Published in final edited form as:
Adapt Behav. 2009 ; 17(3): 213–235. doi:10.1177/1059712309105818.

The Iterated Classification Game: A New Model of the Cultural Transmission of Language

Samarth Swarup and
Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, swarup@vbi.vt.edu

Les Gasser
Graduate School for Library and Information Science, and Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, IL 61820, gasser@illinois.edu

Abstract

The Iterated Classification Game (ICG) combines the Classification Game with the Iterated Learning Model (ILM) to create a more realistic model of the cultural transmission of language through generations. It includes both learning from parents and learning from peers. Further, it eliminates some of the chief criticisms of the ILM: that it does not study grounded languages, that it does not include peer learning, and that it builds in a bias for compositional languages. We show that, over the span of a few generations, a stable linguistic system emerges that can be acquired very quickly by each generation, is compositional, and helps the agents to solve the classification problem with which they are faced. The ICG also leads to a different interpretation of the language acquisition process. It suggests that the role of parents is to initialize the linguistic system of the child in such a way that subsequent interaction with peers results in rapid convergence to the correct language.

1 Introduction

Language is a cultural information system passed from one generation to the next. This basic fact implies that language must have a structure that makes it possible to acquire it relatively easily. From where has this structure come?

The main hypothesis about this question has been that the structure in language is due to some genetically specified linguistic structures in the brain, and thus must have appeared (somehow) through biological evolution (Pinker & Bloom, 1990). There is no doubt that some of the peculiarities of any language must be learned from data, such as the lexicon. Given that, however, there have been a range of answers to how much of the rest of language, particularly syntax, must be innate. This question has been a main theoretical concern in linguistics for at least half a century. Chomsky introduced the idea of a Universal Grammar (UG) that is encoded in the structure of the brain, and is responsible for the general form of human language (Chomsky, 1965). There is a great deal of ongoing debate about the exact nature of this Universal Grammar, and whether it exists as a specific entity at all (Lappin & Shieber, 2007; Edelman & Waterfall, 2007).

Universal Grammar is meant to be a specification of the bias used by humans to learn language. In Chomsky's words, "UG is a theory of the 'initial state' of the language faculty, prior to any linguistic experience." (Chomsky, 1986, p. 3–4). He suggested that the linguistic evidence available to children underdetermines the grammar of the language, a property he called

"poverty of the stimulus". This implies that some of the structure must be encoded as a *linguistic bias* in the brain. Since there can be no learning without a bias (Mitchell, 1980), it is clear that a bias helpful for acquiring natural language must exist in the brains of children. However it is far from clear how language-specific such a bias must be. It is possible that language is acquired through the use of entirely "domain-general" learning biases, i.e. biases which are useful for many cognitive learning problems, not language alone. There have been several attempts to address the poverty of the stimulus argument through statistical learning procedures (e.g. Foraker, Regier, Khetarpal, Perfors, & Tenenbaum, 2007; Perfors, Tenenbaum, & Regier, 2006).

An important point in this regard is that, in practice, learning performance does not depend only on the bias. Before we accept that there might be a language-specific UG, we have to consider the range of factors whose combination might render such a UG unnecessary. These factors are: general cognitive learning biases, existence of constraints between language and the world, learning from parents (who have in turn learned language from their parents and so on), and learning from peers.

Kirby and colleagues have introduced a learning model based on two of these factors: a general cognitive learning bias embodied as a preference for compact representations, and learning from parents. Since their model consists of a sequence of learners (representing generations) each teaching the language to the next learner in turn, it is called the *Iterated Learning Model* (ILM) (Kirby, 1999, 2001; Brighton, 2002; Kirby, 2002; Kirby & Hurford, 2002; Smith, Kirby, & Brighton, 2003; Kirby, Smith, & Brighton, 2004; Brighton, 2005; Brighton, Kirby, & Smith, 2005; Kirby, 2007).

The biggest success of the ILM has been to demonstrate that the process of cultural transmission of language through generations actually can affect the structure of the language. Iterated learning results in the emergence of *compositional* language from a random initial state (which can be thought of as a *holistic* language). A compositional language is one in which utterances can be meaningfully combined, e.g. in English, 'red ball' does not refer to something that is red and something (else) that is a ball; it means that the ball is red. This is actually a very important property of language because it allows the so-called "infinite use of finite means" (von Humboldt, 1972). In other words, it allows a tremendous gain in expressiveness at the expense of only a small increase in cognitive requirements. Kirby et al. have shown that a combination of two factors can result in the emergence of compositionality: a minimum description length (MDL) bias in the learner, and the presence of a "bottleneck" in the transmission from parent to child. The bottleneck is due to the fact that the child is not exposed to every possible valid sentence in the language of the parent, and must therefore infer other valid sentences from the ones that it sees during the learning phase. Indeed this is the essence of learning (as opposed to memorization). Over a number of generations, they show, an initially holistic language is transformed into a compact, compositional language. This model of the emergence of structure in language over a *cultural* timescale strikes a blow against the tenability of an innate, language-specific UG as an explanation for observed linguistic structure, because the emergence of such a UG would have to happen on an *evolutionary* timescale.

Recently, Griffiths and Kalish did an analysis of the ILM and showed that the process of iterated learning actually performs the function of making the learners' bias manifest (Griffiths & Kalish, 2005). They studied a sequence of idealized Bayesian learners, and modeled the process as a Markov chain. They showed that this Markov chain's stationary distribution is the same as the prior distribution (i.e. the bias) of the learners (assuming ergodicity). They concluded, thus, that iterated learning results in the emergence of compositional languages because the learners are biased towards such languages to begin with. In other words, if the learners are *not* biased towards compositional languages, such languages will *not* emerge through the

iterated learning process. Though this does not invalidate the ILM, it is a strong criticism[1]. Since the MDL bias assumed by the ILM is a very general cognitive bias, and not language-specific, the essential point made by the ILM still stands, that the structure *of the language* does not have to be innately specified. However, it seems to move the focus back from the cultural to the evolutionary timescale, to try to understand how a bias favorable for compositional languages came about.

In this article, we create a more realistic model of the generational transmission of language, which shows that by incorporating the other two factors mentioned above: constraints between language and world, and learning from peers, we can actually eliminate the need to include an MDL bias for the emergence of structure in language. Our Iterated Classification Game (ICG) also eliminates two other criticisms of the ILM: that language is, in fact, mainly acquired from peers as opposed to parents (Vogt, 2005b), and that language is *for* something and not just an arbitrary representational system (Vogt, 2005a). It also sheds some light on the poverty of the stimulus debate, by showing that a language that is very difficult to acquire (i.e. takes a long time) through peer interaction alone, can be learned *much* more effectively through a combination of learning from parents and learning from peers. Our model thus offers a new interpretation of the process of language acquisition. In this view, learning from parents serves to appropriately initialize the linguistic system of the child, so that subsequent interaction with peers leads to a very rapid convergence to the correct language.

The rest of this article is organized as follows. We begin by discussing the ILM in detail, including the analysis by Griffiths and Kalish and other criticisms. Then we present the ICG and describe how this model overcomes these criticisms. The ICG is based on an earlier model called the Classification Game (Swarup & Gasser, 2008, To Appear), which we also describe briefly. Then we present a series of experiments that show the effects of the ICG on the emergent language. This is followed by a discussion of the model, what it explains, and what it does not yet explain. This in turn leads to a discussion of possible future work, to conclude the article.

2 The Effect of Cultural Transmission on the Structure of Language

Language is a complex dynamical system proceeding on three different, but interacting, timescales: the capacity for language changes on an evolutionary timescale, which creates the learning mechanisms used to acquire language on an individual timescale, which in turn drive linguistic change on a cultural[2] timescale, which, to come full-circle, creates the fitness landscape for the evolution of the linguistic capacity (Christiansen & Kirby, 2003).

There have been previous attempts to account for the nature of language in terms of individual learning and biological evolution (Pinker & Bloom, 1990, e.g.), without taking into account

[1]Also see the article by Kirby, Dowman, and Griffiths (2007) in this regard.

[2]In this paper (as in most work on the ILM) we use terms such as “cultural”, “cultural evolution” and “cultural transmission” in a very restricted sense. Here, we use the notion of “culture” to imply a collective system of *meanings* bound to a collection of information-based (symbolic) *expressions*, transmitted and perpetuated through social interaction and learning. Similarly, Smith and Kirby (2008, p. 3594) point out that “Individuals acquire their knowledge of language by observing the linguistic behaviour of others, and go on to use this knowledge to produce further examples of linguistic behavior, which others can learn from in turn.”. This approach is consistent with Geertz’s (2004, p. 3) definition of culture as “a historically transmitted pattern of meanings embodied in symbols, a system of inherited conceptions expressed in symbolic forms by means of which men communicate, perpetuate, and develop their knowledge about and attitudes toward life” and with McGrew’s (1998, p. 305) less human-centric definition: “group-specific behavior that is acquired, at least in part, from social influences”. The main idea is that language, being a structured, symbolic, information system is (like culture in general) transmitted and perpetuated through social interaction and learning, and that it “evolves” through those mechanisms as well (Croft, 2008). So, here, “cultural transmission” or “cultural evolution” are collective mechanisms of information-based exchange and learning. See, e.g., (Smith, Kalish, Griffiths, & Lewandowsky, 2008). Clearly, broader issues of cultural evolution such as the emergence of toolmaking or religious practices or the relation of language to these, or the “evolution *of* culture” are not part of our analysis here, and shouldn’t be inferred from our use of the terms above. Our focus, in contrast with anthropologists like Tomasello (1999, e.g.), is on the *effect* of cultural transmission on language structure, and not the cognitive abilities which enable humans to have culture at all.

the cultural timescale. This is clearly problematic since the cultural timescale lies in-between the individual and the evolutionary timescales. It is more justifiable, on the other hand, to give a partial account by ignoring the evolutionary timescale and studying language change over a period of time too short for significant evolutionary change (a period of several generations, say). This is exactly what the Iterated Learning Model does.

A "generation" in the ILM is idealized to consist of a single individual. This individual acquires language from its parent, and passes it on to its child. The individual in the first generation is initialized with a random language, i.e. a random set of associations between signals and meanings. The child of this individual learns the language by observing a fixed number of signal-meaning pairs produced by the parent. This fixed number is called the *bottleneck size*. The child then has a child of its own, and teaches it the language, and so on. Since the bottleneck size is generally less than the total number of signal-meaning pairs in the language, a parent is occasionally called upon to produce an utterance (signal) for a meaning which it did not observe during its own learning phase. In this case, the parent has to *generalize* from its training sample. This generalization depends on the learning bias of the parent. If the bias is not purely random, it can lead to the preference for some particular kind of structure in the language. Over a period of several generations, this can result in a transformation of the language from the initial randomness to a highly structured set of associations, effectively a compositional language.

Iterated learning, thus, exerts a strong structuring force on the language. This force is inversely related to the bottleneck size. If children spend a long time learning from their parents, i.e. the bottleneck size is large, then irregularities in the language can be passed on more easily. As we start reducing the bottleneck size, the children will have to rely more on generalization when it is their turn to teach the language, since they will have encountered fewer examples in their own learning phase. Of course if the bottleneck size is too small, then no language can be passed on stably through generations. Kirby et al. argue, thus, that the bottleneck size is the crucial parameter in determining whether only compositional languages are stable states of iterated learning, and also how many generations are required for the emergence of compositional languages (Smith, Kirby, & Brighton, 2003).

To make the above discussion more specific, we next describe Brighton's (2002, 2005) implementation of the Iterated Learning Model. In his implementation, the signals are strings, and meanings are feature vectors where each feature has a small discrete numerical range. Brighton introduced a formalism called Finite State Unification Transducers (FSUT) for the representation and learning algorithms of the agents. A communication consists of a paired meaning and signal, such as $\langle \{1, 2, 2\}, adf \rangle$. A child agent gets an entire training set of such examples from its parent, where the size of the training set is the bottleneck size.

An FSUT is a finite-state automaton, where the states correspond to progressive partial parses of the given signal-meaning pair. The final, or *accepting*, state of the automaton is reached if and only if the given signal-meaning pair is a valid part of the agent's language. Transitions between the states are marked with single symbols from the signal string, as well as the meaning with which this signal is associated. An agent can use its FSUT to both accept or generate valid sentences in the language. See the article by Brighton (2002) for details of the formalism.

