Connection Science, Vol. 14, No. 1, 2002, 65–84

The cultural evolution of communication in a population of neural networks

KENNY SMITH

Language Evolution and Computation Research Unit, Department of Theoretical and Applied Linguistics, University of Edinburgh, Adam Ferguson Building, 40 George Square, Edinburgh, UK
e-mail: kenny@ling.ed.ac.uk

Abstract. Human language is learned, symbolic and exhibits syntactic structure, a set of properties which make it unique among naturally-occurring communication systems. How did human language come to be as it is? Language is culturally transmitted and cultural processes may have played a role in shaping language. However, it has been suggested that the cultural transmission of language is constrained by some language-specific innate endowment. The primary objective of the research outlined in this paper is to investigate how such an endowment would influence the acquisition of langage and the dynamics of the repeated cultural transmission of language. To this end, a new connectionist model of the cultural evolution of communication is presented. In this model an individual's innate endowment is considered to be a learning rule with an associated learning bias. The model allows manipulations to be made to this learning apparatus and the impact of such manipulations on the processes of language acquisition and language evolution to be explored. These investigations reveal that an innate endowment consisting of an ability to read the communicative intentions of others and a bias towards acquiring one-to-one mappings between meanings and signals results in the emergence, through purely cultural processes, of optimal communication. It has previously been suggested that humans possess just such an innate endowment. Properties of human language may therefore best be explained in terms of cultural evolution on an innate substrate.

Keywords: language, communication, cultural evolution, learning bias.

1. Introduction

Human language is unique among the communication systems of the natural world—it is at least partially culturally transmitted, the relationship between basic lexical tokens and their meanings is arbitrary and those basic lexical tokens are combined to form structured forms which are used to communicate complex structured meanings. How did language come to be as it is and why is it unique?

There is much debate in linguistics about the role played by innate knowledge in determining the syntactic structure of language. Noam Chomsky, the most prominent linguistic of modern times, argues that the poverty of the stimulus available to children during language acquisition forces us to conclude that some of the structure of language must be encoded in an innate Language Acquisition Device (Chomsky 1987).

Connection Science ISSN 0954-0091 print/ISSN 1360-0494 online © 2002 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/0954009021016430 6

Innate knowledge may also play a role in the transmission of vocabulary. While no serious linguistic theory disputes that words are culturally transmitted within a population, it has been suggested that children may come to the task of learning words with some innate endowment. For example, Bloom (1997) suggests that children possess an innate capacity to read the intent of speakers, while Macnamara (1972) suggests that children have an expectation that words refer to whole objects rather than properties of objects.

We know from mathematical models of dual-inheritance systems (such as those of Boyd and Richerson 1985) that if a learner's genetic endowment can influence their acquisition of cultural artifacts then genetic forces can influence a population's culture and a population's culture can in turn influence the forces acting on genetic transmission. How might the types of genetic endowment suggested by Chomsky, Bloom and Macnamara influence the cultural artifact language? How might language influence the genetic endowment of children at subsequent generations?

These are complex questions which need to be tackled piecemeal. The first step is to consider how a certain genetic endowment might influence the acquisition of a communication system. There is a tradition of researchers in the connectionist sciences addressing this question. For example, Elman (1993) considers the role that different maturational schedules have on a neural network attempting to acquire a simple grammar, Batali (1994) investigates the role evolution might play in setting initial connection weights for a neural network attempting to learn a context-free language and Christiansen and Devlin (1997) show that the consistency of recursion within a language influences the learnability of that language by a neural network. A recent paper in this journal (Cangelosi *et al*. 2000) outlines research in which feedforward neural networks are used to investigate the problem of symbol grounding in the acquisition of communication.

While they provide valuable insights, these models do not truly tackle the dynamic nature of repeated cultural transmission. The agents in these models are adapting towards a static target behaviour, an externally-determined language or vocabulary. In a more realistic model of cultural transmission, the target of adaptation shifts as a result of the adaptations made by previous generations and is determined, at least in part, by the dynamics of repeated learning.

The investigation described in Hare and Elman (1995) represents an early attempt to model this kind of cultural dynamic using neural networks. The paper outlines a scenario under which an 'immature' network learns its competence in a task (verb inflection) based on the behaviour of a 'mature' network which has previously undergone training on the same task. This competence then guides the network's behaviour as a mature individual, which is in turn learned from by a new immature network.

The initial inflectional system in Hare and Elman's scenario was, however, externally determined. Can processes of repeated cultural transmission lead to the *emergence* of a system of communication? If so, what property of learners leads to the emergence of that system? This question has been addressed, using neural network-based simulations, by Hutchins and Hazelhurst (1995), Batali (1998), Hazelhurst and Hutchins (1998), Kvasnička and Pospíchal (1999), Livingstone and Fyfe (1999) and Kirby and Hurford (2002). These works establish that, given a particular model of a learner (autoassociative networks in Hutchins and Hazelhurst (1995) and Hazelhurst and Hutchins (1998), feedforward networks in the others) and a particular model of cultural transmission (repeated horizontal transmission within a fixed population in

Hutchins and Hazelhurst (1995), Batali (1998) and Hazelhurst and Hutchins (1998), vertical transmission between non-overlapping generations in Kvasnička and Pospíchal (1999), Livingstone and Fyfe (1999) and Kirby and Hurford (2002)), communication systems which are in some sense optimal emerge. However, these papers fail to explain fully *why* these communications systems emerge—what properties of the learner result in the emergence of communication systems possessing the optimal quality?

A promising approach to addressing precisely this question is outlined in Oliphant (1999). Oliphant investigates how different learning rules influence the development of a vocabulary-like communication system through cultural processes within a population of simple neural networks. While this paper represents a positive development, it suffers from several shortcomings. Firstly, only three possible learning rules are considered. Secondly, the results for those three learning rules are not related to other results in the field, such as those of Hutchins and Hazelhurst (1995), Batali (1998), Hazelhurst and Hutchins (1998), Kvasnička and Pospíchal (1999), Livingstone and Fyfe (1999) and Kirby and Hurford (2002). Thirdly, while it is shown that certain learning rules result in the emergence of optimal communication, the properties of the learning rules that result in this behaviour are not explicitly identified.

The goal of this paper is to investigate how the innate endowment of individuals influences the cultural evolution of communication within a population of such individuals. Specifically, building on the promise of Oliphant (1999), a model of the cultural transmission of a vocabulary-like communication system in a population of simple networks is developed (sections 2 and 4). The relationship between the learning rules used by individuals in such populations and the population's communication system is explored, and the learning rules which result in the emergence of optimal communication in such populations are identified (sections 3 and 5). The key features of these learning rules are described and related to features of other models of the cultural evolution of communication (section 6). Finally, the implications of this research for our understanding of the evolution of human language are considered.

2. The basic model

The basic model consists of two elements: a model of communication systems (section 2.1) and a model of a communicative agent (section 2.2).

2.1. *Communication systems and communication*

A communication system C consists of a *production* function $p(m)$, mapping from unstructured meanings m to unstructured signals s, and a *reception* function $r(s)$, mapping from signals s to meanings m. m and s are selected such that $m \in \mathcal{M}$ and $s \in \mathcal{S}$ where $\mathcal{M} = \{m_1, m_2 \ldots m_{|\mathcal{M}|}\}$ and $\mathcal{S} = \{s_1, s_2 \ldots s_{|\mathcal{S}|}\}$.

