DRAFT ONLY

Methodological Issues in Simulating the Emergence of Language

Bradley Tonkes* and Janet Wiles*†
*School of Computer Science and Electrical Engineering
†School of Psychology
University of Queensland, 4072
Queensland, Australia
{btonkes,janetw}@csee.uq.edu.au

0.1 Introduction

One of the features that differentiates language from other forms of communication is the ability to communicate a greater number of meanings than there are basic signals in the repertoire of the speaker. In human languages syntactic and morphosyntactic constructions allow combinations of simpler elements to express complex meanings. With advances in computational techniques it has become possible to model some of the processes by which populations of communicating agents come to agree on a convention for combining smaller linguistic elements into larger ones.

An interesting idea to emerge from computational studies of this issue is that the dynamics of language transmission (i.e., the process through which speakers of a language teach it to the next generation) may be responsible for some of the phenomena that have been solely attributed to an innate linguistic competence. The proposed hypothesis is that languages adapt to become more easily acquired by their learners and that this process of adaptation may be responsible for some of the observed constraints on cross-linguistic variation (Kirby 1999a). Thus, syntactic conventions are partly determined

by the process of linguistic transmission, rather than simply reflecting an underlying innate grammatical competence.

A major goal of the computational modelling research is to determine the conditions under which syntactic conventions can be established in a population. That is, to determine when language-like systems of communication (i.e., those that combine simpler elements to form larger constructions) can emerge. Humans are alone in their use of structured communication. Which aspects of human brain organisation and human social organisation allowed humans to make the advance beyond the signalling systems found in other species? A secondary goal is to determine the types of structural conventions that can be established in a population. The set of human languages are the only examples that are evident in the real world. However, it is unknown which aspects of this set of languages are inevitable emergent properties and which are idiosyncratically human.

Batali (1998) has demonstrated the emergence of rudimentary syntactic structures in a population of communicating neural networks. Despite the absence of a sophisticated, innate linguistic competence in the networks, Batali was able to show how the population converged on a language with compositional characteristics. Are there other reasons for the emergence of compositional structures? An alternative candidate explanation is the language transmission dynamic itself. Kirby (1999b) has extended Batali's work by showing how compositional structures may be a result of a language transmission dynamic. As languages are passed from one generation to the next, they are filtered through the learning experience. Importantly, the learning experience acts as a bottleneck since a language learner can never observe every sentence in the language. Kirby argues that a consequence of this bottleneck is a pressure for languages to evolve towards forms that are easy generalisable by learners, and presents some intriguing simulations to demonstrate his point.

It is worth emphasising at this stage that the goal of this research is not necessarily to determine the particular properties of humans that give rise to every facet of *human* languages. Humans are amazingly complex beings who cannot be modelled accurately, and it is not feasible to separate those aspects that are important for structured communication from those that are coincidental. Rather, the long-term goal of the research is to establish the conditions under which language-like systems *in general* can emerge and the range of properties that such systems exhibit. While achieving this goal will not inform us of how every feature of human language came to be,

it will shed light on the necessary and sufficient conditions for structured communications systems to emerge, of which human languages are but a subset. It will also be informative about the general properties of structured communication systems, such as whether they are necessarily compositional.

Kirby's simulations, like all computational models, consider an idealised system. Consequently, although Kirby shows that a language-learning evolutionary dynamic is sufficient to evolve a learnable language under a particular set of circumstances, the generality of his results is open to debate. For computational models such as Kirby's, it is important to establish the features of the abstraction that lead to the observed results. That is, we should strive to understand the parts of the abstraction that are required, those which are superfluous, and those that must be constrained to a critical range of values.

In this paper we explore Kirby's simulations in greater detail. Kirby credited his results to the 'learning bottleneck' but didn't examine variations in learners, tasks or parameters. His choice of language learning mechanism was based on a learning algorithm that had been previously used in computational linguistics. The choice of semantic domain was constrained so as to have combinatorial structure. The question we consider is whether the learning bottleneck is the primary factor with a different kind of learning mechanism and a differently structured semantic domain.

In previous work, we have considered communication between a pair of agents that try to communicate a meaning, represented by a value between 0 and 1, using an utterance composed of a sequence of symbols. For each meaning one agent produces an utterance which the other receives and processes back into a meaning. Using this framework, we have shown how a language can evolve to mediate the different computational demands of sender and receiver (Tonkes, Blair & Wiles 1999), and how language evolution can facilitate learning by adapting towards the forms that exploit the weak biases of a general purpose learner (Tonkes, Blair & Wiles 2000). It is this communication task that we incorporate into Kirby's population model in the simulations presented in this chapter.

In section 0.2 we review Kirby's simulations in greater detail and raise issues related to his learning mechanism that we believe are crucial for his results. His learning mechanisms looked for common substrings and inferred generalised rules for generating them. We believe that this assumption is unnecessarily strong, and that a weaker assumption can be tested in an alternative framework. In section 0.3 we present our alternative framework and highlight the similarities and differences to Kirby's, particularly the learner,

the differently structured domain and the parameters. These simulations are performed varying two parameters: the amount of training data supplied to the learners (the size of the bottleneck), and the size of the population. The results of these simulations, presented in section 0.4, reveal that the training corpus size has a significant impact on the communicative accuracy in the population, while changes to the size of the population merely alter the rate at which change occurs. Section 0.5 provides an analysis of why the results vary across changes in these parameters. In section 0.6 we further explore how Kirby's results depend upon experimental conditions, by varying aspects of the learning environment.

0.2 Kirby Revisited

Kirby (2000) presents a compelling demonstration of the emergence of grammar in the absence of any phylogenetic adaptation. A population of ten language users, modeled as context-free grammars, are arranged in a ring so that each individual has two neighbours. Individuals are capable of talking about simple meanings (agent/action/patient tuples) using strings produced from a restricted alphabet of five symbols. While individuals are equipped with a learning mechanism, the initial population has no vocabulary and no grammar. That is, the initial population consists of a mechanism for *acquiring* language, but no language to acquire.

To bootstrap the system, Kirby introduces the notion of random invention: if an individual wants to talk about a particular meaning but has no way of expressing that meaning, it either says nothing or, with small probability, produces a random string. The course of a simulation runs as follows.

1. Replace a randomly chosen individual with a new individual.
2. Produce a corpus of training examples from the utterances produced by the new individual's neighbours.
3. The new individual induces a new grammar based on this corpus. During this training phase, the learner is presented with both the utterance *and its intended meaning.* During normal operation, only the utterance is presented. The learner is thus required to generalise the relationship between utterances and meanings from the subset of observed examples.

4. Return to step (1).

At the start of a simulation run, the training corpora are typically small and contain examples that are more-or-less random. That is, there is no systematic relationship between utterance and meaning. Gradually, the training corpora become larger as each individual's grammar becomes more expressive. After a period of time, individuals start to *regularise* their grammars in a compositional manner using common substrings for common parts of a meaning. For example, a meaning such as (mary, john, likes) may correspond to an utterance such as 'marylikesjohn' while a meaning such as (mary, fred, likes) may correspond to an utterance such as 'marylikesfred.' Eventually, the population comes to use a fully compositional language where every utterance can be broken into subcomponents, each representing a part of the meaning tuple.

