PREFINAL DRAFT: Kirby, S. and Christiansen, M. H. (2003). From language learning to language evolution. In Christiansen, M. and Kirby, S., editors, Language Evolution, pages 272–294. Oxford University Press.

From Language Learning to Language Evolution

Simon Kirby & Morten H. Christiansen

November 4, 2002

1 Introduction

There are an enormous number of communication systems in the natural world (Hauser, 1996). When a male Túngara frog produces "whines" and "chucks" to attract a female, when a mantis shrimp strikes the ground to warn off a competitor for territory, even when a bee is attracted to a particular flower, communication is taking place. Humans as prodigious communicators are not unusual in this respect. What makes human language stand out as unique (or at least very rare indeed, Oliphant, 2002) is the degree to which it is *learned*.

The frog's response to mating calls is determined by its genes, which have been tuned by natural selection. There is an inevitability to the use of this signal. Barring some kind of disaster in the development of the frog, we can predict its response from birth. If we had some machine for reading and translating its DNA, we could read-off its communication system from the frog genome. We cannot say the same of a human infant. The language, or languages, that an adult human will come to speak are not predestined in the same way. The particular sounds that a child will use to form words, the words themselves, the ways in which words will be modified and strung together to form utterances — none of this is written in the human genome.

Whereas frogs store their communication system in their genome, much of the details of human communication are stored in the environment. The information telling us the set of vowels we should use, the inventory of verb stems, the way to form the past tense, how to construct a relative-clause, and all the other facts that make up a human language must be acquired by observing the way in which others around us communicate. Of course this does not mean that human genes have no role to play in determining the structure of human communication. If we could read the genome of a human like we did with the frog, we would find that, rather than storing details of a communication system, our genes provide us with mechanisms to retrieve these details from the behaviour of others.

From a design point of view, it is easy to see the advantages of providing instructions for building mechanisms for language acquisition rather than the language itself. Human language cannot be completely innate because it would not fit in the genome. Worden (1995) has derived a speed-limit on evolution that allows us to estimate the maximum amount of information in the human genome that codes for the cognitive differences between us and chimpanzees. He gives a paltry figure of approximately 5 kilobytes. This is equivalent to the text of *just the introduction* to this chapter.

The implications of this aspect of human uniqueness are the subject of this chapter. In the next section we will look at the way in which language learning leads naturally to language variation, and what the constraints on this variation tell us about language acquisition. In section three, we introduce a computational model of sequential learning and show that the natural biases of this model mirror many of the human learner's biases, and help to explain the universal properties of all human languages.

If learning biases such as those arising from sequential learning are to explain the structure of language, we need to explore the mechanism that links properties of learning to properties of

what is being learned. In section four we look in more detail at this issue, and see how learning biases can lead to language universals by introducing a model of linguistic transmission called the *Iterated Learning Model*. We go on to show how this model can be used to understand some of the fundamental properties of human language syntax.

Finally, we look at the implications of our work for linguistic and evolutionary theory. Ultimately, we argue that linguistic structure arises from the interactions between learning, culture and evolution. If we are to understand the origins of human language, we must understand what happens when these three complex adaptive systems are brought together.

2 From universals to Universal Bias

When a behaviour is determined to a great extent genetically, we do not expect there to be a great deal of variation in that behaviour across members of a species. The whines and chucks a frog makes should not vary hugely from frog to frog within the same species. If there is variation in a behaviour, then this must either be down to genetic variability, or some interaction between the organism's phenotype and the environment that is itself variable.

Clearly the same does not hold for behaviours that are learned. In fact, we can see a learned behaviour as being one in which the environment interacts in a complex way with the phenotype. It is natural, therefore, that human language, involving as it does a large learned component, should exhibit a great deal of variation.

There are currently around 6000 human languages (although this number is dropping fast). Each is distinct in its phonology, lexicon, morphology and syntax. In fact, this severely underestimates cross-linguistic variation. It is likely that no two people have exactly the same linguistic system or *idiolect*. Does this mean that, for every child, the environment is

different? The answer is yes, because the linguistic environment is made up of the set of all utterances that that child hears, and this is unlikely to be the same for any two learners. This in itself makes human language very unusual even compared with other learned behaviours in nature. The environment that the learner interacts with consists of the "output" of that learning by other individuals. We will return to examine the significance of this later.

2.1 Constraints on variation

Linguistics is interested not only in documenting cross-linguistic variation, but also in uncovering the fundamental *limits* on that variation. There a number of obvious universal facts about languages. For example:

Digital infinity Languages make "infinite use of finite means" (Humboldt, 1836, cited in Chomsky, 1965, 8 – see also, Studdert-Kennedy & Goldstein, in Chapter 13). Despite having a fairly small inventory of sounds, language is open-ended, allowing us to produce an unlimited range of utterances.

Compositionality The meaning of an utterance is some function of the meanings of parts of that utterance and the way they are put together. For example, you can understand this sentence by virtue of understanding the meanings of its component words and phrases, and how they are put together. Although we can use non-compositional expressions (for example, the meaning of many idiomatic expressions must be stored holistically), all languages are compositional to a great extent.

Whilst these universals are immediately obvious facts about language, there are other constraints on variation that cannot be uncovered by looking at one or two languages. In order to discover more subtle patterns of variation, typologists categorise languages into a

taxonomy of types. For example, one type might divide the set of languages into those that use prepositions (such as English) and those that use postpositions (such as Hindi). What is interesting is that typically some combination of types are common, whereas others are rare or completely absent in a representative sample of the world's languages. For example:

Branching direction Dryer (1992) has shown that languages with a consistently branching word-order are more common than those with inconsistent branching. This means that languages whose direct object follows the verb are almost certain to have prepositions, for example.