The child agent learns an FSUT from a training set generated by its parent. Learning is neatly partitioned into two phases. In the first phase, called the *memorization* phase, the learner creates an FSUT which is capable of generating and recognizing exactly the signal-meaning pairs in its training set. At this point, the learner is not capable of generating the signal corresponding to any *new*, hitherto unseen, meaning. However there may be some, perhaps coincidental, regularities in the training set. This allows the learner to carry out a second phase, called the *generalization* or *compression* phase. Here, the learner removes redundancies in the FSUT, by

merging certain nodes and edges. This still allows it to generate the signals for the meanings in the training set, but also now allows it to generate signals for some meanings that were *not* in the training set. If the observed regularities in the training set were in fact coincidental, then these new utterances produced by the learner will not match the corresponding utterances its parent would have produced. If, on the other hand, the regularities were true, then the new utterances will match the parent's language closely. This learning procedure is based on the Minimum Description Length (MDL) principle. It is, thus, said to have an MDL *bias*.

Even after this compression phase, the learned FSUT may not have complete coverage, i.e. there might still be meanings for which the agent does not know how to produce signals. Thus, when the agent becomes a parent and has to generate a training set for its child, it may resort to *invention* to produce signals for some meanings. At this point the agent has the option of generating purely random signals for these meanings, or of doing *structured invention*, where it uses the structure of its FSUT to produce a partial signal, which is then completed randomly.

Both the minimization of the FSUT and the structured invention procedure are necessary for the emergence of compositionality in this implementation of the ILM. Brighton (2005) points out that compositionality will not appear if either of these components are removed. The role of the bottleneck, thus, is that it provides the opportunity for an agent to add structure to the language. In each generation some of the randomness in the language is stripped away and replaced with structure through the structured invention procedure. Over a large number of generations, a highly compact and structured language emerges through this iterated learning procedure.

2.1 Critiquing the Iterated Learning Model

Despite its successes, the ILM has faced criticism on at least three fronts. These are,

- that it does not connect language to the world, i.e. the language modeled is not *grounded* or *ecological* (Vogt, 2005a),
- that it models only *vertical* transmission of language, not peer learning (Vogt, 2005b), and
- that it builds in a preference for compositional language through the MDL bias (Griffiths & Kalish, 2005).

We describe each of these in turn.

The languages studied by the ILM have been very simple and *ungrounded*. Harnad (1990, p. 335) originally described the symbol grounding problem as, "How can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our heads?" He suggested that the solution might be to ground symbols in iconic and categorical representations extracted from the environmental input. Subsequent researchers have, however, interpreted the problem in slightly different ways, and have consequently taken different tacks to solving it (Taddeo & Floridi, 2005). There is a significant body of literature which attempts to give an *intra*-generational account of the emergence of grounded language (e.g. Dominey, 2005; De Beule, 2008), which we do not consider here since our concern is with the generational transmission of language. In general, though, it is acknowledged that grounding is a crucial aspect of the language puzzle and that a complete theory of cultural transmission must account for the evolution of meanings as well as symbols (Kirby, 2007).

Vogt, for instance, has adopted an approach called *physical symbol grounding*. In his words, "In this alternative definition, the symbol may be viewed as a structural coupling between an agent's sensorimotor activations and its environment." (Vogt, 2002, p. 429). We adopt a

slightly more general definition, that symbols should have some *ecological relevance*, i.e. they must be grounded in *function* (Swarup, Lakkaraju, Ray, & Gasser, 2006). This means that there is a task the agent has to perform, which often involves building a model of the world, and symbols are tied to the performance of this task. In our examples in this article, the tasks are *classification* tasks, and the model is a function that maps input to classes (which are generally, but not necessarily, just two in number). Symbols, in our case, are names for features that are extracted from the input data and are relevant for classification. This is explained in detail in section 3.1.

Second, each generation in the ILM consists of a *single* individual. Thus language transmission is strictly vertical. It is widely accepted, however, that peer effects dominate in language acquisition. For example, children of immigrants generally adopt the language and accent of their host country, at the expense of their parents' language (Baron, 1992; Harris, 1995, 1998).

One of the claims of the ILM is that iterated learning creates an evolutionary pressure for language to be learnable (Smith, 2006). This is not strictly true, in the sense that the ILM does not demonstrate an *evolutionary* pressure for increasing learnability. An evolutionary pressure is always manifested through variation (creation of alternatives with differing fitness) and selection (fitness-based elimination of some alternatives). These are not present in the ILM since it has only a single agent (and a single language) in each generation.

There have been a few attempts to extend the ILM to overcome these two limitations. We look at two particular examples here. Vogt combined the language game model (Steels, 1996a, 1996b) with the ILM to provide symbol grounding (Vogt, 2005a), and peer learning (Vogt, 2005b).

In his implementation, agents play *discrimination games* (Steels, 1996a) to develop a language. A speaker-hearer pair is presented with a *context* of a few geometrical shapes which differ in their shapes and colors. The speaker selects one of these shapes and describes it to the hearer. Depending on whether the speaker's choice is revealed to the hearer before or after communication, two variants known as the *observational game* and the *guessing game* were studied by Vogt. During the communication process, the agents are also learning to discriminate the objects by creating categories or clusters in a four-dimensional feature space (three color channels, and one "shape" feature). Their language is represented by a grammar, which they also learn during the discrimination games. The precise learning algorithms can be found in the paper (Vogt, 2005a), but the grammar learning is similar in character to Brighton's (2002) FSUT-learning algorithm described earlier, with a merging operation to reduce redundancy in the learned grammar, and a structured invention procedure to create partially structured utterances when presented with a meaning (object) for which the utterance is unknown. Vogt created an iterated learning model by creating populations of parent and child agents, and always choosing speakers from the parent population and hearers from the child population. Note that this means that the transmission is still entirely vertical.

Vogt reports the interesting result that, in his experiments, compositional languages emerge quite rapidly and remain stable when the populations consist of single parents and single children and there is *no* bottleneck on transmission. However, increasing the population size to just three adults and three children results in compositionality being stable only in the guessing game. In the observational game, compositionality appears at first, but then is replaced by a more holistic language. This transition coincides with an *increase* in communication accuracy and coherence (which is a measure of how much the agents have converged in their utterances for the objects). He suggests that this process has to do with the interaction between the conceptual space and the language space, i.e. it is due to the grounding. Compositionality

persisted in the guessing game if a bottleneck was imposed; however if the bottleneck was removed entirely, holistic languages again took over.

In a second study, Vogt included peer learning in the above model, to address the second criticism of the ILM above (Vogt, 2005b). This was done by allowing speakers to be chosen from the child population as well as the adult population. In addition, he also removed the externally imposed bottleneck. He showed that these changes result in compositional languages emerging reliably and stably. This happens due to an *implicit bottleneck* effect. Since children can be speakers, they are often faced with the need to produce utterances for meanings they have not yet encountered. This provides the opportunity to add structure to the language in the same way that an explicit bottleneck does when transmission is strictly vertical. This is an important point, and we shall come back to it in the discussion of the bottleneck in the context of the iterated classification game. Another important point to note about this work is that the structure in the language appears essentially in one generation, rather than appearing gradually over several generations as it does in the ILM. Vogt attributes it to the low complexity of the discrimination task, though it seems that shortening the length of the learning phases would offset this. Vogt did not experiment with shorter learning durations, however, so we cannot say if it would have resulted in a gradual emergence of compositionality.

2.1.1 The Iterated Bayesian Learner—The first two criticisms above are based on the claim that the ILM *leaves too much out*. The third one, however, claims that the ILM *builds too much in*.

Griffiths and Kalish (2005) built an abstract Bayesian model of the ILM, in which agents are Bayesian learners who have some prior distribution over possible languages, and compute a posterior from the data they obtain from their parents. Thus, in generation $n + 1$, the learner estimates the posterior probabilities of the hypotheses (i.e., languages), using Bayes' rule, as follows,

$$p(h_{n+1} \mid x_n, y_n) = \frac{p(y_n \mid x_n, h_{n+1}) p(h_{n+1})}{p(y_n \mid x_n)},$$

where (x_n, y_n) are the data generated by the agent in generation n and presented to the agent in generation $n + 1$. The agent in generation $n + 1$ then samples its posterior to choose a language to communicate to its child.

By summing over the data, it is possible to convert this into a Markov process on the hypothesis space, where the transition probabilities are as follows.

$$p(h_{n+1} \mid h_n) = \sum_{\mathrm{x}} \sum_{\mathrm{y}} p(h_{n+1} \mid \mathrm{x}, \mathrm{y}) p(\mathrm{y} \mid \mathrm{x}, h_n) q(\mathrm{x}),$$

where $q(\mathrm{x})$ is the distribution over the input data and is independent of all other variables. These equations are taken directly from their work (Griffiths & Kalish, 2005).

They showed that the stationary distribution of this Markov process is $p(h)$, which is the prior distribution assumed by the agents! Thus, in a sense, the ILM is not creating a compositional language, beyond making the prior manifest.

It has been suggested that a better model would be for the agents to choose the language corresponding to the *maximum* of their posterior distributions, rather than just sampling from it (Kirby et al., 2007). However, it is only practical to compute the maximum in trivial cases

where the distribution can be maintained explicitly. For any realistic language space, finding the maximum is highly complex. Hence, the agents have to resort to some estimation technique like gradient ascent, which will only get them to a local maximum. Kirby et al. (2007) show that, in the case that the learner does not find the maximum exactly, but is more likely to find a local maximum than to simply sample from the posterior, iterated Bayesian learning will have the effect of amplifying a weak preference for a compositional language into a stronger likelihood that the emergent language after many generations is compositional (technical details can be found in their paper). Their analytical result, however, is still invariant to the presence of a bottleneck (i.e., the bottleneck has no effect on the stationary distribution of languages), and the emergence of compositionality still relies on there being a bias in favor of it.

The implication of this analysis is that the bias is more important than the iterated learning process in the emergence of compositionality, and therefore that the goal of research should be to explain how a bias favoring compositionality arises. Further, since the bias is analogized to intrinsic cognitive constraints, this puts the focus on the biological evolution of cognitive biases, rather than on the cultural process of linguistic transmission.

Our goal here is to demonstrate that this is in fact not the case. We will show this by addressing all of the above three criticisms together. Compositional languages can emerge gradually over generations due to the combination of generational transmission, grounding, and peer learning, without the need to build in a bias for compositionality. This is because grounding and peer learning combine, in our model, to implicitly reduce the complexity of the emergent language.

3 The Iterated Classification Game

The ICG, as the name suggests, consists of a sequence of populations of agents, each playing the classification game. This model addresses the three criticisms above as follows.

- The agents are given a classification task to perform, and the emergent language arises from converting their learned internal representations into symbols that are shared between agents. This ensures that the language is grounded in the classification task.
- Learning is divided into a *parent learning phase* and a *peer learning phase*. In the parent learning phase, speakers are chosen from the adult population and hearers from the child population. In the peer learning phase, both speakers and hearers are chosen from the child population. This ensures that the language transmission is not strictly vertical.
- Agents use neural networks for representation and the backpropagation algorithm for learning, and we demonstrate that this learning system does not have a bias for compositionality. It is the interaction between agents during the peer learning phase that implicitly induces compositionality. The objective function being minimized by the learners is simply the squared error, and does not include a term for the complexity of the representation.

Before we describe the ICG in detail, we describe the classification game itself.

3.1 The Classification Game

The classification game presented here is a slightly simplified version of earlier work (Swarup & Gasser, 2008, To Appear). We suppose that we have a population of agents who are all learning to perform some (the same) two-class classification task. This is a very general category of tasks, where the goal is to learn how to divide a set of objects into two classes (such as whether some substance is edible or not, whether another creature is a threat or not, etc.). Classification tasks have been studied before in the context of language emergence (e.g.

