PREFINAL DRAFT: Kirby, S. and Hurford, J. (1997). Learning, culture and evolution in the origin of linguistic constraints. In Husbands, P. and Harvey, I., editors, ECAL97, pages 493-502. MIT Press.

Learning, Culture and Evolution in the Origin of Linguistic Constraints

Simon Kirby & James Hurford
Department of Linguistics
University of Edinburgh
Edinburgh
Scotland
simon@ling.ed.ac.uk

1 Introduction

One of the interesting challenges facing linguistics today is the explanation of the observed constraints on cross-linguistic variation.[1] Traditional linguistic typology (e.g. [14, 9, 16, 10]) as well as generative theories of language acquisition (e.g. [5, 15]) highlight the fact that the languages of the world appear to fall into a narrowly defined region of the space of logically possible languages. A language typology is a categorisation of some interesting subset of the dimensions along which languages can vary, and language universals are logical statements that relate orthogonal dimensions of such a typology. Although there is a lively debate about what these universal constraints on variation actually are, and what constitutes evidence for them [31], the greatest area of disagreement is clearly how to go about explaining the origins of these constraints [28]. The questions that this conflict of explanation gives rise to go to the heart of modern linguistics.

In this paper we will be examining one aspect of linguistic constraints: the appearance of design. Many attempts at explaining universals have pointed out their fit to the functions of language. Hawkins [17, 18] for example, attempts to explain a whole range of universals relating to word order in terms of the processing load on the human parser. Although this kind of research (known as functionalist explanations in linguistics) is important, we believe they leave the real problem unanswered – how exactly do functional pressures end up being expressed as cross-linguistic constraints on variation? Another influential strand of research (known as formal or innatist linguistics) treats language universals as the direct consequence of the structure of a domain specific language acquisition device (LAD) [5]. Although this bypasses the problem of how the constraints emerge, it fails to explain why the constraints appear to be designed for the purpose of making language easier to parse, for example.

[1]The authors would like to thank Ian Hodson, Bill Turkel, and Rob Clark for their helpful comments and assistance. Of course, they may well disagree with the contents of this paper. All correspondence should be addressed to the first author. This research was supported by ESRC grant R000236551.

In this paper we will argue that the obvious solution to this problem, namely that the LAD has evolved through natural selection to constrain languages to be functional, cannot work. This is true even though just such a constraining LAD would eventually led to a fitter population. Instead, we will show using a computational simulation of evolving and communicating language learners that a *linguistic* selection process means that languages themselves adapt over a historical (cultural) timescale. Surprisingly, this historical adaptation can *enable* or *bootstrap* the evolution of a functional LAD after all. What we are left with is a simple unified explanation of both innate *and* historically emergent constraints on cross-linguistic variation.

2 Phylogenetic functionalism

The Chomskyan LAD is assumed to alleviate the problems of learnability in natural language by severely constraining the search space for the language learner. The Principles and Parameters model, for example, can be thought of as a set of universal absolute constraints on linguistic variation (principles) and a set of finitely variable switches (parameters) that the learner varies in response to input data. The crucial input data to the learner in this view is *trigger experience* in that it triggers the setting of a particular parameter. The exact algorithm that governs parameter setting will be discussed below.

This theory of language acquisition is innatist since the learner is assumed to come equipped with knowledge about the target system at birth — specifically, the principles and the existence of parameters (although not the particular setting of those parameters). This means that language universals are directly explicable in terms of the constraints built into language learners from birth. If these universals look like they are designed to alleviate parsing pressures as Hawkins argues, then we are lead to the conclusion that the innate LAD must somehow be set up in such a way as to be *functional* or *adaptive* for the users of language. More precisely, the LAD in this view is set up in such a way that it constrains humans

from acquiring languages that are dysfunctional in some way, for example by being hard to parse.

The remaining question is how the LAD came to be endowed with these functional constraints. Newmeyer [30] argues that the obvious answer is that the LAD has evolved through a process of natural selection. Fitter individuals are presumably those that are able to receive and transmit linguistic signals most efficiently, and hence it is unsurprising that LADs that lead to linguistic systems that are more communicatively efficient will be selected for. We will refer to this view as *phylogenetic functionalism*.