How can we evaluate the communicative accuracy of such a population? The accuracy of a single communicative event involving a producer P with production function $p(m)$, a receiver R with reception function $r(s)$ and a meaning $m_i \in \mathcal{M}$, $ca(P, R, m_i)$, is defined as

$$ca(P,R,m_i) = \begin{cases} 1 & \text{if } r(p(m_i)) = m_i \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

If $p(m)$ is converted to a probabilistic function $p(s_j \mid m_i)$, which gives the probability of producing signal s_j given meaning m_i, and $r(s)$ is similarly viewed as a probabilistic function $r(m_i \mid s_j)$ then the equation above can be rewritten as

$$ca(P,R,m_i) = \sum_{j=1}^{j=|\mathcal{S}|} p(s_j \mid m_i) \cdot r(m_i \mid s_j) \tag{2}$$

The communicative accuracy of P and R over all meanings, $ca(P, R)$ can then be defined as the average of their communicative accuracy over each meaning $m_i \in \mathcal{M}$, e.g.

$$ca(P,R) = \frac{\sum_{i=1}^{i=|\mathcal{M}|} \sum_{j=1}^{j=|\mathcal{S}|} p(s_j \mid m_i) \cdot r(m_i \mid s_j)}{|\mathcal{M}|} \tag{3}$$

In a population possessing an optimal communication system $ca(P, R) = 1$ for any choice of P and R.

2.2. *Communicative agents*

Communicative agents in the model must be capable of representing such communication systems, modelling production and reception functions of the type outlined above and modifying their behaviour based on observations of systems of the type outlined above.

2.2.1 *Representation.* Agents are modelled using networks consisting of two sets of nodes $\mathcal{N}_M$ and $\mathcal{N}_S$ and a set of weighted bidirectional connections $\mathcal{W}$ connecting every node in $\mathcal{N}_M$ with every node in $\mathcal{N}_S$.

Patterns of activation over $\mathcal{N}_M$ are considered to represent meanings, whereas patterns of activation over $\mathcal{N}_S$ are considered to be signals. Restricting these patterns of activation to contain a single active unit yields $|\mathcal{N}_M|$ orthogonal meaning representations and $|\mathcal{N}_S|$ orthogonal signal representations, suitable for representing sets of unstructured meanings and unstructured signals such as those described above. If *Gi* is the *i*th node from the set $\mathcal{N}_G$ and the activation of node *Gi* is a_{Gi} then the meaning m_i correspopnds to a pattern of activation over $\mathcal{N}_M$ where $a_{Mi} = 1$ and $a_{M(j\neq i)} = 0$. Similarly, the signal s_i corresponds to a pattern of activation over $\mathcal{N}_S$ where $a_{Si} = 1$ and $a_{S(j\neq i)} = 0$. This representational scheme is illustrated in figure 1.

2.2.2. *Retrieval.* Patterns are retrieved from the network using a *k*-winners-take-all strategy. In order to retrieve a pattern of activation over nodes in $\mathcal{N}_S$ based on an input pattern of activation over nodes in $\mathcal{N}_M$ the weighted sum of inputs to node *Si*, q_{Si}, for each $Si \in \mathcal{N}_S$ is calculated according to the formula:

$$q_{Si} = \sum_{j=1}^{j=|\mathcal{N}_M|} a_{Mj} w_{Mj,Si} \tag{4}$$

where $w_{a,b} \in \mathcal{W}$ is the weight of the connection between nodes *a* and *b*. The *k* nodes

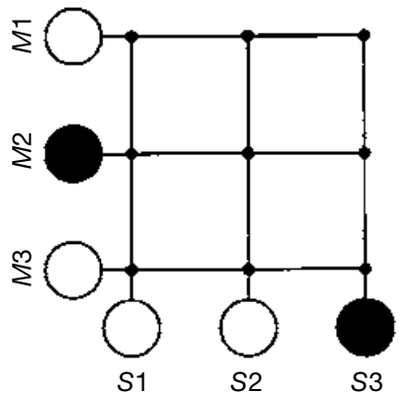

Figure 1. A neural network where $|\mathcal{N}_M| = |\mathcal{N}_S| = 3$. Large filled circles represent nodes with activation of 1, large empty circles represent nodes with activation of 0. The pattern of activation over $\mathcal{N}_M$ therefore represents the meaning m_2 ($a_{M2} = 1$, $a_{M1} = a_{M3} = 0$). Similarly, the pattern of activation over $\mathcal{N}_S$ represents the signal s_3.

in $\mathcal{N}_S$ with the highest values of q then have their activations set to 1, while all other nodes in $\mathcal{N}_S$ have their activations set to 0. If several nodes have equal q a random winner is selected from among them. Patterns of activation over the nodes in $\mathcal{N}_M$ are retrieved based on input patterns of activation over $\mathcal{N}_S$ in exactly the same way. For all simulations outlined in this paper, k =1—retrieved patterns of activation only ever consist of a single active node and ($|\mathcal{N}_G| - 1$) non-active nodes. This ensures that retrieved patterns of activation conform to our representration of meanings and signals outlined above. This retrieval process is illustrated in figure 2.

Retrieving a pattern of activation over $\mathcal{N}_S$ given an input pattern of activation over $\mathcal{N}_M$ corresponds to retrieving the signal associated with a given meaning—*production* of a signal associated with a given meaning. Retrieving a pattern of activation over $\mathcal{N}_M$ given an input pattern of activation over $\mathcal{N}_S$ corresponds to retrieving the meaning

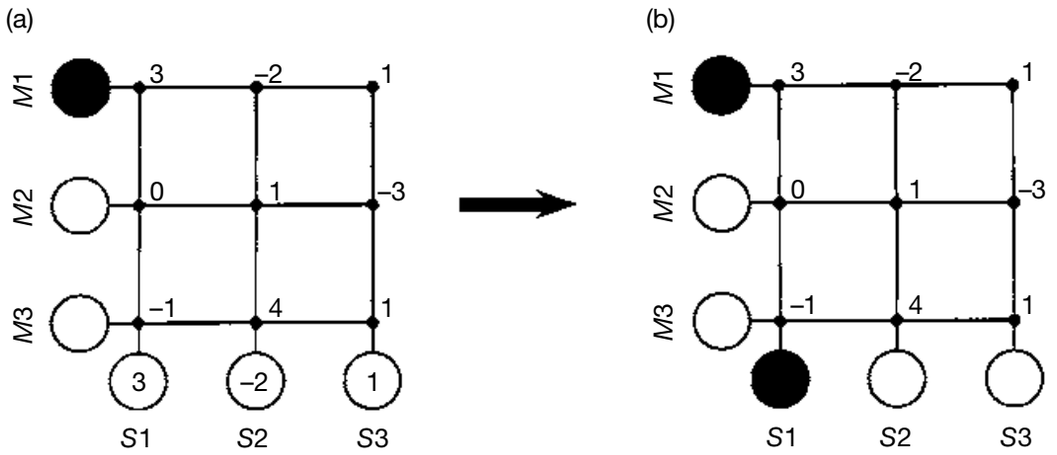

Figure 2. Retrieval of a pattern of activation over $\mathcal{N}_S$ based on a pattern of activation over $\mathcal{N}_M$. As before, large filled circles represent nodes with activation of 1. Connections between nodes are represented by the intersections of connecting lines and have an associated weight. In (*a*), the nodes in $\mathcal{N}_M$ have been set to a pattern of activation, resulting in a pattern of weighted sums of inputs over the nodes in $\mathcal{N}_S$. The numbers in the centre of the nodes in $\mathcal{N}_S$ represent the weighted sums to those nodes. In (*b*) the result of the application of the winner-take-all process is shown—q_{S1} is greater than q_{S2} or q_{S3}, therefore node *S*1 has its activation set to 1 while nodes *S*2 and *S*3 have their activations set to 0.

associated with a given signal—*reception* of a given signal and interpretation of that signal to yield a meaning. Note that the production and reception behaviour of such networks are not necessarily closely related—for example, the network in figure 2 would produce $S2$ when prompted with $M2$, but would interpret $S2$ as meaning $M3$. Using a single network for both production and reception, as opposed to two separate networks, does however allow the possibility of a coupling of production and reception.

