

# Thomas' theorem meets Bayes' rule: a model of the iterated learning of language

Vanessa Ferdinand (v.a.ferdinand@uva.nl)

Willem Zuidema (zuidema@uva.nl)

Cognitive Science Center Amsterdam & Institute for Logic, Language and Computation,  
University of Amsterdam, Plantage Muidergracht 24, 1018 TV, Amsterdam, The Netherlands

## Abstract

We develop a Bayesian Iterated Learning Model (BILM) that models the cultural evolution of language as it is transmitted over generations of learners. We study the outcome of iterated learning in relation to the behavior of individual agents (their biases) and the social structure through which they transmit their behavior. BILM makes individual learning biases explicit and offers a direct comparison of how individual biases relate to the outcome of iterated learning. Most earlier BILMs use simple one parent to one child (monadic) chains of homogeneous learners to study the outcome of iterated learning in terms of bias manipulations. Here, we develop a BILM to study two novel manipulations in social parameters: population size and population heterogeneity, to determine more precisely what the transmission process itself can add to the outcome of iterated learning. Our monadic model replicates the existing BILM results, however our manipulations show that the outcome of iterated learning is sensitive to more factors than are explicitly encoded in the prior. This calls into question the appropriateness of assuming strong Bayesian inference in the iterated learning framework and has important implications for the study of language evolution in general.

## Introduction

The language that you speak is not a product of your mind alone. Of course, in acquiring language, you did bring to task a sophisticated learning apparatus able to discover intricate patterns and track frequency distributions at all levels of linguistic analysis. However, you learned your language from data that had been shaped by countless generations of language learners and users before you. Without this history, your personal language would have looked quite different.

In the field of language evolution, the iterated learning model (ILM) has emerged as a popular formalization of this cultural transmission of language over successive generations. Through a series of models, it has become clear that the iteration of learning may have profound influences on the resulting languages. Many of the defining features of natural language have been shown to be possible results of iteration, including compositional semantics and recursiveness (Kirby, 2002) and the bidirectionality of the sign (Smith, 2002). Moreover, these models have shown that the learnability of language in an important sense comes for free: seemingly arbitrary choices of current learners are correct precisely because they are the same choices as those made by the previous generations (Deacon, 1997; Zuidema 2003). This is an instance of Thomas' theorem: "If men define situations as real, they are real in their consequences" (Thomas & Thomas, 1928).

However, learning in the ILM tradition is operationalized with a variety of adhoc learning algorithms and many questions remain about the role and appropriateness of the built-in

biases. One way to disentangle the role of the learning biases and the role of cultural transmission in iterated learning is to implement ILMs with Bayesian rational agents, whose inductive biases are explicitly coded. The first main BILM result of this kind was that the outcome of iterated learning (i.e. the stationary distribution of the transition matrix describing transition probabilities between all possible languages) exactly mirrors the agents' distribution over priors. This result implied that the learners' biases entirely determine the outcome of iterated learning (Griffiths and Kalish, 2005). However, when agents with a different hypothesis selection strategy are used (i.e. they maximize as opposed to sample from the posterior probabilities over hypotheses), the outcome of iterated learning amplifies the biases and does not mirror the prior exactly (Kirby et al, 2007). This is taken as evidence that cultural transmission is adding something to the outcome of iterated learning.

Given that ILMs are complex, dynamical systems we should look at the outcome of iterated learning as a product of the behavior of individual agents (agent properties) and the ways in which they interact (social properties). In this way, we can make a clear distinction between which aspects of the outcome can be attributed to agent biases and which aspects can be attributed to the transmission process itself. Normally, only manipulations to the agent itself are made (in terms of bias values and hypothesis selection strategies) and any outcomes that do not strictly mirror the bias are attributed to unspecified additions of the cultural transmission process.

In the present research, we have chosen to manipulate two social parameters: population size and bias heterogeneity. In doing so, we find that the outcome of iterated learning no longer mirrors the prior and is sensitive to more biases in the Bayesian agent than are explicitly encoded in the prior. If the prior can no longer be seen as encoding the whole of an agent's bias, this calls into question the appropriateness of the current BILM framework as a remedy for explicitly linking specific properties of the agents to the outcome of iterated learning.