Cangelosi, Greco, & Harnad, 2000). Abstractly, these tasks are studied as the problem of assigning positive and negative labels to a set of *points*. A point is assumed to be obtained by performing a set of measurements on an object (such as evaluating its color, shape, size, etc.), to obtain a set of numbers which form its coordinates. The classification problem, then, is to find a function that takes these numbers, and produces the label for each point. These functions are represented, in our case, by feed-forward artificial neural networks (Haykin, 1998). Each agent is, thus, a single hidden layer feed-forward artificial neural network and is trained to estimate the function using the backpropagation algorithm. When an input is presented to a neural network, propagating it through the first layer of weights results in a vector of activations at the hidden layer. This vector of activations is converted into an utterance by assigning a symbol[3] to each hidden layer node. Arbitrarily, the first hidden layer node is converted into the symbol A, the second into B, and so on. This assignment of symbols to hidden layer nodes is the same for each agent.

Note that this fixed assignment of hidden layer nodes to symbols does not mean that the language is pre-determined. The agents are initialized randomly and, as they learn, they will come to extract features from the inputs. Which hidden layer node comes to represent which feature is determined entirely by the random initialization, the backpropagation algorithm, and the effects of the interaction between agents. Consequently the relationship between symbols and meanings (features extracted from the input) is negotiated through interactions with the environment and with other agents, which is exactly what we desire from a language model. Further, if agents were not interacting with each other, they would typically assign different symbols to meanings, depending on their random initialization. The interaction provides a means for convergence onto a shared symbol-meaning mapping.

The interaction protocol for the classification game is as follows. At each step, we select two agents uniformly randomly from the population. One agent is assigned the role of speaker, and the other is assigned the role of hearer. Both are presented with the same training example. The speaker uses the outputs of its hidden layer, which we call its *encoder*, to generate an utterance in the public language as described above. The hearer converts this utterance into a vector of activations for its own hidden layer, and then generates a label for the given example using its output layer, which we call its *decoder*. The speaker also generates a label for the given example using its own decoder. Both agents are then given the expected label, whereupon they update their neural network weights using the backpropagation algorithm. This entire process is illustrated in figure 1.

Backpropagation for the hearer is slightly tricky, since it is not generating its hidden layer activations using its actual input to hidden layer weights. These weights are therefore updated as follows. First the hearer updates its hidden to output weights in the normal fashion. This results in the generation of a vector of values at the hidden layer, through backpropagation, that represents what the actual hidden layer activations should have been to generate the correct label at the output. This vector is treated as the expected output for the hearer's hidden layer. The hearer then generates its actual hidden layer activations from the input, compares these with the backpropagated hidden layer values, and computes the squared error and the weight updates for the input to hidden weights. The hearer is, in effect, training two levels of perceptrons.

Training is continued with random speaker-hearer pairs, so that, over time, all agents play the role of speaker and hearer a large number of times.

[3]Note the difference between a *symbol* and a *label*: a symbol is a letter of the alphabet assigned to a hidden layer node, and can be a part of an *utterance* (depending on which hidden layer nodes are active at a given time); a label is a category assigned to an input and is generated by the neural network at its output node (and can be 1 or 0).

3.1.1 Learnability vs. Functionality—The classification game sets up a tradeoff between learnability and functionality for the emerging language. Simpler languages are more easily learned, and therefore spread more quickly through the population. More complex representations are more likely to solve the classification problem correctly, and are preferred for that reason. The opposition of these two factors results in the emergence of a representation that is "simple, but not too simple". Thus interaction during learning has the effect of controlling the complexity of the learned solution, which is known as *complexity regularization*. We demonstrate this effect through a simple series of experiments.

We create a population of four agents. Their neural networks have two inputs, four hidden layer nodes, and one output each. The training set has four examples in it, with the inputs represented by two bits each, i.e. the set of training inputs is {00, 01, 10, 11}. The following three experiments will all use these inputs. The goal of the experiments is first to show the effect of the pull for learnability and the pull for functionality separately, and then to show that the combination of these two results in simple and (close to) optimal solutions.

In the first experiment, there is no external task to be learned. We just require the agents to converge on a shared internal representation. We do this by treating the speaker's label, during each interaction, as the true label, i.e. the speaker's label is given to the hearer as the expected label. Since there is no external source of error, this experiment eliminates the functionality factor. Since there is nothing to counter the pull of learnability, we expect the agents to learn trivial mappings, where each example is assigned the same label. This is exactly what happens, and is shown graphically in figure 2(a).

The figure shows the four points corresponding to the four training examples, and the learned hidden layer weights as straight lines. Each hidden layer node can be thought of as a straight line (or a hyperplane in higher dimensions) and the weights on the connections to the node are the coefficients in the equation of that line: $ax + by + c = 0$. A line is assumed to label points on one side of it as positive, and points on the other side as negative. If we rotate any line through 180°, it will lie exactly on top of its original position, but the points previously labeled positive will now be negative, and vice versa. Each agent's decoder (output layer node) will take the labels generated by its encoder (hidden layer nodes) and convert them into a label for each point. Each point in figure 2(a) is labeled negative by all the agents. This is shown in figure 2(a) by coloring the points white.

Each line in the figure is labeled with the symbol corresponding to that particular hidden layer node. We say that the symbol corresponding to a hidden layer node is expressed or uttered if the corresponding straight line is oriented in such a way that the given point is labeled as positive by this line. This happens when the hidden layer node is "active", i.e. has output greater than 0.5.

In the second experiment, the agents have to learn the *xor* task, i.e. the labels provided for the inputs are {0, 1, 1, 0}, respectively. However, in order to eliminate the pull for learnability, we shut off the communication between agents. This means that we just have four neural networks learning the *xor* task independently. This typically results in an overly complex solution, such as the example shown in figure 2(b). The agents were provided with four hidden layer nodes, and this agent uses them all to solve the problem, though only two are required for the minimal solution. This is typical behavior, and it shows that neural networks trained by the backpropagation algorithm do not have an MDL-like bias to reduce complexity. In fact, the problem of developing neural network training algorithms that can control the complexity of the solution has a long history in machine learning (Barron, 1991; Hinton & van Camp, 1993; Hochreiter & Schmidhuber, 1997; Kärkäinnen & Heikkola, 2004, for example), because low complexity solutions have good expected generalization. This is often done by altering the

objective function being minimized to include a complexity term, since without it neural networks are unable to produce low complexity solutions. The objective function being used by our agents is simply the squared error, precisely to avoid the criticism that we are building in a bias for reducing complexity (thereby inducing compositionality).

In the third experiment, we allow the agents to communicate with each other during learning, according to the protocol described earlier. The result of this process is shown in figure 2(c). The agents are able to converge upon the lowest complexity solution almost always. Notice that, in this case, the agents are only using two of the hidden layer nodes. The hyperplanes corresponding to B and D have been pushed aside. The resulting utterances are {ACBD, ABD, ABD, BD}, but B and D are redundant. This is a compositional language, which can be interpreted as follows: the agents treat the default label of the points as 0. A means that the label is 1, and C means negation (i.e. it overrules A). Note that the overly complex representation generated in the previous experiment ({ABC, C, BD, B}) does not have a compositional interpretation. It is a holistic language.

These experiments show that the classification game results in a reduction of complexity due to the interaction between agents, even though they do not have an intrinsic bias for low-complexity solutions. Of course, the *xor* task is very simple. When the task is more complex, the classification game is not guaranteed to find the simplest possible solution, however it does reduce the complexity of the learned solution as compared to vanilla neural networks, even if we increase the number of hidden layer nodes. See (Swarup & Gasser, To Appear) for more details and more experiments.

3.1.2 A game-theoretic perspective—We give a brief overview of a game-theoretic perspective on the classification game, to explain the emergence of low-complexity shared representations. See (Swarup, 2009) for a more complete analysis.

The strategy space for the game consists of the space of possible encoders and decoders. Encoders are collections of k hyperplanes, where k is the number of hidden layer nodes, and decoders are single hyperplanes, since we have restricted the neural networks to have single outputs (though this can easily be generalized to multiple outputs). An equivalent way to characterize the strategy space, given a training set, is as the space of possible encodings, since the choice of encoding is determined by the choice of encoder and decoder. Given an example, the speaker chooses the encoding of it both for the speaker itself and for the hearer. The payoff to the speaker and the hearer can be thought of as the negative of their error on the given example. Thus, maximizing the payoff is equivalent to minimizing the error. Note that, in an interaction, the speaker is essentially unaware of the hearer, since it gets no feedback about the hearer's error on the example. The speaker, thus, chooses its encoding of the given example purely to minimize its own classification error.

Suppose that an agent chooses an internal representation (encoding), υ_1, for a particular positively-labeled example, x^+, and another agent chooses the same internal representation for a particular negatively-labeled example, x^-. We denote this event a *conflict*. A conflict implies that these two agents are guaranteed to make at least one classification error when they interact with each other, because they cannot distinguish x^+ from x^-. This means that a condition for equilibrium in the classification game is,

$$\textit{there must be no conflicts}. \tag{1}$$

When there is a conflict, at least one of the agents much change its encodings in order to increase payoff. Additionally two other conditions need to be satisfied for an equilibrium to be attained.

Encodings must be internally consistent. (2)

This condition states that each agent must choose its own internal representation of the training set in a way that avoids conflicts. This condition is implied by the previous one, but stating it separately allows us to discuss equilibrium selection more clearly. Finally,

an encoding must represent an attainable dichotomy for the decoder. (3)

This means that the chosen encoding must actually render the classification problem solvable for the decoder. For example, if an agent chooses the identity function as its encoder for the *xor* problem, then its decoder will not be able to solve the problem since at least two hyperplanes are required to solve it, as discussed previously. Together, these three conditions specify the Nash equilibria of the classification game. If any of these conditions are violated, there will be an incentive for at least one agent to change its strategy.

Equilibrium selection: Under these conditions, high complexity Nash equilibria are also possible, where each agent chooses a distinct encoding of each point (which would be a holistic language), without creating any conflicts. This is possible if the neural networks have enough hidden layer nodes. However, such high complexity equilibria are never observed, even if we give the neural networks a large number of hidden layer nodes (Swarup & Gasser, To Appear). This is because the dynamic induced by the classification game learning algorithm selects low-complexity equilibria.

Let us suppose that the agents are first allowed to train individually on the training set until convergence. This means that they have arrived at internal representations that do not violate the latter two conditions above. After that, we put them in communication with each other, according to the classification game. It turns out that this still results in the emergence of low-complexity internal representations (that also solve the classification problem). The advantage is that this allows us to ignore the learning transients in the discussion of equilibrium selection. Incidentally, it also shows that there is no implicit bottleneck (Vogt, 2005b) at work here, because, in this case, all the agents are exposed to the entire training set *before* they start communicating with each other.

The key step in the learning process responsible for the emergence of low-complexity equilibria is the training of the hearer's encoder. Note that the hearer uses the encoding generated by the speaker to predict the label and consequently to train its decoder. However, then it proceeds to train its encoder to produce an encoding similar to the encoding provided by the speaker, as described in section 3.1, and illustrated in figure 1. This renders high-complexity equilibria of the sort described above unstable. In fact, any equilibrium where the encodings chosen by the agents are not aligned is unstable, since hearers will update their encoders in this case, even if they do not make an error in classification.

Consider the case where agent *j* has an internal representation of the training set that is less complex than that of agent *k*, due to repetitions. Further, suppose that one of these repeated encodings is in conflict with one of the encodings of agent *k*. This conflict is more likely to be detected when *j* is the speaker and *k* is the hearer than vice versa. This is because the given example has to be exactly the conflicting one when *k* is the speaker, but can be any of the ones which are labeled with the conflicting encoding when *j* is the speaker. This means that *k* (as hearer) is more likely to change its encoding than *j* (when *j* is hearer). This means that *j*'s internal representation is more stable because it is of lower complexity.

Second, consider the case where the encodings chosen by *j* and *k* are *not* in conflict, but not identical either. Suppose that the encoding assigned by *j* to positive point x_1^+ is the same as that assigned by *k* to positive point x_2^+. In this case, if they are presented with x_1^+ and *j* is the speaker, *k* will not make an error, but will still update its encoding to bring it closer to the speaker's. This means that it will acquire the same encoding for both x_1^+ and x_2^+, thus making its encodings simpler.

