Language adapting to the brain: a study of a Bayesian iterated learning model

Vanessa Ferdinand
Institute for Interdisciplinary Studies
University of Amsterdam

and

Willem Zuidema
Institute for Logic, Language and Computation
University of Amsterdam

Abstract

What is the mechanism that translates the individual properties of learners into the properties of the language they speak? This paper will investigate cultural transmission as this mechanism and will take up the Iterated Learning Model as a formal framework in which to address this claim. This model describes language as a special learning problem, where the output of one generation is the input for the next. Previous research has shown that universal properties of human language emerge from the process of cultural transmission. However, particular biases are also necessary to obtain these properties, and the exact interplay between individual biases and cultural transmission is still an open question. In the present research, a computational, Bayesian iterated learning model is constructed to analyze the relationship between learning biases and what additional structure cultural transmission adds to language.

1 Introduction

The language that you speak is not a product of your mind alone. As language is transmitted from person to person and generation to generation, it adapts to the minds it propagates through, and they adapt to it. This makes the evolution of language a special problem, because the output of one learner is the input for the next.

The field of linguistics has made great advances in describing human language. Through the description of language universals and from animal comparative studies, we have a good picture of what human language is, and what it is not. Influential paradigms in 20th century linguistics, such as the generative program, concentrate on in-depth studies of particular languages or the variations and constraints on variation found in the world's languages, to infer what the innate biases of human learners must be. Notwithstanding fundamental differences, most of these research programs have tended to ignore the socio-cultural and historical dimensions of language. Additionally they fail to provide an account of *how* the innate biases of individuals translate to the universals witnessed in the world's languages. This problem of linkage can be overcome by identifying the *mechanism* which translates the properties of individual learners into the properties of human language (Kirby, 1999). This paper will argue in favor of claims that cultural transmission itself, may indeed be this mechanism.

The crucial next step, where linguistics has arguably made much less progress, is to provide a mechanistic explanation of *why* language is the way that it is, and not some other way. This explanation necessitates a description at a level below language itself: What are the constraints that shape language? Where do they come from and how do they interact?

The constraints of language arise from two systems: the embodied cognitive agent and the socio-cultural system in which these agents communicate with one another. The first is the domain of cognitive science and psycholinguistics. Here, the main constraints lie in perception, processing and representation, production. How is the data constrained as it enters the cognitive system, how is it cognitively processed, and how is it constrained as utterances are produced? Some examples of perceptual biases are purely physical and constrained by the human senses, such as the range of sounds one can perceive. Others form with cognitive development, such as the phenomena of categorical perception (Lieberman et al., 1967; Kuhl, 2004). Likewise, production biases are constrained by the physical limitations of human anatomy, such as the frequency range of vocalizations and the degree of motor control we have over our vocal tracts. The processing which mediates what is perceived and what is ultimately produced includes high-level cognitive processes such as reasoning, induction, and learning, each of which come with their own biases. These processes may also be subject to constraints on how linguistic knowledge is represented in the brain; what kind of representations are possible in a network of neurons, and what kinds are not? In short, the constraints which shape a cognitive agent's production of language are both shaped by both biological evolution and individual development, which includes learning from one's environment.

The second system, how cognitive agents interact, has been pioneered by computational modeling and mathematics. The social structure characterizes how the production and perception components of the cognitive agent link up. Particular types of social structures involve different constraints on what kind of access the agents have to the external data that constitutes language. For example, a population with no generational turnover (i.e. no agents are born or die) would conceivably have a very different language than language as we know it. Or for a more intuitive example, if the future of the English language becomes confined to nothing but email communication, its developmental trajectory would be very different than if it remained a spoken language. Therefore, if we want to explain why this new "email English" is the way that it is, and not some other way (like the old spoken English), we would have to describe the constraints of email English – in terms of the constraints of its social system (the network of computers and how this shapes human interaction) and in terms of the cognitive constraints which the new system engages (such as production biases associated with typing).

In this light, language is a complex, dynamical system in its own right. This means that the behavior of the system is a product of both its components (the embodied cognitive agents) and how they interact (their social system). However, as stated in systems theory, no systems have true boundaries, and the borders we impose when we study them are purely artificial constructs (Weisbuch, 1991). There are multiple ways to carve up the systems and their constraints in order to guide our search. The most common delineation, among those who computationally model language evolution, is that language sits at the crux of 3 complex, dynamical systems: biological evolution, cultural transmission, and individual learning. (Christiansen & Kirby, 2003). This tripartite division of these separate, but interacting, systems is misleading because it implies that evolution acts directly on learning as an adaptive system. This view essentially deletes cognition from the picture, because it is the embodied cognitive agent that ultimately roots its high-level process of language induction within the biologically evolved wet-ware that is the true processor of language.

By viewing language as a product of cognitive agents and the cultural transmission system which propagates it, we would expect the constraints of language to be rooted in these two aspects. However, organizing the problem in this way has the side effect of losing any direct linguistic consequences of biological evolution within the embodied cognitive agent, and rightly so. Undoubtedly, the biological endowment which makes us human places hard constraints on the possibilities of ontogenetic development. But the structure of cultural transmission is in the position to place additional constraints on this biological potential, father defining language into its ultimate form, as we witness it in the world. The fact that human language is culturally transmitted is just as universal to our language system as is the shared genetics which makes all humans, human. Though logically, this places both as candidates for the explanatory burden of language universals, the most informative explanation will be the one that cuts language the closest.

So how do the properties of cognitive agents determine the properties of the language they speak, and what does cultural transmission add to this explanation? A good way to proceed with this question is to create a formal framework for testing hypotheses about how cultural transmission mechanistically translates the properties of cognitive agents into the properties of human language, and whether or not the dynamics of this

cultural transmission place additional constraints on the ultimate form of language. This paper will take up one such framework, the iterated learning model, in order to formally address the socio-cultural constraints on human language.

1.1 The Iterated Learning Model

The iterated learning model (ILM) was first formalized for the study of language evolution by Kirby (1998) and provides a framework for the empirical study of cultural transmission and how it effects the information being transmitted. ILMs can be implemented in a variety of ways, but they all contain these fundamental components:

1) A learning algorithm
2) Some form of information which is the input/output of the algorithm
3) Structured transmission of the information, where the output of one learner serves as the input for the next.

Some learning algorithms commonly used in ILMs are symbolic grammar induction algorithms (Brighton & Kirby, 2001), neural networks (Smith, 2002), Bayesian agents (Kalish et al., 2007), and human subjects (Cornish, 2006; Griffiths et al., 2006). The data can be linguistic input or numerical values and the transmission format could be any conceivable social structure, but is commonly kept to a parent-child chain for analytical ease.

Possibly, the first study of this kind was Bartlett's (1932) psychological experiment in "serial reproduction". A subject would be shown a picture, for example a nice sketch of an elk, and then be asked to re-draw it from memory. Then, this copy would be given to another subject to re-draw, and so on. Over the course of this serial reproduction, the information present in the elk would change. The shading would disappear, the complexity of the antlers would diminish, until all that was left was the outline of a cat. Although this is a nice illustration that information can be shaped by the very process of its transmission, Bartlett's stimuli, pictures and stories, were not controlled and therefore do not lend themselves well to empirical study.

The first computational ILMs were developed by Hare & Elman (1995), Batali (1998), and Kirby (1998) as computer programs of agents in a simulated population. Here, agents were simple language-learning algorithms that paired meanings with strings of letters, and one agent would learn its language from another. Their result is that the signal-meaning system became increasingly regular as it passed through more and more agents. The regular structure which emerged was also compositional, where specific letters or letter combinations designated specific parts of the overall meaning, as words do in human language. However, these effects only occurred when there was a transmission bottleneck. This means that the agents cannot pass their language on in totality to the next generation. Humans have an infinite capacity for linguistic expression, however, we can only express a finite amount of linguistic utterances. The transmission bottleneck mirrors this by limiting the number of signal productions to below the number of possible meanings in the meaning space. Only under a specific range of bottleneck size do regularity and compositionality emerge. Many ILM studies which followed these, each using a different learning algorithm and different assumptions regarding the signal-meaning spaces, consistently reported the

same result; the emergence of regularity and compositionality due to the learning bottleneck .

For a concrete example of regularity emerging from a bottleneck, we can look at the English past tense, which has both regular (verb+ed) and irregular (go – went) past tense rules. A regular rule is also a general rule, which is applied every time a language learner uses the past tense of a regular rule. Irregular rules, on the other hand, have to be learned one by one, when the learner comes into contact with the irregular verb. Looking at regular and irregular rules separately, regular rules have a much higher chance of being transmitted to the next generation when the bottleneck is small, because they apply to more verbs and therefore have a higher chance of being produced. Irregular verbs, on the other hand, can only survive over the generations when the verbs they apply to are high frequency verbs (Kirby, 2001). In fact, this is exactly the case with the English past tense; the top 10 most frequent verbs are all irregular. Additionally, it is well-documented historically that low-frequency irregular verbs in English are gradually adopting the regular rule (Lieberman, et al., 2007).

The ILM research demonstrates that, through cultural transmission and the constraint imposed by the bottleneck, the information in language compresses in a self-organizing way (Brighton et al., 2005). Additionally, the language itself adapts to become learnable by the agents which transmit it, and not the other way around (Zuidema, 2003). Agents can only produce what they were able to learn, and when all agents in the population are similar, this makes the task easier for the next agent in the transmission line. Some of the hard claims of ILM proponents are that cultural transmission inevitably leads to regularization, an increase in learnability, and compositionality. In most ILM implementations, no biological criterion of fitness is imposed which selects agents according to the goodness of their language use. Thus, regularization, learnability, and compositionality are all claimed as properties of linguistic evolution, and not biological evolution.

The fact that diverse learning algorithms all produce similar results when iterated, shows that these results are most likely due to the properties of the iteration, and the bottleneck effect, rather than to something inherent in the learning algorithms. However, every learning algorithm has its bias and it is still possible that all of the learning algorithms that were used do share some bias which allows for the emergence of regularity and compositionality. It is possible that some learning algorithms are structured in such a way that they cannot support compositional behavior. In this sense, the bias of a learning algorithm defines what behaviors it can and cannot yield, as well as what behaviors its structure encourages. Smith (2003) carried out a comparative study of the ILM algorithms and determined that they do share two basic biases: a bias toward one-to-one signal-meaning mappings and a bias toward exploiting regularities in the input data. Therefore, these two biases can be seen as two components of the learning algorithm's structure which are necessary for the algorithm to display compositionality. And indeed, these are two biases which human learners likely bring to the task of language induction (Pinker, 1984).

This raises the question of how much the of outcome of iterated learning is determined by cultural transmission and how much is determined by the biases. On the one hand, if the process of cultural transmission completely determines the

outcome of iterated learning, we could expect to see the same results for learning algorithms which have nothing in common. Additionally, we might even expect these properties to hold for data compression algorithms which have no plausibility whatsoever as a cognitive model of language learning. Take, for example, this toy model of an interpolation algorithm which transmits data over a bottleneck of 5 data points (Figure 2). Even here, the function which describes the data increases in regularity and stability with more and more iterations. However, compositionality was not obtained this model. It is even hard to say what compositionality would look like in terms of this model's capabilities. Clearly, all ILMs do not universally yield compositionality. Therefore, we still must need a certain type of bias to obtain compositionally through iterated learning.

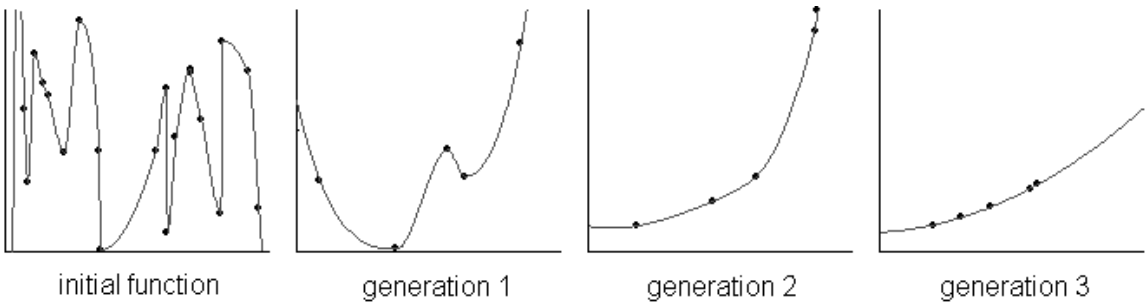

Figure 2

A toy ILM we constructed with a linear interpolation learning algorithm. An initial function is randomly generated. 5 data points, randomly selected from the initial function, serve as input to the first generation. The agent at generation 1 describes these data points with an interpolation function. Next, 5 data points are randomly selected from generation 1's function to serve as input for the next generation, and so on. The function used to describe 5 data points becomes less complex as it is iterated, and will probably stabilize as a linear function.

Unfortunately, in this interpolation model, as with many other learning algorithms, it is difficult to assess exactly what its biases are. What is needed, then, is a model with an explicitly-coded learning bias, so that different outcomes of iterated learning can be attributed to specific manipulations of the bias. Fortunately, Bayesian statistics provides this framework.

1.2 Bayesian Iterated Learning Models

For the readers who are unfamiliar with Bayesian statistics, we will introduce this topic with a practical example:

Picture yourself walking down a street in Amsterdam. Someone bikes past you and you catch a half-second clip of their voice. What language were they speaking? To come to a conclusion, a Bayesian-rational person would take into account three things. First, the candidate languages. For simplicity's sake let's just say you have three hypotheses: Dutch, Arabic, and English. Second, the data: a nice velar fricative. Third, the prior probability: what is the chance someone in that neighborhood would be speaking any of those three languages? The likelihood that Dutch or Arabic would produce a velar fricative is astronomically higher than for English. However, if you are anywhere near the tourist information center, you may just as well conclude than an English-speaker was clearing their throat. Likewise, knowing if you were in a Dutch or a Moroccan neighborhood would break the tie in the data likelihood of the velar fricative. Additionally, the prior knowledge each person brings to an inductive problem can be different. If you happened to be one of those people still in line at the tourist information center, you might think that everyone in Amsterdam speaks Dutch, and therefore you would probably classify most people as Dutch-speakers at first sound-byte.