Kirby deliberately chose the size of the training corpora so that it was highly unlikely that an individual would be exposed to the full set of *(meaning, utterance)* pairs. That is, the only way that an agent could acquire a complete grammar was to generalise from a limited subset of exemplars. Kirby hypothesises that it was this feature of the simulations — the 'learning bottleneck' — that caused the fundamental shift in the languages produced, from non-compositional to compositional.

If meanings and utterances are randomly associated, then there is nothing on which to base a generalisation mechanism. An unobserved association is therefore unlearned. Conversely, with a systematic relationship between meanings and utterances, it is possible to generalise from a limited set of observed exemplars. This dichotomy, Kirby argues, introduces a 'glossogenetic' selection pressure for languages that can be expressed by a few general purpose rules that can be induced from a fewer set of examples. For these languages, it is not necessary to see every *(meaning, utterance)* pair, rather, it is learnable from any subset of exemplars from which the general rules can be derived.

Although there is no phylogenetic adaptation during the course of Kirby's simulations, the model incorporates phylogenetic adaptation implicitly in the design of the individuals' language learning mechanisms. That is, the starting point of the simulations is a population of individuals that are innately endowed with a particular learning mechanism. It seems to us that the chosen induction algorithm is highly biased towards language-like, compositional structures, which is perhaps not surprising given that the algorithm was orig-

inally developed for computational linguistics. Although Kirby highlights the importance of languages themselves being systems that adapt to their human hosts, inherent in his choice of learning algorithm is a strong form of language-specific learning bias.

0.3 Methodology

The design of the simulations in this paper owe much to previous work. The overall dynamic of linguistic interactions, outlined above, is taken from Kirby's (2000) work. The linguistic agents are of the same type used by Batali (1998) and the semantic domain is one that we have used in previous work (Tonkes et al. 2000). While we present here an overview of the simulation design, the interested reader is directed to the original sources for a more in-depth treatment.

0.3.1 Something to talk about

Whereas in Kirby's original simulations, the agents attempted to communicate simple predicates denoting agent, action and patient ('Who did what to whom.'), in our simulations we use a much simpler semantic domain. Meanings are represented as values between 0 and 1, which for simplicity are restricted to 100 values of 0.01 increments (i.e., $0.00, 0.01, 0.02, \ldots, 0.99$). These meanings are numerically related so that is possible to measure the similarity of two meanings by taking their numeric difference (for example, 0.00 is more similar to 0.01 than it is to 0.30). It thus makes sense to introduce the notion of degrees of understanding, rather than deciding that an utterance has been either 'understood' or 'not understood.' To this end, the domain lends itself to a convenient way to measure communicative error, which we will take to be the squared difference between the meaning intended by the sender and the meaning as interpreted by the receiver.[1] The similarity between items in this space is analogous to similarity between real-world items. For example, the similarity between red and pink may be analogous to the similarity between 0.40 and 0.50. However, we are trying to model the

[1]For example, if the sender tries to communicate the meaning 0.45 which the receiver (mistakenly) understands as 0.65, then the communicative error for that interaction is $(0.45 - 0.65)^2 = 0.04$.

similarity structure between items rather than the labels attributed to particular items. The model may thus be interpreted as an abstract conception of the similarity amongst meanings in a semantic domain.

0.3.2 Communicative agents

As noted earlier, our communicative agents are modelled as simple recurrent networks (SRNs) of the same type as those used by Batali (1998, see the enlarged section in Figure 0.1). SRNs (Elman 1990) are a type of neural network that are particularly well suited to sequential tasks such as language processing, where they have demonstrated some impressive results (Elman 1991). SRNs can be trained to associate a sequence of patterns (in this case, an utterance that is a sequence of symbols) with an output pattern (in this case, a meaning value).[2]

In previous work (Tonkes et al. 1999, Tonkes et al. 2000), we have differentiated between senders (those agents that generate utterances from meanings) and receivers (those agents that recreate meanings from utterances). For this chapter we use an alternative approach first introduced by Batali, where the same network is used for both sending and receiving. SRNs are not normally capable of such dual-mode operation, so to achieve the desired behaviour, Batali used networks that were designed to be receivers and applied a special operation to make them capable of sending (for this reason, Figure 0.1 shows only a receiving network). The operation that Batali applied is known as an 'obverter' procedure (Oliphant & Batali 1996). The essential idea is that to communicate some meaning M, an agent searches for an utterance U such that if the agent itself were to hear U, it would interpret it as meaning M. (That is, the agent tries to work out the inverse of its own receive function.) Note that understanding the precise mechanics of the obverter procedure is unnecessary for understanding the remainder of the chapter.

Similar to both Batali and Kirby, the utterances themselves consist of sequences of up to six symbols which are taken from an alphabet of size four. Following common neural network practice, the symbols are represented as four-dimensional binary vectors $[1, 0, 0, 0]$, $[0, 1, 0, 0]$, $[0, 0, 1, 0]$ and $[0, 0, 0, 1]$ which we denote `A`, `B`, `C` and `D` respectively. These vectors are used as the activations for the utterance input units in Figure 0.1.

[2]For the simulations in this chapter, we used the backpropagation-through-time algorithm (Rumelhart, Hinton & Williams 1986) with a learning rate of 0.01 and a momentum term of 0.9.

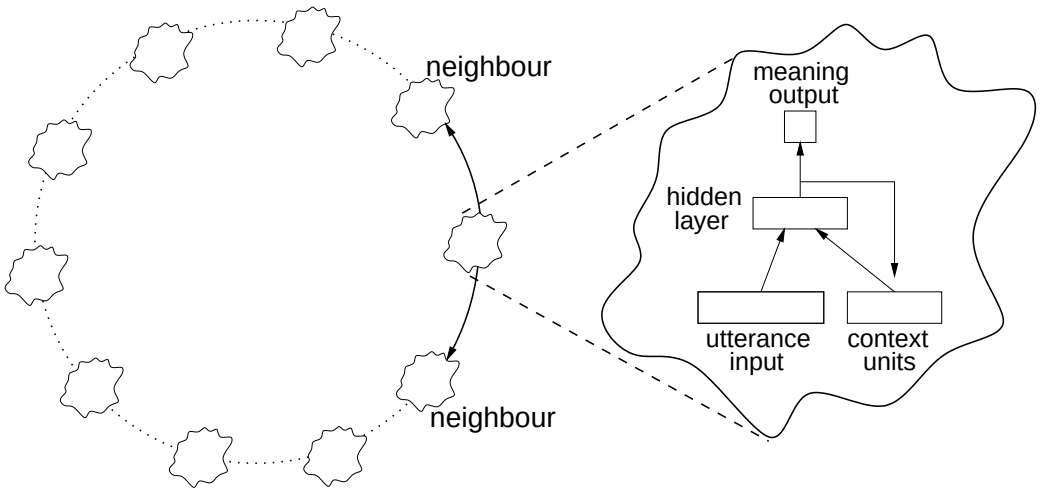

Figure 0.1: A population of communicating agents. Each agent is modeled by a simple recurrent network and can communicate with two neighbours so that the population forms a ring. The operation of the SRN shown in the enlarged section can be described as follows. Each of the blocks represents a set of simple processing units whose activations are determined by both the activations of the processing units in the previous layer and the strengths of the connections between them (also called 'weights'). The process by which activations flow from the utterance inputs to the output is known as propagation. The activations of the utterance inputs are determined externally by an incoming utterance which is received one symbol at a time. The activations of the context units are copied from the activations of the units in the hidden layer after each symbol is processed. These units are used to provide the network with a working memory and are the characteristic feature of the SRN. The hidden units may be viewed as an internal working space. The meaning that the SRN associates with the utterance is read off the single output unit after the final symbol in the utterance has been propagated through the network.