There is a third sort of universal properties that are not as obvious as compositionality, for example, but are typically discovered without looking at a large sample of languages. By very detailed analysis of only a handful of languages, certain regularities become clear that can be used to predict the grammaticality of a variety of constructions. Whereas typology relies heavily on cross-linguistic comparison, evidence for these universals comes mainly from comparing patterns of grammaticality *within* a language. If a large range of seemingly unrelated grammaticality facts can be captured by a simple generalisation, and furthermore, if that generalisation can be applied to other unrelated languages, then this is taken as evidence that the generalisation captures a universal property of language. A classic example of this type of universal is:

Subjacency The subjacency condition limits the distance that an element in a sentence can be moved from its canonical position (though see the next section for a different perspective on this universal). "Movement" is used here to refer to the relationship between the position of *who* in "Mary loves *who*" and in "*Who* does Mary love", for example. The ungrammaticality of sentences such as "Who did Mary tell you when she

had met" is predicted by the subjacency condition. Interestingly, subjacency applies universally, although the precise details of the measurement of "distance" varies from language to language (see Kirby, 1999 for a more detailed discussion.)

2.2 Acquisition as explanation

It is helpful to think of the various aspects of a theory of human language in terms of the sorts of questions that they answer. A major goal of the sorts of cross-linguistic comparison and in-depth syntactic analysis summarised in the previous section is finding an answer to a *what* question:

> ***What*** *constitutes a possible human language?*

Having uncovered some of the features of language — the constraints on possible variation — a second question naturally follows:

> ***Why*** *are human languages the way they are, and not some other way?*

This is the fundamental explanatory challenge for the study of language.

As we argued earlier, it is the unique status of language as a learned communication system that leads to cross-linguistic variation. Given this, it seems sensible to look to learning to explain *why* that variation is constrained in the way that it is. From the viewpoint of Chomskyan generative linguistics (see, e.g., Chomsky, 1986), language universals are determined by the structure of our language learning mechanism, which is in turn determined by our genetic endowment.

In the Chomskyan approach, the characterisation of language universals (i.e., the description of linguistic variation) is replaced by the characterisation of set of constraints on language, termed *Universal Grammar* (UG). Jackendoff (2002) notes that the exact meaning

of UG is somewhat ambiguous in the literature. We use it here to refer to *the prior biases that the language learner brings to the task of learning language* — in some sense, the knowledge of the hypothesis space of the language learner before hearing a single utterance.

Generative linguistics suggests that a correct description of language universals and a correct description of Universal Grammar *are one and the same thing.* In other words, the **what** and the **why** questions can be rolled together. So, for example, if the Subjacency Principle captures linguistic variation correctly, then the Subjacency Principle forms part of UG. Thus, on this account, by careful examination of languages, one may uncover a property of the language-learning part of our brain — the human language acquisition device.

In itself, this is not quite enough to constitute a theory of language learning. UG gives us a model of a child's initial biases relative to language acquisition, but does not tell us how these interact with the child's experience of language to give her knowledge of Hungarian, for example. There are a number of different approaches to language learning in the generative program, but a shared characteristic of these approaches is that they downplay the complexity of the learning task. In fact, the term "learning" is usually avoided in favour of "acquisition". On this view, if all the child is doing is setting a few switches on their already-provided UG rules, then perhaps "learning" is too grand a term, suggesting that the process of picking up knowledge of one's native language is more open-ended than it is.

This may seem like we are merely tinkering with terminology, but in fact opting for "acquisition" over "learning" hides another key feature of the generative approach to explanation. If the constraints on cross-linguistic variation are due to constraints on learning, which in turn are due to our genetic endowment, we have not made any claims that these learning constraints are *specific to language.* After all, a particular genetic predisposition to learn in a particular way may well affect how we develop competency in a range of different domains.

But by removing *learning* from the process of language development, one implicitly allows linguistic constraints to become domain-specific. That is, if language is considered to be acquired rather than learned, the principles of UG can be kept pristine from the influence of other learning biases. From this perspective, it therefore makes sense to assume that linguists can uncover properties of our biology by examining languages alone.

In short, the generative approach contends that what makes a human language is determined by innately-coded domain-specific knowledge. In this chapter, we present a different perspective on language acquisition. Firstly, we will argue that some, perhaps many, of the learning biases we bring to language learning may not be domain-specific. From this perspective, it may be unhelpful to think of the prior knowledge of the child as Universal *Grammar*, with all the language-specific connotations of the term. Perhaps, Universal Bias is a better term. Either way, it should be clear that what we (and many other language researchers) are studying is the set of constraints and preferences that children have and which are brought to bear on the task of language learning.

After showing that we need to be extremely cautious in assuming that constraints on language learning are necessarily language-specific, we go on to show that innateness alone cannot be the determinant of linguistic structure. Our findings suggest that an explanatory theory of language must be *evolutionary* in the widest possible sense.

3 Sequential learning

The appeal to language-specific learning biases seems like a reasonable response to the challenge of explaining universals such as branching-direction and subjacency. How else can we explain universals that appear to be unique to language? A surprising alternative arises

when we look at a computational modelling work involving simple recurrent networks (SRN – Elman, 1990).

3.1 SRNs and language learning

An SRN is a type of neural network that can be used to learn sequences of items. Like many types of network, it consists of three layers of "nodes" wired together with weighted connections. One of these layers, the input layer, represents an element in a sequence that the network is currently attending to. Another layer, the output layer, represents the network's expectation of what the next element in the sequence will be. The weights on the connections between nodes encode the network's knowledge of sequences. The networks can be *trained* on a set of sequences which amounts to tuning the connection weights in such a way that the network is good at next-element prediction. Importantly, the SRN has a way of building up a representation of the structure of a sequence as it goes along from element to element. This representation acts like a memory, and enables the prediction of the next element in a sequence to depend not only on the current element, but also those that have gone before.

SRNs are interesting to us because they can be used to model many aspects of language learning (see Christiansen & Chater, 2001b, for a review) as well as sequential learning (Cleeremans, 1993). While language learning certainly involves more than simply predicting the next word in a sentence, infants do appear to acquire aspects of their language through such statistical learning (e.g., Saffran, Aslin, & Newport, 1996). Importantly, SRN behaviour after training closely resembles that of humans. In particular, the types of sequence that cause difficulty for the networks also seem to be ones that humans find hard. Christiansen & Chater (1999), for example, show that sequences generated by a grammar with multiple center-embeddings are hard for networks to learn correctly. The same sorts of structure are

difficult for us too: consider the difficulty in understanding *balls girls cats bite see fall*, which contains multiple clauses embedded inside each other. Crucially, the SRNs fit the human data without having to invoke external performance constraints as is the case with more traditional non-statistical models.

