\theta_{t+1} \mid x_t) \cdot p_{PA}(x_t \mid \theta_t) && (10)\\
&= \sum_{x_t} p_{LA}(\theta_{t+1} \mid x_t) \cdot \sum_{\theta_t} q(\theta_t) \cdot p_{PA}(x_t \mid \theta_t) && (11)\\
&= \sum_{x_t} p_{LA}(\theta_{t+1} \mid x_t) \cdot p_{PA}(x_t)\\
&\overset{(\star)}{=} \sum_{x_t} \frac{p_{PA}(x_t \mid \theta_{t+1}) \cdot q(\theta_{t+1})}{p_{PA}(x_t)} \cdot p_{PA}(x_t)\\
&= q(\theta_{t+1}) \sum_{x_t} p_{PA}(x_t \mid \theta_{t+1})\\
&= q(\theta_{t+1})
\end{aligned}$$

where $(\star)$ holds by definition of $p_{LA}(\theta_{t+1} \mid x_t)$ and because we use samplers ($\eta = 1$). For maximisers, the proof breaks down at this point.

Similar results hold for the other Markov chains hidden in the iterated learning model (see figure 4). When averaging over interpretations rather than productions, one obtains a Markov chain on the productions:

$$p(x_{t+1} \mid x_t) = \sum_{\theta_{t+1}} p(x_{t+1} \mid \theta_{t+1}) \cdot p(\theta_{t+1} \mid x_t). \tag{12}$$

A proof analogous to eq. 10 shows that this chain converges to the *prior predictive distribution* $p(x) = \sum_\theta p_{\text{PA}}(x \mid \theta) \cdot p(\theta)$ Finally, one could consider a Markov chain over the state space of language-utterance pairs $(\theta, x) \in \Theta \times \mathcal{X}$ with transition probabilities

$$p(\theta_{t+1}, x_{t+1} \mid \theta_t, x_t) = p(\theta_{t+1} \mid x_t) \cdot p(x_t \mid \theta_t). \tag{13}$$

This chain has the joint $p(\theta, x) = p_{\text{PA}}(x \mid \theta) \cdot p(\theta)$ as its stationary distribution. Interestingly, this shows that Bayesian iterated learning implements a *Gibbs sampler*.

Gibbs samplers are often used in Bayesian statistics, whenever it is not possible to work with complicated distributions analytically. *Monte Carlo methods* are work-arounds that collect many samples from the distribution, and approximate the distribution using those samples. To obtain samples, one constructs a Markov chain whose stationary distribution is the distribution of interest. Over time, the visited states will be (correlated) samples from the target distribution. This is the basic idea behind many *Markov Chain Monte Carlo* (MCMC) methods and Gibbs sampling is one of those. It can be used to approximate a joint distribution $p(\theta, x)$ if it is easy to sample from the conditional distributions $p(\theta \mid x)$ and $p(x \mid \theta)$. In every iteration, it fixes one of the variables, say θ_t and samples a new x_{t+1} from $p(x_{t+1} \mid \theta_t)$. Then it fixes x_t and samples θ_{t+1} from $p(\theta_{t+1} \mid x_{t+1})$, and so on. This results in a new sample (θ_{t+1}, x_{t+1}) from the joint after every 'sweep' through the variables. Indeed, this procedure exactly mirrors Bayesian iterated learning with sampling agents, and it follows that the chain in eq. 13 converges to $p(\theta, x)$ (see Griffiths and Kalish 2007 for a longer discussion).

CONVERGENCE TO THE MAXIMUM OF THE PRIOR? What kind of behaviour should one expect in populations of maximisers? This turns out to be a much harder question. There are, to the best of my knowledge, two analytical results — we will return to empirical evaluations in chapter **??** — both suggesting that in populations of maximisers the behaviour is largely determined by the prior, but in a less direct way. First of all, Kirby, Dowman, and Griffiths (2007) analyses the stationary distribution for maximisers ($\eta > 1$) using a constrained set of languages that spread the probability mass uniformly over a (sub)set of utterances.[4] In other words, $p(x \mid \theta)$ is either 0 or equal to a $f(x)$, where the latter does not depend on θ. In that case, the stationary distribution is proportional to $p(z)^\eta$. This implies that cultural evolution results in an exaggerated version of the prior (cf. figure 2).