The communication of a meaning from sender to receiver might proceed in the following manner:

1. The sender decides to communicate a value such as 0.43.

2. Using the obverter procedure outlined earlier, the sender determines the sequence of no more than six symbols (an utterance) that it understands as the best approximation to 0.43. In this example the sender might understand `ACDBA` to mean 0.41, which is the closest approximation it can find.

3. The utterance, `ACDBA`, is sent to the receiver.

4. The activations of the processing units of the receiver are initialised to zero so that there is no memory of previous utterances. The vectors of activation values corresponding to each symbol in the utterance are propagated through the receiver's SRN, one at a time. Each SRN has four utterance input units corresponding to the size of the symbol vectors, five hidden units, and a single output unit used for the interpreted meaning.

5. The meaning as understood by the receiver is read from the activation of the receiver SRN's output, in this case it might be 0.46.

6. The next step depends on whether or not the receiver is being trained.

   - If the receiver is in its learning phase then it is informed of the intended meaning. It then uses the discrepancy between the intended meaning (0.43) and the interpreted meaning (0.46) to update the weights between processing units so that future presentations of the utterance `ACDBA` will tend to be understood as a meaning closer to 0.43. Because of the nature of neural network learning, it may take many presentations of the same training material before the learner makes no errors.
   - If the receiver is not being trained then it is unaware of the intended meaning. It is possible however, for an external observer to measure the squared communicative error, $(0.43-0.46)^2 = 0.0009$.

0.3.3 Population dynamics

A population of networks is arranged in a ring so that each individual has two neighbours (Figure 0.1). Simulations are run for 2500 time-steps. In each time-step of a simulation, the following sequence of events occur.

1. Replace a randomly chosen network by a new network. Set the connection strengths of the network to small random values.

2. Create a training corpus by using the new individual's two neighbours to generate a set of utterances corresponding to a randomly chose set of meanings. The training set contains utterances produced by both neighbours as well as their intended meanings.

3. Train the new network on the training corpus using the process outlined earlier. The entire training corpus is presented to the network 1000 times.

4. Evaluate the *communicative accuracy* of the population in the following way. Every combination of sender and receiver, regardless of location, attempts to communicate the 100 meanings. The squared communicative error for each meaning is summed giving a communicative error score for each (sender, receiver) pair. These scores are then averaged, giving a measure of the average communicative error for the population.

We vary two parameters of the simulations — the size of the training corpus and the size of the population — and consider three variations of these parameters. In the first variation we use a population of size ten and a training corpus of size ten. The second variation increases the size of the training corpus to twenty while keeping the population size at ten. The third variation increases the size of the population to twenty while keeping the training corpus size at ten. We refer to this set of simulations as series 1 and the three combinations of parameter settings as studies 1A (small population, small corpora), 1B (small population, large corpora) and 1C (large population, small corpora). Importantly, the size of the training corpus is chosen to always be significantly less than the size of the full meaning set. Consequently, networks are required to generalise well beyond the examples in the training corpus to communicate about the full set of meanings.

0.3.4 Putting it all together

In this section we briefly describe what happens during a typical run. The initial population of networks are untrained and generally produce uninteresting languages. Networks are unable to produce enough unique utterances to differentiate every meaning. Typically, networks are only able to produce three or four different strings which are reused for many of the 100 meanings. In almost all cases each unique utterance is used for a single contiguous range of meanings. For example, a network may send `DDDD` for meanings with values between 0.00 and 0.35, `DDBD` for meanings with values between 0.36 and 0.65 and `DBBB` for meanings with values from 0.66 to 0.99. Furthermore, the agents in the population disagree on which utterance corresponds to a given meaning. The average communicative accuracy is consequently very poor and agents have little success even in understanding their own utterances. (The degree to which an agent comprehends its own utterances can be tested by taking two copies of the agent, one which acts as sender, the other as receiver, and measuring their communicative error.)

One of the agents is then replaced with a new individual. The new individual is trained on a set of examples produced by its two neighbours. Since the output of the two neighbours is unrelated, the training data for the new network is likely to be a confusing blend. After training, the new network shares some characteristics of the languages produced by its neighbours and is usually able to understand its own utterances. The communicative accuracy of the newly trained network is typically better than the remainder of the population.

After several agents have been replaced and new ones trained, contiguous sections of the population begin to have reasonably high agreement on which utterances to use for which meanings. The consistency is never perfect, but networks do tend towards using similar strings for a given meaning. Often, one contiguous subset of the population will use one convention for a region of the meaning space, while the remainder of the population will use a different convention. For example agents one to five may use `AAAB` to communicate 0.50 while agents six to ten use `DDDC` to communicate the same meaning. At this stage, the vocabulary of the agents expands to around twenty unique utterances. That is, agents are capable of differentiating twenty regions of the meaning space where initially they were able to differentiate only three or four. From this point onwards, the course of the simulation is dependent on the choice of parameters. We elaborate on this point in the next section.

0.4 Base Results

For each of the three combinations of population size and training data parameters, three separate runs of the simulation were performed with different seeds of the random number generator producing different sets of initial weights and different choices of training examples. In all cases, simulations performed under the same parameters yielded qualitatively and quantitatively similar results. The results presented here are based on the communicative accuracy of the populations, averaged across the three trials performed for each set of simulation parameters. The communicative error between a sender and a receiver is determined by the squared error between the meaning intended by the sender and the meaning as understood by the receiver, summed across the 100 possible meanings. The communicative error for the population as a whole is taken to be the average communicative error for every possible combination of sender and receiver. From previous studies, we have determined that a communicative error score of one or less corresponds with acceptable communicative accuracy.

With a small population size and with small training corpora (study 1A), the populations always failed to reach consensus on a language, as shown in Figure 0.2. After a brief initial period where communicative error drops quickly, the error increases again. Throughout the course of a run, the communicative accuracy of the population continues to oscillate, and even during the better periods, the populations fail to communicate with an acceptable degree of error. During the initial improvement in accuracy and during subsequent periods of good performance, individual's languages show a reasonable level of agreement with some other members of the population, and there are easily distinguishable 'families' of languages. The populations that are responsible for the periods of high error show little coherence. Although small subsets of the population (two or three individuals) may use languages that are somewhat similar, there is no consensus amongst the population at large.