But what if the linguistic universals that have been said to require a language-specific UG fall naturally out of the prior biases of an SRN in a similar way?

3.2 SRNs and learning-based constraints

Christiansen and Devlin (1997) set out to explore this possibility. They look at the earlier mentioned branching-direction universal, which captures the fact that some basic word-orders are far more commonly attested in the world's languages than others. There have been a number of suggestions in the literature for how this might be accounted for by a domain-specific account of UG (see, for example Giorgi and Longobardi, 1991). Another influential approach explains word-order universals such as this one by appealing to a particular model of parsing (Hawkins, 1994; Kirby, 1999).

If, however, the pattern of cross-linguistic variation matches the prior bias of an SRN, then we have a strong argument for rejecting a domain-specific or parsing-based explanation. To test this, Christiansen and Devlin (1997) constructed 32 simple grammars each of which exhibited different word-orders, but were in all other aspects identical. The grammars, though simple, posed non-trivial tasks for a neural-network that is learning to predict sequences. For example, they were recursive, allowing sentences where adpositional-phrases were embedded within other adpositional-phrases. They included grammatical number marked on nouns, verbs, and genitive affixes, as well as a requirement that verbs agree with their subjects, and genitives agree with the possessing noun.

Each of the 32 grammars can be used to generate a corpus of sentences. In the Christiansen and Devlin (1997) model, these sentences are actually sequences of grammatical categories and an end-of-sentence marker. So, for example, a possible sequence in one of the grammars that mirrors English word order might be "singular noun + singular verb + preposition + singular noun + singular genitive + singular noun + end-of-sentence marker".[1] In order that these corpora can be used to train SRNs, each grammatical category is arbitrarily assigned one input node and one output node in the networks. So, if node 1 corresponds to a singular noun, and node 2 to a singular verb, then we should expect a network successfully trained on an English-like language to activate node 2 of the output layer after being shown the first word of the sequence above. Actually, things are a little more complicated than this, because a singular verb is not the only thing that can follow a sentence-initial singular noun. For example, a singular genitive affix is also possible. In this case, then, the network should show a *pattern of activation* that corresponds to all the grammatical continuations of a sentence that started with a singular noun in the language that it is trained on.

The ability of a network to correctly predict continuation probabilities after being trained on a corpus can be used to calculate a *learnability score* for a grammar. We can think of this score as reflecting how well each of the 32 grammars fits the prior bias of an SRN sequential learner. The surprising result is that these scores are significantly correlated with the degree of branching-consistency in the grammars used. Furthermore, a direct comparison with the cross-linguistic data (using Dryer's, 1992, statistically controlled database) shows that the prior-bias of an SRN is a good predictor of the number of languages that exhibit each word-order type. Simulations by Van Everbroeck (1999) and Lupyan & Christiansen (2002) involving SRNs learning grammatical role assignments from simple sentences incorporating noun/verb inflection and noun case-markings, respectively, provide further support for

explaining language-type frequencies using learning-based constraints.

This is clearly a striking result. Why postulate a domain-specific constraint if the data that constraint should account for is predicted by a general model of sequential learning? One potential criticism might be that the SRN is so different from the human brain, we can draw no parallels between its learning-bias and humans. Aside from the fact that the same argument could be applied to any model of UG, there are reasons to suspect that the human prior is not so far from the network's.

Christiansen (in preparation; described in Christiansen and Ellefson, 2002) employs an experimental paradigm known as artificial language learning (ALL) to demonstrate this. Two grammars were chosen from the set that the neural networks were trained on: one which exhibited a word-order that was rare cross-linguistically (and which the SRNs found hard to learn), and one which exhibited a more common pattern. The grammars generated "sentences" made up of strings of arbitrary symbols.

In the first phase of the experiment, subjects are exposed to strings generated from one of the grammars. They are asked to read and reproduce the strings, but were not told that the strings conformed to any kind of rule-set or grammar. During the test phase, however, they were told that the strings were part of a language. They were presented with novel strings and asked to say whether they were grammatical or not.

Christiansen found that subjects trained on the grammar with rare word-order were significantly worse at determining grammaticality than those trained on the more common word-order type. This was despite the fact that the word-order of the latter grammar did not have anything in common with English word-order (in fact, the difficult grammar had more features in common with English).

The ALL experiment suggests that the SRN's sequential learning biases mirror those of

humans, and that this non-language-specific bias underpins at least some of the constraints on cross-linguistic variation. Further evidence for this comes from experiments by Christiansen, Kelly, Shillcock, and Greenfield (in preparation) that look at the ability of subjects with agrammatic aphasia to learn sequences in ALL experiments. Agrammatic aphasics are characterised by a reduced ability to deal with grammatical constructions in language. There have also been suggestions (Grossman, 1980) that they have a more general deficit in their ability to deal with sequential learning tasks (see also Lieberman, in Chapter 14). In the ALL experiments, normal subjects were able to determine the grammaticality of strings generated by a simple finite-state grammar whereas the aphasic subjects could not perform above chance. Once again, this provides further evidence for a strong connection between language learning and sequential learning.

What these experiments with ALL and SRNs show us is that we should be careful about ascribing universal properties of language to a domain-specific innate bias. We argue that an explanation that appeals to non-linguistic biases should be preferred where possible. Simplistic all-or-nothing explanations should be avoided, however. Whereas the above results suggests that some language universals may drive from non-linguistic sequential learning biases, others may require language-specific biases for their explanation. We submit, however, that this is an empirical question that cannot be settled a priori.

We end this section with a brief look at another language universal that has been claimed to reflect a domain-specific UG principle: subjacency. Christiansen and Ellefson (2002) used a similar paradigm of SRN and ALL experiments to see if the patterns of grammaticality that are predicted by the subjacency principle actually reflect biases from sequential learning. In particular, they wanted to see if a language with the grammaticality pattern of English would be easier for an SRN to learn than a language with an unnatural grammaticality pattern (i.e.,

one in which subjacency-violating constructions were permitted).

The following types of strings make up the English-like language (English glosses are given for clarity — the networks were trained on category names, and human subjects were presented with arbitrary symbols, as with the word-order experiment):

(1) Everybody likes cats.