A similar conclusion follows from the second result, due to Griffiths and Kalish (2007). They note that maximisers (now $\eta = \infty$) implement a version *Expectation-Maximisation* (EM). This is an iterative algorithm used in models with hidden variables to estimate parameters that are increasingly close to the maximum likelihood estimates, or, in our case, MAP estimates. The trick is to use the current parameters to estimate the *expected* likelihood of the observed and hidden variables, and then update

4 This constraint on languages has a purely mathematical motivation: it is precisely what is needed to factorise the normalising constant in the posterior.

the parameters so that they *maximize* that likelihood. When computing the expectation analytically is intractible, it can be approximated by drawing several samples. The case using a single sample is called *stochastic* EM. Now, suppose, in EM jargon, there are no observed variables, x_t is the latent variable and θ_t the parameter, then stochastic EM in this model amounts to Bayesian iterated learning in a population of maximisers (see Griffiths and Kalish (2007) for details). This characterisation is not as clear-cut as with samplers, but suggests that the stationary distribution over languages will roughly be centred on the maxima of the prior (Griffiths and Kalish 2007).

Convergent controversy

The *convergence to the prior* was the first general result about the long-term behaviour of the iterated learning model. For populations of samplers, the result was crystal clear: starting from any initial distribution, the probability that an agent down the chain would be using language θ is given by the prior probability $p(\theta)$. And this is precisely what we observed in figure 3, which shows the the emergence of compositionality — or rather, the emergence of the prior. The model is an ergodic Markov chain, and over time the probability that a certain language will be used therefore converges to its probability under the stationary distribution, which is the prior. Compositional languages have high probability under that prior, and consequently emerge. Maximisers are much harder to analyse. The probability that a language is used by maximisers seems to be largely determined by the maxima of the prior. Now, what are the implications of all this for cultural language evolution?

BOTTLENECKS AND WEAK BIASES Iterated learning was inspired by the idea that language is a compromise between "the biases of learners, and other constraints acting on language during their transmission" (Smith 2009), originally in the form of a transmission bottleneck. But in the Bayesian models, the bottleneck hardly plays any role. Griffiths and Kalish (2007) conclude that "the emergence of languages with particular properties does not require a bottleneck" (p. 466). Larger bottlenecks do slow down convergence since they imply more faithful transmission and this increases language stability. The Markov chain's walk through the state space consequently slows down, which, somewhat paradoxically, also slows down convergence. But in the long run bottlenecks play no role — at least for samplers. This seems to undermine the idea that compressible languages emerge *because of* cultural transmission. Should we conclude, then, that languages are not shaped by cultural evolution, but primarily by innate constraints? Griffiths and Kalish (2007) conclude that their results "do not indicate which of these explanations is more plausible" (p. 475). There's something for everyone: if the prior captures innate biases, "iterated learning acts as an engine by which these constraints result in universals" (p. 475), but if you prefer the transmission process to actually change the priors, then you "can take heart from our results for learners who use MAP estimation".

Kirby, Dowman, and Griffiths (2007) follow the latter advise. Their paper discusses an iterated learning model with maximisers that have a prior bias towards regular languages. Bottleneck effects can occur in populations of maximisers (Griffiths and Kalish

2007) and the authors accordingly conclude that as the bottleneck tightens in their model, "regularity is increasingly favoured". But there is something peculiar about this conclusion: It seems to hold only because their prior favoured regularity. Had their prior favoured irregularity, irregularity would have been increasingly favoured under a tighter bottleneck.[5] In the Bayesian model, transmission at most amplifies pre-existing biases, which of course can be seen as an effect of cultural transmission. Another conclusion of Kirby, Dowman, and Griffiths (2007) is therefore that processes of cultural evolution can "completely obscure" the *strength* of the bias. A small tendency to favour languages with higher prior probability (i.e. $\eta > 1$) amplifies weak biases and results in strong universals. The strength of the bias has no role, only the ordering of the languages. All in all, it suggests a rather toothless process of cultural evolution. Several researchers therefore started tweaking the assumptions of the model to find out how robust the results are.