Keeping the same population size as for the previous study while increasing the amount of training data presented to new agents (study 1B) significantly improves performance (see Figure 0.3). There is a rapid initial convergence as the population reaches consensus on a language. The languages produced across the population are not identical, however they are sufficiently similar for accurate communication. While the performance of the population remains on average quite good, there are several transient increases in error. During these periods part of the population uses a com-

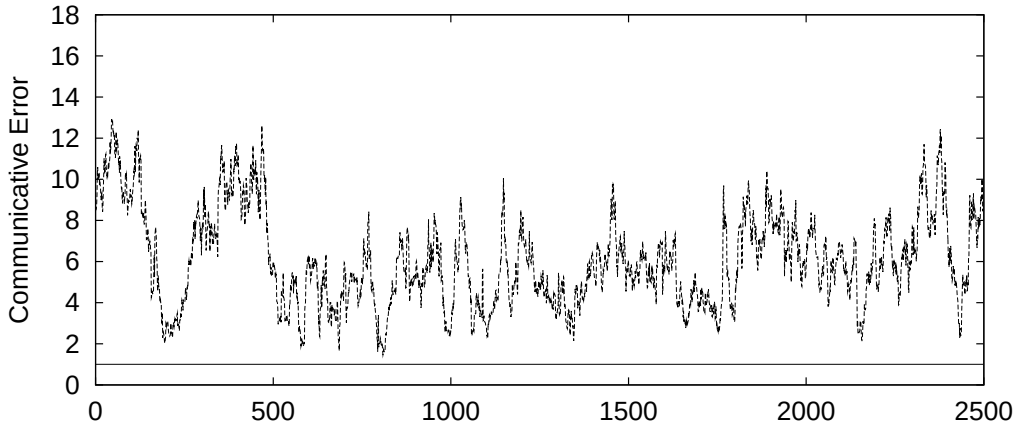

Figure 0.2: Communicative error over time for a population of size ten, using ten examples to train new individuals (study 1A). With these parameters, the population fails to converge on an acceptable language (i.e., one with mean communicative error score lower than threshold at 1.0, shown on the graph).

pletely different language where the population agrees on some regions of the meaning space but not on others. Interestingly, the populations on either side of these transient failures may use languages that are different. That is, following the 'corruption' of the language, the population may reconverge on a different language to the one used previously.

Increasing the population size (study 1C) significantly slows the rate of change of the population (see Figure 0.4). With the larger population size there is a prolonged period before convergence to an acceptable level of agreement. Indeed, for an initial period the communicative error of the population is substantially higher than at the start. In this region the utterances used by some agents for meanings close to zero are the same as those that other agents use for meanings close to one, and vice versa, giving a worse-than-chance error when they attempt to communicate with one another. Furthermore, under these conditions the population remains unstable in the same way as the case above. Running the simulation for more that 2500 generations reveals that after the population converges, the same increases in error occur. Moreover, the periods of increased error are of greater duration than those observed in the smaller populations. A representative example of the types of languages

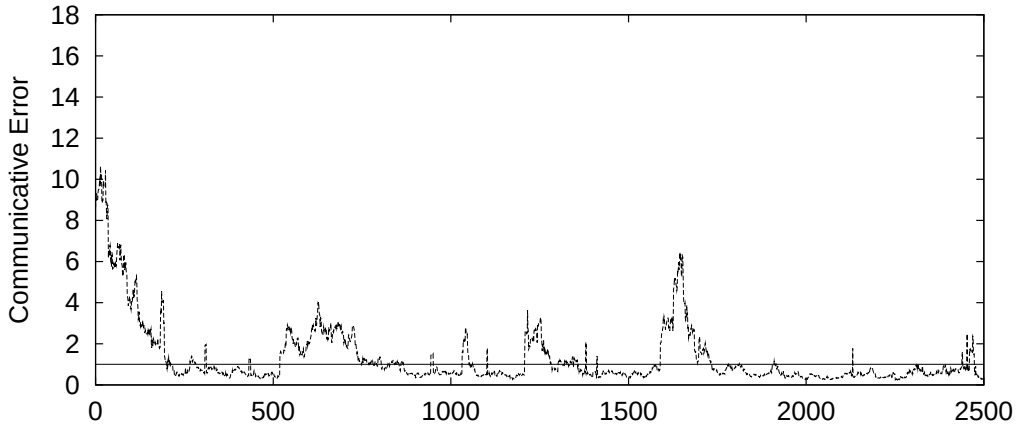

Figure 0.3: Communicative error over time for a population of size ten, using twenty examples to train new individuals (study 1B). Although the population converges to a good language, there are several periods of high error during which two competing languages appear. In these situations the original language may be replaced by a new variant.

found in a population is shown in Table 0.1.

0.5 Analysis of Base Results

From observing the change in the languages of the population over time we conclude that much of the behaviour shown in Figures 0.2, 0.3 and 0.4, and the differences between them can be attributed to one cause. Namely, that if a learner fails to acquire the language of its neighbours, then nothing prevents that individual teaching its poorly formed language to subsequent learners. The most significant factor in the failure of an individual to learn is the data presented to the learner. If the ten or twenty training examples are chosen poorly (for example, if they are all less than 0.5), it is much harder for the learner to successfully generalise to the remainder of the space. Utterances for similar meanings tend to be similar so if an agent knows the utterance associated with a meaning such as 0.78 it is more likely to be able to guess the meaning of the utterance associated with 0.75 that it is to guess the meaning of the utterance associated with 0.10.

As the number of training examples increases, the probability of an in-

Figure 0.4: Communicative error over time for a population of size twenty, using twenty examples to train new individuals (study 1C). The population behaves similarly to that in Figure 0.3 but on a much slower time-scale. If the population is allowed to run beyond the 2500 generations shown here, similar intrusions of rogue languages cause intermittent periods of high error.

adequate sampling of the space diminishes. Hence, the population shown in Figure 0.2 which uses ten training examples is far less stable than the population shown in Figure 0.3 which uses twenty training examples. Other factors, such as the initial connection strengths of the learner may also cause learning failures. However, further simulations (section 0.6.1) indicate that the initial weights do not play as significant a role as the distribution of training data.

The differences in time to convergence between Figures 0.3 and 0.4 can be attributed to greater propagation delays associated with the increase in population size. With a population of size ten, individuals are at most five neighbours away from any other individual. Consequently, the speed with which a change in a language can propagate through the entire population is much greater than with the larger population size (twenty). Once a population forms two (or more) distinct languages it also takes a greater time before one comes to dominate. Assuming that the languages are equally learnable, one comes to dominate only through providing a disproportionate number of examples in the training corpora of new individuals. Since there is random selection of which neighbour provides a training example, language dispersal involves a degree of chance. An increase in population size increases the size

of the region that must be 'conquered', slowing the dispersal process.