By combining your knowledge of the data's likelihood with the prior probability of each hypothesis, you will come to a solution. This solution is the posterior probability of each hypothesis now, *after* you have finished reasoning. Last, you will select your answer in light of these posterior probabilities, choosing the hypothesis with the highest posterior probability, if you're smart.

Thus, the components of a Bayesian inference algorithm are:

1) The hypotheses and the data likelihoods which accompany them
2) The prior probability of each hypothesis: the bias
3) The posterior probability of each hypothesis

As you can see, this is no longer a problem specific to language learning. The investigation of iterated learning in terms of Bayesian agents brings the question of which adds more, biases or cultural transmission, to a new, abstract level. Griffiths & Kalish (2005) were the first to use a Bayesian ILM to address this debate and they found that the outcome of iterated learning was *completely* determined by the prior probability of each hypothesis. Here, this outcome is represented by the proportion that each hypothesis was chosen over the course of the ILM when run to infinity. Clearly, this outcome of iterated learning must be determined analytically. This resulting distribution of hypotheses choices constitutes a stationary distribution, which represents the outcome of iterated learning (Nowak et al., 2001).

The Griffiths and Kalish result showed that the stationary distribution over hypotheses exactly mirrored the prior probabilities of those hypotheses, regardless of specific prior distributions or other parameter manipulations. In particular, manipulating the bottleneck parameter had no effect whatsoever on the stationary distribution. With this, they determined that cultural transmission does not make an independent contribution to the outcome of iterated learning and it is merely a vehicle which

reveals the inductive bias of the learners. However, this result doesn't make much sense given that the bottleneck effect is robust in the many previous ILM simulations.

To counter this claim, Kirby et al. (2007) showed that this result was a consequence of the particular hypothesis choice strategy that was implemented; sampling. An agent that samples randomly chooses a hypothesis, weighted by the posterior probability of each hypothesis. This is known as probability matching in the psychological literature. Conversely, Kirby et al. showed that a Bayesian ILM does not converge to the prior when agents are maximizers, who always choose the hypothesis with the highest posterior probability. Thus, the main question seemed to be whether humans are maximizers or samplers. So, Smith & Kirby (2008) extended their model to include biological evolution and showed that the maximizing strategy is evolutionarily stable over sampling. They concluded that natural selection favors agents whose behavior can be affected by cultural transmission, so that cultural transmission is the primary determiner of linguistic structure. They also asserted that real human behavior probably lies somewhere on a continuum between maximizing and sampling, and should be subject to a more fine-grained analysis.

At first glance, the initial Griffiths & Kalish results could be understood as confirming linguistic nativism; that the ultimate structure of language is determined by our innate biases and nothing else. However, the prior probability in the Bayesian model does not correspond *only* to the learner's innate bias. In this simplistic model of a cognitive agent, the prior represents all properties of the inductive task besides the data itself. Therefore, the prior is everything the agent brings with it to the task; its innate biases, its learned biases, previous domain-specific experience, and even its affective state at the moment of induction.

In light of their own findings, Griffiths & Kalish also propose that ILMs using human subjects can serve as a tool for revealing inductive biases, especially in cases where researchers have little a priori knowledge about what these biases might be (2006). They support their claim in two different experimental tasks where the associated inductive biases are well-established by previous psychological experimentation. In both of these experiments, one in category learning (2006) and another in function learning (2007), the known inductive bias was revealed through iterated learning. However, this method should not be understood as a way to reveal innate biases, for the same reason the prior, as characterized by Bayesian induction, should not be seen as representing only the innate bias. The biases which are revealed by human ILMs are likely to be task-specific, variable with training, and could be subject to priming and context manipulation.

In this paper, we will construct a new Bayesian ILM in order to investigate the different claims about how biases determine the outcome of iterated learning and what cultural transmission adds to this outcome. Section 2 presents our implementation of a Bayesian ILM. This model will investigate the differential behaviors of maximizers and samplers under identical conditions, including how each responds to particular parameter manipulations regarding biases, data likelihoods, population size, and heterogeneity.

Lastly, we would like to add a note about the methodology used in the model analyses. We have chosen to approach this model with an empirical, rather than an

analytical, stand point. By empirically dissecting this model, we are able to provide some deeper insights into the inner dynamics of the Bayesian ILM than some earlier analytical dissects did. Some of the dynamics we have chosen to explore are simply invisible to mathematical descriptions that focus on the limits of model behavior and the cumulative end states of iterated learning when extrapolated to infinity. With this methodology, we will attempt to draw a more complete picture of the mechanisms which drive the model's behavior. Many aspects of the model we will describe in the following section certainly have straightforward analytical solutions which we have not entertained, however our goal here is to set forth a bridge between empirical research on iterated learning systems and their analytical description. Hopefully, this paper will be equally informative for cognitive scientists and mathematicians alike, who may want to continue the work we set forth here.

2 A Bayesian Iterated Learning Model

2.1 Outline

In this section, we will describe the implementation and results of our own model of iterated learning with Bayesian agents. Here, two models are constructed; one with agents who choose their hypothesis by sampling and one with agents who choose by maximizing. Agents use Bayesian inference to produce and induce from data, which is passed between agents across discrete, serially-organized generations. A variety of parameter settings and their effect on the model's behavior will be investigated. This investigation both replicates recent Bayesian ILM results and addresses new hypotheses regarding population size and heterogeneity.

Section 2.2, Model Description, will outline the components and structure of the model and describe the parameters which will be manipulated. Section 2.3, Model Analyses, will describe the analytical tools commonly used in the existing Bayesian ILM literature, to assess model behavior. Here, we will also justify the use of several approximations for these solutions, which are obtained from the model simulations. These experimentally-obtained assessment tools will serve as the basis for this research's model analyses. Section 2.4, Model Results, will describe both the sampler and maximizer model's behavior for a number of parameter manipulations regarding the prior and likelihoods, the bottleneck effect, population size, and heterogeneity of priors. In conclusion, section 2.5 will provide a general discussion of the modeling results.

2.2 Model Description

We will first outline the components of a Bayesian ILM and describe how they are implemented in this model. The implementation and simulations were all carried out in Matlab; the model's code is in Appendix A.

2.2.1 *The Hypotheses*

In this simulation, agents are considered to have a small set of hypotheses about the state of the world, and each of these hypotheses assign different likelihoods to each of

a small set of observations that the agent can make about the state of the world. These hypotheses could represent, for instance, different languages that generate a set of utterances, or different functions that describe a set of data points. However, the exact nature of the hypotheses is left under specified, in order to investigate the general dynamics inherent to Bayesian iterated learning. Thus, the basic properties of the model might be generalizable to a variety of systems where information is culturally transmitted, such as language and function learning, where Bayesian inference serves as a good approximation of the learning mechanism involved. In this model, the hypotheses are set at the beginning of each simulation and all agents have this set of specified hypotheses. For simplicity of analysis, each hypothesis is *completely defined* by the likelihoods it assigns to each observation. Additionally, the number of hypotheses will be restricted to three, and each called H1, H2, and H3 (Figure 2.1). Also, any particular combination of these three hypotheses will be referred to as the "hypotheses structure."

2.2.2 *The Data*

The observations that the agent can make about the state of the world will be referred to as data points. These will also be restricted to three and called d1, d2, and d3 (Figure 2.1). The information that the agents pass between each other is a vector of one or more of these three data points. As will be described in section 2.2.4, the number of data points in this vector defines the "transmission bottleneck."

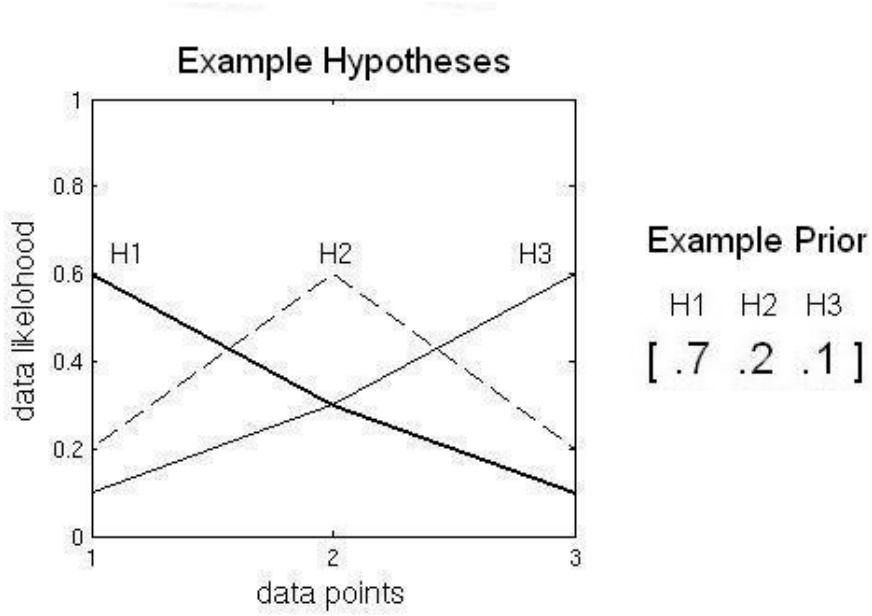

Figure 2.1

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

2.2.3 *The Prior*

The prior probability of each hypothesis is stored in a 3-unit vector, where each entry lists the prior of each hypothesis. The shorthand for and example prior is [.7 .2 .1], showing the prior of H1, H2, and H3 respectively. The difference between the highest and lowest probability create the bias strength. In this example, the bias

strongly favors H1. These probabilities of the prior vector sum to one, indicating that these are the only three hypotheses which can generate or account for the data.

2.2.4 *The Social Structure*

In this model, agents are defined by the process of Bayesian induction from data, hypothesis choice, and data generation. Agents are organized into discrete generations of one or more. When each generation consists of 1 agent, the simulation can be characterized as a Markov chain and is identical to previous ILMs where one adult transmits data to one child. When each generation consists of x agents > 1, each agent will output an equal number of data points into the data vector, and this entire vector will serve as the input to each of the agents in the next generation.

2.2.5 *Bayesian Inference*

Agents both induce from data and produce data according to the likelihood values of their hypotheses. The particular likelihood values of one hypothesis determines the composition of the data string it is likely to produce. For example, H1 will produce d1 70% of the time, d2 20% of the time, and d3 10% of the time. Therefore, a characteristic, 10-sample data string for each hypothesis in figure 2.1 might look like:
H1 → [1 1 1 2 1 2 2 1 3 1]
H2 → [1 2 2 3 2 1 2 2 2 3]
H3 → [3 3 1 2 2 3 3 3 3 2]

When faced with a data string, such as one above, agents use Bayesian inference to decide which hypothesis was most likely to have produced it. Thus, agents use Bayes' Rule (eq. 2.1) to compute the probability that each hypothesis generated the data string:

$$P(h \mid d) = \frac{P(d \mid h)P(h)}{P(d)}$$

Equation 2.1

Here, P(h|d) denotes the posterior probability that a hypotheses could have generated the data in question. This is the outcome of Bayesian induction and is calculated for each hypothesis. P(d|h) is the likelihood value of the data under the hypotheses in question. The data likelihood values for each hypothesis are defined by the hypothesis structure (Figure 2.1). P(h) is the prior probability of a hypothesis. P(d) is the probability of the data averaged over all hypotheses.

Example calculation as implemented in the program:

hypotheses = [.6 .3 .1; .2 .6 .2; .1 .3 .6], prior = [.7 .2 .1], data string = [2 3 3]

1) Calculate likelihood
For data point = 2, the corresponding likelihood values under each hypothesis H1, H2, H3 = [.3 .6 .3]. For data point = 3, the likelihoods are [.1 .2 .6]. Assuming independence, the likelihoods of each element in the data string can be multiplied, to yield the likelihood of the string. Instead of multiplying the probabilities, the log of the likelihoods are added, to make it easier to deal with small numbers. Therefore, the log likelihood of the data string [2 3 3], is calculated as:
log [.3 .6 .3] + log [.1 .2 .6] + log [.1 .2 .6] = [-5.8091 -3.7297 -2.2256].

2) Calculate posterior
posterior = exp (log prior + log likelihood)

First, the log of the prior vector is added to the log likelihood vector of the data string: exp ([-0.3567 -1.6094 -2.3026] + [-5.8091 -3.7297 -2.2256]) = [0.0021 0.0048 0.0108] Last, the posterior vector is normalized, to obtain a probability of 1 that one of the three hypotheses generated the data. This yields: [0.1186 0.2712 0.6102]. This posterior means that H3 is most likely (61%) to have generated the data string. The next likely is H2 at 27% and the least likely is H1 at 12%.

This method of calculating the posterior yields the normalized product of the likelihood and prior and is equivalent to Bayes' rule, above.

2.2.6 *Data Production*
The next step is for the agent to output a new data string. First, a hypothesis is chosen according to the posterior probabilities. Second, the data are generated from the chosen hypothesis.

Hypothesis choice - Maximizing vs. Sampling:
There are a variety of ways in which the hypothesis could be chosen, however in this study we will investigate two cognitively-grounded strategies: maximizing and sampling. Both of these strategies choose between hypotheses according to their posterior probabilities. The maximizer simply chooses the hypothesis with the highest posterior probability. But in the event there is a tie among hypotheses for the highest posterior value, the maximizer randomly chooses between them. The sampler chooses one hypothesis randomly, but weighed by the posterior probabilities.