POPULATION STRUCTURE AND HETEROGENOUS POPULATIONS The population structure was one of the first things addressed. It should be noted that (Griffiths and Kalish 2007) generalised their findings to somewhat different scenario, with finite generations evolving in (discrete or) continuous time (cf. Nowak, Komarova, and Niyogi 2001). In that case the *proportion* $p_t(\theta)$ of the population speaking language θ at time t converges to the prior $p(\theta)$, as can easily be seen. If $\mathrm{p}_t = (p_t(\theta) : \theta \in \Theta)$ and T the transition matrix, these proportions change as

$$\mathrm{p}_{t+1} = \mathrm{T}\mathrm{p}_t, \tag{14}$$

which describes a linear dynamical system with a unique stable equilibrium. The same derivations as eq. 10 show the prior is that equilibrium. However, Niyogi and Berwick (2009) argue that this is an unrealistic model of language evolution as it precludes the possibility of bifurcations. Moreover, language stability cannot be maintained: even if only 0.01% of the population uses a different language, it will spread to a larger share of the population (the prior admitting). As a remedy Niyogi and Berwick (2009) propose an alternative model where agents learn from a mixture of the languages used in the previous generation, not just one. This leads to markedly different nonlinear behaviour with bifurcations and possibility multiple equilibria, which they argue accurately describes historical developments (namely, that English is no longer a 'verb-second' language).

That the behaviour changes in different populations structures was confirmed in several other studies. Smith (2009) similarly considered infinite generations of agents learning from multiple parents. He reports that this precludes convergence to the prior and introduces a dependency on the initial distribution of languages in the population. Ferdinand and Zuidema (2009) draw the same conclusion, but also drop the assumption that all agents share the same innate biases, i.e. that the population is *homogeneous*. In heterogeneous population the convergence to the prior breaks down. Dediu (2009) finds that the strong differences between samplers and maximisers disappears in populations with a different structure or heterogeneity.

The agents in studies such as Ferdinand and Zuidema (2009) are not Bayesian agents in the strict sense that agents assume to be learning from a single language, while in fact the data comes from several sources. Burkett and Griffiths (2010) address this issue

5 I found their PNAS paper is a bit sketchy on the details of their simulations, but these conclusions follow directly from Griffiths and Kalish (2007) and as far as I can see apply equally to Kirby, Dowman, and Griffiths (2007).

in a hierarchical model where agents take into account that they are possibly learning from multiple languages. Accordingly, the convergence to the prior reappears. Very recently, Whalen and Griffiths (2017) extended this to populations with arbitrary network structures, although it should be stressed that agents still learned from a single teacher. Nevertheless, the emerging consensus appears to be that in slightly more complicated population structures (with possibly imperfect Bayesian reasoners) the convergence to the prior can break down and nontrivial cultural effects appear.

LINEAGES AND CUMULATIVE CULTURAL EVOLUTION It is somewhat surprising that the population structure received most criticism, since that aspect of the Bayesian model is perfectly in line with the original iterated learning model. Some other parts, I would argue, are not. First of all, the type of convergence — in language or in probability of using a language — is markedly different. In early iterated learning studies, the population converged to a stable language which could be transmitted faithfully along many generations. In the Bayesian models, nothing of this sort happens. In the simulation of the emergence of compositionality (figure 3) one clearly sees that successive generations can acquire radically different languages: picture English-speaking parents, themselves born to Basque parents, whose children miraculously learned Hungarian.

Transmission in the Bayesian model generally not faithful — indeed, this is necessary for 'convergence' to occur at all. That seems particularly problematic for a model of cultural evolution. Even if transmission shapes languages, it has to be somewhat faithful if one expects any kind of cultural *evolution*. Tomasello (1999) points out that faithful transmission is important because it enables a so called *cultural ratchet*, where cultural innovations are passed on and improved upon by later generations. Cultural evolution, as a result, is *cumulative* and products of cultural evolution consequently reflect their full historical development. If it is not already uneasy that the defining property of a Markov chain is being memoryless, ergodicity certainly conflicts with the idea of cumulative cultural evolution. In an ergodic Markov chain, every 'lineage' is guaranteed to revisit all possible languages infinitely often. That amounts to an infinite reinvention of the wheel — pretty much the exact opposite of cumulative cultural evolution.