0.6 Varying the Learning Environment

Just as in Kirby's simulations we have seen the emergence of co-ordinated, structured communication as a result of the dynamics of linguistic transmission. While not all of Kirby's results have been replicated (which we would not expect given the changes made to Kirby's simulation design), we have seen that one of significant outcomes (structured communication) does replicate with a different learning mechanism and a different semantic domain. We have also see that a successful outcome can be highly dependent on such factors as the size of the population and the amount of training data available to new individuals. In this section we consider alternative aspects of the learning environment that can influence the outcome of language evolution. The analysis of the first series of simulations indicated that part of the reason why populations failed to converge was that a single learner with an idiosyncratic language could corrupt future generations. Kirby explicitly sought to simulate language emergence in the absence of selection pressure to explore the power of glossogenetic adaptation alone. Hence, idiosyncrasies could not be eliminated from a language by a mechanism that removed the poorer speakers from the population. Consequently, the three factors that we vary in series 2–4 are chosen for their potential to either prevent learners from failing, or to stop failed learners propagating their half-formed languages.

It is well understood that failures in neural networks to learn a task can often be attributed to the choice of the initial weights (Kolen & Pollack 1990). In simulation series 2, we repeat the simulations of series 1, but instead of generating the initial weights of new individuals randomly, all new individuals start with the *same* weights. In making this change we allow a language to emerge that is learnable from a specific starting point. This technique has proven successful in other work (Tonkes et al. 2000, Batali 1994).

Another potential cause of learning failure that we have identified is the selection of training data from which new individuals learn. Learners are presented with a set of *(meaning, utterance)* pairs, where the meaning is a value between 0 and 1. If the selection of meanings in the training sample fails to provide sufficient coverage of the full meaning space, then it is much harder for the learner to generalise to unseen examples as they are dissimilar to the previously seen examples. In simulation series 3, rather than training

new learners on different, randomly chosen examples, new learners are trained on the same (randomly chosen) meanings.

In series 4, the variation to series 1 is that we remove the 'neighbourhood' assumption. Instead of using neighbours to provide the training data for new individuals, a 'teacher selection' principle is applied. After every time-step, each individual is given a score based on how well it is understood by the rest of the population (i.e., the portion of error that an individual contributes to the overall error, as plotted in Figures 0.2, 0.3 and 0.4). This score is used to select which networks generate the examples in a training corpus presented to a learner, based on a proportional selection mechanism (the probability of selection is inversely proportional to error). If a network fails to learn the language of its community then it will be unlikely to be selected to provide examples to train new individuals, thus limiting its impact on future generations.

In summary, the simulations of series 1 (section 0.3) are repeated under three different conditions:

1. Using the same set of initial weights for each new learner (series 2: fixed weights).

2. Using the same set of meanings to train each new learner (series 3: fixed examples).

3. Choosing the 'best' networks to generate the training examples for the new learners (series 4: teacher selection).

Again, population size and the training corpus size are varied and the simulations from three different random seeds are repeated under each condition (i.e., we perform studies 2A, 2B, 2C, etc.).

0.6.1 Results of Varying the Learning Environment

In all cases, the three repetitions of a condition yielded quantitatively similar results. However, across the conditions, the results varied radically. In the 'fixed weights' condition, the populations rapidly achieved reasonably low communicative error (see Figure 0.5). This effect may be attributed to the fact that all members of the initial population were identical (having the same, unadjusted connections). However, as in the original series of simulations, the population was unable to maintain this low degree of error and the error fluctuated markedly.

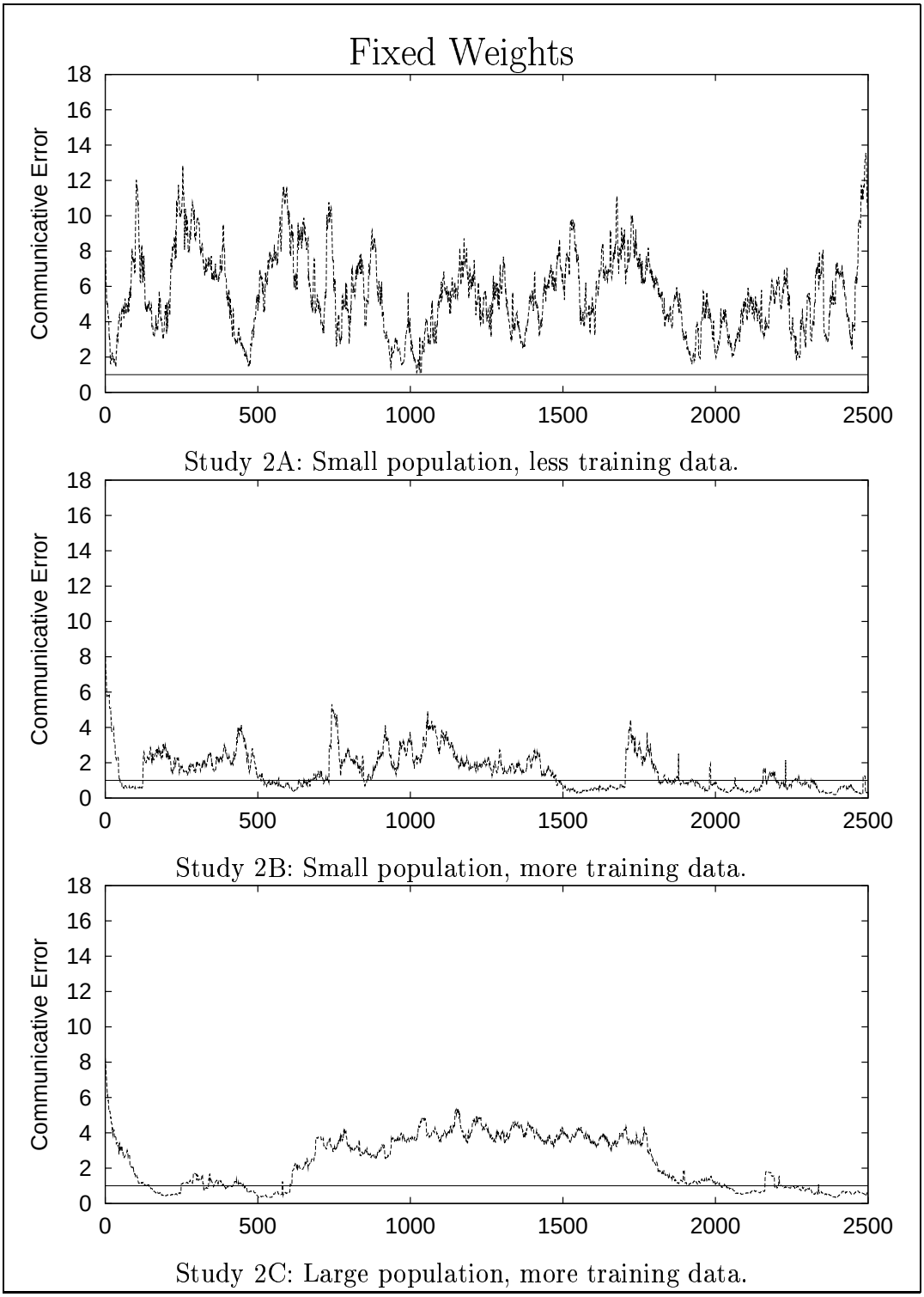

Figure 0.5: Communicative error of populations over time when new individuals always start from the same initial weights (series 2). Since all individuals are originally identical, the population converges quickly. However, as in section 0.4 the population frequently departs from an established convention.