Example Posterior Vector

H1	H2	H3
0.12	0.27	0.61

Table 2.1

According to the posterior values in table 2.1, the maximizer will choose H3. The sampler will have a 12% chance of choosing H1, a 27% chance of choosing H2, and a 61% of choosing H3. These different strategies are implemented separately, creating

two Bayesian iterated learning models which differ only in the respect of hypothesis choice. This leads to characteristic differences in the dynamics of each model, which will be addressed in the Analysis section.

Data choice:
Data is generated from the chosen hypothesis according to the likelihood values of that hypothesis. Assuming the agent has chosen H3, each data point in the output string will be randomly generated, but weighted according to the likelihood of each data point under H3. Therefore, given the likelihood values of H3 = [.1 .3 .6], data point 1 has a 10% chance of being generated, data point 2 a 30% chance, and data point 3 a 60% chance. The next 3-value data string might look something like this: [3 2 3].

2.2.7 *Iteration*
Cultural transmission 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.

2.2.8 *Model Parameters*
A variety of parameters can be manipulated to investigate the dynamics of the system. These manipulations will be used to compare and contrast the dynamics specific to the Maximizer (MAP – maximum *a posteriori*) and Sampler models. The following manipulations that will be investigated in the present research are as follows:

1) The prior.
2) Homogeneity and heterogeneity of the agents' priors. Each agent in a population greater than 1 can be assigned a different set of priors. This is the only parameter which can be manipulated heterogeneously in the population. The remaining manipulations below hold for all agents in the simulation.
3) The hypotheses structure (the likelihood values of each hypothesis).
4) The bottleneck. How many data samples each generation produces.
5) Population size. Usually kept to 1 in the homogenous simulations and 2 in the heterogeneous simulations.

2.3 Model Analyses

2.3.1 *Overview of Assessment Methods*
Each model has a unique dynamical fingerprint. Understanding why two models work differently is understanding how their dynamics differ. Each parameter manipulation can potentially change the dynamics of the model, and depending on the properties of the model, certain manipulations can change the dynamics in a different, but systematic way. Therefore, in order to characterize each model's dynamical fingerprint, we are looking for features that are invariant to specific parameter manipulations as well as changes in the dynamics 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, or Q matrix (Nowak et al., 2001), for each model. This matrix gives the probabilities that each hypotheses will lead to itself or any other hypotheses in the next generation. In essence, all probable trajectories that an ILM might take are wrapped up in this matrix. From the Q matrix, we can also derive the stationary distribution, which is the stable outcome of iterated learning (Griffiths & Kalish, 2005; Kirby et al., 2007).

In the following sections, both the Q matrix and stationary distribution will be explained in detail, for readers who may be unfamiliar with these terms. Additionally, we will justify the use of certain experimental approximations of these two analytical tools. These approximation heuristics are readily obtainable from iterated learning simulations and are especially valuable when the computational requirements of the analytical solutions is high or simply not feasible.

The next section will walk through the analytical calculation of a couple Q matrices, as applied to the iterated learning model. Because the Q matrix defines the model's dynamics, it is important to note, during the calculation process, how each model component comes into play. These seemingly minute details will have important consequences for understanding the mechanism behind the dynamics in later analyses.

2.3.2 *Analytical and Experimental Q Matrix Calculations*

If the agent in one generation has hypothesis 1, then what's the probability that the agent in the next generation will have hypothesis 1, 2, or 3? These probabilities are displayed in the transition matrix (or Q matrix). In the example Q matrix below (Table 2.2), when an (parent) agent in one generation produces data from H1, then the probability that that data will lead the (child) agent of the next generation to choose H1 is 80%. Since parent H1 can produce data that best supports H2 or H3, then "miscommunications" occur, leading the child to induce H2 or H3 each 10% of the time.

Example Q Matrix

Q matrix		child		
		H1	H2	H3
parent	H1	0.8	0.1	0.1
	H2	0.1	0.8	0.1
	H3	0.1	0.1	0.8

Table 2.2

Analytical Q matrix for Sampler with bottleneck of 1:

The following will show the analytical calculation of the Q matrix for a Sampler model with a bottleneck of 1 data sample per generation. All calculations in this section will use the following prior and data likelihood values:

	Priors
hypothesis 1	0.7
hypothesis 2	0.2
hypothesis 3	0.1

	Data Likelihoods		
	data 1	data 2	data 3
hypothesis 1	0.8	0.1	0.1
hypothesis 2	0.1	0.8	0.1
hypothesis 3	0.1	0.1	0.8

Table 2.3
Prior and data likelihood values used for all calculations in section 2.2.2

Beginning with cell (H1, H1), we want to know how often parent H1 will produce each possible data string, and how often each of those data strings will lead to the child choosing H1. For a bottleneck of 1, there are just three possible data strings: [1] [2] and [3]. As defined by the data likelihood values of each hypothesis, a parent with H1 will produce d1 with p = 0.8, d2 with p = 0.1, and d3 with p = 0.1. Next, the probability that the child will choose H1 from each of the three data points is defined by the child's computed posteriors (table 2.4) and their hypothesis choice strategy, sampling. All posterior probabilities are computed with Bayes' rule as outlined in section 2.1.4.

Posterior Values

	H1	H2	H3
[1]	0.9492	0.0339	0.0169
[2]	0.2917	0.6667	0.0417
[3]	0.4118	0.1176	0.4706

Table 2.4

Therefore, when the sampler receives data string [1], it will choose H1 with p = .95, H2 with p = .03, and H3 with p = 0.02. To find out how often parent H1 will lead to child H1, we must multiply the probability that each data string leads to child H1 by the likelihood of that data string being generated by parent H1. Thus, the probability of parent H1 leading to child H1 is (0.9492*0.8) + (0.2917*0.1) + (0.4118*0.1) = 0.8297

Q matrix

		child		
		H1	H2	H3
parent	H1	0.8297	0.1056	0.0648
	H2	0.3695	0.5485	0.0821
	H3	0.4535	0.1641	0.3823

Table 2.5
Analytically-calculated Q-matrix for Sampler with bottleneck of 1

Q matrix

		child		
		H1	H2	H3
parent	H1	0.8231	0.1102	0.0667
	H2	0.3685	0.5422	0.0893
	H3	0.4539	0.1665	0.3796

Table 2.6
Experimentally-calculated Q-matrix for Sampler with bottleneck of 1

Experimental Q matrix for Sampler with bottleneck of 1:
For comparison, table 2.6 shows an experimentally-calculated Q matrix for the same prior and likelihood values. The experimental calculation was obtained from the model by setting the parent to one hypothesis, allowing it to generate a 1-sample data string, and simply tallying how many times the child arrived at each hypothesis over

10,000 runs. As evidenced in this comparison, and other trial calculations, this method of experimentally calculating the Q matrix reliably approximates the analytical solution.

Analytical Q matrix for MAP with bottleneck of 1:
All the steps above for the Sampler are the same for the Maximizer (MAP) except for the way the posteriors enter the equation. As opposed to the samplers, which choose their hypotheses with the probability of each hypotheses posterior probability, the MAP simply chooses the hypothesis with the highest posterior probability. Going back to the posteriors (Table 2.4), data string [1] will always lead to H1, [2] will always lead to H2, and [3] will always lead to H3. So, multiplying the probability that the parent produces each data string times the probability it will be induced under each hypotheses, simply yields the data likelihoods as defined by each hypothesis (Table 2.7).

Q matrix		child		
		H1	H2	H3
parent	H1	0.8	0.1	0.1
	H2	0.1	0.8	0.1
	H3	0.1	0.1	0.8

Table 2.7
Analytically-calculated Q-matrix for MAP with bottleneck of 1

Q matrix		child		
		H1	H2	H3
parent	H1	0.8062	0.0948	0.099
	H2	0.1016	0.7983	0.1001
	H3	0.1038	0.0954	0.8008

Table 2.8
Experimentally-calculated Q-matrix for MAP with bottleneck of 1

Experimental Q matrix for MAP with bottleneck of 1:
Again, for comparison, Table 2.8 shows that the experimentally-calculated Q matrix closely approximates the analytical Q matrix.

Q matrix calculations for Sampler with bottleneck of 2:
As more data samples are allowed, computing the analytical solution becomes quite cumbersome. This is because the data likelihood and posteriors of all possible data strings must be calculated. For a bottleneck of 2, there are 6 (order-independent) data strings. Below are the new data likelihoods (Table 2.9) and the posterior values (Table 2.10) for every possible data string.

Data Likelihoods

	H1	H2	H3
[1 1]	0.64	0.01	0.01
[2 2]	0.01	0.64	0.01
[3 3]	0.01	0.01	0.64
[1 2] or [2 1]	0.16	0.16	0.02
[1 3] or [3 1]	0.16	0.02	0.16
[2 3] or [3 2]	0.02	0.16	0.16
sum	1	1	1

Table 2.9

Posterior Values

	H1	H2	H3
[1 1]	**0.9933**	0.0044	0.0022
[2 2]	0.0515	**0.9412**	0.0074
[3 3]	0.0959	0.0274	**0.8767**
[1 2] or [2 1]	**0.7671**	0.2192	0.0137
[1 3] or [3 1]	**0.8485**	0.0303	0.1212
[2 3] or [3 2]	0.2258	**0.5161**	0.2581

Table 2.10

These new likelihoods (Table 2.9) are obtained by multiplying the likelihood values of the data points in question, as defined by each hypothesis (Table 2.3). For example, H1 produces data point 1 with p=.8, so producing it twice has the probability of p=.64. All probabilities in each column sum to one because they cover all possible data strings.

Again, the posterior values (Table 2.10) are computed with Bayes' rule. When the Sampler receives string [1 1], it will choose H1 with a 99% probability, and H2 and H3 with less than 0.5% probability each. To get each entry of the Q-matrix, all the likelihoods of each string being induced under each hypothesis must be multiplied by the likelihoods that each string is produced at all and then these values are summed. So, for parent H1 going to child H1, this value is the sum of the likelihood that each string is produced by parent H1 times the probability it will be induced as child H1: (.64*.9933)+(.01*.0515)+(.01*.0959)+(.16*.7671)+(.16*.8485)+(.02*.2258) = .9002 Table 2.11 shows the analytical Q matrix and Table 2.12 shows the experimental Q matrix for comparison.

Q matrix		child		
		H1	H2	H3
parent	H1	0.9002	0.0627	0.037
	H2	0.2197	0.7209	0.0594
	H3	0.2591	0.1188	0.6221

Table 2.11
Analytically-calculated Q-matrix for Sampler with bottleneck of 2

Q matrix		child		
		H1	H2	H3
parent	H1	0.9004	0.0635	0.0361
	H2	0.2218	0.7178	0.0604
	H3	0.2588	0.1166	0.6246

Table 2.12
Experimentally-calculated Q-matrix for Sampler with bottleneck of 2

Q matrix calculations for MAP with bottleneck of 2:
Again, calculations for the MAP differ from the sampler in terms of hypothesis choice. To obtain the analytical Q matrix, only the likelihoods of strings with the maximum posterior will be summed under each hypothesis. These are the values in bold in Table 2.10. Here, strings [1 1], [1 2], [1 3] will always lead to H1. Strings [2 2], [2 3] will always lead to H2, and string [3 3] will always lead to H3. Therefore, the probability that the data from H1 will lead to H1 in the next generation is .64 (for [1 1]) + .16 (for [1 2]) + .16 (for [1 3]) = .96. Table 2.13 is the resulting analytical Q matrix and Table 2.14 is an experimentally-calculated Q matrix for comparison.

Q matrix		child		
		H1	H2	H3
parent	H1	0.96	0.03	0.01
	H2	0.19	0.80	0.01
	H3	0.19	0.17	0.64

Table 2.13
Analytically-calculated Q-matrix for MAP with bottleneck of 2

Q matrix		child		
		H1	H2	H3
parent	H1	0.9600	0.0289	0.0111
	H2	0.1900	0.8018	0.0092
	H3	0.1900	0.1674	0.6439

Table 2.14
Experimentally-calculated Q-matrix for MAP with bottleneck of 2

2.3.3 *Analytical and Experimental Stationary Distribution Calculations*
The Q matrix summarizes the potential for transition dynamics in the system which it describes. But what can this dynamical fingerprint tell us about the outcome of iterated learning? 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 the Q matrix. This distribution is known as the stationary distribution and serves as an idealized shorthand for the "outcome of iterated learning." As demonstrated by Griffiths & Kalish (2005) and Kirby et al. (2007), the stationary distribution is proportional to the first eigenvector of the Q matrix. Therefore, the stationary distribution is easily determined for each model, by normalizing the first eigenvector of its analytically-calculated Q matrix.

In an experimental run, the relative frequency of all chosen hypotheses are also entirely determined by the Q 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 Q 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. Below are the stationary distributions for each of the analytical Q matrices from the previous section (Table 2.15). For comparison, next to each is the normalized hypothesis history of a corresponding simulation run of 10,000 generations. The normalized hypothesis history is a reliable, experimental approximation of the stationary distribution.

Stationary Distribution Approximations

	Analytical Stationary Distribution			Normalized Hypotheses History			Posterior Mean		
	H1	H2	H3	H1	H2	H3	H1	H2	H3
S1	0.70	0.20	0.10	0.6946	0.2040	0.1014	0.6992	0.2019	0.0989
S2	0.70	0.20	0.10	0.7095	0.1939	0.0966	0.7083	0.1954	0.0962
M1	0.33	0.33	0.33	0.3310	0.3327	0.3363	0.5497	0.2726	0.1777
M2	0.83	0.15	0.03	0.8167	0.1553	0.0280	0.7768	0.1665	0.0567

Table 2.15

Normalized Hypothesis History approximates the analytical stationary distribution for both the Sampler and MAP model. The posterior mean is only a reliable approximation for the Sampler model. **S1** = Sampler with bottleneck of 1, **S2** = Sampler with bottleneck of 2, **M1** = MAP with bottleneck of 1, **M2** = MAP with bottleneck of 2.