2.2.3. *Storage.* In order to store the association between patterns of activation over $\mathcal{N}_M$ and $\mathcal{N}_S$ the activations of the nodes in $\mathcal{N}_M$ and $\mathcal{N}_S$ are set to the required values and the weights of the connections in $\mathcal{W}$ are adjusted according to some weight-update rule W. If we assume that W must only adjust connection weights based on local information and that all patterns of activation will be binary, W can be specified by the 4-tuple $(\alpha\ \beta\ \gamma\ \delta)$, where the value in α specifies how the weight of connection $w_{i,j}$ should be adjusted when $a_i = a_j = 1$, the value in β specifies how $w_{i,j}$ should be adjusted when $a_i = 1$ and $a_j = 0$, the value in γ specifies how $w_{i,j}$ should be adjusted when $a_i = 0$ and $a_j = 1$ and the value in δ specifies how $w_{i,j}$ should be adjusted when $a_i = a_j = 0$. While weights could be adjusted in many ways we will restrict ourselves here to the simplest case where α, β, γ and δ must take integer values in the range $(-1, 1)$. This yields a range of $3^4 = 81$ possible weight-update rules.

Given our interpretations of patterns of activations of $\mathcal{N}_M$ and $\mathcal{N}_S$ this storage process will be interpreted as the process of learning the association between a meaning and a signal in a meaning-signal pair $<m, s>$ according to some rule W. The storage process is illustrated in figure 3.

3. Acquisition of an optimal system

We now have a model of communication, a model of an agent and processes of production, reception and learning. The first question to be addressed is to ask whether individual agents, in isolation, can acquire an optimal communication system. To this end an unambiguous set of meaning-signal pairs $\mathcal{A} = \{<m_1, s_1>,$

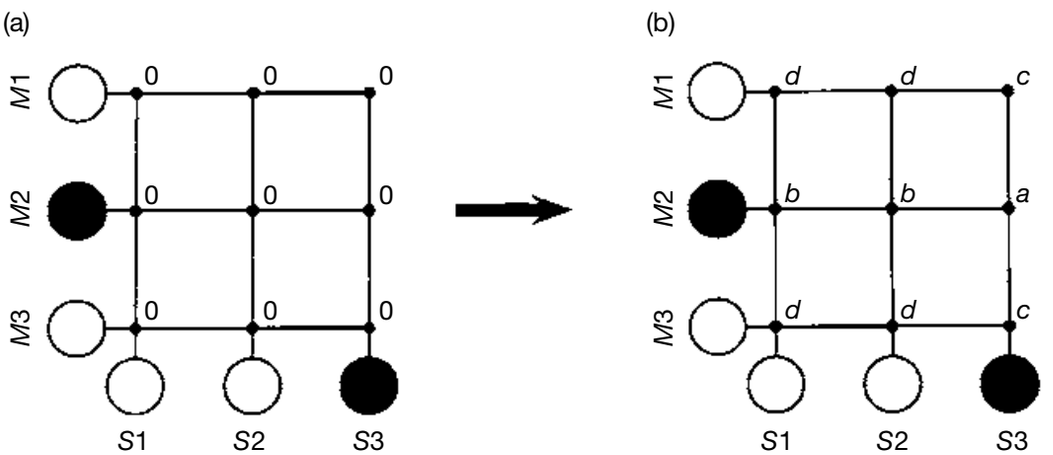

Figure 3. Storage of the meaning-signal pair $<m_2, s_3>$ using the weight-update rule $W = (a\ b\ c\ d)$. In (a), the nodes in $\mathcal{N}_M$ and $\mathcal{N}_S$ have been set to the pattern of activation representing m_2 and s_3. All connections have weight 0. In (b) the result of the application of the storage process is shown—all connections now have weights of a, b, c or d, depending on the activations of the nodes they connect.

$<m_2, s_2> \ldots <m_{10}, s_{10}>$ was constructed. Agents using each of the 81 possible weight update rules were then trained on $\mathcal{A}$, by storing each meaning-signal pair in $\mathcal{A}$ in their network. The agents were then evaluated to see if they had successfully acquired an optimal communication system based on exposure to the unambiguous set of meaning-signal pairs $\mathcal{A}$. Agents were judged to have acquired an optimal system, if, for every $<m_i, s_i> \in \mathcal{A}$ both

1 production of the signal associated with m_i always[1] resulted in s_i being produced, i.e. $<m_i, s_i>$ can be reproduced in production
2 *and* reception of s_i always resulted in the interpretation m_i, i.e. $<m_i, s_i>$ can be reproduced in reception, meaning that the agent would communicate optimally with itself or another agent using the same weight-update rule exposed to $\mathcal{A}$.

The 81 weight-update rules can therefore be classified according to a [+/– learner] feature. Thirty-one of the 81 possible weight-update rules were judged to be capable of acquiring the optimal communication system and were classified as [+ learner]. The remaining 50 weight-update rules were classified [– learner].

4. The iterated learning model

As discussed in section 1, many connectionist models of the acquisition of communication end with an experiment of the sort outlined in the preceding section, where the ability of individual agents to acquire a system with a predefined structure is investigated. This type of static analysis fails to take the essentially dynamic nature of repeated cultural transmission into account. The model outlined in section 2 can be extended to allow the study of the dynamics of repeated cultural transmission, with the aim of identifying the circumstances under which optimal communication emerges through purely cultural processes. While there are numerous potential factors which will influence this matter, the one which will concern us here is that of *learning bias*—which weight-update rules result in the emergence of optimal communication through purely cultural processes, and why?

A model of cultural transmission is presented in section 4.1. In section 4.2 some of the assumptions of this model are discussed.

4.1. *Cultural transmission*

The results of repeated cultural transmission can be investigated using an iterated learning model (Kirby 2001, Brighton 2002). In an iterated learning model agents require their competence through learning from observations of the external behaviour of other agents. This competence is then used to generate external behaviour which is observed in turn by other agents. In the case of our model, the culturally-transmitted behaviour of interest is linguistic behaviour. Agents acquire their linguistic competence based on the linguistic behaviour of other agents and use this acquired linguistic competence to produce linguistic behaviour.

What constitutes the linguistic behaviour that agents are required to acquire their competence from? External signals are clearly an observable aspect of linguistic behaviour. However, as with other iterated learning models dealing with the cultural evolution of linguistic behaviour, we assume that learners are able to observe meanings in addition to signals, at least during the learning period—immature agents

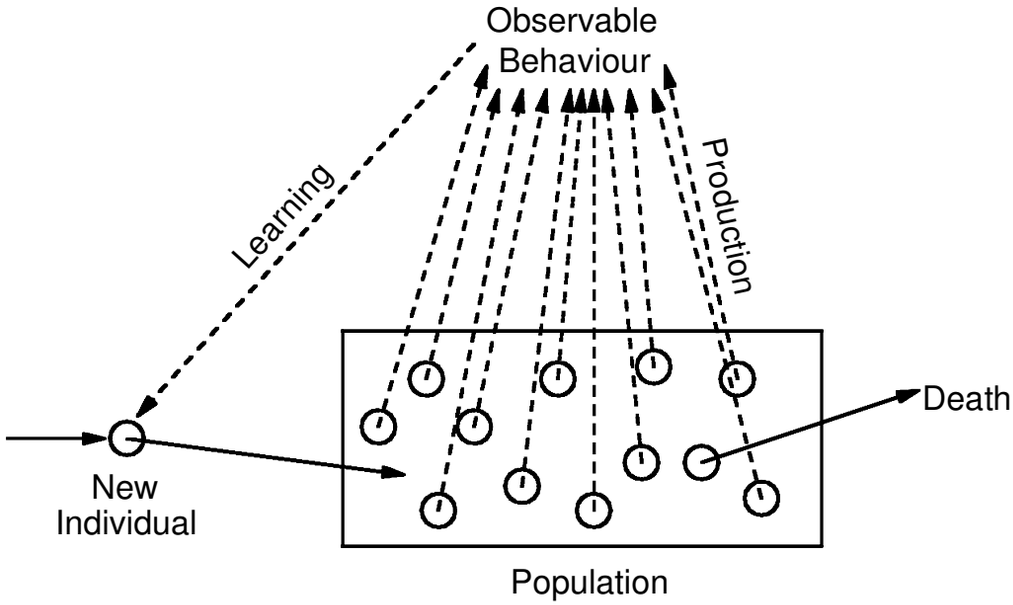

Figure 4. The population-level iterated learning model. At every iteration an individual (represented by a circle) is removed and a new individual joins the population, after observing and learning from the behaviour of the remaining members of the population.

can identify the communicative intentions of other agents, as well as their signals, and observe and learn from meaning-signal pairs. The relationship between this assumption and the symbol grounding problem is discussed in section 4.2.