## Model description

In this section we outline the components of a Bayesian ILM and describe how they are implemented in our model. The components are the standard ones from any Bayesian inference model (hypothesis space, data, prior, posterior) combined with those from iterated learning (the learners or 'agents', data production, population structure). The implementation and analysis were all carried out in Matlab (for details, see Ferdinand & Zuidema, 2008).

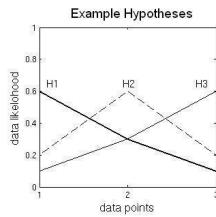


Figure 1: Graph of hypotheses  $[.6\ .3\ .1; .2\ .6\ .2; .1\ .3\ .6]$  and example prior vector  $[.7\ .2\ .1]$ . Each hypothesis' shape is entirely determined by the likelihoods it assigns to the data.

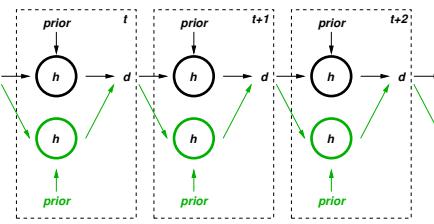


Figure 2: Bayesian iterated learning, in the monadic (black) and polyadic (+green/grey) condition. When all priors are equal, the chain is called homogeneous, otherwise it is heterogeneous.

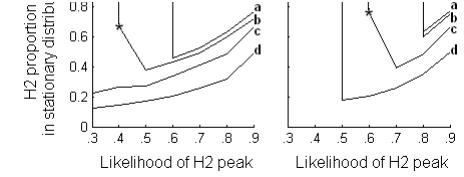


Figure 3: Hypothesis structure and priors affect stationary distribution of monadic maximizers.  $b=1$ ; hypotheses  $[\alpha (1-\alpha)/2 (1-\alpha)/2; (1-\alpha)/2 x (1-\alpha)/2; (1-\alpha)/2 (1-\alpha)/2 \alpha]$ , with  $x$  on horizontal axis and  $\alpha = .33$  in curve (a),  $.4$  in (b),  $.6$  in (c) and  $.8$  in (d).

**Hypotheses:** agents are assumed to consider a fixed set of hypotheses about the state of the world. Each of these hypotheses assigns different likelihoods to each of a fixed set of possible observations (they are thus probability distributions over *data*,  $P(d|h)$ ). These hypotheses could represent, for instance, different languages that generate a set of utterances, or different functions that describe a set of data points. The exact nature of the hypotheses is left unspecified, but the basic properties of the model should be generalizable to a variety of systems where information is culturally transmitted (including not only language, but also, e.g., music and bird song). In this paper, the hypotheses are set at the beginning of each simulation and are the same for all agents. In our examples, we will use a set of hypotheses of size three:  $h \in H = \{h_1, h_2, h_3\}$ .

**Data:** The observations that an agent can make about the state of the world will be referred to as data points. Data points are from a fixed set of possible values and in our model we restrict the size of this set to three:  $D = \{d_1, d_2, d_3\}$ . In linguistic terms, a data point can be interpreted as any piece of evidence (a “trigger”) that the target language has a particular property. Different hypotheses assign different likelihoods to different data points. Multiple data points may be organized into a string, where the likelihood of the string is the product of the likelihoods of the points. Data in our model is any such string of length  $b \geq 1$ :  $d \in D^b$ .

**Prior probability:** The *prior* is a probability distribution over *hypotheses*, defining the probability  $P(h)$  assigned to each hypothesis before any data is seen. Thus, the prior models the inductive bias of the learner. An example shorthand we will use to note the prior is  $[.7\ .2\ .1]$ , where the prior probabilities for  $h_1$ ,  $h_2$  and  $h_3$  are  $.7$ ,  $.2$  and  $.1$ , respectively<sup>1</sup>.

<sup>1</sup>Note that in Bayesian modelling the choice of priors is typically constrained by a set of principles that assure the outcome is not trivially steered by the modeller (e.g., priors must be uninformative) and that calculations remain feasible (e.g., priors must be conjugate to the likelihood function). In our model, we allow any choice of prior over any set of hypotheses; this is ‘unorthodox’ yet

Agents perform inference by combining their prior beliefs and observations of data using Bayes’ rule:

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)} \quad (1)$$

where  $P(h|d)$  denotes the posterior probability that a hypothesis could have generated the data in question and  $P(d) = \sum_{h \in H} P(h)P(d|h)$  is the probability of the data averaged over all hypotheses.