An example of the way phylogenetic functionalism works is given by the Subjacency Condition [33]:

Subjacency condition No rule can relate X, Y in the structure

$$\ldots X \ldots [\alpha \ldots [\beta \ldots Y \ldots$$
or
$$\ldots Y \ldots] \beta \ldots] \alpha \ldots X \ldots$$

where α, β are bounding nodes.

This is an example of a cross-linguistic universal principle that operates to constrain the distance over which a rule (typically movement of an element in a structure) can operate. The definition of *bounding node* is an example of a parameter, since it has been shown to vary from language to language within certain limits. In English, the bounding nodes are IP and NP which leads to the ungrammaticality of the sentences below in which *who* needs to be directly related to its trace over two intermediate bounding nodes:[2]

(1) *Paul phoned the singer who$_i$ we recorded $_{NP}$[the song which$_j$ $_{IP}$[t_i sang t_j]]

(2) *Who$_i$ did $_{IP}$[Paul tell you when$_j$ $_{IP}$[he had phoned $t_i t_j$]]

It has been pointed out that the subjacency condition tends to rule out sentences in which the distance between the *wh*-element and its co-indexed gap is long [3]. There is a pressure on the human parser to keep this distance at a minimum for reasons of memory load. The phylogenetic functionalist would say that this parsing pressure leads to the biological selection of a language acquisition device that had some way of eliminating the worst *wh*-extractions from the range of possible languages, hence the subjacency condition becomes part of our innate LAD.

[2]The details of this analysis are unimportant the crucial features of the Subjacency Condition are that it constrains the number of possible ways in which a "moved" element can be interpreted and it can be formulated in a universal fashion (as a principle) even though it varies from language to language (via different possible parameter settings).

3 Glossogenetic functionalism

An alternative explanation for the origin of particular language universals can be termed glossogenetic functionalism[3] [21, 23, 24, 20]. In this approach, the constraints on variation are not assumed to arise directly from the structure of our innate language learning mechanism. Instead, the universals emerge over a historical/cultural timescale from the process of language acquisition *and* use.

This type of explanation relies on the principle that language learner does not necessarily converge on the same grammatical system as the adults in the population. Crucially, the triggering experience that the learner uses will not accurately reflect the linguistic competence of the adults because it is filtered through the "arena of use" [20]. There are various pressures that operate during communication that will have a selective effect on the different linguistic variants that are being transmitted from generation to generation.

In earlier work [23, 25, 22] Kirby has shown that the selective effect of the parser in the cycle of acquisition and use can give rise to language universals of the sort that typologists observe cross-linguistically. It appears that languages adapt to aid their own survival over time. More correctly, proportions of competing variants in a language change over time through differential selection in the arena of use and this gives rise to a pattern of cross-linguistic variation that shows the characteristic "appearance of design" that we have been talking about. We will show an abstract version of this process at work in the simulation later in this paper, but for the moment it is worth having a look at a more concrete example that can be explained very simply in terms of glossogenetic functionalism.

A well known language universal[4] relates to the ordering in the string of branching and non-branching constituents:

> "**Branching Direction Theory** (BDT): ...a pair of elements X and Y will employ the order XY significantly more often among VO languages than among OV languages if and only if X is a nonphrasal category and Y is a phrasal category." [11, p.89]

Hawkins [18] shows how a series of universals very similar to this one appear to be a response to parsing pressures. His theory of parsing complexity includes a measure of the distance between categories in the string

[3]The term glossogenetic is used in contrast with the ontogenetic and phylogenetic timescales. It is the timescale over which languages change.

[4]This is a *statistical* language universal. It does not have the same *absolute* status as the Subjacency Condition, for example, but the universal is still a statistically significant statement about cross-linguistic distribution.

which construct dominating tree structure. Simplifying somewhat, the longer the distance between non-branching nodes in a tree the higher the parsing complexity of that tree. The most efficient tree structures will be those which order non-branching nodes on the same side of branching nodes throughout the structure. The BDT states that languages which generate such tree structures will be more common than those that do not.