Additionally, for the Sampler only, the average of all agents' posterior values serve as a good approximation for the stationary distribution. This is because the hypotheses are chosen according to the exact proportions of the posterior vector. For the MAP, posterior mean can not be used as an approximation heuristic. MAP dynamics are not tied to the exact values of the posterior, because agents only respond to the maximum. Table 2.15 shows the posterior mean of the same simulation runs.

2.3.4 *Summary*

The Bayesian ILM of the present research can be used to experimentally determine the internal dynamics and associated stationary distribution of both Sampler and MAP models, and over a wide variety of parameter combinations. Determining the Q matrices and stationary distributions through experimental calculations and simulation heuristics provide a good alternative to computing the analytical solutions, which becomes increasingly cumbersome as the bottleneck or population size or increases. Additionally, the simulations will allow the investigation of certain parameter combinations, such as multi-agent populations with heterogeneous biases, which do not have straightforward analytical solutions.

2.4 Model Results

This section will describe the differences between the MAP and Sampler given the parameter manipulations described earlier. First it will cover replicated aspects of

previous Bayesian ILMs. Last, it will address new findings for multi-agent populations with heterogeneous and homogeneous biases.

2.4.1 *Basic Sampler Behavior of 1-agent, 1-sample simulations*

Griffiths & Kalish (2005) showed that the stationary distribution of the Sampler always mirrors the prior. This was confirmed in our model for a 1-agent population. Over all combinations of priors and hypotheses structures tested, the Sampler model's stationary distribution mirrored the prior. However, this was not the case for multi-agent populations, and will be addressed in section 2.4.4.

2.4.2 *Basic MAP Behavior of 1-agent, 1-sample simulations*

Kalish et al. (2007) find that the MAP's dynamics are effected by the prior, data likelihoods (aka: hypothesis structure), and noise. However, it is not understood exactly how the likelihoods affect the dynamics. Because our model does not investigate the effect of noise on the model's behavior, it is more readily apparent which aspects of the dynamics are due to the prior and which are due to the hypothesis structure. The following explanations of MAP behavior in terms of hypotheses structure and bias influence are novel and were informed by simulations with the present model.

From the Q matrix calculations in the previous section, it is clear that the Q matrix values of a 1-sample simulation *are* the data likelihood values for each hypothesis. This leads to consistent patterns in the stationary distribution for particular types of hypotheses structures. Overall, the hypotheses structures investigated in this model can be broken down into two main categories; canonical and asymmetrical. Canonical hypotheses structures are ones where each hypothesis is defined by the same set of data likelihood values, but shifted so that each hypothesis' peak is over a different data point. Examples of canonical hypotheses are in Table 2.16, *a*-*e*. Within a canonical hypotheses structure, each hypothesis has identical probabilities of transitioning to every other hypothesis and therefore, when there is no prior bias, each hypothesis is equally represented in the stationary distribution.

Asymmetrical hypotheses structure occurs when each of the hypotheses are not composed of the same values, and therefore have more complex transition probabilities among themselves. Examples of asymmetrical hypotheses are in Table 2.16, *f-j*. Figure 2.1 is also an asymmetrical hypotheses structure. The stationary distributions of this category of hypotheses are difficult to predict, however we have identified some general trends in the dynamics. Though, these trends may only hold for this particular model's implementation, with an equal number of hypotheses as data points. The first concerns their relative peak height . The hypothesis with the highest peak likelihood value will be represented with the highest proportion in the stationary distribution. Likewise, the hypothesis with the lowest peak will be represented the least. The second concerns their relative overlap. When all hypotheses have peaks with equal likelihood values, but one has higher extreme likelihoods than the two hypotheses, as does H2 in example *f*, it will be represented with the greatest proportion in the stationary distribution. These relationships regarding hypothesis overlap and relative likelihood values probably have straightforward analytical solutions and are open points for further analyses.

Hypotheses Structure Effect on Stationary Distribution

	Canonical Hypotheses			Normalized Hypotheses History		
	H1	H2	H3	H1	H2	H3
a	.8 .1 .1	.1 .8 .1	.1 .1 .8	0.3250	0.3360	0.3390
b	.7 .15 .15	.15 .7 .15	.15 .15 .7	0.3387	0.3352	0.3261
c	.6 .2 .2	.2 .6 .2	.2 .2 .6	0.3337	0.3269	0.3394
d	.5 .25 .25	.25 .5 .25	.25 .25 .5	0.3352	0.3310	0.3338
e	.4 .3 .3	.3 .4 .3	.3 .3 .4	0.3347	0.3304	0.3349

	Asymmetrical Hypotheses			Normalized Hypotheses History		
	H1	H2	H3	H1	H2	H3
f	.6 .3 .1	.2 .6 .2	.1 .3 .6	0.2810	0.4295	0.2895
g	.8 .1 .1	.2 .6 .2	.2 .2 .6	0.4976	0.2525	0.2499
h	.6 .2 .2	.1 .8 .1	.2 .2 .6	0.2535	0.5062	0.2403
i	.4 .3 .3	.1 .8 .1	.3 .3 .4	0.1985	0.6061	0.1954
j	.4 .3 .3	.25 .5 .25	.1 .3 .6	0.2168	0.3801	0.4031

Table 2.16

Differences in normalized hypothesis history for the two categories of hypotheses; canonical and asymmetrical. All results above were calculated with an unbiased prior.

Because these relationships have to do with the entire hypotheses structure, the effect that one hypothesis' likelihoods has on the stationary distribution always depends on its context, which is the other two hypotheses. This makes for a difficult analysis. Figure 2.2 shows the manipulation of just one hypothesis, H2, in 4 different contexts, and with an unbiased prior. Here, H2's peak is slowly raised from likelihood value 0.33 (flat/no peak) to 0.9, as shown on the x-axis. The context hypotheses structures are displayed in the columns of graphs at the sides (these graphs display the hypotheses structure as introduced in Figure 2.1). The left column shows a snapshot of the hypotheses structure in order for lines *a*-*d* at x = 0.3. The right column shows the structures at x = 0.9. For *a*, H1 and H3's peaks = 0.33 (flat), *b* peaks = 0.4, *c* peaks = 0.6, and *d* peaks = 0.8. The y-axis shows the proportion of H2 in the normalized hypothesis history. It is clear to see that raising the peak of H2, raises its proportion in the hypothesis history. However, the higher the context hypotheses, the lower the proportion of H2. Additionally, the gray line at y = 1/3 marks the point where all hypotheses are level in the hypothesis history. All hypotheses structures found at the intersection with this line are the canonical forms; where the H2 peak and context peaks are the same height.

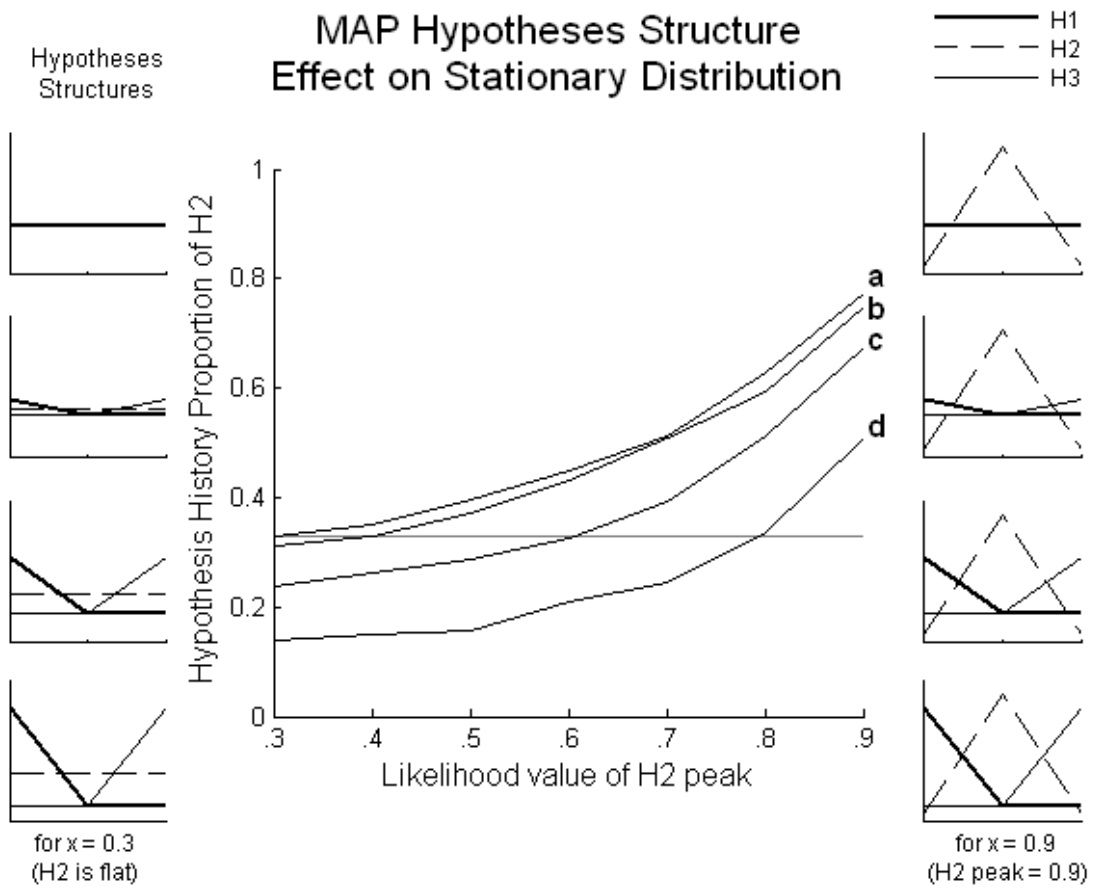

Figure 2.2

Proportion of H2 in the MAP stationary distribution as a function of H2's hypothesis peak in 4 different hypotheses structures. Peaks of context hypotheses H1 & H3 in $a = 0.33$ (flat), $b = 0.4$, $c = 0.6$, $d = 0.8$. Prior is unbiased.

The picture becomes even more complex when a bias is introduced. Figure 2.3 shows the same center graph as above, but for 3 different prior biases in favor of H2. The underlying dynamics remain the same, but the bias adds an additional layer of complexity. When the maximum prior probability is higher than the maximum likelihood value, the hypothesis which the bias favors becomes 100% represented in the stationary distribution, meaning this is the only hypothesis which an agent is able to choose. This is because the posteriors of all data strings will be maximum under the hypothesis which the bias favors. When the maximum prior probability is equal to the maximum likelihood value (indicated by the stars), the H2's proportion in the stationary distribution is raised considerably. But when the maximum prior probability is less than the maximum likelihood value, there is no change to the stationary distribution. Therefore, no manipulation to the bias, when in this range, will affect the stationary distribution. To summarize Figure 2.3, the MAP hypothesis structure plays a considerable role in shaping the system's dynamics, but when the prior is high enough, these dynamics are overridden by the bias and all agents choose the hypothesis that has the highest prior probability.

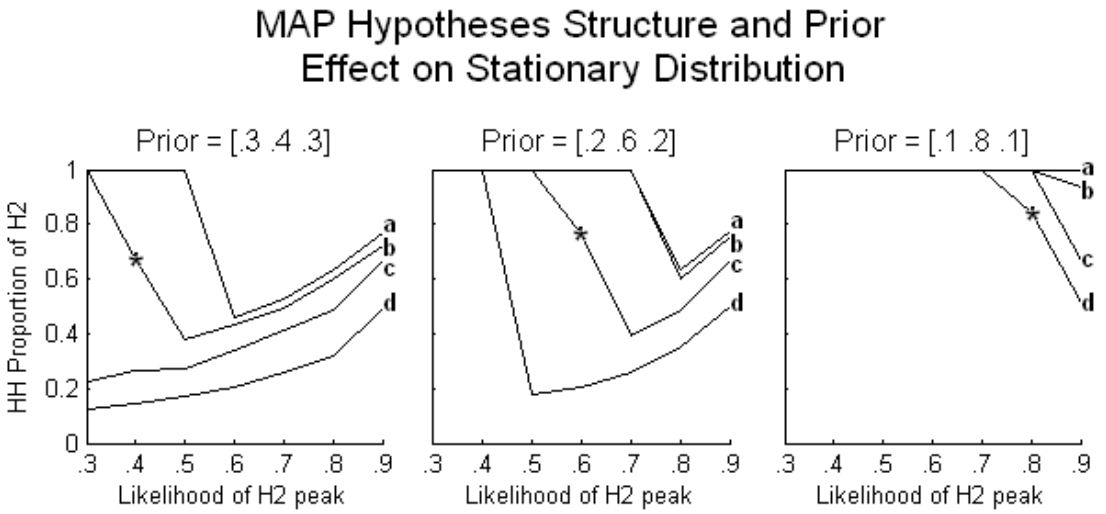

Figure 2.3
Hypotheses structure effect on the MAP stationary distribution, with added effects from prior biases.

Table 2.17 directly visualizes this threshold for line *c* (hypotheses = [.6 .2 .2; .2 .6 .2; .2 .2 .6] of the middle graph in Figure 2.3. Here, the posterior values are given for all possible data strings [1], [2], and [3]. The location of the maximum posterior values (in bold) are what determine the MAP hypothesis choice. Across this threshold, these posterior maxima make a shift, thus shifting the outcome of iterated learning for this model. When the prior value is anywhere lower than the H2 peak, as in prior = [.205 .59 .205], the dynamics remain completely determined by the hypotheses structure. However, nudging the prior up to [.2 .6 .2], which is the same level of the H2 peak, the posteriors move to favor H2 because the MAP is now faced with 2 maximum posterior values for 2 of the data strings and will choose them each 50% of the time. Finally, as soon as the prior bias for H2 exceeds the H2 likelihood peak, as in [.195 .61 .195], all posterior maxima are located under H2. At this point, all agents in the simulation will choose H2 for all possible data strings.