In stark contrast, the populations in the 'fixed examples' condition took longer to converge in each case but showed a remarkable degree of stability (see Figure 0.6). Although there are some increases in error after the population has apparently converged, the error remains low. Surprisingly, there is no significant difference in the accuracy of the networks when the amount of training data is varied.

The populations in the 'teacher selection' condition demonstrate yet another pattern of error (see Figure 0.7). Again, the population rapidly attains a reasonable degree of communicative accuracy (low error). However, any increases in error are very short-lived, far more so than in the original simulations. With a small population and a small amount of training data (study 4A) the population is still unstable, but is much better on average than in the original simulations (Figure 0.2). Even with a larger size, the population very quickly arrives at a point of low error and tends to remain there, despite the occasional increases in error.

0.6.2 Analysis of Learning Environments

Again we performed an analysis of the changes in the population by observing the changes in the languages generated by each population. Apart from the initial improvements in communicative accuracy, the results of the 'fixed weights' populations are effectively the same as in section 0.4, indicating that the choice of initial weights is largely irrelevant. Conversely, the performance of populations in the 'fixed examples' condition suggest that the choice of training data is of vital importance. In this condition, only a single training corpus is generated. The probability that this particular corpus is unrepresentative of the meaning space is small, as it is for networks trained in the original simulations. In the original simulations 2500 different corpora are generated, one for each learner. The probability that some of these corpora are unrepresentative of the meaning space far exceeds the probability that the single corpus in the later simulations is unrepresentative. If, by chance, the single corpus was chosen poorly, we might expect that the population might never be successful. The results of the 'fixed weights' and 'fixed examples' simulations lead us to hypothesise that the populations evolve languages to a point where they are reliably learnable regardless of the initial weights of a network, and that only poorly chosen training samples prevent individuals from learning.

Populations in the 'teacher-selection' condition successfully reduced the

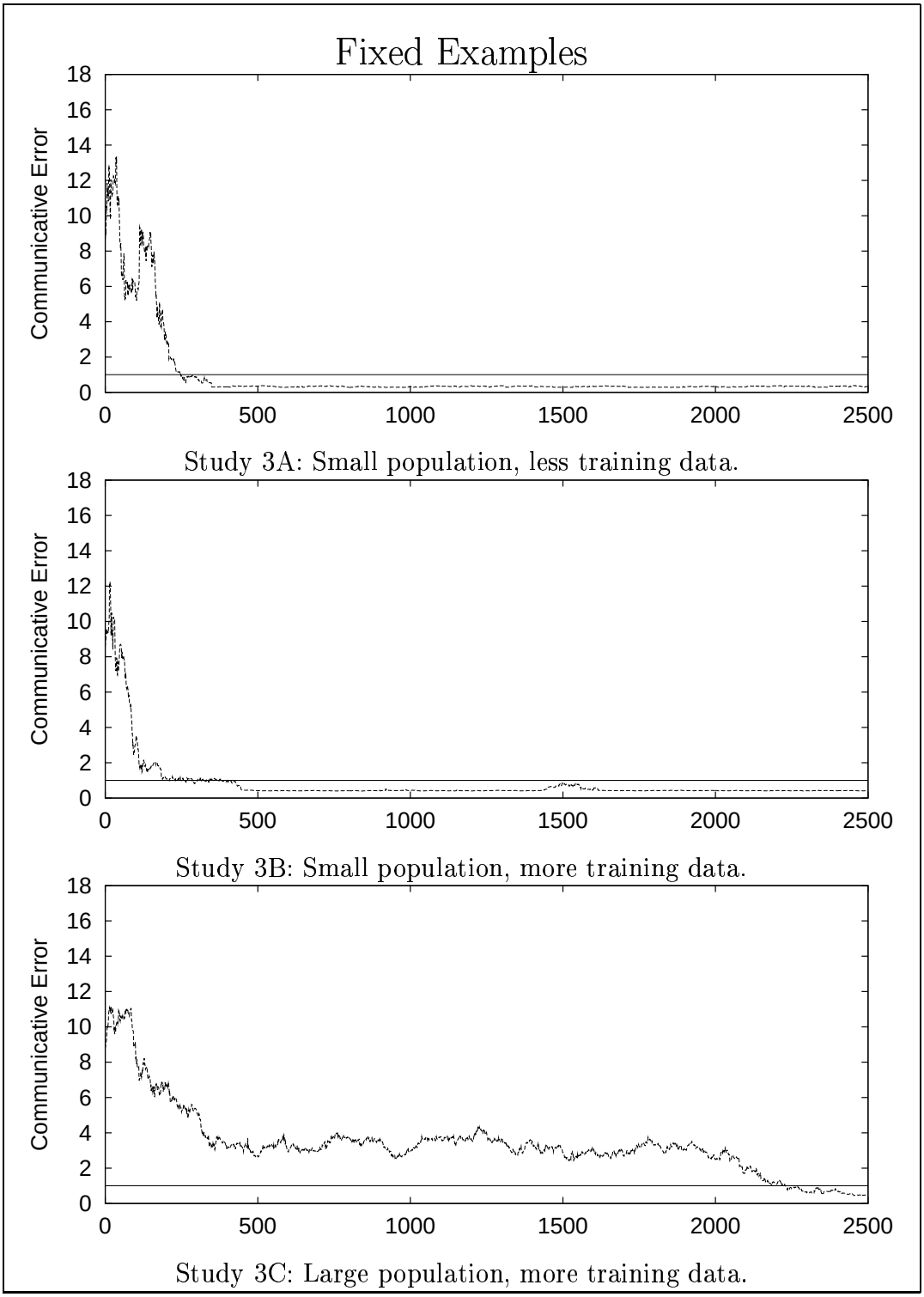

Figure 0.6: Communicative error of populations over time when new individuals are always trained on the same set of meanings (series 3). In all cases, the population is much more stable than its counterpart in the original simulation. Convergence is still slow for larger populations.