Conclusions

The first iterated learning models suggested that languages primarily pick up systematicity during cultural transmission. Simulations showed how compositional structure accumulated in initially unstructured languages when a bottleneck pressured the languages to become more compressible. However, the learning algorithms that generalise a few observations to a full language implement all kinds of implicit biases, and possibly provide an implicit pressure towards compositional structures. To make general claims about the interaction of cultural processes and innate biases, the two need to be separated clearly. Bayesian iterated learning models did precisely that, but were also shown to *converge to the prior*. That meant that the probability that a certain language would be used, is after a while completely determined by the biases of the learners, independent of the initial conditions. In populations of maximisers, the relation is less

transparent and the shape of the prior (its maxima in particular) largely appears to determine the outcome of cultural evolution.

The Bayesian iterated learning model moved the explanatory load from the cultural process to the prior biases of the learners. However, the strong conclusions were in several studies shown to break down in more complicated populations. The Bayesian model moreover results in an arguably unrealistic model of cultural evolution, with no language stability, nor any cumulative effects. Despite these shortcomings, the field made significant progress due to the work of Griffiths and Kalish. Explicitly encoding the biases of the learners made studies of the interactions between the 'nature' and 'nurture' of language much more principled, and moreover resulted in cognitively better motivated agents. The focus on analytic results regarding the long-term behaviour brought further transparency to the somewhat opaque conclusions suggested by simulations alone — irrespective of whether the results are ultimately convincing.

In sum, combining the criticism and benefits, I would draw up the following list of desiderata for a model that aims to show that cultural processes can shape the evolution of language (in arbitrary order):

(D1) **Explicate biases.** The biases of the agents should be explicitly specified in the model.

(D2) **Strategies.** The model should explore a wide range of strategies, such as sampling or MAP strategies.

(D3) **Analysable.** The model should be amenable to analytical scrutiny, and it should ideally be possible to draw general conclusions about long-term behaviour.

(D4) **Nontrivial cultural effects.** The model should exhibit non-trivial cultural effects, which might for example result in lineage-specific evolution: different runs resulting in different outcomes of cultural evolution.

(D5) **Robustness to population structure.** The model should exhibit behaviour that is fairly robust to changes in population structure.

(D6) **Language stability.** The model should result in a 'reasonable' degree of language stability. Reasonable, since languages are never perfectly stable (see also Kirby 2001).

(D7) **Empirically testable.** The model should give an empirically plausible, mechanistic explanation of cultural evolution, which is further testable against empirical linguistic findings (predating the lab).

The list is no doubt incomplete, but mainly serves as a guide to what I will address in this thesis, most notably in chapter **??**.

Bibliography

Brighton, Henry (2002). "Compositional Syntax From Cultural Transmission". In: *Artificial Life* 8.1, pp. 25–54. DOI: 10.1162/106454602753694756.

Burkett, David and Thomas L. Griffiths (2010). "Iterated learning of multiple languages from multiple teachers". In: *The evolution of language: Proceedings of EvoLang*, pp. 58–65.

Chomsky, Noam (1986). *Knowledge of language: Its nature, origin, and use*. Convergence. New York: Praeger.

Cornish, Hannah (2011). "Language Adapts: Exploring the Cultural Dynamics of Iterated Learning". PhD thesis. University of Edinburgh.

Culbertson, Jennifer and Simon Kirby (2016). "Simplicity and specificity in language: Domain-general biases have domain-specific effects". In: *Frontiers in Psychology* 6.JAN, pp. 1–11. DOI: 10.3389/fpsyg.2015.01964.

Dediu, Dan (2009). "Genetic biasing through cultural transmission: Do simple Bayesian models of language evolution generalise?" In: *Journal of Theoretical Biology* 259.3, pp. 552–561. DOI: 10.1016/j.jtbi.2009.04.004.

Ferdinand, Vanessa and Willem Zuidema (2009). "Thomas' Theorem meets Bayes' Rule: a Model of the Iterated Learning of Language". In: *Proceedings of the 31st Annual Conference of the Cognitive Science Society.* Austin, Texas, pp. 1786–1791.