Posterior values under different priors

data string	Prior = [.205 .59 .205]			Prior = [.2 .6 .2]			Prior = [.195 .61 .195]		
	H1	H2	H3	H1	H2	H3	H1	H2	H3
[1]	**0.44**	0.42	0.15	**0.43**	**0.43**	0.14	0.42	**0.44**	0.14
[2]	0.09	**0.81**	0.09	0.09	**0.82**	0.09	0.09	**0.82**	0.09
[3]	0.15	0.42	**0.44**	0.14	**0.43**	**0.43**	0.14	**0.44**	0.42

Table 2.17
Hypotheses = [.6 .2 .2; .2 .6 .2; .2 .2 .6]

Basic Sampler vs. Maximizer Conclusion:
For the Sampler model, the most salient determiner of the dynamics is the prior. Although the transitions in the Q matrix are not trivially determined, the stationary distribution derived from the Q matrix exactly mirrors the prior, despite manipulations to the hypotheses structure. The MAP model's dynamics, on the other hand, are most saliently determined by the data likelihood values. For 1-agent, 1-sample simulations, the Q matrix exactly mirrors the data likelihood values as defined by each hypothesis, and standard calculus should be able to predict the stationary distribution. When the

hypotheses structure is canonical, then the probability of an agent choosing any given hypothesis in the stationary distribution is equal. When the hypotheses structure is of various, asymmetrical combinations, the stationary distribution reflects each of them differently. A prior bias adds yet more to the MAP dynamics, but only when it is stronger than the likelihood values.

2.4.3 *The Bottleneck Effect*

The number of data points that are transmitted between generations constitute the learning bottleneck. The bottleneck size, therefore, equals the number of data samples in the data string. Varying the bottleneck size directly effects the transmission dynamics. When the bottleneck is large, there is a much higher probability that the proportion of data samples in the data string faithfully reflects the likelihoods of hypothesis it was generated from. This leads to greater fidelity of transmission; where each generation usually chooses the same hypothesis as the generation before it. When very little data is transmitted over each generation, transmission fidelity is much lower, yielding many transitions between hypothesis choices within the simulation run. Transmission fidelity is directly visible in the diagonal axis of the Q matrix. A high probability of each hypothesis leading to itself equals high transmission fidelity and a lower number of transitions in the simulation run. As the bottleneck increases, transmission fidelity increases until it reaches 100% and Q matrix diagonals are all equal to 1. Depending on the strength of the bias and the distinctiveness of the hypothesis peaks, this increase occurs at different speeds (Figure 2.4). However, this rate does not seem to be affected by hypothesis choice strategy. All models will eventually reach 100% transmission fidelity at a certain bottleneck size.

In Figure 2.4, the transmission fidelity index used here is the average of the values on the diagonal of the Q matrix. This indicates the probability, for any randomly-chosen hypothesis, that the child will choose the same hypothesis. When the index reaches 1, this means that all diagonal values in the Q matrix are 1. In this case, miscommunication is impossible and every generation will have the hypothesis of the previous generation. Here, the outcome of iterated learning will be solely determined by the initial data. Therefore, the hypothesis which the initial data best supports will be the hypothesis that all generations will choose.

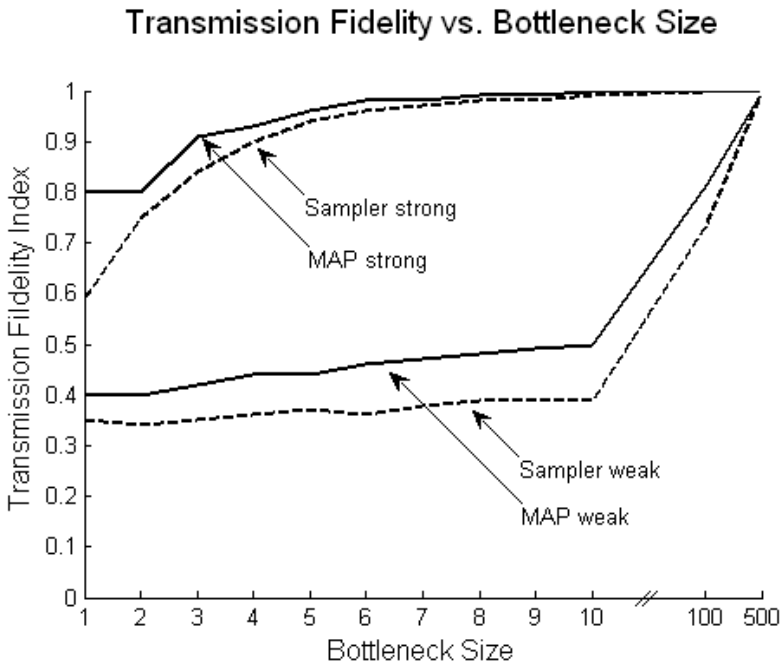

Figure 2.4

Increase in transmission fidelity is slower for models with weak biases and likelihoods. Behavior is more determined more by these factors than by hypothesis choice strategy.
Strong: prior = [.7 .2 .1] and hypotheses = [.8 .1 .1; .1 .8 .1; .1 .1 .8]
Weak: prior = unbiased and hypothesis = [.4 .3 .3; .3 .4 .3; .3 .3 .4]

For a finite number of generations, all simulations will appear to display complete transmission fidelity when the bottleneck is wide enough. This will occur when the probability of miscommunications (non-diagonal cell values) make it unlikely that they will appear within the given number of generations. For example, if one particular miscommunication has a probability of 0.01, it will usually not occur in a simulation of with less than 100 generations, but it likely to occur several times in a simulation of 10,000 generations.

For infinite generations, on the other hand, complete transmission will never occur as long as the hypotheses overlap and there exists some probability of transitioning from one hypothesis to another. But for finite runs, the practical appearance of complete transmission fidelity is determined by the combination of the prior and hypotheses structure. When hypotheses have small overlap and a strongly-biased prior, less data samples are needed to unequivocally indicate which hypothesis distribution they were generated from. In this case, complete transmission fidelity will occur at smaller bottleneck sizes (Figure 2.4, “MAP strong” and “Sampler strong”). However, for hypotheses with more overlap and weaker biases, complete transmission fidelity will occur at larger bottleneck sizes (Figure 2.4, “MAP weak” and “Sampler weak”).

Bottleneck effect differences between MAP and Sampler:
For both the MAP and Sampler, transmission fidelity increases as the bottleneck widens. The Sampler's stationary distribution continues to mirror the prior, over all bottleneck sizes and priors tested. This confirms that the bottleneck has no effect on the outcome of iterated learning for the Sampler model. However, it does effect the internal dynamics of transmission and may well have an effect on the outcome of iterated learning over finite time spans. The MAP's stationary distribution, on the other hand, continues to be affected both by the likelihoods and priors, but changes non-monotonically as the bottleneck widens. Though the MAP's transmission fidelity steadily increases, the dynamics reflected by the stationary distribution are surprisingly unstable (Figure 2.5). Interestingly, this instability only occurs with asymmetrical hypotheses structures, where the slightest asymmetry leads to wildly different stationary distributions for each bottleneck size. For canonical hypotheses structures, all hypotheses continue to be equally represented in the stationary distribution. Unfortunately, the cause of this strange behavior has not been obtained.

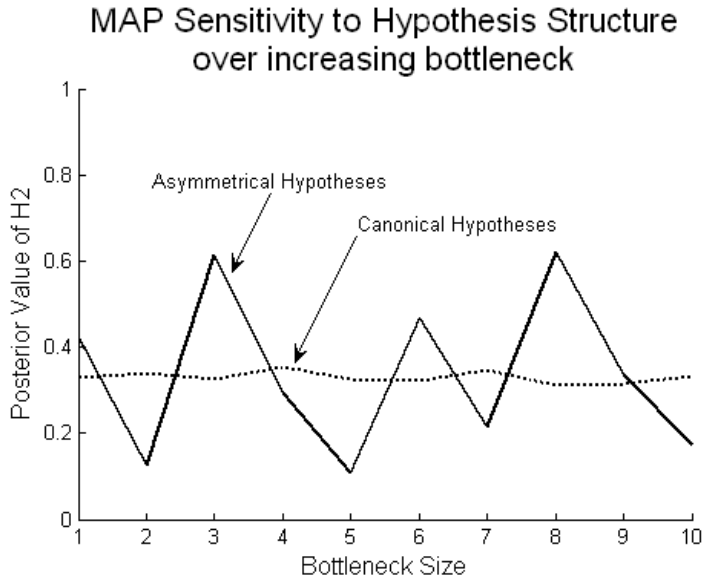

Figure 2.5
MAP posteriors non-monotonically vary as bottleneck widens. The y-axis shows the proportion of H2 in the experimentally-calculated stationary distribution (normalized hypothesis history). Prior = unbiased, Canonical Hypotheses = [.6 .2 .2; .2 .6 .2; .2 .2 .6], Asymmetrical Hypotheses = [.6 .3 .1; .2 .6 .2; .1 .3 .6]

Bottleneck and data variance issues:
Because these simulations are confined to a finite number of generations, the experimentally-derived stationary distributions are less reliable under larger bottlenecks. This is directly due to the increase in transmission fidelity. Under small bottlenecks, the high number of transitions in the simulation ensure that the resulting distribution in the normalized hypothesis history reflects the true stationary distribution. When transmission fidelity increases, the variation between simulation runs also increases, and thus more generations (or multiple runs) are needed to obtain a reliable approximation of the stationary distribution. If the simulation could be run an infinite number of generations, then the normalized hypothesis history would *be*

the stationary distribution, and transmission fidelity would have no effect. However this is impossible. For all of the data in the present report, all simulations were run for 10,000 generations. At this setting, variance begins to become a problem with Q matrix and stationary distribution approximation around a bottleneck of 6-10. Above this level, multiple runs must be averaged to gain a more complete picture of the model's dynamics.

2.4.4 *Population Size*

When the population consists of multiple agents, which dynamics found in the single-agent models hold, and which do not? And what is the outcome of iterated learning when this population has heterogeneous biases? The remaining sections will answer, in terms of this model, these new questions regarding population size and heterogeneity.

In this model, each agent in a multi-agent population sees the same data string, separately calculates their posterior values, chooses their own hypothesis, and generates their own data. The data from each agent 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 parameter is set to any number x, all generations have x population members. When the number of data samples is set to y, each agent in the population produces y number of data samples, yielding a bottleneck size of $x*y$.

The multi-agent and single-agent configurations differ in one respect: the data string that is passed between generations is not stochastically generated from one unified agent, but from many. This has different consequences for the MAP and the Sampler models. For a homogeneous, multi-agent MAP 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 (and this is also the case for the Samplers). However, all the MAP agents will choose the same hypothesis (Table 2.18), because this choice is based on the maximum value of their identical posteriors. The only exception to two MAP agents choosing different hypotheses based on the same data string 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 MAP agents producing y samples, is equivalent to one MAP agent producing $x*y$ samples (Table 2.18). Therefore, MAP dynamics due to population size are identical to the dynamics due to the bottleneck (see section 2.4.3). However, due to the implementation of the multi-agent model, where all agents produce equal an equal number of data samples, only even-numbered bottleneck sizes can be investigated for population sizes greater than 1. Therefore, the non-monotonic variance in the MAP model (referring back to figure 2.5) is less apparent in these cases.

Normalized Hypothesis History

	MAP		
	H1	H2	H3
samples = 4	0.6352	0.2488	0.1160
population = 4	0.6304	0.2561	0.1135

	Sampler		
	H1	H2	H3
samples = 4	0.6977	0.1918	0.1105
population = 4	**0.8104**	**0.1254**	**0.0642**

Table 2.18

Population size does not add new dynamics for MAP, but for Samplers it does – the stationary distribution no longer mirrors the prior. Prior [.7 .2 .1], hypotheses [.8 .1 .1; .1 .8 .1; .1 .1 .8], 10,000 generations.

Hypotheses Choice of Multi-agent Sampler vs. MAP

		Sampler			MAP		
		H1	H2	H3	H1	H2	H3
Hypotheses Chosen	agent 1	**7609**	**1579**	**812**	8234	1496	270
	agent 2	**7674**	**1570**	**756**	8234	1496	270

Table 2.19

MAP agents choose the same hypothesis, whereas Samplers do not.
Prior [.7 .2 .1], hypotheses [.8 .1 .1; .1 .8 .1; .1 .1 .8], 10,000 generations.

For a homogeneous, multi-agent Sampler model, the dynamics are markedly different. Because samplers choose their hypotheses weighted by their posteriors, a homogenous population will not choose the same hypotheses each generation (Table 2.19). Therefore, the data samples do not come from the same set of likelihood values. This has interesting implications concerning the perfect Bayesian rationality of the agents. In the case of the MAP, the agents have all the possible sets of likelihoods that the data could be generated from, already given to them as their hypotheses. When a string of data is generated from a set of likelihoods which the agents 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 in 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-population of Samplers, the stationary distribution no longer mirrors the prior (Figure 2.6).

Kalish et al. (2007) mathematically show that their single-agent results can be generalized to multi-agent populations, where the stationary distribution will continue to mirror the prior. However, this proof would require, in practice, that each Sampling agent 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. Although perfect Bayesian rationality is a simple assumption for mathematical analyses of ILMs, the practicality of maintaining this assumption is dubious for actual model implementations, let alone for actual humans.