The process of iterated learning requires a model of population turnover. In this paper we use a gradual population turnover model, illustrated in figure 4. At every time-step a single agent is selected at random and removed from the population. The remaining members of the population produce some observable behaviour, in the case of our model sets of meaning-signal pairs. A new individual arrives and learns based on observations of the population's observable behaviour, then enters the population. The process then repeats.

More formally, the iterated learning model consists of an initialization process and an iteration process:

Initialization: Create a population of N agents,[2] each using the weight-update rule W and possessing communication system L.

Iteration:

(1) Select an agent at random from the population and remove it.
(2) For every remaining member of the population, generate a set of meaning-signal pairs by applying the network production process to every $m \in \mathcal{M}$. Noise is added to each meaning-signal pair[3] with probability p_n.
(3) Create a new agent with connection weights of 0 who uses weight-update rule W.
(4) The new agent receives e exposures to the population's communicative behaviour.[4] During each of these e exposures the new agent observes the complete set of meaning-signal pairs of a randomly selected member of the population and updates their connection weights according to the observed meaning-signal pairs and their weight-update rule W.
(5) The new agent joins the population. Return to (1).

Each pass through the iteration process will be termed a *cohort*. Note that the random removal of agents from the population means that there is no selection based on communicative ability. The fact that every individual begins its life with a weight-update rule and initial set of connection weights suggests some kind of innate endowment of these components. It is our goal to investigate the impact of this innate endowment on the communicative behaviour of the population. However, every agent begins life with the *same* endowment—there is no possibility of genetic variation within the population. The emergent behaviour of the population will therefore be determined by the dynamics resulting from the iterated cultural transmission of communcation systems among populations of individuals with a common genetic endowment.

4.2. *The environment and internal representations*

In this model, as with most other iterated learning models, there is no notion of an environment outwith the agents—meanings and signals in the model are arbitrary agent-internal representations. This model assumes that there is some shared, stable mapping between external situations and internal representations of those events—indeed, communication would be impossible without an external manifestation of internal representations of signals. However, this mapping is not the focus of this paper. An account of language as a mapping from aspects of the environment to other aspects of the environment must account for two additional mappings (see figure 5):

- the mapping between states of the environment representing situations to be communicated and internal representations of those states (meanings in this model);
- the mapping between states of the environment representing communicative alterations of the environment and internal representations of those states (signals in this model).

The nature of the mapping between environment and meaning forms a key part of the symbol grounding problem (Harnad 1990). The mapping between environment and signal corresponds to a mapping between strings of phonemes and articulatory

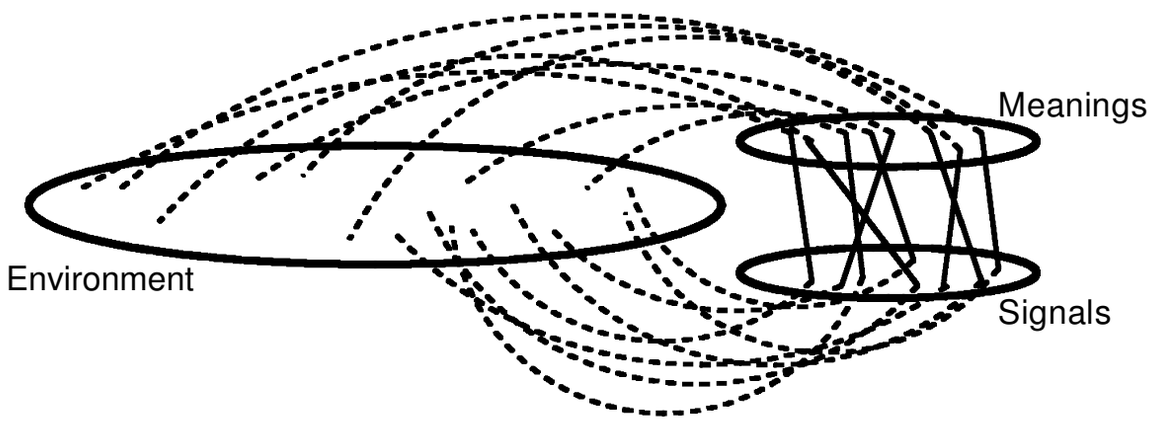

Figure 5. In a complete model we must account for the mappings between three spaces (ellipses)—the mappings between the environment space and the internal representational spaces (dashed lines) in addition to the mapping between internal representational spaces (solid lines).

movements. In the model described here these mappings are not under consideration. For our purposes it is sufficient to assume that all agents share mappings, possibly based on categorization, between the environment and meanings and between the environment and signals. While the mapping for the purposes of this model is assumed to result in unstructured internal representations, it should be noted that structured internal representations corresponding to a perception of structure in the environment are a prerequisite for syntactically-structured language (Brighton 2002, Smith in preparation).

Given the absence of an environment the assumption that learners observe meaning-signal pairs is unavoidable. In a fuller model, the more reasonable assumption could be made that learners are exposed to an environment which includes a state being communicated about and a set of articulatory gestures intended as a communicative alteration to the environment. The learner then has the task of identifying the communicatively relevant state and the communicative alteration, representing both internally and then learning the mappings between the internal representations and possibly the mappings between the internal representations and the environment. Given that in the current model there is no environment, this is not possible. We could simulate the difficulty of the environment-meaning and environment-signal mappings by adding noise to the observed meaning-signal pairs. However, it is more appropriate to leave the nature of the external–internal mappings to models which are explicitly designed to deal with them and focus here on the learning of the internal mapping.

There is a body of evidence which suggests that children have various strategies for mapping from the environment to internal representations of relevant parts of the environment. Much of this points to the importance of joint attention and intentional inference. Studies by Baldwin (1991, 1993b) show that infants cannot learn words for toys simply by hearing the word for the toy while attending to the toy. The child must witness an intentional agent direct their attention to the toy while naming it. Under these circumstances the infant will learn the word for the toy, even if there is a delay between witnessing the intentional agent directing their attention at the toy and being able to attend to the toy directly themselves (Baldwin 1993a).

The development of this external–internal mapping in conjunction with the development of an internal meaning-signal mapping has been modelled computationally. Neural network models show that genetic evolution can lead to the formation of internal representations which correspond to a categorization of the environment (Cangelosi and Parisi 1998). These internal representations may form the basis of an innate (as in Cangelosi and Parisi 1998) or learned (Cangelosi 1999) communication system. Hazelhurst and Hutchins (1998) show that the negotiation of ritualized shifts of joint attention subserves the emergence of a learned communication system. Symbolic computational models demonstrate that shared mappings from the environment to interal representations of meanings can emerge through individual learning, both with explicit feedback (Steels and Kaplan 1999) and without (Smith 2001). These models and the data from real language acquisition outlined above give some hope that an integrated model, of the type depicted in figure 5, will be achievable.

5. Maintenance and construction of optimal systems

In addition to the static tests of acquisition outlined in section 3, two sets of tests were carried out on populations of agents using each of the 81 possible weight-update rules—tests for maintenance of an optimal communication system in a population over time (section 5.1) and construction of an optimal communication system in a population over time (section 5.2). These experiments give greater insight into the dynamics arising from repeated cultural transmission and a fuller understanding of the properties of learning rules than that afforded by static tests alone.