The glossogenetic functionalist would say that the parsing pressure for consistent branching direction leads to the *linguistic* selection in the arena of use of variant word orders that are consistent with the branching direction of the rest of the language over those that are inconsistent. So, if there are examples of prepositional phrases and postpositional phrases in the input data given to a learner with a language whose verb precedes its object, then the prepositional phrases are more likely to be parsed successfully. This eventually leads to the loss of the postpositional phrases in VO languages.

4 Elements of the model

In order to test the validity of the two types of functionalism outlined above and make explicit what exactly these theories involve, we have constructed a simple idea model that incorporates the necessary features to model all the processes that are involved in both approaches. Ultimately we wish to explore the different interactions between learning (using an algorithm with partial innate constraints), cultural transmission (through an arena of use involving linguistic selection), and biological evolution of the innate learning constraints (based on the communicative success of the individuals).

4.1 Representation of grammar

The representation of the mature grammatical competence of the individuals in the simulation is borrowed from a paper by Turkel [34] as are many of the other details of the implementation — Turkel's model is similar to the one presented in this simulation, but does not include any real cultural/linguistic transmission or a model of functional pressures.[5]

We simply encode a grammar as a string of 1s and 0s. In the results reported here, every individual's competence is an 8 bit string. This leads to 256 *logically possible languages*. Of course, we expect to show that the actually occurring languages will not be evenly distributed in this space.

4.2 Representation of LAD

The LAD, again following Turkel who borrows from Clark [7], is coded in the genome as a string of genes each of which has three possible alleles: 0, 1 or ?. Wherever there is a 1 or 0 allele, the resulting LAD will only be able to acquire grammars with the same symbol in the corresponding position. These alleles can be seen as coding for different possible *principles*. The ? allele, on the other hand, corresponds to a *parameter*. The resulting LAD will be able to acquire grammars with either a 1 or a 0 in the same position as the ? in the genome.

The most constrained LAD, then, is one whose genotype consists of no ? genes. Such an LAD will only ever acquire one language. In fact, the LAD will not learn at all, since the language is fully innate. The least constrained LAD is one with solely ? genes — all parameters. Such an LAD could *in principle* learn any one of the 256 logically possible languages. In this extreme, there are no innate constraints on variation.

4.3 Utterances as triggers

As Clark and Roberts [8] do, we will treat each utterance in the simulation as a *trigger* for a particular subset of the set of possible grammars. To take a concrete example, the first Hungarian sentence below could potentially provide the learner with evidence that she is hearing a sentence produced from a language with locative case endings, but cannot even in principle trigger the setting of a pro-drop parameter (a parameter that specifies that subjects can be lacking in main clauses). The second sentence provides the opposite triggering experience.

(3) Én is vagyok bárban
I also be(1SG) bar+in
'I am in the bar too'

(4) Egy pohár sört kérek
a glass beer+ACC want(1SG)
'I want a glass of beer'

We accordingly code utterances as a string of 1s, 0s and $*$s. Each 1 or 0 potentially triggers the acquisition of a grammar with the same digit in the corresponding position. Each $*$ carries no information about the 'target' grammar. With our example above, imagine that locative case-coding languages have a 1 in the first position of our grammar coding, and pro-drop languages have a 0 in the second position of the coding. The first sentence above would be represented as $< 1, *, \ldots >$. The second would be $< *, 0, \ldots >$.

In the simulation results reported here, the individuals produce utterances randomly which are consistent with their grammars and only provide evidence for one digit of their grammars. In other words, each trigger will have seven $*$s and one digit.

4.4 The Trigger Learning Algorithm

The next basic element of our model is an algorithm for parameter setting. An algorithm that has been discussed

[5] Turkel's goal was to show the plausibility of a partial biologisation of learned parameter settings.

in the literature recently [32, 6, 35, 12] is the Trigger Learning Algorithm of Gibson and Wexler [13]. We employ this algorithm in our simulation for simplicity, but for the simulation runs presented here little relies on this choice. The reason for this, and general issues relating to parameter setting are discussed later.