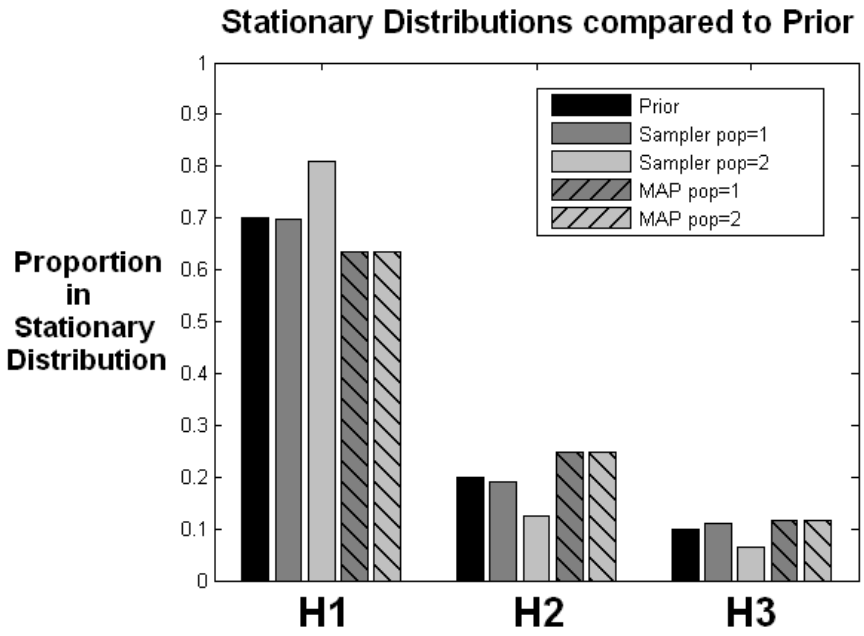

Figure 2.6

The MAP 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 MAP and Sampler models with: prior [.7 .2 .1], hypotheses [.8 .1 .1; .1 .8 .1; .1 .1 .8], 10,000 generations.

Additionally, some systematic variance was observed for the multi-agent Sampler model in regard to manipulations of the likelihood structure. Figure 2.7 shows that the stationary distribution mirrors the prior less and less as the hypotheses 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. Additionally, increasing the population size systematically amplifies the effect of the likelihoods on the Sampler's stationary distribution (Figure 2.8).

Figure 2.8 shows that the stationary distribution reflects hypotheses structure in the absence of a prior bias. For the canonical hypotheses structure *a*, the stationary distribution remains flat despite changes in population size. This is similar to the MAP behavior given canonical hypotheses under different bottleneck sizes. Also like the MAP model, the Sampler is differentially sensitive to asymmetrical hypotheses structures, however the relationships are in the opposite direction. Here, the highest peaked hypothesis is the lowest in proportion in the Sampler's stationary distribution and the lowest peaked hypothesis is the most represented.

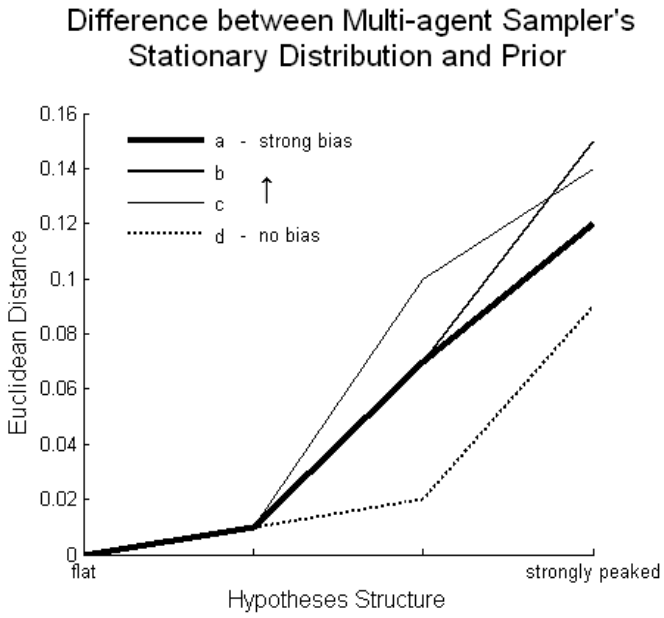

Figure 2.7

Strong biases and peaked hypotheses lead the sampler away from converging to the prior.. a = prior [.3 .3 .3] b = prior [.6 .2 .2] c = prior [.7 .2 .1] d = prior [.8 .1 .1], Population = 2.

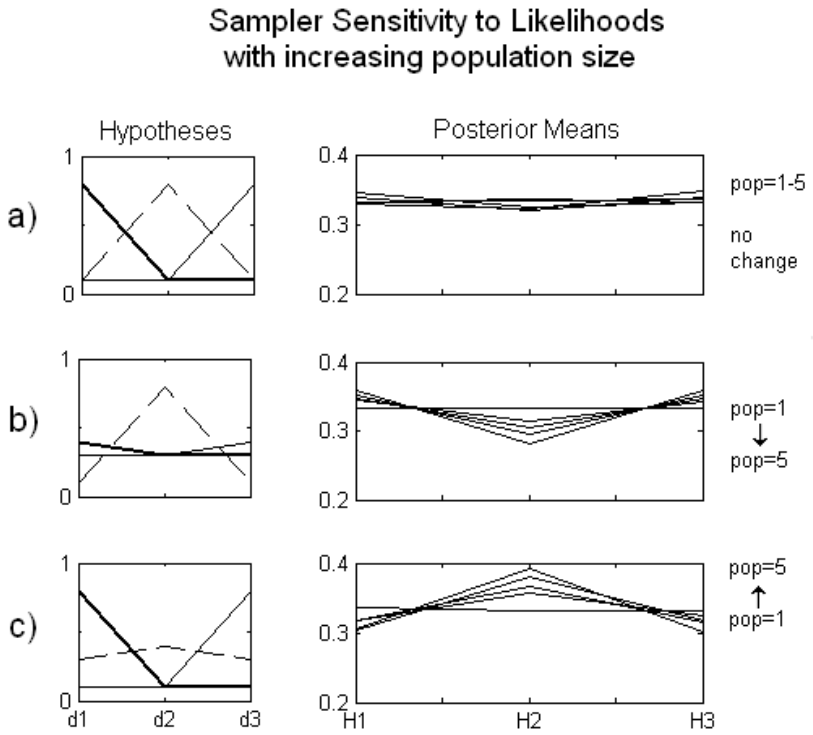

Figure 2.8

Population size amplifies Sampler sensitivity to hypotheses structure. a = hypotheses [.8 .1 .1; .1 .8 .1; .1 .1 .8] b = hypotheses [.4 .3 .3; .1 .8 .1; .3 .3 .4] c = hypotheses [.8 .1 .1; .3 .4 .3; .1 .1 .8]. Prior = unbiased. Population sizes 1 to 5.

2.4.5 *Heterogeneity*

A heterogeneous ILM was implemented by taking a multi-agent model and assigning different prior vectors to each of the agents. This model, therefore, is the most complex of all models constructed. For this reason, only 2-agent heterogeneous populations will be used as examples in this section.

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 (Figure 2.10 and 2.12). This conforms to the general tradeoff between the likelihoods and prior in Bayesian induction; the more data that is seen, the less the effect of the prior on the posterior distribution over hypotheses. Because the behavior of both models is based on the posterior values, increasing the amount of data which the agents share produces increasingly similar posterior values, despite differences in agents' priors. In the following analyses, convergence is measured by the Euclidean distance between each agent's normalized hypotheses history vector.

Since agents' behavior is converging, the natural question is, to what? To structure this question more, we decided to investigate whether or not the converged behavior of a heterogeneous ILM (where agent *x* and agent *y* each differ in their prior bias) is just an average of the behavior of one homogeneous run with agent *x* and one with agent *y*. It turns out, this is not a simple question. First, it is difficult to determine exactly what the true average of agent *x* and agent *y*'s behavior is, due to the variation among runs inherent in the simulation. Additionally, as discussed previously, the stationary distribution of a particular model changes as a function of bottleneck or population size. Therefore, we cannot just average the stationary distribution of agent *x* and agent *y* for comparison to a 2-agent simulation composed of agent *x* and agent *y*. Instead, we should match for the number of samples in the data string. For the Sampler, it is established that manipulating bottleneck size does not effect the stationary distribution: it will continue to mirror the prior. However, manipulating the population size slightly effects the stationary distribution away from mirroring the prior. Although, this effect isn't noticeable at a population of 2 for relatively flat hypotheses and a mildly biased prior. Therefore, the average behavior for Sampler agent *x* and Sampler agent *y* can be done with some confidence – by just averaging the priors of the two agents – but only when hypotheses are relatively flat and the bias is weak.

Heterogeneous Sampler behavior:

So, the question for the heterogeneous Sampler model is: Does the converged behavior of a 2-agent heterogeneous Sampler model come to the average of their priors? The answer appears to be yes. Figure 2.9 shows the convergence of a 2-agent, heterogeneous Sampler ILM and Figure 2.10 shows the difference (measured in Euclidean distance) between each agent's hypothesis history, for the data in Figure 2.9. For clarity, Figure 2.9 does not plot the entire stationary distribution, but only H1's proportion in the stationary distribution. The gray line indicates what H1 should be if the convergence reflects a trivial average of individual agent behavior. The heterogeneous behavior seems to converge to this trivial average. But this is difficult to tell with certainty because, by the time the agents' behavior converges completely, the variance between runs (due to the increasing bottleneck size, see section 2.4.3) is too high to determine what the true convergence values are.

Looking at the last reliable run from Figure 2.9, at bottleneck 16, the normalized hypotheses history of each agent are displayed in Table 2.20. Here, it is clear that the convergence behavior is settling around the average of both agent's priors.

	Average of Converging Behavior		
	H1	H2	H3
agent 1	0.43	0.21	0.36
agent 2	0.37	0.21	0.42
average	**0.40**	**0.21**	**0.39**

Average of both Priors		
0.4	**0.2**	**0.4**

Table 2.20
Hypotheses History of each agent at bottleneck = 16, from Figure 2.9, and the average of their priors.

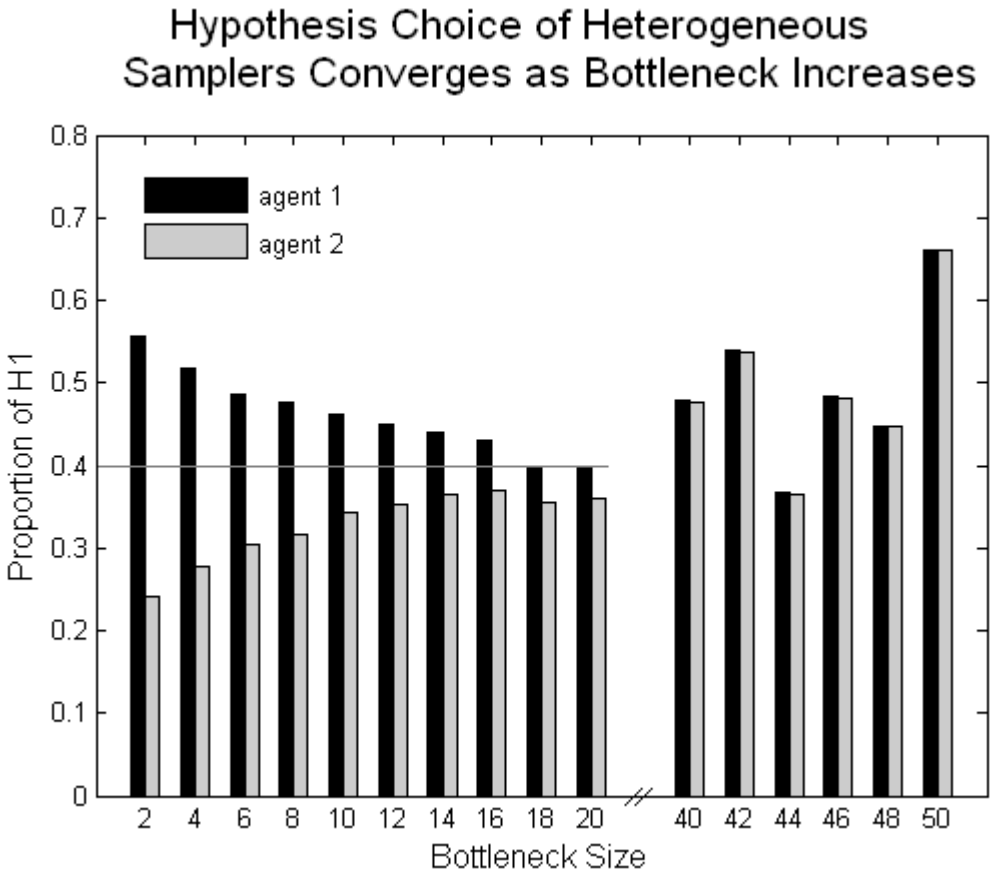

Figure 2.9
The behavior of agents with different priors converges when they are allowed to share more and more data. Variation of individual runs is still a problem over large bottleneck sizes. Population = 2, Prior agent 1 = [.6 .2 .2], Prior agent 2 = [.2 .2 .6], Hypotheses = [.6 .2 .2; .2 .6 .2; .2 .2 .6].