(2) Who (did) Sara like?

(3) Sara heard (the) news that everybody likes cats.

(4) Sara asked why everybody likes cats.

(5) What (did) Sara hear that everybody likes?

(6) Who (did) Sara ask why everybody likes cats?

The unnatural language looks like this (note the last two examples, which are ungrammatical in English since they violate the subjacency principle):

(7) Everybody likes cats.

(8) Who (did) Sara like?

(9) Sara heard (the) news that everybody likes cats.

(10) Sara asked why everybody likes cats.

(11) *What (did) Sara hear (the) news that everybody likes?

(12) *What (did) Sara ask why everybody likes?

As with the SRN simulations, and ALL experiments for word-order, human subjects and neural networks behaved the same way. The unnatural language was significantly harder for subjects to learn, and resulted in significantly greater network error, than the English-like language.

These two string sets obviously do not capture all the possible distinctions between a subjacency-obeying language and a hypothetical subjacency-violating language. There are many more predictions about grammaticality that follow from the principle. Nevertheless, the learning biases reflected in this language universal is not likely to be completely domain-specific.

4 Iterated learning and the origins of structure

In the previous section we examined one approach to answering the question: "why are languages the way they are". We argued that the range of cross-linguistic variation is ultimately determined by our innately given learning biases, many of which may not be specific to language (see figure 1).

There is something missing from this picture, however. We have looked at the way in which particular types of learning respond to particular sets of data. We have not said anything about where this data comes from. In most models of learning, the data is considered to be given by some problem domain. In Christiansen's simulations, for example, the training data is provided by the experimenter.[2] It is the ability of the networks or experimental subjects to learn the data, and the limitations they exhibit, that is of interest.

4.1 The ILM

Kirby (2000) has suggested that we need to look more carefully at where the training data comes from if we want a truly explanatory account of the structure of language. What makes language unique in this regard is that the data that make up the input to learning *is itself the output of that same process* (see figure 2). This observation has lead to the development of a model of language evolution — the *Iterated Learning Model* (ILM) — that builds this in directly (Kirby and Hurford, 2002).

The ILM is a multi-agent model that falls within the general framework of situated cognition (Brighton, Kirby, and Smith, in press). It treats populations as consisting of sets of individuals (agents), each of which learns its behaviour by observing the behaviour of others (and concequently contributes to the experience of other agents' learning). We can contrast this approach with the idealisations of the homogenous speech community of Chomsky (1965).

The ILM is evolutionary in the sense that the linguistic behaviour of agents is *dynamic*. It is not predetermined, but emerges from the process of repeated use and acquisition from generation to generation. This is not biological evolution taking place over the lifetime of a species, rather the process of linguistic transmission operating on a historical/cultural timescale.

Iterated learning is discussed in depth in Kirby (1999), where it is used to demonstrate how biases (such as the bias for consistent branching direction) actually lead to language universals, and why some classes of universal (specifically, implicational or hierarchical ones) can only be explained from the viewpoint of iterated learning (for possible psychological and connectionist underpinnings of this view, see Christiansen, 1994). In the following section, we will summarise recent research that suggests some of the more fundamental properties of language arise from the process of repeated acquisition and use.

4.2 The origins of compositionality

A striking property of human language that sets it apart from most other communication systems is the way in which the meaning of an utterance is composed of the meanings of parts of that utterance.[3] This *compositional* nature of language is so fundamental to the syntax of human language, it is rarely considered a target of explanation. Yet, it is compositionality that gives language its open-ended expressivity (consider how difficult it would be if you had to give every communicatively relevant situation a unique, atomic name).

Whilst compositionality is rife in language, it is not everywhere equally. For example, idiomatic expressions are not as clearly compositional. The sentence "John bought the farm" can be read compositionaly as being information about someone called John buying a farm. In some dialects of English, however, it can be taken to mean that John has died. This second meaning of "bought the farm" is holistic, rather than compositional.

Similarly, if we look at the morphology of a lot of languages, we can see both holism and compositionality. In the English past tense, there are regular verbs, such as *walk/walked*, and irregulars, such as *go/went*. The former express their past tense compositionally through the addition of the affix *-ed*, whereas the latter are holistic.

Where did compositionality come from and what maintains it in language? If language was once holistic (as Wray, 1998, suggests), by what mechanism did this change? Kirby (2000) suggests that the answer to these questions lies in understanding the process of iterated learning.

To test this, Kirby and others (Batali, 1998; Batali, 2002; Brighton and Kirby, 2001; Brighton, 2002; Kirby and Hurford, 2002; Kirby, Smith, and Brighton, 2002; Tonkes, 2002) have implemented the ILM as a computational simulation. Modelling situated cognition, and

dynamical systems in nature more generally, using computational techniques is an increasingly popular methodology (see, e.g., Kirby, 2002, for a review of the ways in which it has been used in evolutionary linguistics). Computer modelling gives us an easy way to uncover the relationship between the components of a complex system — in this case individual learners — and the emergent outcomes of their interactions.

The simulation models are typically made up of these components:

1. A population of agents.

2. A space of possible signals (usually strings of symbols).

3. A space of possible meanings (usually some kind of structured representation, such that some meanings are more similar to each other than others).

4. A production model. This determines how, when prompted with a meaning, an agent uses its knowledge of language to produce a signal.

5. A learning model. The learning model defines how an individual agent acquires its knowledge of language from observing meaning-signal pairs produced by other agents.[4]

Within this broad framework, there is a lot of variation in the literature. In particular, a range of different learning models have been employed; from symbolic approaches such as grammar induction (Kirby and Hurford, 2002) to connectionist approaches such as SRNs (Batali, 1998). A key point in common between the various ILM simulations is that the language of the population persists only by virtue of its constant use and acquisition by individual agents.

These simulations are usually initialised with a random language. That is, the agents initially produce purely random strings of symbols for every meaning that they are prompted

with. Excepting some highly improbable set of random choices, this language will be purely holistic. That is, there will be nothing in the structure of the strings of symbols that corresponds to the structure of the meanings being conveyed.