**Hypothesis choice:** Once the posterior probabilities are calculated, we still need a strategy to choose one particular hypothesis  $h$ , from which to generate the data. We consider the same two strategies as in much earlier literature: (i) the *maximizer* simply chooses the hypothesis with the maximum posterior probability (MAP) or, if there are multiple maximums, a random one from the set of MAP-hypotheses; (ii) the *sampler* chooses one hypothesis randomly, but weighted by the posterior probabilities.

**Data Production:** Once a hypothesis is chosen, the next step is to generate new data using that hypothesis. For instance, assuming the agent has chosen  $h_1$ , each data point in the output string will be randomly generated, but weighted according to the likelihood of each data point under  $h_1$ . The number of data points in this string defines the “transmission bottleneck”  $b$ . A characteristic string of size  $b = 10$  under hypothesis  $h_1$  of figure 1 would be  $< d_1 d_1 d_2 d_1 d_2 d_2 d_1 d_3 d_1 >$ .

**Population structure:** Agents are organized into discrete generations of one or more individuals (figure 2). If the number of agents per generation is exactly 1, the model resembles previous ILMs; we will call this condition *monadic*. We will also consider larger populations, a condition labeled *polyadic*. Because each generation only learns from the previous, the model can – regardless of

very relevant for some of our results, as one reviewer notes, but it is consistent with our view that it is the reality of human biology that ultimately determines which priors and hypotheses are appropriate in linguistic modelling, and not the mathematical convenience for the modeller.

the population size – be characterized as a *Markov chain*<sup>2</sup>, where the population at time  $t$  (after learning) is characterized as a vector of hypotheses  $(\vec{h}_t)$ , and for each pair of hypotheses  $\langle h_t, h_{t+1} \rangle$  the probability of transitioning from one to the other is given by  $\sum_d P(d|h_t) \times P(h_{t+1}|d)$ .

Cultural transmission (iteration) is modeled by using each generation’s output data string as the next generation’s input data string. All agents in one generation produce the same number of data samples, which are all concatenated into the output data string for that generation. The likelihood of a data string is invariant to the order of the data samples it contains. Each agent has no way of knowing the number of agents which produced the data string or which data came from which agent. Additionally, each generation has an identical composition of agents as the generation before it. If a population is heterogeneous, then the same set of heterogeneous agents occur in each generation.

## Results

We will study this model numerically while varying the prior, hypothesis structure, bottleneck size, hypothesis choice strategy, population size, and population heterogeneity in terms of different priors per agent. In this section we will first briefly describe how we monitor the behavior of the model. We then show our model can replicate aspects of previous Bayesian ILMs. We will then address new findings for multi-agent populations with heterogeneous and homogeneous biases.

### Monitoring model behaviour

We are looking for features that are invariant as well as changes in the dynamics of the model that can be causally attributed to specific changes in parameter settings. A concrete representation of a “dynamical fingerprint” can be obtained by constructing a transition matrix  $\mathbf{Q}$  for each model. This matrix gives the probabilities that any of the hypotheses will lead to itself or any other hypothesis in the next generation. In linguistic terms, the transition matrix is the idealized equivalent of a theory of (generational) *language change*: it gives the probability that a language in a particular state at generation  $t$  will change into any of a set of possible states at generation  $t+1$  (cf. Niyogi, 2002).

If an ILM simulation could be run for an infinite amount of time, the relative frequency of each chosen hypothesis would settle into a particular distribution that is determined entirely by these transition probabilities. This distribution is known as the *stationary distribution* and serves as an idealized shorthand for the “outcome of iterated learning”. The stationary distribution is proportional to the first eigenvector of the transition matrix (Griffiths & Kalish, 2005). In some cases, we can therefore easily determine the stationary distribution analytically. In linguistic terms, the stationary distribution is the idealized equivalent of *linguistic typology*, a

<sup>2</sup>(Note that for large populations and large sets of hypotheses the transition matrix is of course too big to calculate exhaustively, but identifying a system as a Markov chain can help us decide on the right approximations to study such systems in simulations.

theory of language universals and variation: it gives the relative frequencies with which we should see particular language variants, when very many ( $N \rightarrow \infty$ ) independent cultural evolution chains have run for a very long time ( $t \rightarrow \infty$ ).