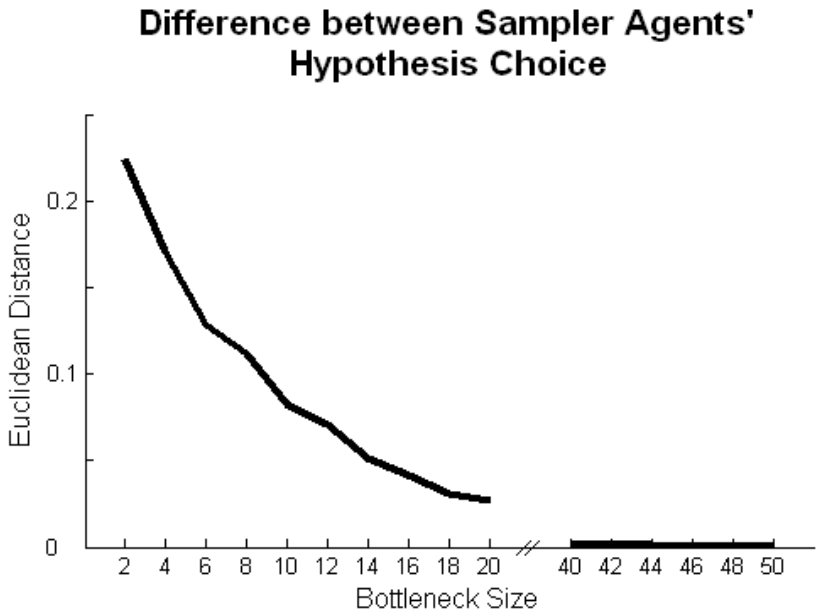

Figure 2.10
Euclidean distance between the agents' hypotheses histories, of which H1 is graphed in Figure 2.9.

In section 2.4.4 we discussed some parameter combinations, for the multi-agent Sampler, which did not result in a stationary distribution that mirrored the prior. Namely, strongly-peaked hypotheses structures plus strong biases. Some of these combinations were also analyzed in the heterogeneous model, however, the normal variance of each run subsumed the difference between the averaged priors and the average outcome of the multi-agent runs which did not mirror the prior. Therefore, it is impossible to determine a difference in convergence when comparing these two conditions.

Heterogeneous MAP behavior:
As we've demonstrated in previous sections, the dynamics of the MAP model are more complex than that of the Sampler, and the heterogeneous models are no exception. The MAP model behaves very differently over an increasing bottleneck depending on whether the hypotheses structure is canonical or asymmetrical (refer back to Figure 2.5). Due to the unexplained, non-monotonic variance over bottleneck size of MAP models with asymmetrical hypotheses, we will restrict the heterogeneous MAP analyses to canonical hypotheses structures.

To determine whether the MAP model convergence of agent *x* and agent *y* is a trivial average of each of agent *x*'s and agent *y*'s normal stationary distribution, we need to know what agent *x* and agent *y*'s normal behavior is. For the MAP, there is no difference in dynamics whether the models are matched for population size or bottleneck. Therefore, the stationary distribution of a 1-agent *x*, bottleneck = 2 simulation and the stationary distribution of a 1-agent *y*, bottleneck = 2 simulation can be averaged to represent a trivial convergence state. For all the MAP models tested here, their convergence state does not conform to this average exactly. Figure 2.11 shows convergence in MAP hypotheses choice behavior, again just for H1's proportion in the stationary distribution. The analytically-determined stationary

distribution for single-agent, bottleneck = 2 model for agent 1 is [.83 .15 .03] and agent 2 is [.03 .15 .83]. The average of these vectors is [.43 .15 .43]. However, the hypothesis history of the heterogeneous model does not settle upon this average. Due to the variation over runs, it is difficult to determine what the actual convergence state is. However, it seems that the heterogeneous model is converging somewhere off of this average, rather then homing in on it as the Sampler model did.

Figure 2.12 shows some additional simulations, where each graph shows a simulation with a different set of agent priors. These also seems to converge somewhere off of the trivial average. Recall that MAP stationary distributions are differentially sensitive to whether the maximum prior value is higher or lower than the maximum hypotheses peak value (refer back to Figure 2.3). This differential sensitivity is also confirmed in the heterogeneous MAP model. In figure 2.12 and 2.13, models (a) and (b) have a bias which exceeds the maximum likelihood values of the hypothesis structure, whereas the bias in models (c) and (d) do not. The convergence behavior for these two sets of models are qualitatively different.

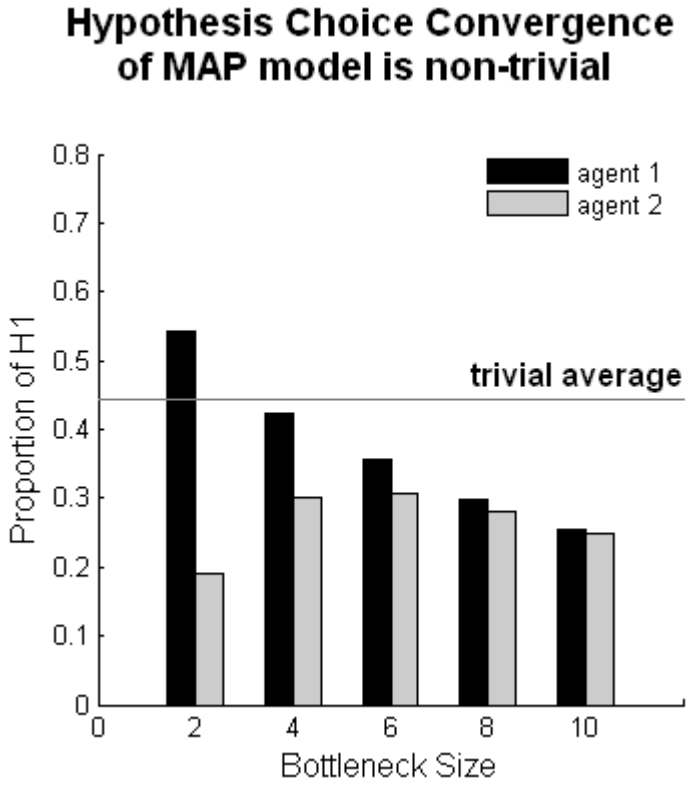

Figure 2.11

MAP model shows signs of converging to something other than the average behavior of appropriately matched single-agent, homogeneous models (represented by the horizontal line). Hypotheses = [.8 .1 .1; .1 .8 .1; .1 .1 .8], prior agent 1 = [.7 .2 .1] and prior agent 2 = [.1 .2 .7]

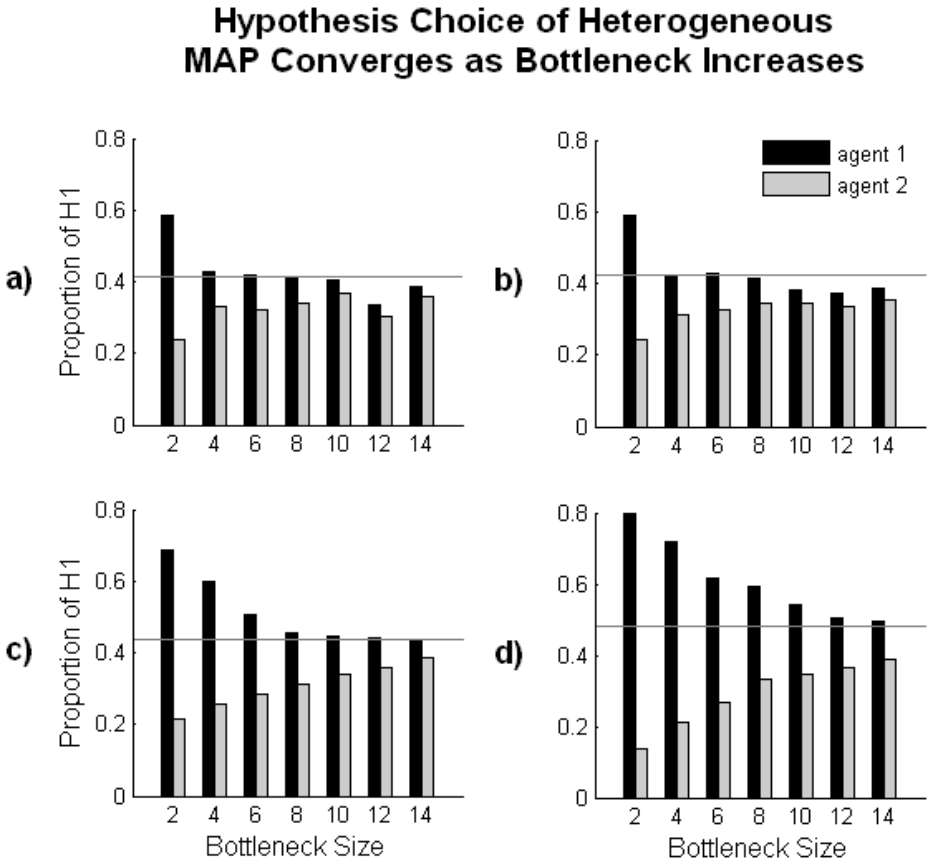

Figure 2.12

MAP convergence for 4 simulations, each where agents have a different set of priors. a-d: Hypotheses = [.6 .2 .2; .2 .6 .2; .2 .2 .6], a: prior agent 1 = [.4 .3 .3] and prior agent 2 = [.3 .3 .4], b: prior agent 1 = [.59 .205 .205] and prior agent 2 = [.205 .205 .59], c: prior agent 1 = [.6 .2 .2] and prior agent 2 = [.2 .2 .6], d: prior agent 1 = [.8 .1 .1] and prior agent 2 = [.1 .1 .8]

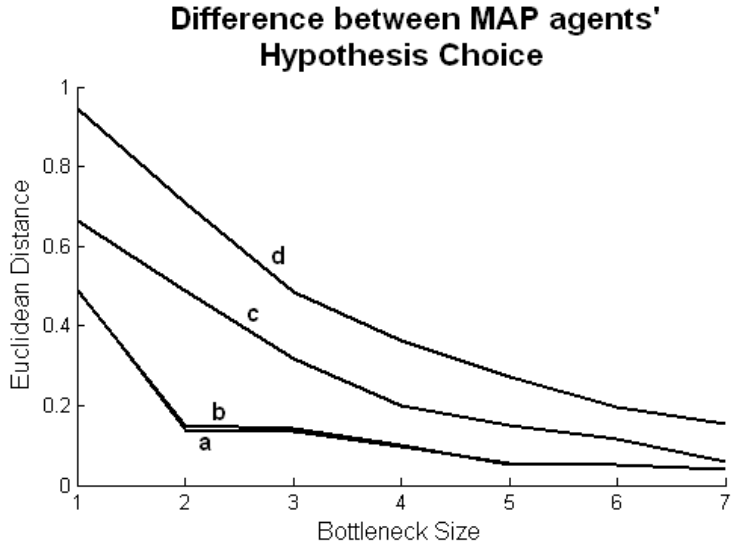

Figure 2.13

Euclidean distance between the agents' hypotheses histories, of which H1 is graphed in Figure 2.12.

Summary Heterogeneous ILM:
For both the MAP and Sampler models, the behavior of 2 agents with heterogeneous biases converges as a function of the bottleneck size. The more agent's share each other's data, the more they choose the same hypotheses as each other. At very large bottlenecks, agents choose the exact same hypothesis throughout the simulation, showing that a sufficiently strong data likelihood value can override agent's differences in prior biases. However, the inherent variance of a finite simulation over high bottlenecks, makes it impossible to obtain an exact distribution of hypotheses that is being converged to. It appears that the Sampler model tends toward convergence to a trivial average of the agents' priors. The MAP model seems to converge slightly below this level. However, overall the behavior of the Sampler and the MAP in the heterogeneous model are qualitatively very similar. we've probably shied away from addressing the true complexity of convergence by limiting myself to 2-agent, bottleneck = 2 simulations, canonical hypotheses structures, and symmetrical prior sets. Undoubtedly, such simulations will yield more variance in behavior according to more fine-grained categories of parameter settings. However, these simulations are only a first step in addressing bias heterogeneity in a Bayesian ILM.

2.5 Model 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 single-agent Sampler model's stationary distribution always mirrors the prior bias of the agent (Griffiths & Kalish, 2005), and a MAP model's stationary distribution is determined both by the prior and the data likelihood values (Kalish et al., 2007). Also, for a range of parameters, the prior has no effect on the MAP 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 MAP models (Kalish et al., 2007). Last to note, the normalized history of all hypotheses choices over the course of the simulation yielded the same solution as analytically calculating the Q matrix, as in Nowak et al. 2001. 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 MAP and Sampler were obtained. The focus of all previous research with Bayesian ILMs is the on 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 hypotheses correspond to what we call the canonical form in our analyses. Otherwise, the rest of the literature does not manipulate the data likelihoods of their model. My analyses of both canonical and asymmetrical hypotheses structure, and in the absence of noise, sheds new light on *how* the likelihoods effect the stationary distribution, by way of determining the posterior values, which determine the transition probabilities of the Q matrix, which yields a particular stationary distribution. There is a great deal of complexity inherent in the nature of the hypotheses overlap, where different hypotheses structures can be shown to determine the outcome of iterated learning just as much as the prior bias does. These complexities are also strongly influenced by the bottleneck, showing that hypotheses overlap is responsive to the pressures of cultural transmission. However,

in the case of asymmetrical hypotheses, this sensitivity to the bottleneck is surprisingly unstable.

The results of Smith & Kirby (2008) demonstrate that the MAP strategy of hypothesis choice is evolutionarily stable over that of samplers. However, the MAP parameters for which this result was proven, it seems, were derived from a canonical hypotheses structure and for the range of priors which are unaffected by bias strength (refer back to Figure 2.3). The result is consistent behavior of the Smith & Kirby's MAP model over the bias values they selected. My results show that this is a subset of MAP behavior and that unstable behavior is easily obtained for the right relationship between hypotheses and priors. So, perhaps MAP would not be the evolutionary stable strategy in all cases. Or additionally, perhaps it can be shown that a certain range of MAP parameters is evolutionary stable over other MAP parameter sets. Knowledge of this kind would help guide the right choice of MAP parameter sets to use for iterated learning simulations, rather than convention (i.e. assuming one un-manipulated set of canonical hypotheses).