In the simulations, these intial random languages are typically highly *unstable*. A snapshot of the language of the population at one point in time will tend to tell you very little about what the language will be like at a later point. It is easy to see why. Unless a learning agent is exposed to the output of a particular speaker *for every possible meaning*, there is no way in which that learner will be able to reproduce accurately the language of that speaker.

In some simulations, holistic languages can be made to be stable, but only if the learners hear so much data that they are guaranteed to observe a speaker's utterance for every possible meaning. However, this is a highly unrealistic assumption. Much is made in the language acquisition literature of the "poverty of the stimulus" (see Pullum and Scholz, 2002, for a review), suggesting that learners are exposed to an impoverished subset of the total language. Aside from this, in reality the range of possible meanings is open-ended, so complete coverage cannot be assumed.

In the initial phases of the simulations, then, the language of the population tends to vary widely and change rapidly. The striking thing is that eventually some part of the language will stabilise, being passed-on faithfully from generation to generation. How long the simulations run before this happens depends on a variety of factors relating to the particular simulation's population dynamics (we return to this briefly in the next section). Nevertheless, for a broad range of different simulation models with different learning models and so on, these pockets of stability always seem to occur.

As the simulations continue, more and more of the language increases in stability until eventually signals corresponding to the entire meaning-space are passed on reliably from

generation to generation without the learner being exposed to the whole language. These final languages invariably use a compositional system to map meanings to strings (and *vice versa*).

How are these languages stable, and why are they compositional? The initial holistic languages are unstable by virtue of the poverty of the stimulus, which acts as a *bottleneck* on the transmission of language. In the early stages of the simulation, the population is essentially randomly searching around the space of possible meaning-string pairs, driven by the learners' failure to generalise (non-randomly) to unheard meanings. At some point, a learner will infer some form of non-random behaviour in a speaker and use this to generalise to other meanings. In the first instance, this inference of "rule-like" behaviour will actually be ill-founded (since the speaker would have been behaving randomly). Nevertheless, this learner will now produce utterances that are, to a small extent, non-random.

Languages that are generated by a learner who has generalised are themselves generalisable by other learners. The key point is that the aspects of the language that are generalisable in this way are more stable. This is because, by definition, generalisations can be recreated each generation without the need for complete coverage in the training sample. In other words, because a generalisation can be used for a range of meanings, and can be learnt by exposure to a subset of those meanings, generalisations can pass through the learning-bottleneck more easily. As Hurford (2000) puts it: "social transmission favours linguistic generalisation".

As a result of this differential in stability between random and non-random parts of the linguistic system, the movement towards a language that is made up of generalisations is inevitable. The meanings in the model have internal structure, as do the signals. What is learnt is the mapping between these two spaces. Generalisations about this mapping have to utilise the fact that the spaces are structured. It is unsurprising, therefore, that

a compositional system of mapping is inevitable. In a compositional language, the internal structure of a meaning to be conveyed is non-randomly related to the structure of the string that conveys that meaning.[5]

These results link compositional structure in language to the impoverished stimulus that a learner faces. Poverty of the stimulus arguments are usually used to suggest the need for a strongly constraining prior provided "in-advance" by our biology. For us, the restricted richness of input is actually the engine that drives the evolution of language itself. Language adapts to aid its own survival through the differential stability of generalisations that can be transmitted through the learning bottleneck.

Further evidence that this type of evolution through iterated learning explains some aspects of language structure is given by Kirby (2001). Returning to the fact that language is *not* 100% compositional, Kirby wonders whether the ILM might explain not only why compositionality emerges, but also why some kinds of holism persist.

In this simulation, the frequency with which particular meanings are used by the agents is not uniform. In other words, some meanings turn up more frequently than others (in contrast to earlier work where every meaning has an equal chance of making it through the bottleneck). This is clearly a more realistic assumption, mirroring the fact that the real world is "clumpy" — that some things are common, others rare. With this modification in place, the result is a language that utilises both compositional structure, and holistic expressions.

The simple language in the simulation can be compared to a morphological paradigm. The infrequently used parts of the paradigm tend to be compositionally structured, whereas the more frequent parts are holistic.[6] The simulation appears to be an appropriate model for language since there is a clear correlation between frequency and irregularity in morphology (the top-ten verbs in English by frequency are all irregular in the past-tense).

Once again, we can understand these results in terms of the pressure on language to be transmitted through a learning bottleneck. Frequently used expressions may be faithfully transmitted even if they are idiosyncratic simply by virtue of the preponderance of evidence for them in the data provided to the child (or to the learning agents in the simulation). Infrequent expressions, on the other hand, cannot be transmitted in this way and must instead form part of a larger paradigm that ensures their survival even if they do not form part of the linguistic data observed by the learner. This also explains why languages vary in the degree to which they admit irregulars. A completely regular paradigm is perfectly stable, so there is no reason to expect that language must have irregulars (for example, Chinese is noted for the rarity of irregularity). On the other hand, if for reasons such as language-contact, or processes of phonological erosion, irregulars make their way into a language, the pressure to regularise them will be strongest in the low-frequency parts of the system.

5 Implications and conclusion

We have argued that human languages vary because they are learned. The key to understanding the constraints on this variation (and hence the universal structural properties of language) lies in a characterisation of the properties of the language learner. The generativist approach to language equates language universals with domain-specific, innate constraints on the learner — essentially drawing a direct relationship between universals and UG.

We agree that it is important to give an account of the initial state of the language learner: the Universal Biases. However, it is premature to assume that such an account necessarily has to be domain-specific. Results from artificial language learning experiments and neural-network models show that some universals can be explained in terms of biases relating to

sequential learning in general.

The results from the Iterated Learning Model have further implications for an explanatory account of linguistic structure. Typically, the relationship between prior (i.e., innate) bias and linguistic variation is assumed to be transparent and one-to-one. That is, the "problem of linkage" (Kirby, 1999) is left unsolved. Iterated learning is an *evolutionary* approach to the link between bias and structure, in that it looks at the dynamical system that arises from the transmission of information over time. It no longer makes sense to talk about language structure as being purely innately coded. Instead, learning bias is only one of the forces that influence the evolutionary trajectory of language as it is passed from generation to generation. Other factors include the number of utterances the learner hears, the structure of the environment, social networks, population dynamics, and so on.[7] If we are right, we cannot infer the nature of the learning biases purely through linguistic analysis. In some sense, the learning biases act like the environment within which languages themseleves adapt (Christiansen, 1994), and without understanding the process of adaptation we cannot be sure any particular theory of language acquisition makes the correct predictions.