In an experimental run, the relative frequency of all chosen hypotheses are also entirely determined by the transition matrix, but because a run contains a finite number of transitions, it represents one actual trajectory of transitions, from a larger set of probable trajectories under that matrix. However, when a large number of transitions can be recorded in a simulation (by setting the number of generations sufficiently high), then a tally of the actual hypotheses chosen by the agents over the course of the simulation closely approximates the analytical stationary distribution. This is the *normalized hypothesis history* (HH); it is a reliable, experimental approximation of the stationary distribution and is reported in most of our results.

### Monadic iterated learning

Griffiths & Kalish (2005) showed, both analytically and numerically, that the stationary distribution of iterated learning of *monadic* (i.e., a single agent per generation) samplers mirrors the prior, independently from hypothesis structure or bottleneck size. In our numerical model we easily replicate this result; for all choices of priors, hypotheses and bottlenecks we find that the (approximation) of the stationary distribution (approximately) matches the prior (table 1).

	analytical			numerical		
	$h_1$	$h_2$	$h_3$	$h_1$	$h_2$	$h_3$
S1	.7	.2	.1	.6946	.2040	.1014
S2	.7	.2	.1	.7095	.1939	.0966
M1	.33	.33	.33	.3310	.3327	.3363
M2	.83	.15	.03	.8167	.1553	.0280

Table 1: Analytically and numerically determined stationary distributions for 4 conditions of the monadic BILM. S marks sampler, M maximizer, and the 1 and 2 give bottleneck sizes. Prior = [.7 .2 .1], hypotheses as in figure 1; 10000 iterations.

One interpretation of this result is that, in the end, the features of the languages of the world are entirely determined by the biases of the learner – but we will see that this interpretation is incorrect. (Kirby, Dowman, & Griffiths, 2007) emphasized that in the maximizer condition, iterated learning will not converge to the prior. They describe the maximizer stationary distributions as amplifying the structure of the prior, although they are careful to note that the exact distribution will depend on the hypothesis structure. We find that it is in fact easy to choose parameters such that the stationary distribution is entirely determined by the hypothesis structure and prior play no role, as shown in the third line of table 1. In figure 3 we show an example of the complex interaction between priors and hypothesis structure under the maximizer condition.

## Polyadic iterated learning

In our model, each agent in a *polyadic* run sees the same data string, separately calculates its posterior values, chooses its own hypothesis, and generates its own data. The data from all agents of the same generation are then concatenated into one unified data string, which is given to the next generation as their input. When the population size is set to any number  $n$  and the number of data samples is set to  $m$ , the bottleneck size is  $b = n \times m$ .

The polyadic and monadic configurations thus differ in one respect: the data string that is passed between generations is not stochastically generated from one single hypothesis, but from multiple. This has different consequences for the maximizer and the sampler models. We find that polyadic maximizers behave almost exactly as monadic maximizers. Polyadic samplers, however, show markedly different behavior. In particular, samplers in this condition do not converge to the prior. This seems to be in contradiction with the mathematical result in Kalish et al. (2007), who show that their single-agent results can be generalized to multi-agent populations, where the stationary distribution will continue to mirror the prior.

It is worthwhile working out the reasons for this difference. First consider a homogeneous, polyadic maximizer model. The behavior of all agents in the population is identical: because all agents receive the same data string and have identical priors and hypotheses, the posterior of all agents will be the same. Maximizers will all choose the same hypothesis, because this choice is based on the maximum value of their identical posteriors. The only exception is when there are multiple maximum values in their posterior. In this case, they each choose one of the maximum value hypotheses randomly, with equal weight. This situation generally only arises when there is no bias in the prior values (to help diversify the posterior values). Aside from this exception, multiple maximizers producing  $m$  samples, is equivalent to a single maximizer producing  $n \times m$  samples. Therefore, maximizer dynamics due to population size are identical to the dynamics due to a larger bottleneck.