5.1. *Maintenance of an optimal system*

The first question to be addressed using the iterated learning model is whether a population of agents possessing a weight-update rule W can maintain an optimal system over time in the presence of a small degree of noise. Recall from the description of the iterated learning model given in section 4 that the agents in the initial population use some predefined communication system L. For the experiments outlined in this section, the initial population's sets of weights $\mathcal{W}$ were constructed such that the $p(m)$ of the initial L generated the set of meaning-signals pairs $\mathcal{L} = \{<m_1, s_1>, <m_2, s_2> \ldots <m_{10}, s_{10}>\}$. Iterated learning models were run with each of the 81 possible learning rules, with noise introduced with probability $p_n = 0.05$. Populations were defined as having *maintained* the initial optimal system if the population's communicative accuracy remained above 0.95 for every cohort of a run.[5] Weight-update rules were classified as [+ maintainer] if the optimal system was maintained for each of ten 2000-cohort runs.

The populations exhibited four typical patterns of behaviour, illustrated in figure 6. Populations (*a*), (*b*) and (*c*) in figure 6 have failed to maintain the optimal system and can therefore be classified as [– maintainer], although population (*a*) in figure 6 exhibits a more rapid decrease in communicative accuracy than populations (*b*) and (*c*). Unsurprisingly, all 50 populations using weight-update rules with the [– learner] feature followed the pattern of (*a*) and can therefore be classified [– learner, – maintainer]. Of the remaining 31 weight-update rules, 13 resulted in the type of pattern exemplified by populations (*b*) and (*c*) and can be classified as [+ learner, – maintainer] and 18 resulted in patterns similar to that of population (*d*) in figure 6 and can be classified as [+ learner, + maintainer].

5.2. *Construction of an optimal system*

Finally, the 81 weight-update rules were examined to see whether they resulted in the emergence of optimal communication systems from random behaviour when placed in the context of the iterated learning model. In the previous section the initial population's communication system, L, was optimal. In the iterated learning models outlined in this section L has maximum entropy—every $m \in \mathcal{M}$ is associated with every $s \in \mathcal{S}$ with equal probability, $|\mathcal{M}| = |\mathcal{S}| = 10$. This was achieved by setting the connection weights of every individual in the initial population to 0. Unlike the previous section, cultural transmission is noise-free—$p_n = 0$ (although results show that similar behaviour occurs with $p_n > 0$). Iterated learning models were run for each of the 81 possible learning rules. A population was defined as having *constructed* an optimal system if the population's communicative accuracy reached 1.0. Weight-update rules were classified [+ constructor] if optimal systems were constructed in each of ten 2000-cohort runs.

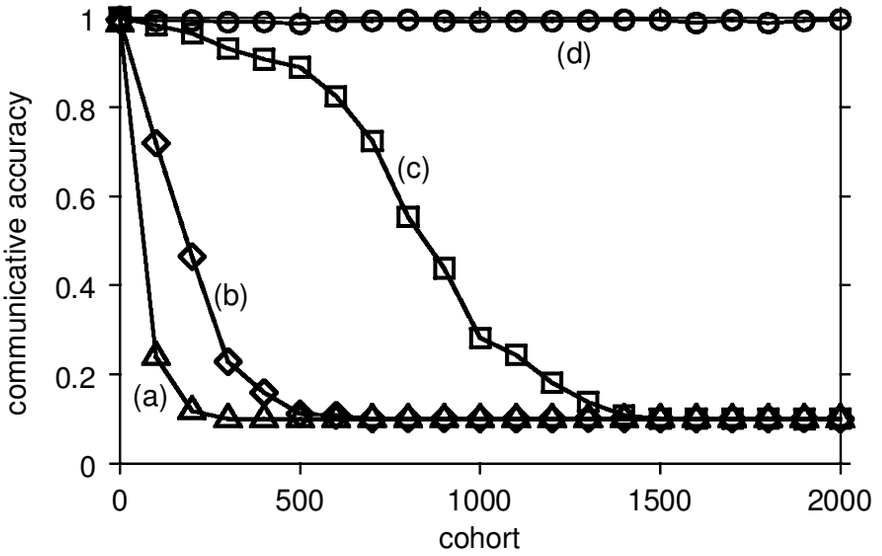

Figure 6. Populations of agents using the 81 learning rules exhibit four patterns of behaviour when attempting to maintain an optimal system. This figure plots the communicative accuracy over time of single populations exhibiting these patterns of behaviour: rapid collapse to chance levels of communicative accuracy, as in (*a*); less rapid collapse to chance levels of communicative accuracy, as in (*b*) and (*c*); maintenance of the optimal system, as in (*d*).

The populations exhibited three typical patterns of behaviour, of which populations (*a*), (*b*) and (*c*) in figure 7 are representative examples. The populations which fit the pattern exemplified by (*a*) in figure 7 have clearly failed to construct an optimal system and in fact persist at the random level of performance for $|\mathcal{M}| = |\mathcal{S}| = 10$. All of the weight-update rules which were classified as [– maintainer] follow this pattern and can be classified as [– constructor].

Populations behaving similarly to population (*b*) in figure 7 are clearly performing above the random level, but have not constructed an optimal system as defined above.

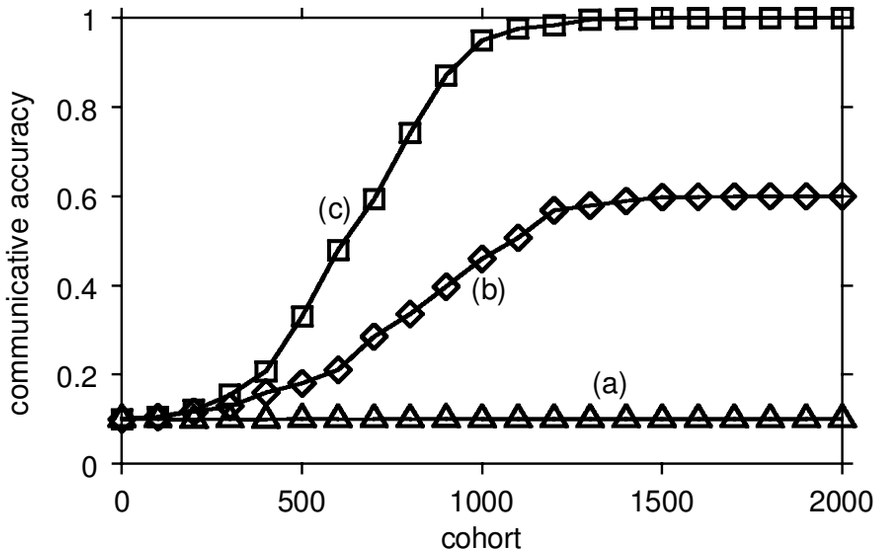

Figure 7. Populations of agents using the 81 learning rules exhibit three patterns of behaviour when attempting to construct an optimal system: failure to construct an optimal system and chance-level communicative accuracy, as in (*a*); failure to construct an optimal system, but levels of communicative accuracy significantly above chance, as in (*b*); construction of an optimal system, as in (*c*).

In fact, as suggested for a more limited case by Oliphant (1999), the level of communicative accuracy in these populations hovers around the level we would expect given a random assignment of signals from $|\mathcal{S}|$ to meanings from $|\mathcal{M}|$ with replacement

$$ca \approx \left(1 - \left(1 - \frac{1}{|\mathcal{S}|}\right)^{|\mathcal{M}|}\right) \tag{5}$$

The reasons for this level of performance will be made clear in section 6. Nine of the 18 weight-update rules which were classified [+ maintainer] fit this pattern and can be classified as [– constructor].

Populations fitting the pattern exemplified by population (*c*) in figure 7 have clearly succeeded in constructing an optimal system from random behaviour and can be classified as [+ constructor]. Nine of the 18 weight-update rules which were classified as [+ maintainer] fit this pattern.

5.3. *The classification hierarchy*

The three tests outlined above divide the 81 weight-update rules into four groups, summarized in table 1.

The fact that all weight-update rules which are [+ constructor] are [+ maintainer] and all rules which are [+ maintainer] are [+ learner] suggests a hierarchy of weight-update rules, summarized in figure 8.