So far in this chapter we have said little about the biological evolution of the language learner, instead focussing on the evolution of language itself as it passes from learner to learner. Pinker and Bloom (1990), in their classic paper on language evolution, take a complementary stance by arguing that evolution by natural selection is the key explanatory mechanism (see also Pinker, in Chapter 2). The logic of their argument runs as follows:

1. The principle features of language appear to be adapted to the task of communication

2. These features arise from a domain-specific, innate language faculty.

3. The only known mechanism that can deliver a biological faculty with the appearance

of design is evolution by natural selection.

Obviously, if our conclusions are right, we cannot use this logic for at least some of the properties of language since they are neither innate nor domain-specific. The question of whether the language faculty is well adapted to communication would be the subject of another chapter, but Kirby (1999) argues that in fact *dis*functionality may be the hallmark of innate aspects of language.

Does biological evolution have any role to play in an explanatory account of language structure? Despite our reservations about the adaptationist approach of Pinker and Bloom, we agree it must (how else can we understand human uniqueness?). To our **what** and **why** questions, we add a third:

***How** did language come to exhibit the structure that it does?*

It makes little sense to mount an explanation in terms of natural selection for the aspects of learning that are not language specific. We could, however, try to uncover evolutionary pressures that shaped more general biases such as those that underpin sequential learning (see e.g., Conway & Christiansen, 2001). Much structure in biology is the result of *exaptation* rather than *adaptation* (although a combination of both forces is likely to be common). Perhaps the mechanisms for sequential learning evolved earlier in our species history and more recently were employed to assist language learning (see Lieberman, in Chapter 14). Perhaps the relevant biases are *spandrels* arising as a bi-product of other aspects of our biology. Currently, there is no way of telling.

Another possible line of enquiry might be to look at what the evolutionary prerequisites for iterated learning itself might be. The simulation models mentioned in the previous section take a lot of things for granted. There are three principle assumptions:

1. Agents have structured representations of the world.

2. Learners have some way of infering the meaning of a particular signal. At least some of the time, they can mind-read.

3. Speakers are inclined to communicate about an open-ended range of topics.

By systematically varying the representations of meanings the agents in the ILM have access to, we are able to see under which circumstances structured mappings between meanings and signals will emerge. Brighton (2002) shows, using mathematical models, that compositional languages have a stability advantage over holistic ones when agents have the capacity to form representations of the environment that are multi-dimensional. In other words, when they are highly tuned to structural similarities between different environmental states. The question of how learners have access to these intended meanings (at least some of the time) is tackled by researchers working on negociation of communication that is grounded in the environment (see, e.g., Cangelosi, 1999). For example, there is a growing body of research exploring learned communication between robots (Steels and Kaplan, 2002) that may provide answers to these questions. Finally, we need to understand why humans communicate to such a degree. Lessons may be learned here from the ethology literature, research on other, closely related, primates, and work looking more carefully at the selective benefits of human language (see, e.g., Dunbar, in Chapter 12).

It is also possible that learning mechanisms may undergo biological adaptation after they have been employed for a particular task. Briscoe (Chapter 16) looks at this possibility in some detail (though from the viewpoint of a language specific UG). We can consider an extension to the iterated learning model that also allows for genetic specification of different learning biases which may lead to co-evolution of language and language learners (see figure

3). This provides a potential framework for understanding the evolution of language learning mechanisms that may include both domain-general and domain-specific components.

A complete theory of language evolution will necessarily be multi-faceted. We should not expect a single mechanism to do all the work. Ontogenetic development, cultural transmission and evolution of prior biases may all conspire to give rise to human language. As we begin to understand the central role of learning in these three interacting complex dynamical systems, we will get closer to a fundamentally evolutionary understanding of linguistic structure.

Further readings

In this chapter, we have shown how a number of different computational techniques can be used to explore the concept of innate learning biases. This approach to language evolution is rapidly gaining in popularity as complex simulations become more practical. A review of this literature, looking in particular at so-called "artificial life" models is Kirby (2002). Cangelosi and Parisi (2002) and Briscoe (2002) are two collections of recent research into this area.

For some of the work we describe, simple recurrent neural networks act as models of learning. See McLeod, Plunkett & Rolls (1998) for a hands-on introduction to using neural networks – including the SRN – to model cognitive behavior. Cleeremans (1993) present SRN simulations of sequential learning. Christiansen & Chater (2001a) provides a comprehensive introduction to connectionist modelling of many different kinds of psycholinguistic phenomena.

We use the term "bias" throughout this chapter to describe the preferences a learner has before the linguistic experience of that learner is taken into account. A technical discussion about bias from the perspective of statistics and the relationship between bias and learning mechanisms can be found in Mitchell (1997). This book is also a excellent textbook on various approaches to modelling learning more generally.

We talk about both "language universals" and "learning biases" when discussing what evolutionary linguistics needs to explain. The former term is used is typically used in the field of linguistic typology (see, e.g., Croft, 1990), whilst the latter within generative approaches are associated with the notion of a universal grammar (see, Jackendoff, 2002, for an accessible introduction). Newmeyer (1998) reviews a fundamental divide between "functional" and "formal" approaches in linguistics that is commonly associated with the typological and

generative perspectives.

Notes

[1]To see that this is like English word-order, consider a sentence like "the man sat on the cat's mat". It should be noted that one of the simplifications in the grammars used is that determiners are ignored.

[2]Though see Christiansen & Dale (in press) for simulations in which the learning biases of the SRN forces languages to change over generations of learners to become more easily learnable.

[3]We note, however, that during normal language comprehension each word in a sentence may not contribute to its overall meaning – sentence interpretation appears to be at least partially underspecified. For a review of the relevant literature from psycholinguistics and computational linguistics, see Sanford & Sturt (2002).