Another novel result 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. The result is that the hypothesis with the highest prior becomes amplified in the stationary distribution, and some sensitivity to the likelihood structure emerges. This result is due to the fact that Samplers, according to this model's multi-agent implementation, can no longer be classified as perfect Bayesian reasoners (refer back to section 2.4.4).

The population manipulation was also informative in terms of this paper' ultimate question; what does cultural transmission add? Since population size, in part, defines the social structure and transmission dynamics of an ILM, any manipulation to population size that affects the stationary distribution can be taken as evidence for cultural transmission "adding" something. Referring back to Figure 2.6, it is clear that cultural transmission adds additional dynamics in the case of the Sampler, but not in the MAP. Interestingly enough, the existing literature claims the reverse: cultural transmission adds nothing to the Sampler model, but does to the MAP model.

In summary, this section has shown that the particular dynamics which were previously thought to differentiate the Sampler and MAP models, may not be as clear cut as they previously seemed. It is clear that the parameters which encode manipulations to the cultural transmission system (bottleneck and population size) affect *both* the Sampler and MAP models. It is also clear that non-convergence to the prior cannot be taken as evidence against the Sampling strategy or support for MAP strategy.

One suggestion for future Bayesian ILM research would be to investigate models where agents are no longer perfect Bayesian reasoners, by giving agents heterogeneous hypotheses structures. If each generation of agents do not have the exact same hypotheses structures as the previous generation, then agents will not be able to calculate the optimal posterior probabilities over hypotheses. It is certainly the case that humans are not perfect Bayesian reasoners, because we do not have complete knowledge of the exact likelihoods involved in the processes of our environment, but rather we learn these probabilities and construct our own hypotheses, imperfectly, through experience. Another suggestion would be to explore hypothesis choice strategies that are a mix between Sampling and MAP behavior. It is more likely that human behavior can be better approximated by a strategy that lies on the continuum between sampling and maximizing, rather than one at either of these extremes. Both of these suggestions should yield results which further inform us about the outcome of iterated learning in human populations.

3. Conclusions

Overall, the present research has demonstrated a wider variety of behavior than has been previously obtained in iterated learning models, yielding new insights into the complex interplay between individual biases and the cultural transmission of language.

The debate over how much innate biases vs. cultural transmission determine the outcome of iterated learning, seems to be somewhat reconciled. The present modeling results demonstrated that Samplers in a population larger than 1 do not converge to the prior and are sensitive to manipulations in the data likelihoods. Thus, when an ILM does not converge to the prior, this can neither be taken as evidence against the sampling strategy, nor for the MAP strategy. Additionally, both models are sensitive to dynamics imposed by cultural transmission; the MAP model is sensitive to bottleneck size, and the Sampler model to population size. Thus, we cannot expect that either of these models will simply converge to the prior within ILMs that more realistically approximate human social systems.

The last important points regard the use of Bayesian inference as a model of human cognition. First, it may not be the case that the prior fully specifies the bias for the Bayesian inference algorithm once its adapted into an agent within a cultural transmission system. Additions to the Bayesian inference algorithm, such as a hypothesis choice strategy, must be implemented so that this algorithm can output data for other agents in the simulation. These additions are probably building in additional biases to the agents behavior, as is apparent in the behavioral differences between Samplers and Maximizers. This raises doubt to the claim that Bayesian ILMs are the solution to previous ILM confounds, where the learning algorithms had implicit and incomparable biases.

Second, we would certainly want to know how much the behavior of the models change when agents are no longer perfect Bayesian reasoners. Relaxing this assumption could give a better account of human behavior in iterated learning. It could be quite interesting, for researchers in cognitive science, to investigate computational ILMs where agents are heterogeneous in respect to their hypotheses structures, because an in-depth study of the hypotheses component of Bayesian

inference could provide a formal framework to investigate the representational constraints of individual cognitive agents, and how they affect the transmission of language.

The work presented in this paper has attempted to synthesize the findings of a computational ILM and an iterated learning experiment with human subjects. As hopefully demonstrated in the present research, this combination of methodologies can provide us with deeper insights into explaining the structure of human language, as rooted both in the biases of individual cognitive agents, and the system of cultural transmission in which the interact.

Acknowledgments

Thanks to Federico Sangati, Sara Ramezani and Emily Morgan for technical help, support and comments. We gratefully acknowledge funding from an Amsterdam Merit and HSP Huygens scholarship to VF and a VENI research fellowship to WZ from the Netherlands Organization for Scientific Research (NWO, project nr. 639.021.612). This paper is based on chapter 1 and 2 of Ferdinand (2006).

Appendix A

Bayesian ILM code for the MAP agent
Matlab

```
%% Parameters:
N_gen = 10000; % number of generations
N_pop = 2; % must enter same number of rows in prior matrix as N_pop
N_hyp = 3;   % number of hypotheses
N_sam = 1; % number of samples
N_sampop = N_sam*N_pop; % number of samples, totaled over agents
N_dat = 3;   % number of data-values (assuming data-values range from 1 to N_dat)

%% Main program loop

%% initialize
posteriorhistory = zeros(N_gen,N_pop,N_hyp);
hypotheseshistory = zeros(N_gen,N_pop);
hyphist = zeros(N_pop,N_hyp);
hyphistnorm = zeros(N_pop,N_hyp);
posterior = zeros(N_pop,N_hyp);
agents_posterior = zeros(N_pop,N_hyp);
agents_hypothesis = zeros(N_pop,1);
data_each_agent = zeros(1,N_sam);
posteriormean = zeros(N_pop,N_hyp);
summary = zeros(N_pop,N_hyp);
likelihood = zeros(1,N_hyp);
prior = zeros(N_pop,N_hyp);
prior = log ([.8 .1 .1; .1 .1 .8]); % must enter #rows=N_pop
hypotheses = log ([.6 .2 .2; .2 .6 .2; .2 .2 .6]);
data = zeros(1,N_sampop); % data is a vector. each agent's output follows in chunks
data = random('unid',N_dat,[1,N_sampop]),

%% iterate
for generation=1:N_gen,

    %calculate posterior
    for a=1:N_pop,
    likelihood = [0.0 0.0 0.0]; %resets likelihood to zeros, each loop
        for i=1:N_sampop, likelihood = likelihood +
                transpose(hypotheses(:,data(1,i))); end;
        agents_posterior(a,:) = logBayesRule(prior(a,:),likelihood);
    end;
    agents_posterior; %matrix of all agents posteriors for this generation

    posteriorhistory(generation,:,:) = agents_posterior;
    for a=1:N_pop,
    posteriormean(a,:) = sum(posteriorhistory(:,a,:)) ./ N_gen;
    end;

%Maximizer

    %choose hypothesis
    for a=1:N_pop,
        %randomize order hypotheses are evaluated to be max or not, because
        %if there are multiple identical max values, max() always returns the first
one, biasing towards h1, then h2.
        maxtest = transpose(randsample(3,3));
        [value,position] = max(agents_posterior(a,:));
        if agents_posterior(a,maxtest(:,1)) == max(agents_posterior(a,:));
           hstar = maxtest(:,1);
        elseif agents_posterior(a,maxtest(:,2)) == max(agents_posterior(a,:));
           hstar = maxtest(:,2);
        elseif agents_posterior(a,maxtest(:,3)) == max(agents_posterior(a,:));
           hstar = maxtest(:,3);
```

```
        else 'I cant program';
        end;
        agents_hypothesis(a,:) = hstar;
        hypotheseshistory(generation,a) = hstar;
    end;
    agents_hypothesis; %matrix of all agents' chosen hypothesis for this generation

    %generate data
    data = [];
    for a=1:N_pop, data_each_agent =
              randsample(1:N_dat,N_sam,true,exp(hypotheses(agents_hypothesis(a),:)));
    data = [data data_each_agent];
    end;

end;

%create hyphist
for a=1:N_pop,
    for h=1:N_hyp,
        for g=1:N_gen,
            if hypotheseshistory(g,a) == h;
                hyphist(a,h) = (hyphist(a,h))+1;
            else hyphist(a,h) = hyphist(a,h);
            end;
        end;
    end;
end;

for a=1:N_pop,
    hyphistnorm(a,:) = hyphist(a,:) ./ sum(hyphist(a,:));
end;
```

Bayesian ILM code for the Sampler agent hypothesis choice
Matlab

```
%choose hypothesis
    agents_hypothesis = [];
    for a=1:N_pop,
        agents_hypothesis(a) = randsample(1:N_hyp,1,true,agents_posterior(a,:));
        hypotheseshistory(generation,a) = agents_hypothesis(a);
    end;
    agents_hypothesis; %matrix of all agents' chosen hypothesis for this generation

    %generate data
    data = [];
    for a=1:N_pop, data_each_agent =
       randsample(1:N_dat,N_sam,true,exp(hypotheses(agents_hypothesis(a),:)));
    data = [data data_each_agent];
    end;
```

References

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge: Cambridge University Press.

Batali, J. (1998). Computational simulations of the emergence of grammar. In Hurford, J. R., Studdert-Kennedy, M., Knight, C. (Eds.), *Approaches to the Evolution of Language: Social and Cognitive Bases,* pages 405-426. Cambridge: Cambridge University Press.

Brehmer, B. (1971). Subjects' ability to use functional rules. *Psychonomic Science*, 24, 259-260.

Brehmer, B. (1974). Hypotheses about relations between scaled variables and in the learning of probabilistic inference tasks. *Organizational Behavior & Human Decision Processes*, 11, 1-27.

Brighton, H. (2002). Compositional Syntax From Cultural Transmission. *Artificial Life, 8*(1).

Brighton, H. & Kirby, S. (2001). The survival of the smallest: stability conditions for the cultural evolution of compositional language. In Kelemen, J. & Sosik, P. (Eds.), *ECAL01*, pages 592-601. Springer-Verlag.

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

Busemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In Lamberts, K. & Shanks, D. R. (Eds.), *Knowledge, concepts, and categories: Studies in cognition*, pages 408-437. Cambridge: Cambridge, MA: MIT Press.

Christiansen, M. & Kirby, S. (2003). Language Evolution: Consensus and controversies. *Trends in Cognitive Science*, 7(7), 300-307.

Cornish, H. (2006). *Iterated Learning with Human Subjects: an Empirical Framework for the Emergence and Cultural Transmission of Language*. Unpublished Masters thesis, School of Philosophy, University of Edinburgh, U.K.

Ferdinand, V. (2008). *How learning biases and cultural transmission structure language: Iterated learning in Bayesian agents and human subjects*. Unpublished Master's thesis, Institute for Interdisciplinary Studies, University of Amsterdam

Flaherty, M. & Kirby, S. (2008). Iterated language learning in children (abstract). In Smith, A. D. M., Smith, K., & Ferrer I Cancho, R. (Eds.), *Proceedings of the 7th International Conference (EVOLANG7)*, pages 425-426. World Scientific.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, 29, 737-767.

Griffiths, T. L. and Kalish, M. L. (2005). A Bayesian view of language evolution by iterated learning. In Bara, B.G., Barsalou, L., and Bucciarelli, M. (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society,* pages 827-832. Erlbaum, Mahwah, NJ.

Griffiths, T. L., Christian, B. R., & Kalish, M. L. (2006). Revealing Priors on Category Structures Through Iterated Learning. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

Hare, M., & Elman, J. L. (1995). Learning and morphological change. *Cognition*, 56, 61-98.

Hurford, J. R., (2000). Social transmission favors linguistic generalization. In Knight, C., Studdert-Kennedy, M., Hurford, J. R. (Eds.), *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form,* pages 324-352. Cambridge: Cambridge University Press.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). *Psychological Review*, 111(4), 1072-1099.

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

Kirby, S. (1998). Language evolution without natural selection: From vocabulary to syntax in a population of learners. Unpublished manuscript.

Kirby, S. (1999). Function, Selection, and Innateness: the Emergence of Language Universals. Oxford university Press.

Kirby, S. (2000). Syntax without Natural Selection: How compositionality emerges from vocabulary in a population of learners. Unpublished manuscript.

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

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

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

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, 5(11), 831-843.

Lieberman, A. M., et al. (1967). Perception of the Speech Code. *Psychological Review*, 74, 431-61.

Lieberman, E., Michel, J., Jackson, J., Tang, T., & Nowak, M. (2007). Quantifying the evolutionary dynamics of language. *Nature*, 449, 713-716.

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

Pinker, S. (1984). *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.

Scott-Phillips, T. C., Kirby, S., & Ritchie, G. R. S. (2008). Signalling signalhood and the emergence of communication (abstract). In Smith, A. D. M., Smith, K., & Ferrer I Cancho, R. (Eds.), *Proceedings of the 7th International Conference (EVOLANG7)*, pages 497-498. World Scientific.

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

Smith, K. (2003). Learning biases and language evolution. In Kirby, S. (Ed.) *Language Evolution and Computation (Proceedings of the Workshop on Language Evolution and Computation, 15th European Summer School on Logic, Language and Information)*.

Smith, K., & Kirby, S. (2008). Natural selection for communication favors the cultural evolution of linguistic structure. In Smith, A. D. M., Smith, K., & Ferrer I Cancho, R. (Eds.), *Proceedings of the 7th International Conference (EVOLANG7)*, pages 283-290. World Scientific.

Vogt, P. (2003). Iterated learning and grounding: from holistic to compositional languages. Unpublished manuscript.

Weisbuch, G. (1991). Complex systems dynamics: an introduction to automata networks. *Santa Fe Institute Studies in The Sciences of Complexity Lecture Notes*, vol. 2.

Zuidema, W. (2003). How the poverty of the stimulus argument solves the poverty of the stimulus argument. In Becker, S., Thrun, S., & Obermayer, K. (Eds.) *Advances in Neural Processing Systems 15*. Cambridge, MA: MIT Press.