It is possible that some of these updates cause new violations of conditions 2 and 3, which the agents will correct when they are in the role of speaker (since, being unaware of the hearer, the speaker only acts to reduce its classification error).

The net result of this process is the emergence of a shared low-complexity internal representation, which nevertheless has low error on the classification problem. These representations are the compositional languages.

3.2 The ICG protocol

We create an iterated version of the classification game as follows. We generate an initial population of *N* randomly initialized adult agents and train them individually on the task for a short period of time. Then we generate another population of *N* child agents who are also randomly initialized, and where each child agent is associated with a unique adult agent (its parent). Each parent then plays the classification game with its child, but the parent is always the speaker and the child is always the hearer. This is known as the *parent learning phase*.

After the parent learning phase, the adult population is discarded. The child agents now interact with each other repeatedly, in random speaker-hearer pairings, playing the classification game with each other. This is known as the *peer learning phase*.

After this phase, the children are assumed to have become adults. They now have children of their own (one per adult agent), and the process is repeated. Each *generation*, thus, except the first, consists of a parent learning phase and a peer learning phase. We run experiments for multiple generations, until a stable language emerges.

In the next section we show the results of this process through experiments.

4 Experiments

The *xor* task is a little too simple to demonstrate emergence over multiple generations, so we extend it to create a slightly more difficult task. The inputs now have 12 bits, and output is the *xor* of 3 of them[4]. Since a neural network does not have any in-built notion of adjacency or sequence, we can, without loss of generality, set the output to be the *xor* of the first 3 bits. This means that the output is 1 if there is an odd number of 1s in the first 3 bits of the input, otherwise it is 0. The remaining 9 bits are noise, i.e., they are irrelevant to determining the label.

We create a population of 10 new agents in each generation, i.e. 10 neural networks, with 10 hidden layer nodes each. This is a small population size, but it is sufficient to show the results we need. Since there are 10 parents (the learners from the previous generation) that teach the language to the 10 new agents (children), effectively there are 20 agents in each generation except the first. Increasing the population size beyond this does not result in a qualitative difference in the results. The agents actually only need 3 hidden layer nodes to be able to

[4]The *xor* of more than 2 bits is also known as the *odd-parity* function. It is achieved by taking the *xor* of the first two bits, and then the xor of this result with the next bit, and so on. This results in an output of 1 if there are an odd number of 1s in the bit sequence.

perform the classification task accurately, as we will see, but the extra hidden layer nodes serve to provide the capacity for generating more complex languages.

We create a training set by choosing 1000 input vectors and the corresponding labels randomly from the 4096 possible. Once chosen, this training set is fixed and is used for each generation. Thus each generation sees the same 1000 examples. The parent learning phase is set to be 100,000 interactions (time steps) long, and the peer learning phase is set to be 900,000 interactions long. Thus a generation is 1 million interactions long. These numbers, for the lengths of the two learning phases, are chosen to be somewhat realistic in the sense that an agent's peer-learning phase is set to be about 10 times as long as its parent-learning phase. If we assume that children enter the peer learning phase once they start talking, then the parent-learning phase would last until about age 2, and the peer learning phase would last up to about age 20, at which point they are assumed to become parents. The system is actually not very sensitive to the relative lengths of the peer and parent learning phases, though the peer learning phase needs to be longer than the parent learning phase. The actual number of interactions that determine the length of the learning phase depend on the particular learning problem and on the amount of data available. One time step corresponds to one interaction between a randomly chosen parent and its child. Thus, since there are 10 agents and 1000 input vectors, it means that each child agent gets to see each input vector 10 times in the parent learning phase, on average. Similarly, each agent gets to see each input vector 90 times in the peer learning phase, 45 times as speaker, and 45 times as hearer, on average.

The learning rate for the neural networks is set to 0.06, and each child population is initialized with random weights in the range [−0.5, 0.5].

We evaluate four error measures on the child population after every 10,000 time steps. The classification error on all possible 4096 samples by speakers is called simply *speaker error*. The classification error by speakers on the training set of 1000 samples is known as the *speaker training error*. Additionally, for each hearer, we choose a random speaker (a different one for each hearer) and then evaluate the hearer's errors, given this random speaker's utterances, on all possible samples and on the training set, which we call *hearer error* and *hearer training error* respectively. The speaker error measures tell us how well the population is doing on the task, on average. The hearer errors tell us how well-converged are the representations learned by the population, on average. We choose only *one* random speaker for calculating hearer error for computational efficacy, which can lead to an overestimate of the true hearer error if there is a poor speaker in the population. However, in practice, it does not affect the results much.

4.1 Measuring compositionality

Finally, to evaluate the complexity of the emergent languages, we measure the average entropy of the population's representations as follows. For each speaker, i, we calculate the probability of utterance, p_s, of each symbol, s, (where a symbol is one letter of the alphabet, corresponding to one hidden layer node), and calculate the entropy as,

$$H(i) = -\sum_{s} p_s \log_2 p_s, \tag{4}$$

where we assume $0 \log_2 0 = 0$. p_s is simply the number of training examples for which s appears in the utterance generated by the agent, divided by the total number of training examples. We then find the average entropy of the population by taking the average of $H(i)$ over all i. This gives us a measure of the complexity of the language of the population. The higher the complexity, the more *holistic* the language. Conversely, the lower the complexity, the more *compositional* the language.

In general, for a given number of meanings to be expressed, there will be more than one language capable of expressing them. Importantly, these languages will not all have the same complexity. Suppose we define the complexity of a language to be its information entropy, as above. Then a holistic language, which assigns a unique utterance to each meaning, will have higher complexity than a compositional language, which assigns unique utterances to a subset of the meanings and expresses the rest using combinations of these basic utterances. If we assume a uniform distribution over meanings, it is easy to see that the set of utterances generated by the holistic language will have the highest entropy. The set of utterances generated by the compositional language will have lower entropy because of repetitions. Conversely, for a given complexity (entropy) value, a compositional language will be more expressive than a holistic language.

At the other extreme, zero complexity languages are the ones that are referentially useless. They correspond to saying nothing at all, or saying the same thing at every instant. In other words, they carry no information. These can also be interpreted as holistic languages. Therefore, in the following experiments, the entropy and the classification error need to be taken into consideration together. A language whose complexity is too low will have high classification error, and a zero complexity language will have the same error as random classification. Ideally we would like the languages that evolve to be in the region that is “just right”, i.e., where the language that evolves is both easily learnable and just complex enough for the given classification task. The optimal complexity depends on the particular classification task, and is not practical to compute in general. We will see in the following experiments that the languages that emerge in the ICG have substantially lower complexity, in addition to achieving zero classification error, than the corresponding version of the ILM.

In addition, we also present the emergent languages (not just their entropy), in figure 5 and table 1, so that we can see that the language that emerges in the ICG is in fact compositional (and nearly optimal in this particular case).

4.2 Results

The next three experiments are designed to show the following:

- the gradual emergence of a stable, compositional, and close to optimal linguistic system over the course of a few generations in the ICG, where each generation of agents is able to learn the right language very quickly,
- the long time-scale required for the emergence of the same linguistic solution in a population of agents playing the classification game, without iteration through generations, which demonstrates the large speed-up obtained in the previous experiment, and
- the high complexity of the resulting language in the absence of peer interaction, which makes the system equivalent to the ILM.

Together, the three experiments will show that the ICG is able to resolve all the three problems with the ILM, while also showing that cultural transmission through generations results in a speedup in learning in each generation that is analogous to the speedup in the face of the poverty of the stimulus observed empirically in child language acquisition. This is discussed in more detail in section 5.

The results of the first experiment are shown in figure 3, over 10 generations of the ICG. The system is initialized by creating a population of 10 agents who are allowed to train individually (i.e., without interaction) on the training set for 100,000 time steps. They then become the parents in the first generation. Since the child population is always initialized randomly, the beginning of a generation shows up as a spike in the entropy of the representation. This serves

to mark the boundaries between generations nicely. The learning curves (speaker and hearer errors) lie mostly on top of each other, which means that the agents are generalizing well.

In the space of a few generations, four in this case, a stable linguistic system emerges. Each child population begins with high error on the task due to the random initialization, but through parent and peer interaction, quickly learns to solve the problem. Further, the representation that emerges is stable across generations, as shown by the entropy curve, which rises to the same level in each generation and then does not waver. The flattening of the entropy curve shows that the *structure* of the language stabilizes in a few generation (which is the important aspect under consideration). However, we have checked that the lexicon (i.e., the set of particular symbols used by the agents) also remains stable once the structure stabilizes. It turns out that these learning and entropy curves are extremely stable and representative of the behavior of the system, since they show the average over a population. Across runs, the only variation occurs in the number of generations required for the emergence of a stable linguistic system. For the task presented here, this number is generally small, both because the task is quite simple and because the population only consists of 10 agents. Increasing either task complexity or population size results in a larger number of generations before the emergence of a stable system. By running the system out to several generations after the emergence of the final language, we see that this emergent linguistic system remains stable.

The results of the second experiment, in figure 4, show the same measures for a single population of 10 agents playing the classification game, which is equivalent to a single generation of only peer learning. We see that it takes a very long time to converge to a solution, approximately 3.2 million time steps. In contrast, we can see from figure 3 that once a stable iterated system emerges in the ICG, each generation converges on the solution in less than 200,000 time steps. This is a more than 15-fold speedup, which shows that a combination of parent and peer learning is much more effective than peer learning alone.

The table in figure 5 shows the utterances produced by one of the agents in the 10^{th} generation of the first experiment. Since only the first 3 bits matter for the label, we show only these in the first column. In the second column we show the utterances produced by the agent. The agents have learned to ignore the noise bits, so they generate the same utterance for all the inputs that have the same first 3 bits. Further, some symbols are redundant as they are generated for all the inputs. These are like the symbols B and D in figure 2(c). The third column, therefore, shows the information-carrying symbols, which are actually used in decision-making by the agents. This is close to an optimal solution, since only three symbols (i.e. three hyperplanes) are needed to accomplish the task. Eliminating either C or H would make it optimal. The diagram in figure 5 shows a schematic view of how the hyperplanes corresponding to the symbols partition the 3-dimensional subspace of the input that is relevant to determining the label.

In the third experiment, we create a single agent in each generation. In addition, this time we eliminate the peer learning phase. Thus, each agent passes on its language to its own child in each generation, making the system equivalent to the ILM. The resulting learning and entropy curves are shown in figure 6.

We see that the learning curves go to zero in a few generations, as in the ICG. Note that the generations are much shorter now, since they only consist of the parent learning phase. The entropy curve, in this case, settles at a much higher value than in figure 4. This indicates that the language found by the agents has much less compositionality than in the ICG. This happens because the agents are unable to find the optimal solution, and fit their solution to the noise bits also. It is worth noting, though, that a stable linguistic system *does* emerge, as indicated by the flattening of the entropy curve in figure 6.

We can examine the emergent language after it stabilizes, as before. A sample is shown in table 1. Again, we only show the bits relevant for classification in the first column. However, since the agents are varying their utterances according to the remaining 9 bits also, we present a sample of utterances generated by an agent in the 20^{th} generation for each set of equivalent (w.r.t. classification) inputs. These are shown in the second column of table 1. From the utterances the agent generates for each equivalence class, we can extract the common set of symbols, which are actually relevant to classification. These are shown in the third column. We see that the agent is using all 10 possible symbols (i.e. all 10 hidden layer nodes) for classification and communication, since there are no symbols that are either absent or repeated for all possible inputs. Further, even the set of symbols being used for classification is more complex than the set being used for classification in the ICG (compare the right columns of the table in figure 5 and table 1). The complexity of the entire set of utterances generated by the agent in the ILM is, of course, even higher because it generates symbols in response to the noise bits also.

5 Discussion

The simulations show that by connecting languages to tasks, and by including a peer learning phase, we are able to address all of the criticisms of the ILM.

- The emergent language is grounded because it is relevant to the classification task, as shown in figure 5.
- Agents are no longer just learning passively from their parents. They take an active role in the peer learning phase (are speakers as well as hearers).
- They do not have an innate bias for inducing compositional languages, as demonstrated in figure 2(b), and also by the entropy curves in figure 6. Rather, compositionality is induced by peer learning.