[4]The fact that the learners are given meanings as well as signals seems unrealistic. Ultimately, simulations of the process of iterated learning will need to enrich the model with *contexts*. Whereas meanings are private and inaccessible to learners, contexts are public and may allow the inference of meanings. See, e.g., (Steels, Kaplan, McIntyre, and Van Looveren, 2002; Smith, 2001) for discussion of these fascinating extensions to the model.

[5]The stability of particular representations in an iterated learning scenario is related to their compressibility (see Brighton, 2002; Teal and Taylor, 1999 for discussion).

[6]An interesting by-product of the introduction of frequency biases to the meaning space is the removal of a fixed end-point to the simulations. The language in this model is always changing - but not so much that speaker-to-speaker intelligibility is degraded. This is another way in which these simulation results seem to mirror what we know about language more

accurately. Further research on the dynamics of language change in these models is needed. For example, see Niyogi and Berwick (1997) and Briscoe (2000) for discussion of iterated learning and the oft-noted logistic time-course of change.

[7]We do not describe them here for lack of space, but a comparison of different ILM simulations (e.g., Batali, 1998 and review in Hurford, 2002) reveals the importance of *vertical* versus *horizontal* transmission in the population. When learners mainly learn from adults, the language changes relatively slowly and may take many generations to stabilise on a structured system. If, on the other hand, there is a lot of contact between learners, structure can emerge very rapidly. It has been suggested that modelling work may provide insights into the very rapid emergence of languages like Nicaraguan Sign Language (Ragir, 2002).

References

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*, pp. 405–426. Cambridge: Cambridge University Press.

Batali, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In E. Briscoe (Ed.), *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, pp. 111–172. Cambridge: Cambridge University Press.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life 8*(1), 25–54.

Brighton, H. and S. Kirby (2001). The survival of the smallest: Stability conditions for the cultural evolution of compositional language. In J. Kelemen and P. Sosik (Eds.), *Advances in Artificial Life (Proceedings of the 6th European Conference on Artificial Life)*. Heidelberg: Springer-Verlag.

Brighton, H., S. Kirby, and K. Smith (in press). Situated cognition and the role of multi-agent models in explaining language structure. In D. Kudenko, E. Alonso, and D. Kazakov (Eds.), *Adaptive Agents*. Springer.

Briscoe, E. (2000). Evolutionary perspectives on diachronic syntax. In S. Pintzuk, G. Tsoulas, and A. Warner (Eds.), *Diachronic Syntax: Models and Mechanisms*. Oxford: Oxford University Press.

Briscoe, E. (Ed.) (2002). *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge: Cambridge University Press.

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*, Number 1674 in Lecture notes in computer science. Springer.

Cangelosi, A. and D. Parisi (Eds.) (2002). *Simulating the Evolution of Language*. Springer.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1986). *Knowledge of Language*. Praeger.

Christiansen, M.H. (1994). *Infinite Languages, Finite Minds: Connectionism, Learning and Linguistic Structure*. Ph. D. thesis, University of Edinburgh.

Christiansen, M.H. (in preparation). Cognitive constraints on word order universals: Evidence from connectionist modeling and artificial grammar learning. Manuscript in preparation.

Christiansen, M.H. & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. Cognitive Science, 23, 157-205.

Christiansen, M.H. & Chater, N. (Eds.) (2001a). Connectionist psycholinguistics. Westport, CT: Ablex.

Christiansen, M.H. & Chater, N. (2001b). Connectionist psycholinguistics in perspective. In M.H. Christiansen & N. Chater (Eds.), Connectionist psycholinguistics (pp.19-75). Westport, CT: Ablex.

Christiansen, M.H. & Dale, R. (in press). The role of learning and development in the evolution of language: A connectionist perspective. In D. Kimbrough Oller and U. Griebel (Eds.), The evolution of communication systems: A comparative approach. The Vienna Series in Theoretical Biology. Cambridge, MA: MIT Press.

Christiansen, M.H. and J. Devlin (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*, pp. 113–118. Lawrence Erlbaum Associates.

Christiansen, M.H. and M. Ellefson (2002). *Linguistic Adaptation without Linguistic Constraints: The role of sequential learning in language evolution*, Chapter 16, pp. 335–358. Oxford University Press.

Christiansen, M.H., L. Kelly, R. Shillcock, and K. Greenfield (in preparation). Artificial grammar learning in agrammatism.

Cleeremans, A. (1993). *Mechanisms of implicit learning: A connectionist model of sequence processing.* Cambridge, MA: MIT Press.

Conway, C.M., & Christiansen, M.H. (2001). Sequential learning in non-human primates. Trends in Cognitive Sciences, 5, 539-546.

Croft, W. (1990). *Typology and Universals.* Cambridge: Cambridge University Press.

Dryer, M. (1992). The Greenbergian word order correlations. *Language 68*, 81–138.

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

Giorgi, A. and G. Longobardi (1991). *The syntax of noun phrases: configuration, parameters and empty categories.* Cambridge University Press.

Grossman, M. (1980). A central processor for hierarchically structured material: evidence from Broca's aphasia. *Neuropsychologia 18*, 299–308.

Hauser, M. D. (1996). *The Evolution of Communication.* Cambridge, MA: MIT Press.

Hawkins, J. A. (1994). *A performance theory of order and constituency.* Cambridge University Press.

Humboldt, W. v. (1836). *Über die Verschiedenheit des Menschlichen Sprachbaues*. Berlin.

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

Hurford, J. R. (2002). Expression/induction models of language evolution: dimensions and issues. In E. Briscoe (Ed.), *Linguistic Evolution Through Language Acquisition: Formal and Computational Models*. Cambridge: Cambridge University Press.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press.

Kirby, S. (1999). *Function, selection and innateness: the emergence of language universals*. Oxford: Oxford University Press.

Kirby, S. (2000). Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners. In C. Knight, M. Studdert-Kennedy, and J. R. Hurford (Eds.), *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, pp. 303–323. Cambridge: Cambridge University 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(2)*, 102–110.

Kirby, S. (2002). Natural language from artificial life. *Artificial Life 8*, 185–215.

Kirby, S. and J. R. Hurford (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. Springer Verlag.

Kirby, S., K. Smith, and H. Brighton (2002). Language evolves to aid its own survival. In preparation.