For a homogeneous, polyadic sampler model, the dynamics are very different. Because samplers choose their hypotheses weighted by their posteriors, a homogeneous population will not choose the same hypotheses. Therefore, the actual data the next generation receives, may come from individuals with different hypotheses. This implies that the data is a sample from the *product* of the distributions defined by the hypotheses of those individuals. The actual probability distribution over possible data strings is therefore not among the set of hypotheses entertained by our learners!

This has interesting implications concerning the *Bayesian rationality* of the agents. In general, when a string of data is generated from a set of likelihoods which learners are not explicitly given, then they are not longer perfect Bayesian reasoners. This is exactly the case with a multi-population of samplers. When a data string is generated from 2 different

hypotheses, these probabilities do not conform to the likelihoods as defined by any of their hypotheses. The result is, for a multi-agent population of samplers, the stationary distribution no longer mirrors the prior (figure 4).

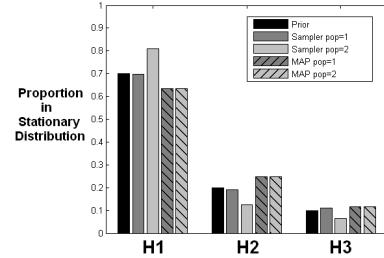


Figure 4: The maximizer model's stationary distribution is invariant to population size. For samplers, population size does affect the dynamics and the stationary distribution no longer mirrors the prior. Stationary distributions for populations 1 and 2, for maximizer and sampler models with: prior [.7 .2 .1], hypotheses [.8 .1 .1; .1 .8 .1; .1 .1 .8], 10,000 generations.

Hence, the mathematical proof of (Griffiths & Kalish, 2005) requires, in practice, that each sampler is given a new set of hypotheses, for each corresponding population size, where each hypothesis represents the combined likelihood set for each possible combination of hypotheses that the agents of the population may have when outputting into the data string. This set is different when learners learn from 1, 2, 3, ..., 100 individuals, but the proof requires a perfect hypothesis space.

The stationary distribution of polyadic samplers are a function of priors, hypotheses and population size. Figure 5 (left) shows that the stationary distribution mirrors the prior less and less as the hypothesis structure becomes strongly peaked and the prior more biased. However, for a combination of relatively flat hypotheses and weakly biased priors, the stationary distribution still mirrors the prior. Furthermore, increasing the population size systematically amplifies the effect of the likelihoods on the sampler's stationary distribution (figure 5, right). The figure also shows that the stationary distribution reflects hypotheses structure in the absence of a prior bias.

## Heterogeneous iterated learning

Finally, we implemented a heterogeneous ILM by taking a polyadic model and assigning different prior vectors to each of the agents (figure 2). This model, therefore, is the most complex of all models constructed. The main result is that heterogeneous agents' hypotheses choices converge as they are allowed to share more and more data, despite having fixed and different priors from each other. In the case of heterogeneous samplers, we find that the stationary distribution is simply the average of the stationary distributions of the corresponding homogeneous sampler runs. In the case of maximizers, however, the heterogeneous stationary distribution is not a simple average, and depends on all hypothesis structures and priors (details in Ferdinand & Zuidema, 2008).

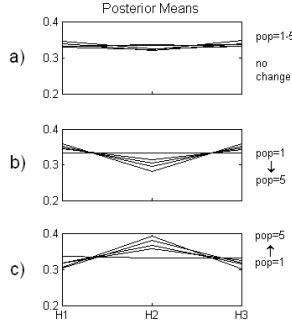


Figure 5: Population size amplifies sampler sensitivity to hypotheses structure. Hypotheses: a [.8 .1 .1; .1 .8 .1; .1 .1 .8], b [.4 .3 .3; .1 .8 .1; .3 .3 .4], c [.8 .1 .1; .3 .4 .3; .1 .1 .8]; prior = unbiased. Population sizes n=1 to 5.