6. The key bias

What is it about the particular assignment of –1s, 0s and 1s to the four conditions α, β, γ and δ that makes one weight-update rule incapable of learning an optimal communication system whereas another weight-update rule is capable of constructing such a system from random behaviour in the context of iterated cultural transmission? There is in fact a clear pattern relating the properties of weight-update rules to the assignment of actions to values in the (α β γ δ) 4-tuple. This bias is best described in terms of the one-to-one nature of mappings between meanings and signals.

As defined in section 2.1, in an optimal communication system $r(p(m)) = m$ for all $m \in \mathcal{M}$. This requires that:

1 Each $m \in \mathcal{M}$ should be expressed by a distinct $s \in \mathcal{S}$, i.e. $p(m)$ should be a one-to-one function.

Table 1. The number of weight-update rules of each particular complete classification, from the sample of 81.

Classification	Number
[– learner, – maintainer, – constructor]	50
[+ learner, – maintainer, – constructor]	13
[+ learner, + maintainer, – constructor]	9
[+ learner, + maintainer, + constructor]	9

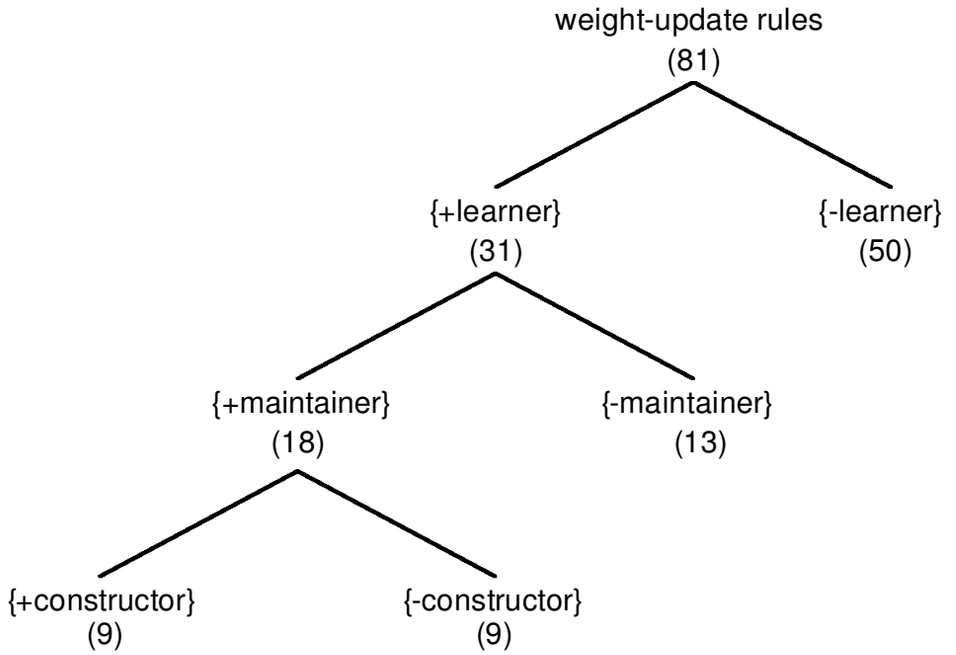

Figure 8. The hierarchy of weight-update rules. Read from the top, each node places additional restrictions on the properties of the weight-update rules. The numbers possessing each feature are given in parentheses at each point in the tree.

2 Each $s \in \mathcal{S}$ should map back to a single $m \in \mathcal{M}$ such that $p(m) = s$, i.e. $r(s)$ should be a superset of the inverse of $p(m)$.

Given the (approximately) bidirectional nature of the networks and assuming $|\mathcal{S}| \geq |\mathcal{M}|$, point 1 above proves to be crucial in determining which weight-update rules are [+ constructor], which are [+ maintainer, – constructor] and which are [+ learner, – maintainer, – constructor]. Weight-update rules which are [+ constructor] are biased in favour of a one-to-one $p(m)$, those which are [+ maintainer, – constructor] are neutral with respect to the one-to-one nature of $p(m)$ and those which are [+ learner, – maintainer, – constructor] are biased in favour of a many-to-one $p(m)$.

6.1. *The [+ constructor] bias*

Is there any pattern of assignment of values to conditions in the weight-update rule specification ($\alpha\ \beta\ \gamma\ \delta$) that characterizes rules which are [+ constructor] but not rules which are [– constructor]? Yes.

$$\text{A weight-update rule is [+ constructor] iff } \alpha > \beta \wedge \delta > \gamma$$

Why does this pattern of weight changes result in the construction of optimal systems from random behaviour? Consider a network where $|\mathcal{N}_M| = |\mathcal{N}_S| = 2$ using the weight-update rule (*a b c d*). Prior to learning, all the connection weights in $\mathcal{W}$ are 0. If we represent $\mathcal{W}$ as a matrix with the value in row *i* and column *j* representing the weight of the connection between nodes *Mi* and *Sj* then its initial weights will be

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

If this network is exposed once to the meaning m_1 (recall from section 2.2.1. that for this meaning $a_{M1} = 1$, $a_{M2} = 0$), paired with the signal s_1 (similarly, $a_{S1} = 1$, $a_{S2} = 0$), its weight matrix will be

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

For rules which are [+ constructor] $a > b$. This means that if our simple network uses a [+ constructor] rule it will correctly produce s_1 to communicate m_1, due to the winner-take-all retrieval procedure.

For [+ constructor] rules, $d > c$. In the context of our simple network, this means that if the network uses a constructor rule it will automatically prefer to use the signal s_2 to communicate meaning m_2, despite the fact that it has only been trained to associate m_1 with s_1. This is the crucial property of [+ constructor] rules—they are biased in favour of acquiring one-to-one mappings between meanings and signals. What consequences does this bias have in the context of iterated cultural transmission?

Only communication systems which conform completely to the biases of learners will be stable over iterated cultural transmission—communication systems which partially conform to learner biases will be less likely to be acquired than systems which conform more fully to the learner biases, and will therefore be filtered out of the population over time. This differential retention of communication systems resulting from learner biases can be termed *cultural selection*. The [+ constructor] bias in favour of one-to-one mappings between meanings and signals results in many-to-one mappings being filtered out of the population. Eventually, through the process of iterated learning, the population converges on a shared one-to-one mapping between meanings and signals—an optimal communication system is constructed.

6.2. *The [+ maintainer] bias*

Can the [+ maintainer] property also be explained in terms of allocations of actions to the (α β γ δ) weight-update rule specification? First, is there any pattern which uniquely identifies the [+ maintainer, – constructor] rules? Yes.

A weight-update rules is [+ maintainer, – constructor] iff $\alpha > \beta \wedge \delta = \delta$

Once again consider a network where $|\mathcal{N}_M| = |\mathcal{N}_S| = 2$ using the rule (a b c d) exposed once to m_1 paired with s_1. As before, the resultant weight matrix is

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

As for [+ constructor] rules, for [+ maintainer, – constructor] rules $a > b$. This means that if our simple network uses a [+ maintainer, – constructor] rule it will correctly produce s_1 to communicate m_1.

For [+ maintainer, – constructor] rules $d = c$. This means that, unlike [+ constructor] rules, the network using a [+ maintainer, – constructor] rule will be equally likely to

express m_2 using s_1 and s_2, due to their equal weights in the network. [+ maintainer, – constructor] rules are therefore neutral with respect to one-to-one mappings. This explains both the ability of populations of agents using such rules to maintain optimal systems in the context of the iterated learning model and the behaviour of these populations as they attempt to construct optimal systems.

[+ maintainer, – constructor] rules can maintain an optimal system in the presence of noise. The initial optimal system is, by definition, a one-to-one mapping between meanings and signals. Given the neutrality of [+ maintainer, – constructor] rules to the one-to-one nature of mappings, such optimal systems can be acquired in the presence of noise, provided the noise is not sufficient to drown out the one-to-one nature of the mapping.