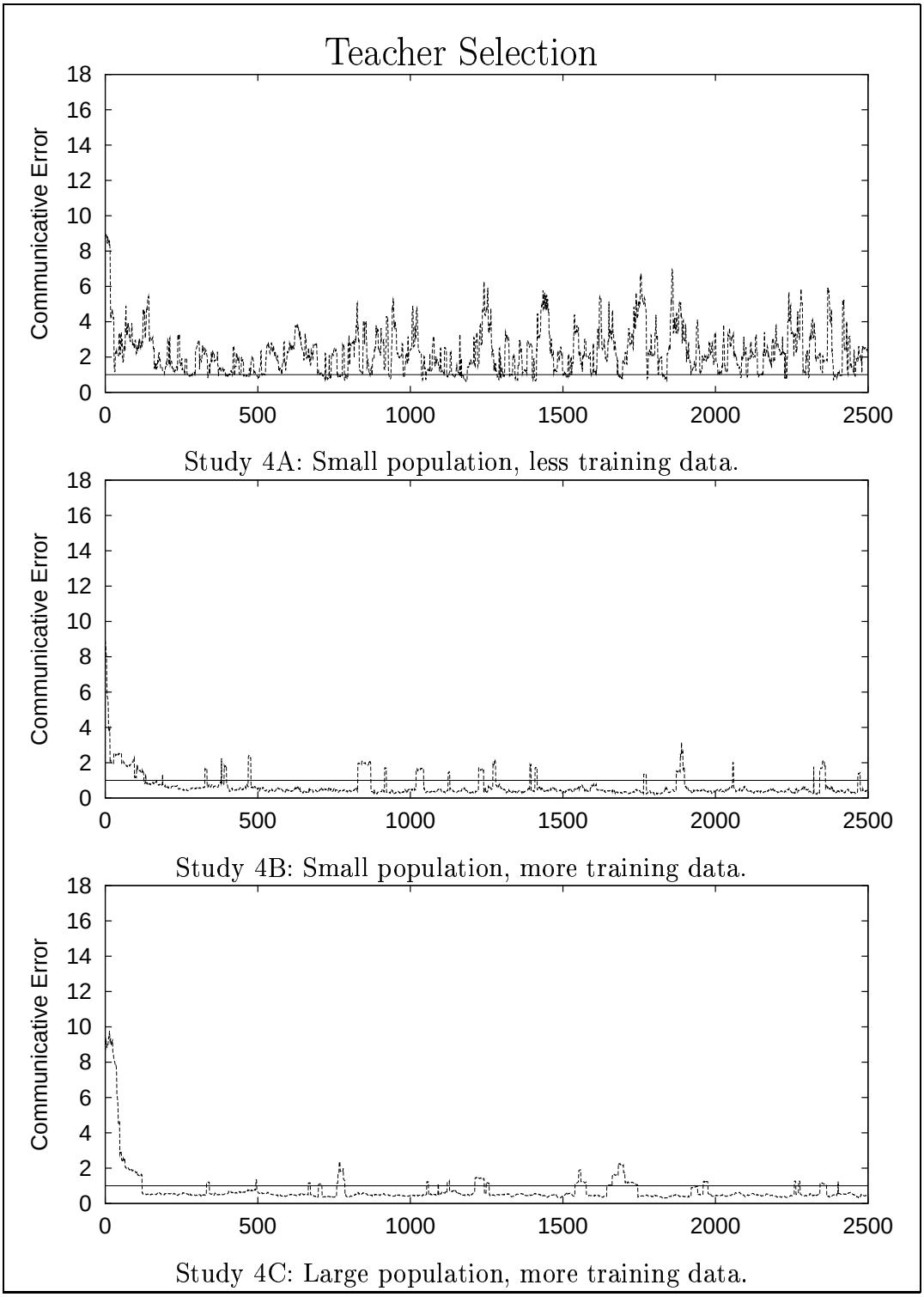

Figure 0.7: Communicative error of populations over time when new individuals are taught by the better communicators in the population (series 4). Convergence is rapid, even for larger populations. Periods of higher error tend to be transient.

influence of rogue learners. The impact can be best seen from the length of time that any population experiences high error. Particularly with a population size of ten and a training corpus of size twenty (study 4B, the middle graph in Figure 0.7), the length of periods of increased error closely follow the expected lifespan of an individual (ten time-steps on average). This observation suggests that while a rogue learner may lower the communicative error of the population, it does not pass its incompatible language to future generations. Communicative accuracy is thus restored once the rogue learner leaves the population. The effect is much less clear with the smaller training corpus since the probability of multiple successive failed learners is considerably higher. Increases in the number of inconsistent networks in the population increases the probability of further inconsistent networks, hence the instability in this case.

0.7 Discussion and Conclusions

In this final section we consider what correspondences can be drawn between the framework of these studies and characteristics of human language learners and environments. Simulations of populations of communicating simple recurrent networks showed that in favourable circumstances, languages could emerge in the absence of phylogenetic adaptation (section 0.4).

Our results demonstrate that one of Kirby's major findings — that a structured communication system can emerge from the dynamics of language transmission — has a generality beyond his original domain. While the kinds of language structures that emerge in our simulations are significantly different to those that emerged from Kirby's simulations, such a result should not be unexpected. The agents will employ the most appropriate structures for their respective communication tasks. Given the structure of the languages produced in our simulations, the results of the simulations may be used to refute claims that classical compositional syntactic structures are the only viable form of linguistic structure. Thus, human languages exhibit compositional structure not because it is the only valid alternative, but because other constraints on human communicative needs (such as the similarity structure of meanings as represented in the human mind) necessitate compositionality.

The effect of manipulating the two parameters in the simulations — population size and training corpus size — suggests some interesting implications for human languages. The results showed that populations converged on

languages regardless of the population size, although time to convergence was slowed by the larger population. Conversely, increasing the size of the training corpus (which can be viewed as increasing a learner's exposure to language, perhaps by increasing the critical period) vastly improved the success of populations. While it is not possible to state categorically that the same would be true of human populations, the results suggest an interesting hypothesis for the emergence human languages.

The modifications to the learning environment made in section 0.6 are also suggestive of the desirable conditions for language emergence. The first modification (fixed weights) may be viewed as analogous to a very weak genetic endowment of linguistic knowledge. This modification proved unsuccessful at improving the communicative accuracy of the population. In the second modification (fixed examples), the learning environment is consistent for every individual — every learner has the same set of experiences. With this environment, populations were far more successful at accurate communication. It is not outlandish to suggest that for humans, there is some degree of commonality between learning environments, although no two humans will share the exact same set of experiences.

Preventing failed learners from acting as teachers was also effective in maintaining the language of a population, but still required that learners were given sufficient training data. This condition introduced a selection mechanism, something which Kirby deliberately avoided adding. However, in populations where learners can fail, and then corrupt future learners, our simulations show that some kind of selection mechanism is important to maintain population stability. Such a mechanism may be manifested in a real-world situation by the direction of a learner's attention away from speakers with impaired language abilities.

Although Batali (1998) also used neural networks in his simulations, he did not include any generational component, instead using a static population. In his model, the agents in the population communicate amongst themselves until a consensus is reached. Consequently, after the first round of 'negotiations,' agents are no longer naive about the language of the community, making it difficult to look at changes in the language due to selection pressure for (naive) learnability. Batali also used a different semantic domain and his population lacked any kind of spatial organisation. However, it is interesting to note that Batali's populations were successful in producing basic combinatorial language structures despite the lack of an explicit 'learning bottleneck' — the very mechanism to which Kirby ascribes the success of his

simulations. One possible explanation for this disparity is that the learning mechanism itself may provide an implicit bottleneck. One feature of neural networks is their tendency to generalise based on similarity. Consequently, it is much easier for a neural network to learn a regular language than an irregular one; it may even be the case that a neural network will be *unable* to learn some irregular forms. In a series of negotiations, it would thus be expected that the more easily learnable forms (i.e., the regular languages) would persist — networks would compromise on the easier forms. By contrast, in Kirby's simulations, learners were not able to 'forget' associations between meanings and utterances: once a learner acquired an association, it remained for life. Thus, Kirby's learners lack the implicit bottleneck since they will always succeed at finding a grammar that is consistent with the training data.