## Discussion

This Bayesian ILM both replicated the general properties of previous existing Bayesian ILMs and provided new results regarding multi-agent populations and bias heterogeneity. The replications are that a monadic sampler's stationary distribution always mirrors the prior bias of the agent (Griffiths & Kalish, 2005), and a maximizer's stationary distribution is determined both by the prior and the hypothesis structure (Kalish et al., 2007). Also, for a range of parameters, the prior has no effect on the maximizer stationary distribution (Smith & Kirby, 2008), but above this threshold, iterated learning amplifies this bias (Kirby et al., 2007). Additionally, a strong bottleneck effect was observed, with the general effect of increasing transmission fidelity of both the sampler and maximizer models (Kalish et al., 2007). All of these replications attest to this particular implementation as a valid Bayesian iterated learning model.

Throughout the replication work, new insights into the role of data likelihoods for both the maximizer and sampler were obtained. The focus of all previous research with Bayesian ILMs is the on the prior and how its manipulations affect the stationary distribution. Kalish et al. (2007) manipulated the degree of hypotheses overlap, as well as noise level, but it appears their hypothesis structures were all symmetrical. Smith & Kirby (2008) demonstrate that the maximizer strategy of hypothesis choice is evolutionarily stable over that of samplers. However, the maximizer parameters for which this result was proven, it seems, were again derived from a symmetrical hypothesis structure and for the range of priors which are unaffected by bias strength. The result is consistent behavior of the Smith & Kirby's maximizer model over the bias values they selected. Our results show that this is a subset of maximizer behavior and that unstable behavior is easily obtained for the right relationship between hypotheses and priors. So, perhaps maximizer would not be the evolutionary stable strategy in all cases, and a certain range of maximizer parameters is evolutionary stable over other maximizer parameter sets. Knowledge of this kind would help guide the choice of max-

imizer parameter sets to use for iterated learning simulations.

Our main result, however, was obtained from a manipulation in population size. By just increasing the population size to 2, the sampler model's stationary distribution does not strictly mirror the prior. One response to this finding could be to claim that we have made a mistake by not giving our agents the right set of hypotheses. Indeed, in a Bayesian rational analysis it is the task of the modeller to choose hypothesis structure and prior carefully such that it objectively represents the uncertainty the learner faces. The learner doesn't have to actually use any of the Bayesian apparatus; it might rely on completely different heuristics that have somehow (e.g., by natural selection) been optimized to *approximate* the Bayesian rational solution (i.e., it is a *normative* model). However, if this line of attack is chosen, it should apply to all cases where iterated learning does not converge to the prior (and we listed many). In the stationary distribution, the objective prior probability for learners is the stationary distribution. A Bayesian rational analysis of iterated learning that predicts non-convergence to the prior is thus inconsistent.

The alternative response is to acknowledge that the prior probabilities and the hypothesis structure (i.e., likelihoods) together instantiate the subjective *biases* of the individual agents. The Bayesian inference in that case should be interpreted as a *descriptive* model, and the components of the Bayesian apparatus must somehow correspond to cognitively real processes. We have seen that convergence to the prior only occurs in the exceptional case that (i) learners have perfect knowledge of the possible sources of the data they are learning from, (ii) they carry out the Bayesian probability calculation perfectly and (iii) they then *sample* from the posterior. In all other cases – which include human language learning, we suggest – predicting stationary distributions is in fact a difficult task that depends on many variables. The results in this paper form a further contribution to developing the computational tools for that task.

Hence, Bayesian Iterated Learning does not actually give us what it was intended to: an explicit and complete quantification of an agent's bias. When we look at the cultural transmission system as a sum of the agents' behavior and the way they interact (the social structure) we see the prior does not quantify the whole agent's bias. We see that not only does the hypothesis choice strategy affects agent behavior and therefore the outcome of iterated learning, but hypothesis structure does the same (since we show an effect of different hypothesis structures on the stationary distribution). If we attribute these components to a cognitive agent, they should all fall under the agent's bias (innate or learned) if the bias is everything that the learner brings to the task - so everything besides the data itself.

Moreover, the particular dynamics which previously thought to differentiate the sampler and maximizer models, may not be as clear cut as they seemed and non-convergence to the prior cannot be taken as direct evidence against the sampling strategy or support for the maximizer strategy.