Recall from section 5.2 and figure 7 that, when provided with an initially random system, populations of agents using [+ maintainer, – constructor] rules converge on the level of communicative accuracy one would expect given a random assignment, with replacement, of signals to meanings. This can be explained in terms of the neutrality of [+ maintainer, – constructor] rules to the one-to-one nature of mappings. The initial population's random behaviour, when taken as a whole, will embody a random assignment of signals to meanings. This random assignment will become shared among the population through the process of iterated learning. While [+ constructor] agents remove the many-to-one elements of the initial random system, [+ maintainer, – constructor] agents do not—the population's eventual communication system will embody the same number of many-to-one mappings as the initial random behaviour.

What then of the [+ maintainer] property in isolation from the [+/– constructor] feature? This can be captured thus

$$\text{A weight-update rule is [+ maintainer] iff } \alpha > \beta \wedge \delta \geq \gamma$$

The fact that rules which are [+ constructor] are always [+ maintainer] is captured by this statement, as is the fact that it is possible to be [+ maintainer, – constructor].

6.3. *The [+ learner] bias*

The pattern of assignments of actions to the weight-update rule specification ($\alpha\,\beta\,\gamma\,\delta$) that characterizes rules which are [+ learner] is

$$\text{A weight-update rule is [+ learner] iff } \alpha + \delta > \beta + \gamma$$

or, in simple terms, in order to be able to acquire an optimal communication system you must make stronger associations between units which tend to have matching activations than between units which tend to have conflicting activations. Note that the $\alpha > \beta \wedge \delta \geq \gamma$ constraint on [+ maintainer] rules guarantees that all such rules are also [+ learner].

Why are rules which are [+ learner, – maintainer, – constructor] unable to maintain or construct optimal communication systems? As we might expect, such weight-update rules are biased *against* one-to-one mappings between meanings and signals and in favour of many-to-one mappings. This immediately rules out construction of the one-to-one mappings characterizing optimal systems, and also maintenance of such systems. Any many-to-one mappings introduced by noise will be preferentially

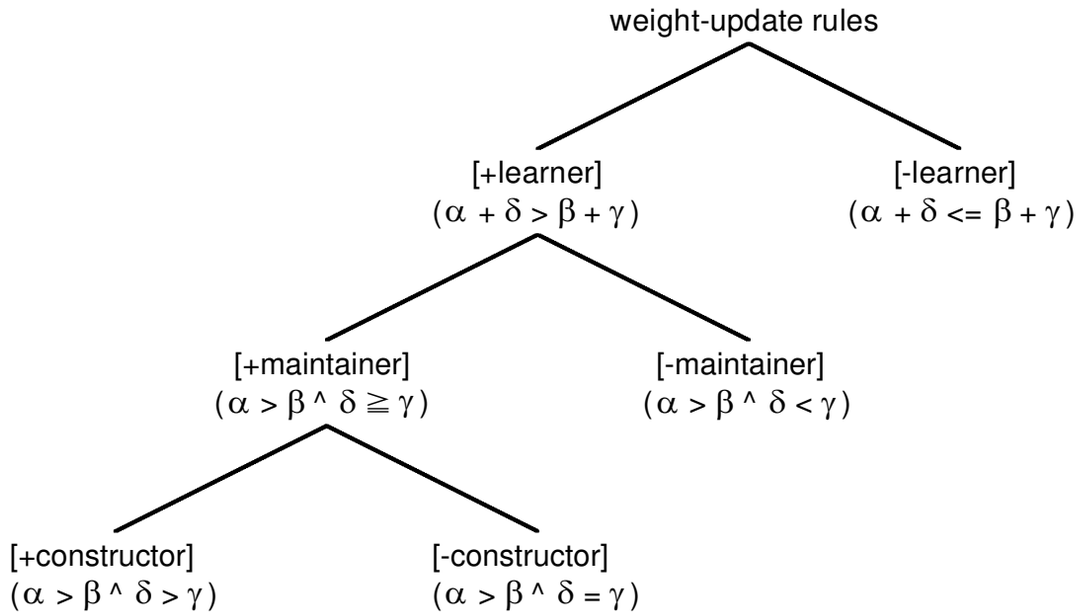

Figure 9. The hierarchy given in figure 8, expressed in terms of restrictions on possible values in each condition of weight-update rules.

acquired by [+ learner, – maintainer, – constructor] agents and will spread through populations of such agents, resulting in the type of decrease in communicative accuracy seen in figure 6.

6.4. *Summary of the key bias*

The weight-update rule hierarchy given in figure 8 is re-presented in figure 9 in terms of the constraints on the values of the weight-update rules. Each terminal node of the tree has a bias, summarized in table 2.

6.5. *The key bias in other models*

Does this key bias appear in other computational models of the cultural evolution of communication? Specifically, does it appear in the neural network models of Hutchins and Hazelhurst (1995), Batali (1998), Hazelhurst and Hutchins (1998), Kvasnička and Pospíchal (1999), Livingstone and Fyfe (1999) and Kirby and Hurford (2002)?

The models of Hutchins and Hazelhurst (1995) and Hazelhurst and Hutchins (1998) can be treated separately from the other models, which all share a common model of a learner. Hutchins and Hazelhurst use autoassociator networks to model

Table 2. A summary of the learning biases of each particular combination of features. Weight-update rules which are classified as [– learner, – maintainer, – constructor] cannot be said to have a learning bias as they cannot learn.

Classification	Bias
[– learner, – maintainer, – constructor]	NA
[+ learner, – maintainer, – constructor]	Favours many-to-one mappings
[+ learner, + maintainer, – constructor]	Neutral
[+ learner, + maintainer, + constructor]	Favours one-to-one mappings

communicative agents in both cases, with patterns of activation over the hidden layer being interpreted as signals. Autoassociative networks must develop a distinct pattern of activation over the hidden layer for every input–output pair (input–output pairs are equivalent to meanings as defined here, although Hutchins and Hazelhurst consider them to be visual stimuli) in order to succeed in the autoassociator task. Interpreting the hidden-layer patterns of activation as signals therefore builds in a one-to-one bias of the type identified as crucial for developing an optimal communication system.

Batali (1998), Kvasnička and Pospíchal (1999), Livingstone and Fyfe (1999) and Kirby and Hurford (2002) all use feedforward networks mapping from signals to meanings to model agents. The feedforward activation process in these networks therefore corresponds to $r(s)$. In order to derive $p(m)$ from non-reversible networks a somewhat ad-hoc reversal process is used. Briefly, if the feedforward network has learned to associate a set of inputs $\mathcal{S}$ with a single output m, the reversal process deterministically selects a signal $s \in \mathcal{S}$ when prompted to produce a signal for m—the network's $p(m)$ is a subset of its $r(s)$.

These networks adjust their connection weights using the backpropagation procedure on the basis of signal-meaning pairs, which are derived from observed meaning-pairs. From observing systems with a many-to-one $p(m)$ agents attempt to learn to associate multiple meanings with a single signal. However, such one-to-many mappings are unlearnable by feedforward networks. Many-to-one mappings from meanings to signals are therefore culturally unstable. One-to-many mappings between meanings and signals are learnable but unstable due to the reversal process. The only culturally stable mapping is therefore a one-to-one mapping between meanings and signals. These networks therefore encode the same bias as the [+ constructor] agents described in this paper.

7. Conclusions

Investigation of the properties of a range of weight-update rules for simple networks, both in isolated learning tasks and in the context of the iterated learning model, reveals a hierarchy of such rules. This hierarchy can be described in terms of restrictions on the possible assignment of actions to conditions in a weight-update rule. Further investigation of these restrictions reveals that the bias of the rules with respect to the one-to-one nature of meaning-signal mappings is crucial. Assuming the capacity to acquire an observed mapping (i.e. ignoring the [– learner] rules), a bias against one-to-one mappings results in failure to maintain an optimal communication system over time in the presence of noise. Neutrality with respect to the one-to-one nature of meaning-signal mappings gives the ability to maintain an optimal system in the presence of a degree of noise, but inability to construct an optimal system from initially random behaviour. A bias in favour of one-to-one mappings between meanings and signals results in cultural selection in favour of such mappings and the emergence of optimal communication through purely cultural processes.