To provide a comparison between Kirby's explicit bottleneck, and the hypothesised implicit bottleneck of the neural network learner, we ran a control study which repeated the first series of simulations (described in section 0.3), without removing individuals from the population. Instead, an individual was chosen to be given additional (learning) exposure to the language of its neighbours as in Batali's simulations. With small populations and small training corpus sizes, the population quickly reached a communicative error score of around one. The languages of these populations were still unstable, although not to the same extent as the population shown in Figure 0.2. Increasing the amount of training data received in each round resulted in a much more stable population. Even though populations in this condition periodically disagreed, such events were not as catastrophic as those in Figure 0.3. With a large population and large training corpora, populations were slow to attain reasonable communicative accuracy, much as in Figure 0.4, though the initial period of very high error was much shorter.

These results, although they are only preliminary, suggest that Kirby's explicit learning bottleneck may not be necessary. Certainly, they indicate that the role of the bottleneck is not as straightforward as Kirby described. Of course, in the case of human languages there clearly is such a bottleneck between generations of learners. Further work may help to determine whether this bottleneck plays a fundamental role, or is merely incidental to the course of language emergence. What seems plausible is a relationship between the implicit bottleneck of the learning mechanism, and the explicit bottleneck in Kirby's simulations.

The major contribution of this chapter is to broaden our ideas of when

structured communication systems emerge (and are stable) and when they do not. The chapter also considers the *types* of language structures that emerge from a given situation. Human languages are the only natural example of symbolic structured communications systems that we have. It is difficult to establish the causes for such unique phenomena. Computational models allow us to construct a variety of communication systems and to explore the conditions under which language-like systems can emerge. By examining the conditions under which language does, and does not emerge, we can draw conclusions about the significant aspects of the human environment that led to the evolution of human languages. The long-term goal is to deduce the general principles behind the emergence of language and properties of those languages. The work presented in this chapter represents a small step towards that goal.

0.8 Key further readings

Much of the simulation design used in this chapter was taken from the first three papers.

Batali, J. (1998). Computational simulations of the emergence of grammar, *in* J. R. Hurford, C. Knight & M. Studdert-Kennedy (eds), *Approaches to the Evolution of Language*, Cambridge University Press, Cambridge, England, pp. 405–426.

Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners, *in* C. Knight, J. R. Hurford & M. Studdert-Kennedy (eds), *The Evolutionary Emergence of Language: Social function and the origins of linguistic form*, Cambridge University Press, Cambridge, England.

Tonkes, B., Blair, A. D. & Wiles, J. (2000). Evolving learnable languages, *in* S. A. Solla, T. K. Leen & K. R. Müller (eds), *Advances in Neural Information Processing Systems 12*, MIT Press, pp. 66–72.

Elman, J. L. (1991). Distributed representations, simple recurrent networks and grammatical structure, *Machine Learning* **7**: 195–224. The classical introduction to simple recurrent networks with applications for grammatical processing.

Steels, L. (1997). The synthetic modeling of language origins, *Evolution of Communication* **1**(1): 1–34. A taxonomy of problems in evolutionary linguistics to which computational approaches have been applied, and a review of the range of different computational approaches.

References

Batali, J. (1994). Innate biases and critical periods: Combining evolution and learning in the acquisition of syntax, *in* R. Brooks & P. Maes (eds), *Proceedings of the Fourth Artificial Life Workshop*, MIT Press, pp. 160–171.

Batali, J. (1998). Computational simulations of the emergence of grammar, *in* J. R. Hurford, C. Knight & M. Studdert-Kennedy (eds), *Approaches to the Evolution of Language*, Cambridge University Press, Cambridge, England, pp. 405–426.

Elman, J. L. (1990). Finding structure in time, *Cognitive Science* **14**: 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks and grammatical structure, *Machine Learning* **7**: 195–224.

Kirby, S. (1999a). *Function, Selection, and Innateness*, Oxford University Press.

Kirby, S. (1999b). Learning, bottlenecks and the evolution of recursive syntax, *in* E. J. Briscoe (ed.), *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, Cambridge University Press.

Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners, *in* C. Knight, J. R. Hurford & M. Studdert-Kennedy (eds), *The Evolutionary Emergence of Language: Social function and the origins of linguistic form*, Cambridge University Press, Cambridge, England.

Kolen, J. F. & Pollack, J. B. (1990). Back-propagation is sensitive to initial conditions, *Complex Systems* **4**(3): 269–280.

Oliphant, M. & Batali, J. (1996). Learning and the emergence of coordinated communication. Submitted.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning internal representations by error propagation, *in* D. E. Rumelhart & J. L. McClelland (eds), *Parallel Distributed Processing: Explorations in the microstructure of cognition*, MIT Press, chapter 8, pp. 318–361.

Tonkes, B., Blair, A. D. & Wiles, J. (1999). A paradox of neural encoders and decoders, or, why don't we talk backwards?, *in* B. McKay, X. Yao, C. S. Newton, J. H. Kim & T. Furuhashi (eds), *Simulated Evolution and Learning*, Vol. 1585 of *Lecture Notes in Artificial Intelligence*, Springer.

Tonkes, B., Blair, A. D. & Wiles, J. (2000). Evolving learnable languages, *in* S. A. Solla, T. K. Leen & K. R. Müller (eds), *Advances in Neural Information Processing Systems 12*, MIT Press, pp. 66–72.

Table 0.1: The utterances used by a neighbourhood of a population for a subset of the meaning space. This small sample shows two competing language forms. Where the first three agents use strings beginning with `B` for meanings with low numerical values, the other three agents use strings beginning with `D`. Note that agent 4 shows some similarities to both families. This example also demonstrates that even within one language 'family' there is significant variability.

Concept	Agent 1	Agent 2	Agent 3	Agent 4	Agent 5	Agent 6
0.00	BBBBBB	BBBBBB	BBBBBB	DDDDDD	DDDDDD	DDDDDD
0.01	BBBBBB	BBBB	BBBBBB	DDDDDD	DDDDDD	DDDDDD
0.02	BBBBBB	BBB	BBB	DDDDD	DDDDDD	DDDDDD
0.03	BBBBBB	BB	BBB	DDDD	DDDDD	DDDDDD
0.04	BBBBBB	BB	BB	DDDB	DDDD	DDDDDD
0.05	BBBBBB	BB	BB	DDD	DDD	DDDDDD
0.06	BBBBBB	B	BB	DDD	DDD	DDDDDD
0.07	BBBBBB	B	B	DDD	DD	DDDDDD
0.08	BBBB	B	B	DDB	DD	DDDDDD
0.09	BBB	B	B	DDB	DD	DDDDDD
0.10	BBB	B	B	DDB	DD	DDDDDD
0.11	BB	B	B	DD	DD	DDDDDD
0.12	BB	B	B	DD	D	DDD
0.13	BB	BDB	B	DD	D	DDD
0.14	BB	BDB	B	DD	D	DD
0.15	BDD	BDB	B	DD	D	DD
0.16	BDD	BDB	BDB	D	D	DD
0.17	BD	BDB	BDB	D	D	DD
0.18	BD	BD	BD	D	D	DD
0.19	BD	BD	BD	D	D	D
0.20	B	BD	BD	DB	D	D
0.21	B	BD	BD	DB	D	D