Thus we see the emergence of a structured language over a period of generations, through cultural learning alone. The ICG offers a different interpretation of the cultural learning process, however. It seems, here, that the role of parent learning is to appropriately initialize the language system of the children, so that they are able to quickly find the “correct” language through interaction. Without this initialization, a population takes an extremely long time to converge on a solution through peer learning alone, as demonstrated in figure 4. The massive speedup that is obtained through combining parent learning with peer learning in the ICG (over peer learning alone) is an interesting result in itself. Vogt (2006,fig. 3) shows a similar result, though he does not show how long it takes for a single generation to learn the language through peer learning alone. This speedup presents an interesting analogy to the poverty of stimulus argument. According to these experiments, a language that takes very long to learn through interaction alone can be learned much faster through a combination of parent and peer learning. Smith, Brighton, and Kirby (2003) have suggested that the poverty of stimulus paradox is solved by the emergence of structured languages through iterated learning. They show that the language that emerges through iterated learning is one which can be learned from sparse input. In their case, however, it is not clear that there is a difference between the time to learn the *final* language with and without iteration. Note that we are not judging *learnability* of a language, but how quickly it is *actually learned.* This is an important distinction. Learnability is always an “order-of” statement (i.e. it ignores constant factors), such as “this language can be learned in exponential time”. Judging learnability is independent of whether the language is a product of iteration or not. However, the point is that actual learning performance depends on the constant factors also. What we are showing is that a language can be learned much faster in practice through a combination of parent and peer learning (after a few generations of iteration), than through peer learning alone. Since the language is the same in both cases, its learnability is also the same, but the actual time taken to learn differs (greatly, as we show).

This is very relevant to the Poverty of Stimulus (PoS) debate, since that is about the actual time taken by children to acquire language, not about the abstract learnability of natural language. PoS says that children do not have enough time to learn language (based on arguments about the complexity of natural language), and the ILM response is simply to say that they *do* have enough time (based on arguments about the simplicity of languages emerging from ILM experiments). Our response to PoS is stronger, we believe, because it shows that even if it seems like language is too complex to learn by children alone, it is still possible to learn it much more quickly through a combination of parent and peer learning.

5.1 The Bottleneck Question

In Brighton's instantiation of the ILM, the bottleneck size, *b*, is a key parameter. This is because when *b* is small, the agents are faced with hitherto unseen meanings more often, and consequently have to generalize or *invent* more often. This offers the opportunity to add structure to the language, and results in the gradual simplification and structuring of the language over generations. The smaller the bottleneck size, the quicker a structured language emerges.

To investigate the role of the bottleneck in the ICG, we try varying the number of examples available to a parent-child pair in the parent learning phase. The simulation is set up as follows. We choose 1000 of the 4096 possible examples to make up the training set as before. In the parent learning phase, we only provide a randomly chosen subset of *b* examples, where *b* is the bottleneck size. In the peer learning phase, all 1000 examples are available to the population. In the parent learning phase for the next generation, we choose another random subset of *b* examples from the training set, and so on. We vary *b* from 100 to 500, in steps of 100, and run each simulation 10 times. Each simulation of the ICG was run for only 5 generations because experiments have shown that the populations always converge to a stable language in about 4 generations. The main finding is that the complexity of the resulting language is not sensitive to the bottleneck size, nor is the number of generations required for the emergence of a stable linguistic system sensitive to bottleneck size. A stable system emerges almost always in about four generations for this learning problem.

We also compared these experiments with the case where there is no peer learning, i.e. the ILM version of the ICG. Each simulation of the ILM was run for 20 generations, even though the population actually converges to a stable language in the space of just a few generations. Again, when we vary the bottleneck size as before, there isn't a noticeable difference in the complexity of the resulting final language. We conclude that the final language complexity in our model of the ILM is not sensitive to the bottleneck size. This result is analogous to the bottleneck-invariance in Iterated Bayesian Learning (Kirby et al., 2007), where the stationary distribution over languages is not affected by bottleneck size[5]. For smaller bottleneck sizes, we do see a reduction in complexity over generations, i.e. the language that emerges in early generations has a higher complexity that gets reduced over time. An example, for $b = 100$, is shown in figure 7. Thus, for small bottleneck sizes, we see a *trajectory* that is similar to the ILM of Brighton, where the complexity of the language reduces over generations (though the effect is much smaller in our implementation).

The comparison of the complexity of the resulting final languages for the ILM and the ICG is shown in figure 8. We see from the figure that the entropy of the resulting language, after the system has stabilized, is more or less constant for the various bottleneck sizes. Further, the error bars are much tighter for the ICG than for the ILM. In this sense, the ICG is more stable.

[5]Incidentally, this implies that the final language complexity in Iterated Bayesian Learning is also not sensitive to bottleneck size, since it is the expected complexity over the stationary distribution.

In the ILM of Brighton and in the Iterated Bayesian Learner of Griffiths and Kalish, in contrast to the above experiment, reducing the bottleneck size makes the population converge *faster* because the prior distribution has a greater effect on the learner in each generation. This seems a little counter-intuitive. It suggests that if children learn very little language from their parents, the language will become structured very quickly, whereas if they spend a lot of time learning language from their parents, their language will remain unstructured. The odd thing about this is that it seems like the ideal solution would be not to learn language from parents at all, but to create it anew in each generation.

In the ICG, on the other hand, if the agents spend little time learning from their parents, i.e., if we reduce the length of the parent learning phase, they have to spend a lot of time interacting with their peers to induce structure in their language. There is no "free lunch" when it comes to creating a structured language. In this sense, we believe that the ICG offers a more realistic picture of the cultural transmission of language.

It turns out that there is evidence for both kinds of phenomena. Verb regularization is a good example of children inducing additional structure in their language early in learning, and losing it on further learning. In this case, increasing the period of acquisition corresponds to *reducing* the amount of structure in the language, as predicted by the ILM. However, it is not clear if this observation applies to language as a whole.

In the development of the Nicaraguan Sign Language, for example, it has been observed that each generation learns from the previous and then induces additional structure through peer interaction, i.e. the complexity has been increasingly gradually over generations, which more closely matches the phenomenology of the ICG. This is worth remarking upon in more detail.

5.2 The Nicaraguan Sign Language

The Nicaraguan Sign Language (NSL) has emerged over the last approximately 30 years (Kegl & Iwata, 1989; Senghas, Kita, & Ozyurek, 2004; Senghas, 2003). It started with the establishment of a school for special education in Managua in 1977. Deaf enrollment in the school has increased gradually since then. Before this period, deaf children had little interaction with people outside their families, and especially, had no interaction with other deaf children. They generally had some system of "homesign" that they would use to communicate with members of their families, but not a fully developed sign language. Once some deaf children began to interact with each other in and out of the school, they began to develop a more mature and fully-formed sign language. Each new cohort of children that joined the school learned the existing signing system from their predecessors and then expanded upon it, making it more complex and language-like. This has resulted in the youngest signers being the most fluent, while the oldest signers (since the language is only about 30 years old, its history can be traced back to the earliest generation still) are the least fluent.

The phenomenology of the ICG is remarkably similar. The initial population, which is allowed to train on the task individually for a short while at first, is like the initial cohort at the school, who had only simple homesign systems they had developed themselves to rely on. As soon as the first peer learning phase occurs in the ICG, the complexity of the language drops to nearly zero, and then increases gradually until it stabilizes. This is similar to the gradual increase in the complexity of the NSL that has been reported. Since the learning task we have used in the ICG is quite simple, the emergent language stabilizes quite quickly. The NSL is understandably more complex and has taken longer to emerge. However, the important point to note is that in the ICG, as in the NSL, the final structure of the language does not emerge in a single generation. Vogt (2005b) has also modeled the NSL using a combination of iterated learning and language games, as described in section 2.1, but in his model structure appears in the language in a single generation and then remains stable.

The NSL is the best (certainly best studied) example available to us of the emergence of a new language over generations (The Al-Sayyid Bedouin Sign Language (Sandler, Meir, Padden, & Aronoff, 2005), is another). It deserves much closer analysis and modeling, and and this presents an interesting direction for future research.

6 Conclusion

We have shown that cultural transmission of language does have an important role to play in the emergence of structure in language. This phenomenon can be demonstrated without having to build in any preference for a structured language into the agents. Instead, interactions with peers during learning result in simplification, and hence structuring, of the language.

We have reviewed the Iterated Learning Model, and some important critiques of it, and have shown how all of these are addressed in our model. We link the emerging language to a task that the agents have to perform, via the classification game model. This provides the grounding for the language. This grounding is essential because it provides the pull of functionality to counter the tendency for populations to simplify the language to make it more learnable. This tradeoff is reminiscent of the joint least effort model of Ferrer i Cancho and Solé (2003). They showed that jointly minimizing the entropy for the speaker and the hearer results in a linguistic system with a scaling structure very similar to natural language. The hearer's "effort", in our model, can be conceived of as the attempt to interpret the speaker's utterance in a functional way, i.e. to deduce the label. We show that this tradeoff results in a language that is close to optimal with respect to the task at hand. Exploring the connections between this model and the least effort model presents an interesting opportunity for future research.

We have split the acquisition phase into an idealized parent learning phase and an idealized peer learning phase. In the real world, of course, the distinction between parents and peers is blurred. Essentially when a child starts speaking, it can be said to have entered a peer learning phase, since now it can take on the roles of both speaker and hearer. This peer learning process is essential in our model for inducing structure in the language of the child. If we eliminate the peer learning phase, we see that the resulting language is significantly more complex than necessary.

Finally, and perhaps most importantly, we have demonstrated that the agents do not have a built-in preference for structured language, or anything else that could be interpreted as a Universal Grammar. If trained individually on the task, they inevitably end up with a very unstructured (i.e., holistic) language. The peer interaction takes on the role played by the minimum description length bias and the structured invention procedure in the ILM. This results in a much stronger argument for the effects of cultural transmission on the structure of language.

In effect, the Iterated Learning Model and the Classification Game each solve a problem for the other. The ILM addresses the long convergence times of the CG, and the CG addresses the biasing problem for the ILM. There is much room for future work. We have presented a game-theoretic perspective on the complexity regularization phenomenon in the classification game, but much work can still be done along these lines, such as working out the precise computational mechanism that enables agents to find low-complexity equilibria. In addition, we need to investigate more closely the amount of speedup that is obtained through the combination of parent and peer learning over peer learning alone, and its significance to the poverty of stimulus argument. We also believe that the ICG can be used to create a much more compatible model of the emergence of Nicaraguan Sign Language than has been accomplished earlier.

Acknowledgments

We thank our external collaborators, members of the Network Dynamics and Simulation Science Laboratory (NDSSL) at Virginia Tech, members of the Language Evolution and Distributed Information Systems (LEADS) group at the University of Illinois at Urbana-Champaign, and the anonymous reviewers for their suggestions and comments. Material in this paper is based upon work supported by the U.S. National Science Foundation under Grants No. IIS-0205346, CNS-062694 and CNS-0831633. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. S. S. has also been partially supported by HSD Grant SES-0729441, CDC Center of Excellence in Public Health Informatics Grant 2506055-01, NIH-NIGMS MIDAS project 5 U01 GM070694-05, and DTRA CNIMS Grant HDTRA1-07-C-0113.

References

Baron, NS. Growing up with language: How children learn to talk. Reading, MA: Addison-Wesley; 1992.

Barron, AR. Complexity regularization with application to artificial neural networks. In: Roussas, G., editor. Nonparametric functional estimation and related topics. Boston, MA, and Dordrecht, The Netherlands: Kluwer Academic Publishers; 1991. p. 561-576.and Dordrecht

Brighton H. Compositional syntax from cultural transmission. Artificial Life 2002;8(1):25–54. [PubMed: 12020420]

Brighton, H. Linguistic evolution and induction by minimum description length. In: Werning, M.; Machery, E., editors. The compositionality of concepts and meanings: Applications to linguistics, psychology and neuroscience. Frankfurt: Ontos Verlag; 2005.