What can this simple model tell us about the evolution of communication in general and human language in particular? This model suggests two necessary preconditions for the emergence of optimal communication through cultural processes:

(1) the capacity to read, at least to some extent, the communicative intentions of others.
(2) possession of a bias in favour of one-to-one mappings between meanings and signals.

It should be noted that these preconditions are perhaps not sufficient for the cultural evolution of optimal communication—there may be other preconditions which do not feature in this model. Notwithstanding this caveat, does any species meet these two criteria?

Human infants come to the language-learning task equipped with a sophisticated ability to judge the communicative intention of others (e.g. Baldwin 1991, 1993a, b, Bloom 1997) and an apparent bias in favour of one-to-one mappings between objects and words (the principle of contrast (Clark 1988)). This model suggests that the iterated application of these biases should result in the cultural emergence of a near-optimal, or at least effective, learned communication system. The uniqueness of human language suggests that no other species possesses these biases, although a fuller understanding of all necessary preconditions may shed more light on this issue.

Where might these biases in humans come from? It is tempting to conclude that such an endowment must have evolved through natural selection to facilitate communication or language. However, preliminary results from a model based on the model outlined here, where genetic evolution of weight-update rules occurs alongside cultural transmission of communication systems, suggest that the evolution of appropriate learning biases may not be straightforward (Smith 2001).

In addition to its learned and symbolic nature, human language is unique in possessing syntactic structure. While previous work (e.g. Kirby 2001, Brighton 2002) suggests that this structure may be due to cultural dynamics rising from the poverty of the stimulus available to language learners, the results of the model outlined in this paper suggests that learner biases may also have a role to play. Is a bias in favour of one-to-one mappings between (parts of) meanings and (parts of) signals a necessary precondition for the cultural evoluton of syntax? As this paper illustrates, the combination of associative networks and the iterated learning model is a powerful technique for investigating such questions.

Acknowledgements

Thanks to Prof. Jim Hurford, Dr Simon Kirby, Henry Brighton, Andrew Smith and the anonymous reviewers for their helpful comments on this work.

Notes

1 The term ‘always’ has to be introduced to account for the stochastic nature of the behaviour of some networks, resulting from multiple nodes in the network receiving the same weighted sum of inputs on presentation of a pattern. In practice, ‘always’ was reduced to ‘for every one of 1000 trials’.

2 $N = 100$ for all iterated learning models outlined in this paper. However, different values on N yield qualitatively similar results.

3 In order to add noise to a meaning-signal pair $<m_i, s_j>$, s_j is replaced with a randomly-selected $s_k \in \mathcal{S}$, where $k \neq j$.

4 $e = 3$ for all iterated learning models outlined in this paper. Once again, different values of e yield qualitatively similar results.

5 The population’s communicative accuracy was estimated by evaluating every individual’s average communicative accuracy as both producer and receiver with two randomly selected partners according to the measure $ca(P, R)$ given in section 2.1 and averaging over all individuals in the population.

References

Baldwin, D.A., 1991, Infants’ contribution to the achievement of joint reference. *Child Development*, **62**: 875–890.

Baldwin, D.A., 1993a, Early referential understanding: Infants' ability to recognise referential acts for what they are. *Developmental Psychology*, **29**: 832–843.

Baldwin, D.A., 1993b, Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language*, **20**: 395–418.

Batali, J., 1994, Innate biases and critical periods: Combining evolution and learning in the acquisition of syntax. In R. Brooks and P. Maes (eds) *Artificial Life 4: Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems* (Redwood City, CA: Addison-Wesley), pp. 160–171.

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

Batali, J., in press. The negotiation and asquisition of recursive grammars as a result of competition among exemplars. In E. Briscoe (ed.) *Linguistic Evolution through Language Acquisition: Formal and Computational Models* (Cambridge: Cambridge University Press).

Bloom, P., 1997, Intentionality and work learning. *Trends in Cognitive Sciences*, **1**: 9–12.

Boyd, P., and Richerson, P.J., 1985, *Culture and the Evolutionary Process* (Chicago, IL: University of Chicago Press).

Brighton, H., 2002, Compositional syntax from cultural transmission. *Artificial Life*: **8**.

Cangelosi, A., 1999, Modelling the evolution of communication: from stimulus associations to grounded symbolic associations. In D. Floreano, J.-D. Nicoud and F. Mondada (eds) *Advances in Artificial Life: Proceedings of the 5th European Conference on Artificial Life* (Heidelberg: Springer-Verlag), pp. 654–663.

Cangelosi, A., Greco, A., and Harnad, S., 2000, From robotic toil to symbolic theft: grounding transfer from entry-level to higher-level categories. *Connection Science*, **12**: 125–148.

Cangelosi, A., and Parisi, D., 1998, The emergence of a 'language' in an evolving population of neural networks. *Connection Science*, **10**: 83–97.

Chomsky, N., 1987, *Knowledge of Language: Its Nature, Origin and Use* (Dordrecht: Foris).

Christiansen, M., and Devlin, J., 1997, Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations. In M. Shafto and P. Langley (eds) *Proceedings of the 19th Annual Cognitive Science Society Conference* (London: Lawrence Erlbaum Associates), pp. 113–118.

Clark, E., 1988, On the logic of contrast. *Journal of Child Language*, **15**: 317–335.

Elman, J., 1993, Learning and development in neural networks: The importance of starting small. *Cognition*, **48**: 71–99.

Hare, M., and Elman, J. L., 1995, Learning and morphological change. *Cognition*, **56**: 61–98.

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

Hazelhurst, B., and Hutchins, E., 1998, The emergence of propositions from the co-ordination of talk and action in a shared world. *Language and Cognitive Processes*, **13**: 373–424.

Hutchins, E., and Hazelhurst, B., 1995, How to invent a lexicon: the development of shared symbols in interaction. In N. Gilbert and R. Conte (eds) *Artificial Societies: the Computer Simulation of Social Life* (London: UCL Press).

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

Kirby, S., and Hurford, J.R., 2002, The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi and D. Parisi (eds) *Simulating the Evolution of Language* (London: Springer-Verlag), pp. 121–147.

Kvasnička, V., and Pospíchal, J., 1999, An emergence of coordinated communication in populations of agents. *Artificial Life*, **5**: 319–342.

Livingstone, D., and Fyfe, C., 1999, Modelling the evolution of linguistic diversity. In D. Floreany, J.-D. Nicoud and F. Mondada (eds) *Advances in Artificial Life: Proceedings of the 5th European Conference on Artificial Life* (Heidelberg: Springer-Verlag), pp. 704–708.

Macnamara, J., 1972, The cognitive basis of language learning in infants. *Psychological Review*, **79**: 1–13.

Oliphant, M., 1999, The learning barrier: Moving from innate to learned systems of communications. *Adaptive Behavior*, **7**: 371–384.

Sampson, G., 1997, *Educating Eve: The 'Language Instinct' debate* (London: Cassell).

Smith, A., 2001, Establishing communication systems without explicit meaning transmission. In J. Kelemen and P. Sodik (eds) *Advances in Artificial Life: Proceedings of the 6th Europeanm Conference on Artificial Life* (Heidelberg: Springer-Verlag), pp. 381–390.

Smith, K., 2001, The importance of rapid cultural convergence in the evolution of learned symbolic communication. In J. Kelemen and P. Sosik (eds) *Advances in Artificial Life: Proceedings of the 6th European Conference on Artificial Life* (Heidelberg: Springer-Verlag), pp. 637–640.

Steels, L., and Kaplan, F. 1999, Collective learning and semiotic dynamics. In D. Floreano, J.-D. Nicoud and F. Mondada (eds) *Advances in Artificial Life: Proceedings of the 5th European Conference on Artificial Life* (Heidelberg: Springer-Verlag), pp. 679–688.