Goodman, Noah D. and Joshua B. Tenenbaum (2016). *Probabilistic Models of Cognition*. v2.

Griffiths, Thomas L., Kevin R. Canini, et al. (2007). "Unifying rational models of categorization via the hierarchical Dirichlet process". In: *Proceedings of the 29th annual conference of the cognitive science society*, p. 323328.

Griffiths, Thomas L. and Michael L. Kalish (2005). "A Bayesian view of language evolution by iterated learning". In: *Proceedings of the 27th annual conference of the cognitive science society*, pp. 827–832.

— (2007). "Language Evolution by Iterated Learning With Bayesian Agents". In: *Cognitive Science* 31.3, pp. 441–480. DOI: 10.1080/15326900701326576.

Griffiths, Thomas L., Charles Kemp, and Joshua B Tenenbaum (2008). "Bayesian models of cognition". In:

Hurford, James R (2000). "Social Transmission Favours Linguistic Generalization". In: *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*.

Kirby, Simon (2001). "Spontaneous evolution of linguistic structure - An iterated learning model of the emergence of regularity and irregularity". In: *IEEE Transactions on Evolutionary Computation* 5.2, pp. 102–110. DOI: 10.1109/4235.918430.

— (2017). "Culture and biology in the origins of linguistic structure". In: *Psychonomic Bulletin & Review* 24.1, pp. 118–137. DOI: 10.3758/s13423-016-1166-7.

Kirby, Simon, Mike Dowman, and Thomas L. Griffiths (2007). "Innateness and culture in the evolution of language." In: *Proceedings of the National Academy of Sciences of the United States of America* 104.12, pp. 5241–5245. DOI: 10.1073/pnas.0608222104.

Kirby, Simon, Tom Griffiths, and Kenny Smith (2014). "Iterated learning and the evolution of language". In: *Current Opinion in Neurobiology* 28, pp. 108–114. DOI: 10.1016/j.conb.2014.07.014.

Kirby, Simon and James R Hurford (2002). "The Emergence of Linguistic Structure: An Overview of the Iterated Learning Model". In: *Simulating the Evolution of Language*. London: Springer London, pp. 121–147. DOI: 10.1007/978-1-4471-0663-0_6.

Kirby, Simon, Kenny Smith, and Henry Brighton (2004). "From UG to Universals: Linguistic adaptation through iterated learning". In: *Studies in Language* 28.3, pp. 587–607. DOI: 10.1075/sl.28.3.09kir.

Kirby, Simon, Monica Tamariz, et al. (2015). "Compression and communication in the cultural evolution of linguistic structure". In: *Cognition* 141, pp. 87–102. DOI: 10.1016/j.cognition.2015.03.016.

Niyogi, Partha and Robert C Berwick (2009). "The proper treatment of language acquisition and change in a population setting". In: *PNAS* 106.25, pp. 10124–10129.

Nowak, Martin a., Natalia L. Komarova, and Partha Niyogi (2001). "Evolution of universal grammar." In: *Science (New York, N.Y.)* 291.5501, pp. 114–8. DOI: 10.1126/science.291.5501.114.

Perfors, Amy et al. (2011). "A tutorial introduction to Bayesian models of cognitive development". In: *Cognition* 120.3, pp. 302–321. DOI: 10.1016/j.cognition.2010.11.015.

Smith, Kenny (2002). "The cultural evolution of communication in a population of neural networks". In: *Connection Science* 14.1, pp. 65–84. DOI: 10.1080/09540090210164306.

— (2009). "Iterated learning in populations of Bayesian agents". In: *Cogsci Society Conference*, pp. 697–702.

Tomasello, Michael (1999). *The cultural origins of human cognition*, p. 248.

Whalen, Andrew and Thomas L. Griffiths (2017). "Adding population structure to models of language evolution by iterated learning". In: *Journal of Mathematical Psychology* 76, pp. 1–6. DOI: 10.1016/j.jmp.2016.10.008.

Zuidema, Willem (2003). "How the Poverty of the Stimulus Solves the Poverty of the Stimulus". In: *Advances in Neural Information Processing Systems 15* 15, p. 51.