Brighton, H.; Kirby, S.; Smith, K. Cultural selection for learnability: Three principles underlying the view that language adapts to be learnable. In: Tallerman, M., editor. Language origins: Perspectives on evolution. Oxford: Oxford University Press; 2005.

Cangelosi A, Greco A, Harnad S. From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. Connection Science 2000;12(2):143–162.

Chomsky, N. Aspects of the theory of syntax. Cambridge, MA: MIT Press; 1965.

Chomsky, N. Knowledge of language: Its nature, origin, and use. Westport, CT: Praeger Publishers; 1986.

Christiansen MH, Kirby S. Language evolution: Consensus and controversies. Trends in Cognitive Science 2003 July;7(7):300–307.

Croft W. Evolutionary linguistics. Annu. Rev. Anthropol 2008 October;37:219–234.

De Beule, J. The emergence of compositionality, hierarchy, and recursion in peer-to-peer interaction; Proceedings of the 7th international conference on the evolution of language (evolang); 2008.

Dominey PF. From sensorimotor sequence to grammatical construction: Evidence from simulation and neurophysiology. Adaptive Behavior 2005;13(4):347–361.

Edelman S, Waterfall H. Behavioral and computational aspects of language and its acquisition. Physics of Life Reviews 2007;4:253–277.

Ferrer i Cancho R, Solé RV. Least effort and the origins of scaling in human language. PNAS 2003;100:788–791. [PubMed: 12540826]

Foraker, S.; Regier, T.; Khetarpal, N.; Perfors, A.; Tenenbaum, JB. Indirect evidence and the poverty of the stimulus: The case of anaphoric *One*; Proceedings of the 29th annual conference of the cognitive science society; 2007.

Geertz, C. Religion as a cultural system. In: Banton, M., editor. Anthropological approaches to the study of religion. Routledge: 2004.

Griffiths, TL.; Kalish, ML. A Bayesian view of language evolution by iterated learning; Proceedings of the 27th annual conference of the cognitive science society; 2005.

Harnad S. The symbol grounding problem. Physica D 1990;42:335–346.

Harris JR. Where is the child's environment? A group socialization theory of development. Psychological Review 1995;102(3):458–489.

Harris, JR. The nurture assumption: Why children turn out the way they do. Free Press; 1998.

Haykin, S. Neural networks: A comprehensive foundation. 2nd ed.. Prentice Hall: 1998.

Hinton, G.; van Camp, D. Keeping neural networks simple; Proceedings of the international conference on artificial neural networks; Amsterdam: 1993. p. 11-18.

Hochreiter S, Schmidhuber J. Flat minima. Neural Computation 1997;9(1):1–42. [PubMed: 9117894]

Kärkäinnen T, Heikkola E. Robust formulations for training multi-layer perceptrons. Neural Computation 2004;16(4):837–862.

Kegl, J.; Iwata, G. Lenguaje de Signos Nicaragüense: A pidgin sheds light on the “creole?” ASL. In: Carlson, R.; DeLancey, S.; Gilden, S.; Payne, D.; Saxena, A., editors. Proceedings of the fourth annual meeting of the pacific linguistics conference; Eugene: University of Oregon, Dept. of Linguistics, Oregon; 1989.

Kirby, S. Learning, bottlenecks and infinity: a working model of the evolution of syntactic communication. In: Dautenhahn, K.; Nehaniv, C., editors. Proceedings of the aisb’99 symposium on imitation in animals and artifacts. 1999.

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

Kirby, S. Learning, bottlenecks and the evolution of recursive syntax. In: Briscoe, T., editor. Linguistic evolution through language acquisition: Formal and computational models. Cambridge University Press; 2002.

Kirby, S. The evolution of meaning-space structure through iterated learning. In: Lyon, C.; Nehaniv, C.; Cangelosi, A., editors. Emergence of communication and language. Springer Verlag; 2007. p. 253-268.

Kirby S, Dowman M, Griffiths TL. Innateness and culture in the evolution of language. PNAS 2007 March 20;104(12):5241–5245. [PubMed: 17360393]

Kirby, S.; Hurford, J. The emergence of linguistic structure: An overview of the iterated learning model. In: Cangelosi, A.; Parisi, D., editors. Simulating the evolution of language. London: Springer Verlag; 2002. p. 121-148.

Kirby S, Smith K, Brighton H. From UG to universals: Linguistic adaptation through iterated learning. Studies in Language 2004;28(3):587–607.

Lappin S, Shieber SM. Machine learning theory and practice as a source of insight into universal grammar. J. Linguistics 2007;43:393–427.

McGrew WC. Culture in nonhuman primates? Annu. Rev. Anthropol 1998;27:301–328.

Mitchell, TM. The need for biases in learning generalizations. Rutgers University; 1980. (Tech. Rep. No. CBM-TR-117)

Perfors, A.; Tenenbaum, JB.; Regier, T. Poverty of the stimulus? A rational approach; Proceedings of the 28th conference on the cognitive science society; Vancouver: 2006. p. 663-668.

Pinker S, Bloom P. Natural language and natural selection. Behavioral and Brain Sciences 1990;13:707–784.

Sandler W, Meir I, Padden C, Aronoff M. The emergence of grammar: Systematic structure in a new language. PNAS 2005;102:2661–2665. [PubMed: 15699343]

Senghas A. Intergenerational influence and ontogenetic development in the emergence of spatial grammar in Nicaraguan Sign Language. Cognitive Development 2003;18(4):511–531.

Senghas A, Kita S, Ozyurek A. Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. Science 2004;305(5691):1779–1782. [PubMed: 15375269]

Smith, K. Cultural evolution of language. In: Brown, K., editor. The encyclopedia of language and linguistics. 2nd ed.. Elsevier; 2006. p. 315-322.

Smith K, Brighton H, Kirby S. Complex systems in language evolution: the cultural emergence of compositional structure. Advances in Complex Systems 2003 December;6(4):537–558.

Smith, K.; Kalish, ML.; Griffiths, TL.; Lewandowsky, S., editors. Phil. Trans. R. Soc. B. Vol. 363. 2008 Nov. Theme issue on cultural transmission and the evolution of human behaviour.

Smith K, Kirby S. Cultural evolution: Implications for understanding the human language faculty and its evolution. Phil. Trans. R. Soc. B 2008 November 12;363:3591–3603. [PubMed: 18801718]

Smith K, Kirby S, Brighton H. Iterated learning: A framework for the emergence of language. Artificial Life 2003;9(4):371–386. [PubMed: 14761257]

Steels, L. Emergent adaptive lexicons. In: Maes, P., editor. From animals to animats 4: Proceedings of the fourth international conference on simulation of adaptive behavior; Cambridge, MA: The MIT Press; 1996a.

Steels L. A self-organizing spatial vocabulary. Artificial Life 1996b;2(3):319–332. [PubMed: 8925502]

Swarup S. The classification game: Complexity regularization through interaction. Submitted. 2009

Swarup, S.; Gasser, L. Simple, but not too simple: Learnability vs. functionality in language evolution; Proceedings of the 7th conference on the evolution of language; Barcelona, Spain: 2008 Mar 11–15.

Swarup S, Gasser L. The classification game: Combining supervised learning and language evolution. Connection Science. (To Appear).

Swarup, S.; Lakkaraju, K.; Ray, SR.; Gasser, L. Symbol grounding through cumulative learning. In: Vogt, P.; Sugita, Y.; Tuci, E.; Nehaniv, C., editors. Symbol grounding and beyond: Proceedings of the third international symposium on the emergence and evolution of linguistic communication (eelc06). Rome, Italy: 2006 Sep.

Taddeo M, Floridi L. Solving the symbol grounding problem: a critical review of fifteen years of research. Journal of Experimental and Theoretical Artificial Intelligence 2005;17(4):419–445.

Tomasello M. The human adaptation for culture. Annu. Rev. Anthropol 1999 October;28:509–529.

Vogt P. The physical symbol grounding problem. Cognitive Systems Research 2002;3(3):429–457.

Vogt P. The emergence of compositional structures in perceptually grounded language games. Artificial Intelligence 2005a September;167(1–2):206–242.

Vogt P. On the acquisition and evolution of compositional languages: Sparse input and the productive creativity of children. Adaptive Behavior 2005b;13(4):325–346.

Vogt, P. Cumulative cultural evolution: Can we ever learn more?. In: Nolfi, S., et al., editors. From animals to animats: Proceedings of the 9th international conference on simulation of adaptive behavior (sab); Rome, Italy: Springer Verlag; 2006. p. 738-749.

von Humboldt, W. Linguistic variability and intellectual development. Philadelphia: Pennsylvania University Press; 1972. (Originally published 1836 by the Royal Academy of Sciences of Berlin, “Uber die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwickelung des Menschengeschlechts”)

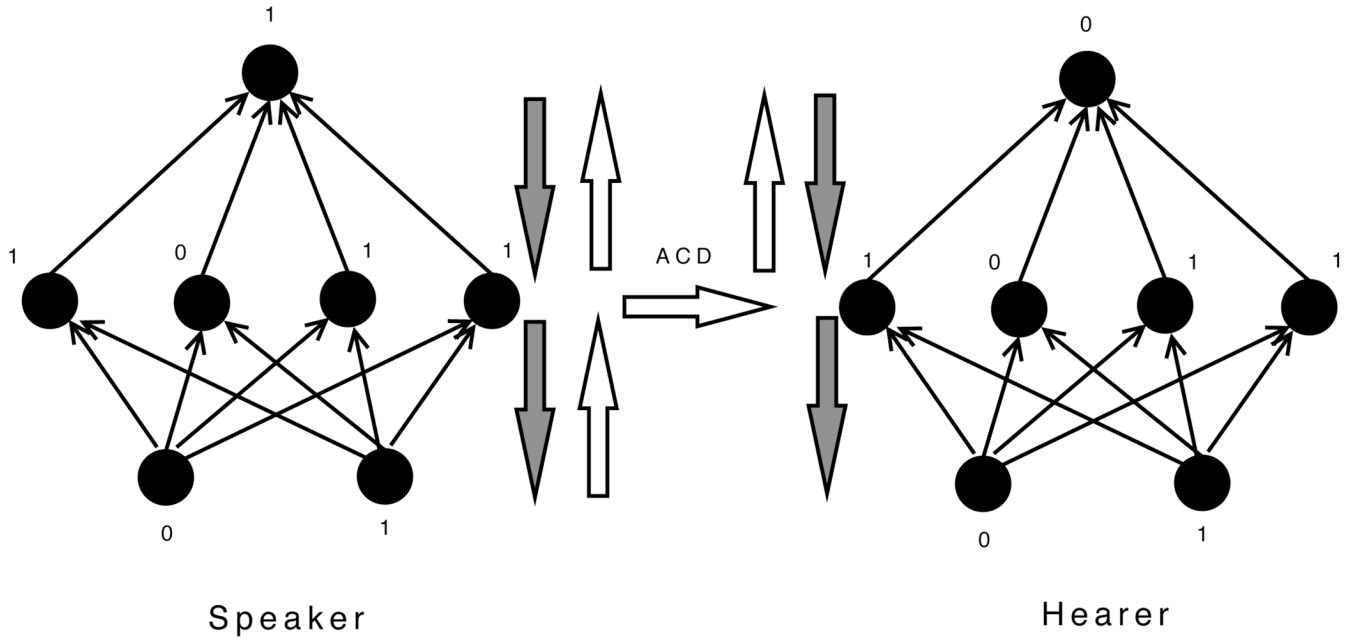
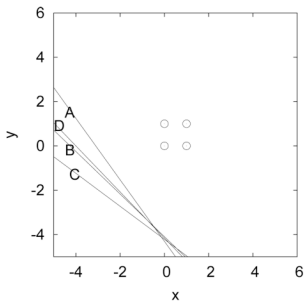
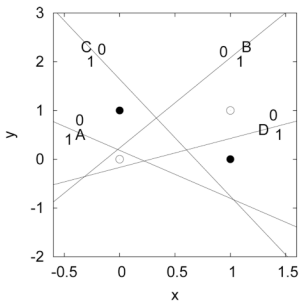


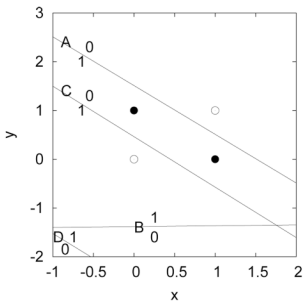
Figure 1.
Speaker-hearer interaction in the classification game. The speaker's internal state (1011) is converted into a public utterance (ACD) by mapping the active hidden layer nodes to letters of the alphabet. The white arrows show the direction of signal flow, and the gray arrows show the direction of error flow in the system.