These results are bad news when trying to interpret iterated learning experiments with human and animal subjects, aimed at connecting their learning biases to the outcomes of iteration (Kalish, Griffiths, & Lewandowsky, 2007; Kirby, Cornish, & Smith, 2008)

manipulation	monadic		polyadic		heterogen.	
	S	M	S	M	S	M
prior ( $b$ is small)	+	-	-	-	-	-
prior ( $b$ is large)	+	+	+	+	+	+
hypothesis structure	-	+	+	+	+	+
population size			+	-	+	+
homo/heterogeneity					-	+

Table 2: Summary of manipulations and effects. S and M mark samplers and maximizers. A ‘+’ indicates that the (average) stationary distribution will change when the given manipulation is applied.

However, on a more positive note, we find that the stationary distributions in the many different conditions – monadic or polyadic, samplers or maximizers, homogeneous or heterogeneous – are selectively responsive to particular manipulations of the iterated learning regime. In table 2 we give a summary of our results; the table indicates, for instance, that if a population is assumed to be homogeneous, then *change in population size* is the relevant manipulation to distinguish between maximizers and samplers. Similarly, in monadic iterated learning, a change in hypothesis structure should only result in a different stationary distribution when learners are maximizers.

## Conclusions

Children learn their native language efficiently and accurately from very noisy data provided by earlier generations. Understanding how they accomplish this task, and why it is so difficult to mimick their learning in computer models, is a key challenge for the cognitive sciences. In recent years, two approaches for solving bits of the puzzle – Bayesian inference and iterated learning – have yielded many new insights. In a series of papers, Griffiths, Kalish and co-workers have combined the strengths of Bayesian modelling and iterated learning (Kalish et al., 2007; Kirby et al., 2007), and developed a mathematical framework for understanding the dynamics of cultural transmission and the roles of learning and innateness in shaping the languages we see today. The best known result from this work so far, is that “iterated learning converges to the prior” (Griffiths & Kalish, 2005).

In this paper, we have extended this work by investigating the outcomes of iterated learning under a number of manipulations, including an increase in population size and heterogeneity of biases. However, by detailing several apparent counterexamples to the “convergence to the prior” results, our results also point to some caveats in the interpretation of those earlier results and of Bayesian models of cognition more generally. Perhaps due to the mathematical sophistication of such

models, these caveats are often overlooked. In particular, a *Bayesian rational analysis* is not applicable here. In such an analysis, we (i) take the *learning problem* as fixed, and (ii) calculate the ‘rational’ strategy that would allow children to solve it in an optimal way, and (iii) assume that the actual learning strategy approximates the rational one. The latter (iii) is a very strong assumption in general, but particularly problematic in iterated learning where (i) is violated.

A consistent interpretation of Bayesian Iterated Learning must therefore take the BI as an accurate description of individual learning. The ILM takes (i) the *learning strategy* as fixed, and studies (ii) how the learning problem – the language – changes over generations, and, while it becomes better fitted to the learning strategies, (iii) gives rise to language universals and variation. However, under this interpretation, Bayesian rationality cannot be assumed and convergence to the prior is rather the exception than the rule. This leaves the original linguistic challenge – how are language universals and learner biases connected? – still unsolved, although we have progressed yet another step in developing the theoretical and experimental paradigms for addressing that question and established that cultural transmission really adds something to the explanation.

## References

Deacon, T. (1997). *Symbolic species, the co-evolution of language and the human brain*. The Penguin Press.

Ferdinand, V., & Zuidema, W. (2008). *Language adapting to the brain: a study of a Bayesian iterated learning model*. (Tech. Rep. No. PP-2008-54). University of Amsterdam.

Griffiths, T. L., & Kalish, M. L. (2005). A Bayesian view of language evolution by iterated learning. In *Proc. of the 27th annual conference of the cognitive science society*.

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

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: formal and computational models*. Cambridge University Press.

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

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *PNAS*, 104(12), 5241-5245.

Niyogi, P. (2002). Theories of cultural evolution and their applications to language change. In T. Briscoe (Ed.), *Linguistic evolution through language acquisition: formal and computational models*. Cambridge University Press.

Smith, K. (2002). The cultural evolution of communication in a population of neural networks. *Connection Sci.*, 14(1), 65-84.

Thomas, W., & Thomas, D. (1928). *The child in america: Behavior problems and programs*. New York: Knopf.

Zuidema, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15 (proceedings of nips'02)* (p. 51-58). MIT Press.