Lupyan, G.& Christiansen, M.H. (2002). Case, word order, and language learnability: Insights from connectionist modeling. In Proceedings of the 24th Annual Conference of the Cognitive Science Society (pp. 596-601). Mahwah, NJ: Lawrence Erlbaum.

McLeod, P., Plunkett, K. & Rolls, E.T. (1998). Introduction to connectionist modelling of cognitive processes. Oxford: Oxford University Press.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.

Newmeyer, F. J. (1998). *Language Form and Language Function*. Cambridge, MA: MIT Press.

Niyogi, P. and R. Berwick (1997). A dynamical systems model of language change. *Linguistics and Philosophy 17*.

Oliphant, M. (2002). Rethinking the language bottleneck: Why don't animals learn to communicate? In K. Dautenhahn and C. L. Nehaniv (Eds.), *Imitation in Animals and Artifacts*, Complex Adaptive Systems, pp. 311–325. MIT Press.

Pinker, S. and P. Bloom (1990). Natural language and natural selection. *Behavioral and Brain Sciences 13*, 707–784.

Pullum, G. K. and B. C. Scholz (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review 19*(1-2).

Ragir, S. (2002). Constraints on communities with indigenous sign languages: clues to the dynamics of language origins. In A. Wray (Ed.), *The Transition to Language*. Oxford: Oxford University Press.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926-1928.

Sanford, A.J. & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. trends in Cognitive Sciences, 6, 382-386.

Smith, A. D. M. (2001). Establishing communication systems without explicit meaning transmission. In J. Kelemen and P. Sosík (Eds.), *Advances in Artificial Life: Proceedings of the* 6^{th} *European Conference on Artificial Life*, Number 2159 in Lecture Notes in Artificial Intelligence, pp. 381–390. Heidelberg: Springer-Verlag.

Steels, L. and F. Kaplan (2002). Bootstrapping grounded word semantics. In E. Briscoe (Ed.), *Linguistic Evolution through Language Acquisition: formal and computational models*, pp. 53–73.

Steels, L., F. Kaplan, A. McIntyre, and J. Van Looveren (2002). Crucial factors in the origins of word-meaning. In A. Wray (Ed.), *The Transition to Language*. Oxford, UK: Oxford University Press.

Teal, T. and C. Taylor (1999). Compression and adaptation. In D. Floreano, J. D. Nicoud, and F. Mondada (Eds.), *Advances in Artificial Life*, Number 1674 in Lecture Notes in Computer Science. Springer.

Tonkes, B. (2002). *On the Origins of Linguistic Structure: Computational models of the evolution of language*. Ph. D. thesis, University of Queensland.

Van Everbroeck E (1999). Language type frequency and learnability: A connectionist appraisal. In: Proceedings of the 21st Annual Cognitive Science Society Conference, pp 755-760. Mahwah, NJ: Lawrence Erlbaum.

Worden, R. (1995). A speed limit for evolution. *Journal of Theoretical Biology 176*, 137–152.

Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language and Communication 18*, 47–67.

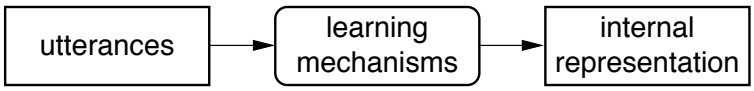

Figure 1: Language learning. The child's learning mechanisms take linguistic data (utterances) and generates some kind of internal representation of a language. The range of possible languages are determined by the structure of the learning mechanisms (i.e., its prior biases).

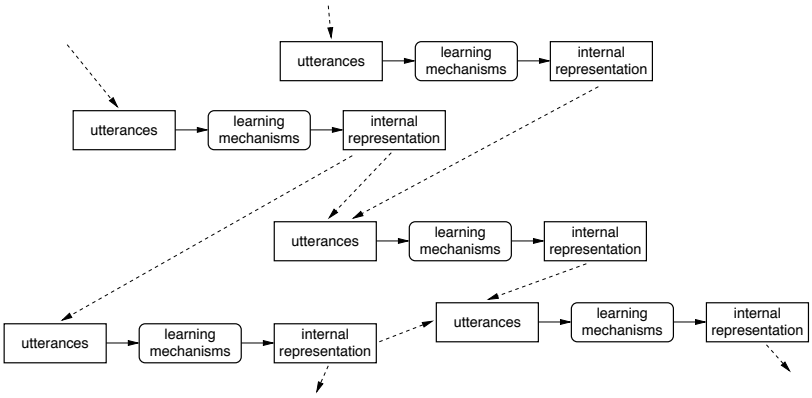

Figure 2: Iterated learning. The input to learning is the product of the acquired language of others. There are now two dynamical systems that contribute to the range of possible languages: individual learning, and social/cultural transmission

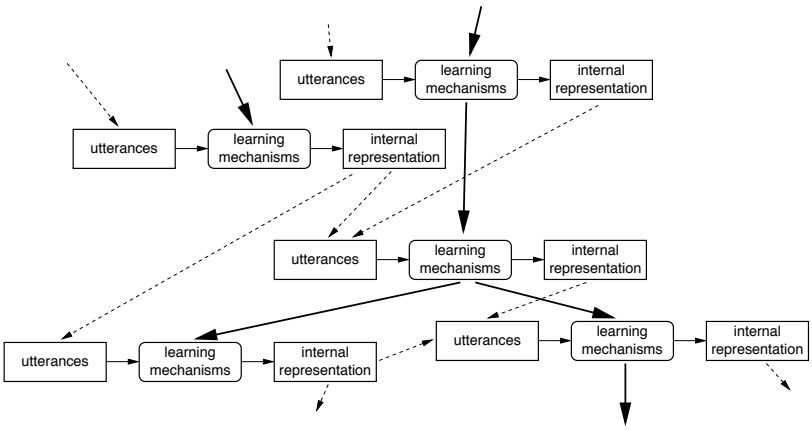

Figure 3: Evolutionary iterated learning. The cognitive mechanisms (and therefore prior biases) of an individual language learner are provided through genetic transmission, which necessarily involves selection. The structure of language arises from the interaction of three systems of information transmission.