(a) Communication but no external task.



(b) Task but no communication.



(c) Task and communication.

Figure 2.

The regularizing effect of the classification game is due to the push-pull between learnability and functionality. Figure 2(a) shows the effect of learnability alone which results in a representation of minimal complexity. The resulting language shown here is {C, C, C, C} because hyperplane C happens to be oriented in a manner that labels all the points 1, while the other hyperplanes are oriented in the other direction. Figure 2(b) shows the effect of functionality alone, which results in an overly complex representation. The resulting language shown here is {ABC, C, BD, B}. Finally, figure 2(c) shows the effect of combining the two forces, through the process of the classification game, which results in a tradeoff that achieves optimal complexity in this simple example. The resulting language shown here is {ACBD, ABD, ABD, BD}

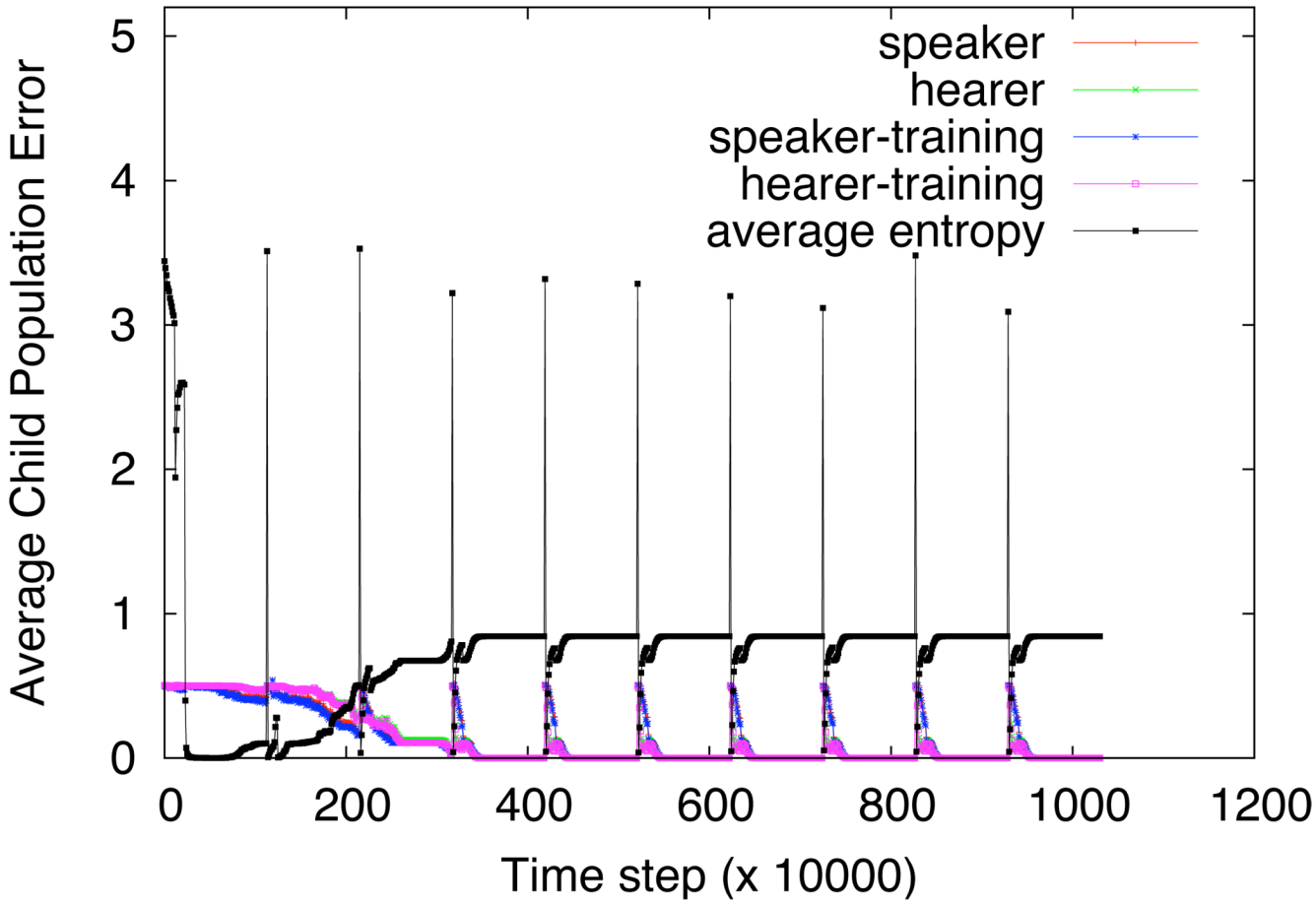


Figure 3.
Learning curves on the 12-inputs-first-3-xor task. We see that the error drops to zero in the space of a few generations. Over the same span, the complexity of the language increases to a stable value. Thereafter the linguistic system of the population stabilizes, and each generation is able to achieve this stable state very quickly.

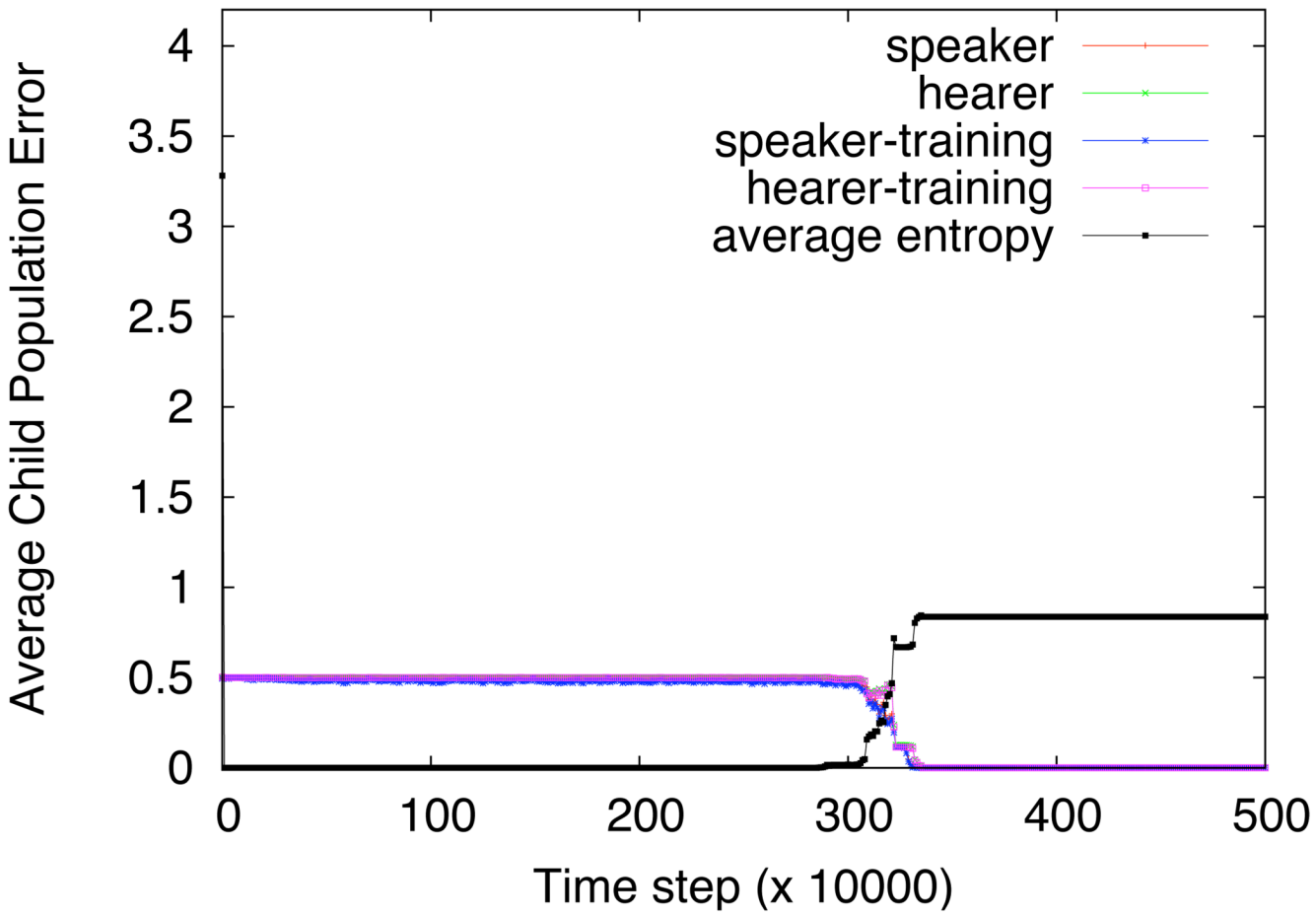


Figure 4.
Learning curves on the 12-inputs-first-3-xor task for the classification game, i.e., a single generation of peer learning. We see that the population takes a *much* longer time to converge.

First 3 input bits	Utterance	Relevant symbols
000	ACDFIJ	CF
001	ACDFGIJ	CFG
010	ACDFGIJ	CFG
011	ADFGHIJ	FGH
100	ACDIJ	C
101	ACDFIJ	CF
110	ACDFIJ	CF
111	ACDFGIJ	CFG

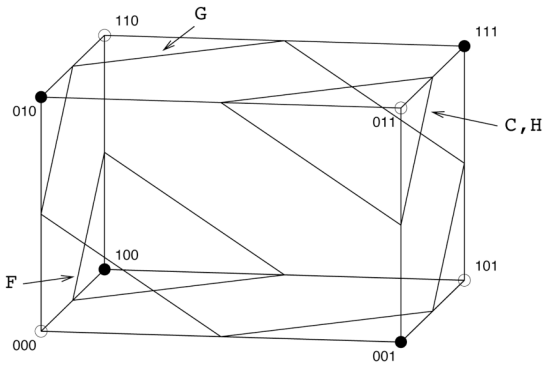


Figure 5.

The middle column shows the utterances produced by an agent in the 10^{th} generation of the ICG on the input data. Only the first three bits of the inputs are shown in the left column, because the rest are noise. The agent has learned to ignore the noise, as evidenced by the fact that it generates the same utterance for all inputs that have the same first 3 bits. Some symbols in the utterances are redundant (A, D, I, and J), since they are generated in response to all the inputs. The relevant, i.e. information-carrying, symbols are shown in the right column. The partitioning of the space by the symbols is shown schematically in the figure on the right.

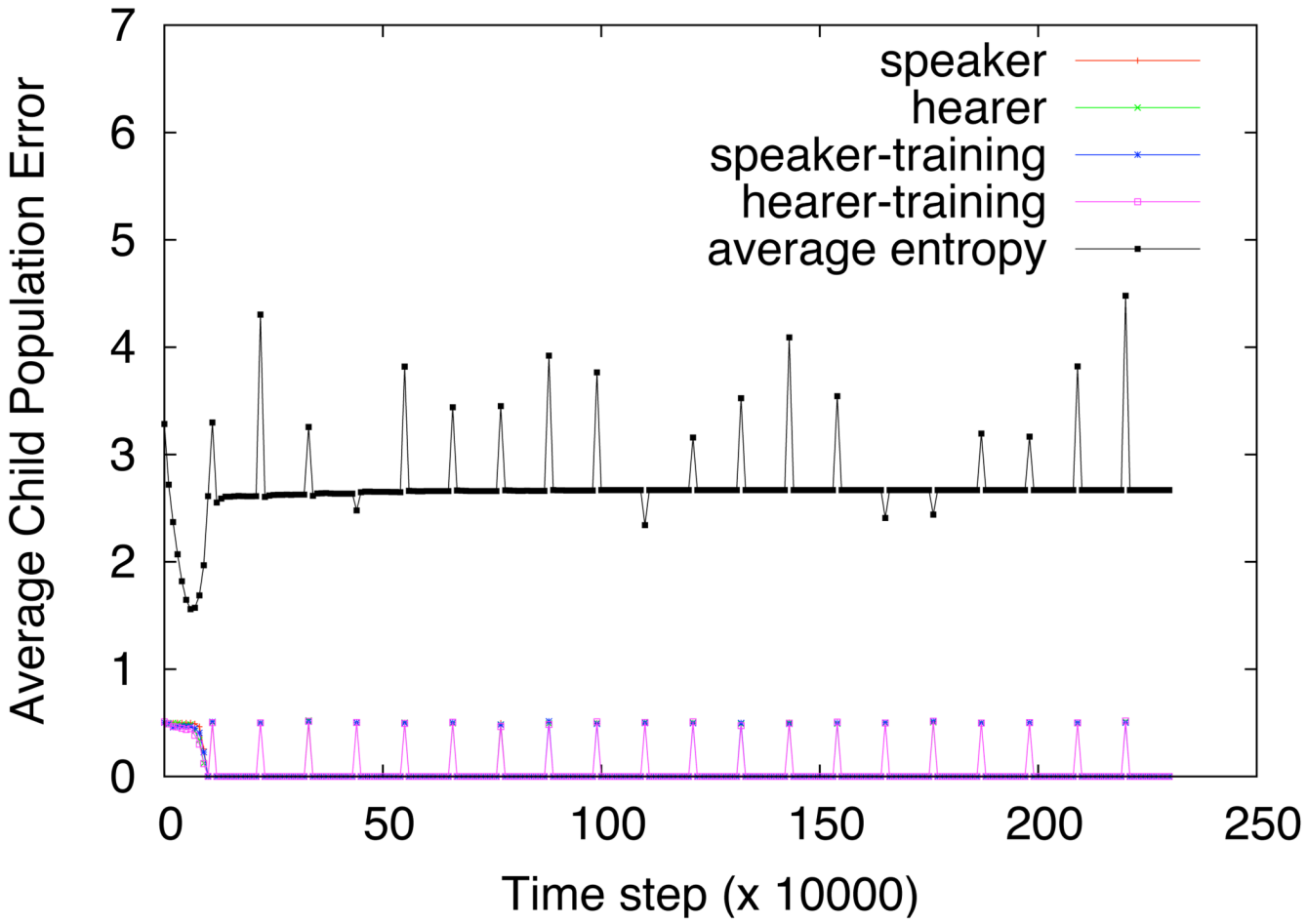


Figure 6.
Learning curves from running the ICG without peer learning, which makes it equivalent to the ILM. We see that although the agents find a correct solution, the resulting complexity is higher than that in the ICG (fig. 3).

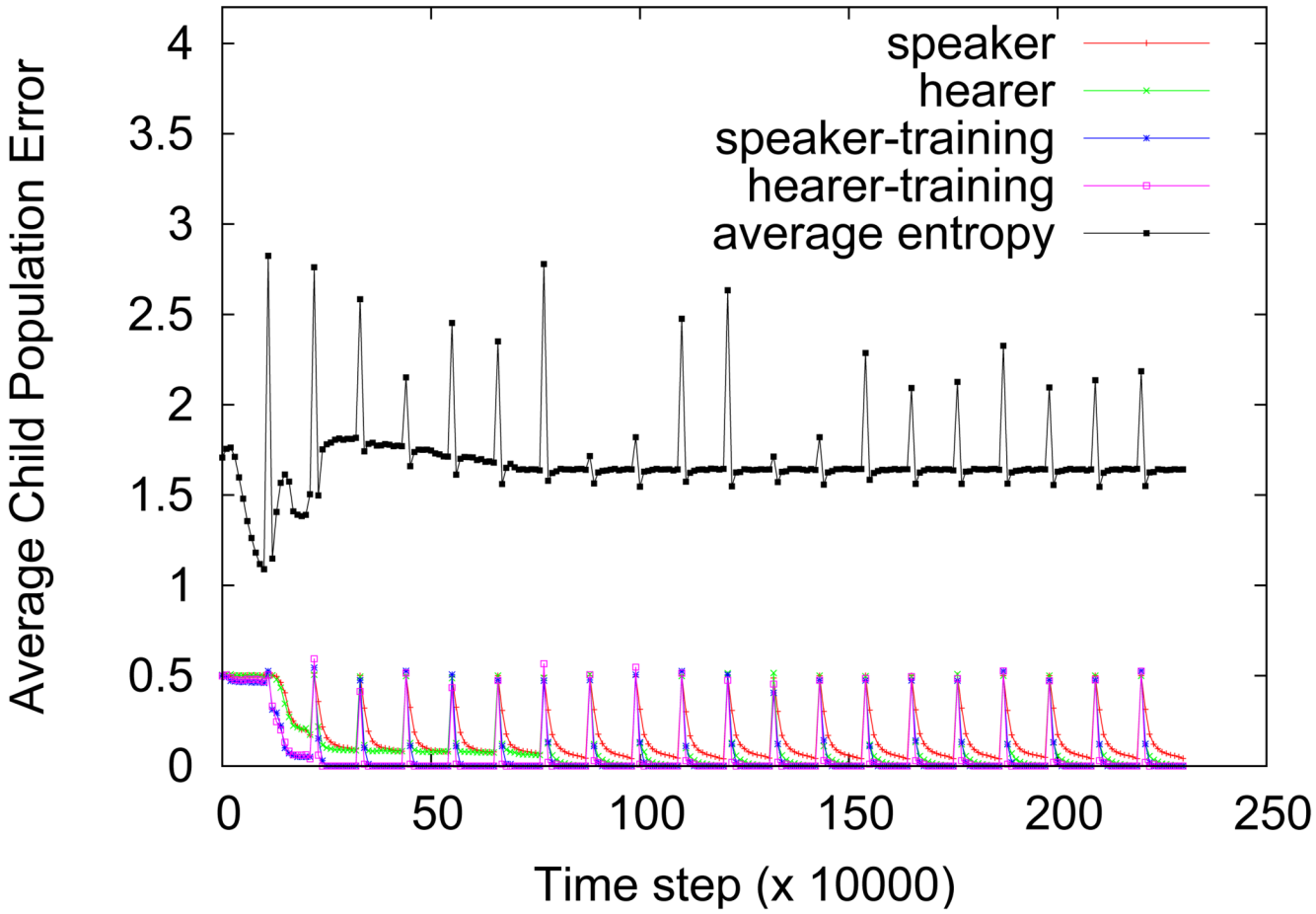


Figure 7.
Learning curves for the ILM, with bottleneck size = 100. We see that the emerging language has a higher complexity in early generations than in the later generations. This effect disappears as the bottleneck size is increased, as can be seen in figure 6, for example. The final complexity of the emergent language remains much higher than that obtained in the ICG, as can be seen in the comparison in figure 8.

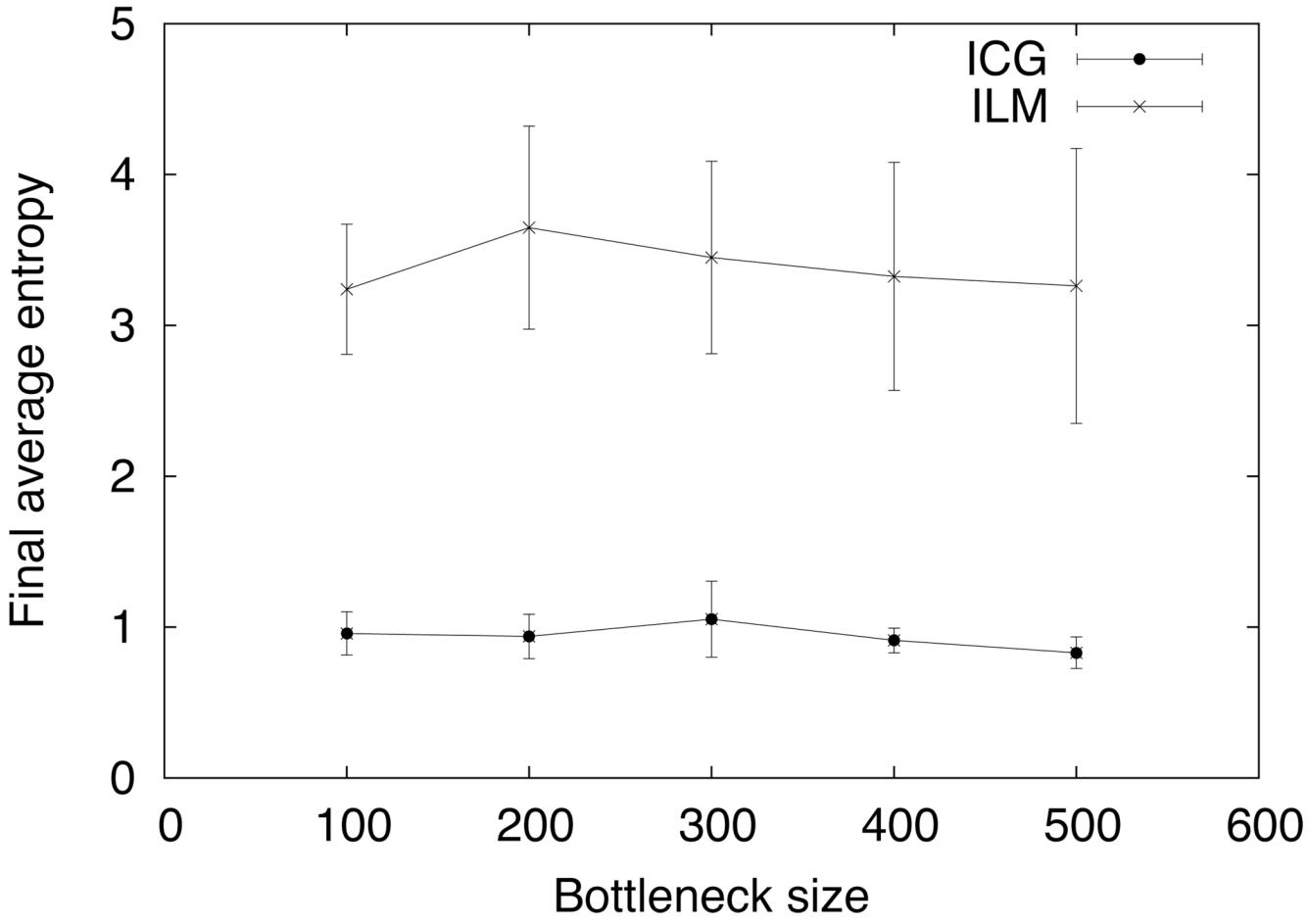


Figure 8.
A comparison, between the ILM and the ICG, of the average final entropy, for various bottleneck sizes. The error bars show one standard deviation, from 10 runs. We see that the average final entropy does not vary much with bottleneck size for either model, and is significantly higher for the ILM than for the ICG.

Table 1

The ILM results in the emergence of a stable language of higher complexity, because it produces utterances in response to the noise bits in the input also. The middle column shows some sample utterances produced by an agent in the 20^{th} generation in response to inputs whose first 3 bits (the only ones relevant for the task) are shown in the left column. The right column shows the symbols in the utterances that are relevant to the classification task. Note that there are no symbols that are generated for all inputs, thus no symbols are redundant. Comparing with the table in figure 5 shows these have higher complexity (i.e., are less compositional) than the solution found by the ICG.

First 3 input bits	Sample utterances	Relevant symbols
000	BCDEFHIJ, BCEFHI, CHIJ, BCEGHI	CHI
001	ABCDEFGHIJ, ABCHIJ, ACFHIJ	ACHIJ
010	ABCDEFGHIJ, ABCHI, ABFHIJ, ABEHI	ABHI
011	ABFIJ, ABIJ, ABCDEFIJ	ABIJ
100	ABCEFHIJ, ACEFHI, ACEHI	ACEHI
101	AEFIJ, ABCDEFIJ, ACFIJ, ACEFI, ACEIJ	AI
110	ABCDEFIJ, ABIJ, ABCEFI, ACFIJ	AI
111	ABFJ, ABCDEFJ, ABCEFGIJ	AJ