The Trigger Learning Algorithm (TLA) Given an initial set of values for n binary-valued parameters, the learner attempts to syntactically analyze the incoming sentence S. If S can be successfully analyzed, then the learner's hypothesis regarding the target grammar is left unchanged. If, however, the learner cannot analyze S, then the learner uniformly selects a parameter P (with probability $1/n$ for each parameter), changes the value associated with P, and tries to reprocess S using the new parameter value. If analysis is now possible, then the parameter value change is adopted. Otherwise, the original parameter value is retained.

4.5 Linguistic selection

So far we have made no reference to the role of communicative function in our model. All utterances in all languages have an equal status in the formulation given above. As mentioned earlier, Kirby [23] has modelled the role of communicative function in the cycle of acquisition and use as adjusting the probabilities of a variant being taken up as part of the learner's trigger experience. Robert Clark [6] shows how this idea of linguistic selection can be built into a modified version of the TLA.

In the original formulation of the TLA, a new parameter setting is only taken up if analysing the input is possible with the new setting and not with the old setting. In the simulations in this paper there is a certain probability (that can be varied from run to run) that the criteria for taking up a new parameter setting will be based not on the absolute analysability of the trigger but on its parsability. The procedure on receiving a trigger is therefore:

Modified TLA If the trigger is consistent with the learner's LAD:[6]

1. If the trigger can be analysed with the current grammar, score the parsability of the trigger with the current grammar.
2. Choose one parameter at random and flip its value.
3. If the trigger can be analysed with the new grammar, score the parsability of the trigger with the new grammar.
4. With a certain pre-defined frequency carry out linguistic selection (a), otherwise (b):
   (a) If the trigger can be analysed with the new grammar, and the new grammar's parsability score is higher than that of the current grammar, or the trigger cannot be analysed with the current grammar, adopt the new grammar.
   (b) If the trigger cannot be analysed with the current grammar, and the trigger can be analysed with the new grammar, adopt the new grammar.
5. Otherwise keep the current grammar.

[6] This clause is here simply to save work, since it is possible that no combination of parameter settings will be able to analyse the trigger. In other words, the trigger is outwith the constraints imposed by the learner's LAD.

4.6 Natural selection

In order to implement natural selection we need some way of assessing the communicative success of individuals *after* learning. We use the concept of a *critical period* [27, 26] during which learning occurs, which is followed by a period of continued language use, but no grammatical change. As a simplifying assumption we measure the communicative fitness of individuals after this critical period.

The fitness can be based on either the transmission ability of an individual, the reception ability of an individual, or a combination of both. Each individual is involved in a certain number of random communicative acts, for half of which he is the hearer and half the speaker. The individual's transmission ability is scored on the basis of how many of the utterances spoken were analysable by the hearer, and how parsable those utterances were. The individual's reception ability is scored similarly on the utterances heard. If speakers and hearers are drawn from the same linguistic community (see the next section for more details), then this procedure can be used to test both the success of learning, and the "functionality" of the individual's grammar.[7]

Fitness measurement

1. For each utterance heard, with a certain pre-defined frequency carry out (a), otherwise (b):
   (a) If the utterance can be analysed, measure the utterance's parsability score, and with a probability proportional to that score increase reception fitness.
   (b) If the utterance can be analysed, increase reception fitness.
2. For each utterance produced, with a certain predefined frequency carry out (a), otherwise (b):

[7] Notice this looks similar to the linguistic selection in the previous section, but has no impact on the transmission of language from generation to generation.

(a) If the utterance can be analysed by the hearer, measure the utterance's parsability score, and with a probability proportional to that score increase transmission fitness.

(b) If the utterance can be analysed by the hearer, increase transmission fitness.

5 Layout of the model

In the previous section we quickly reviewed the six central elements of the model: grammars, LADs, utterances, parameter setting, linguistic selection, and natural selection. Figure 1 shows how these various components fit together in our simulation.

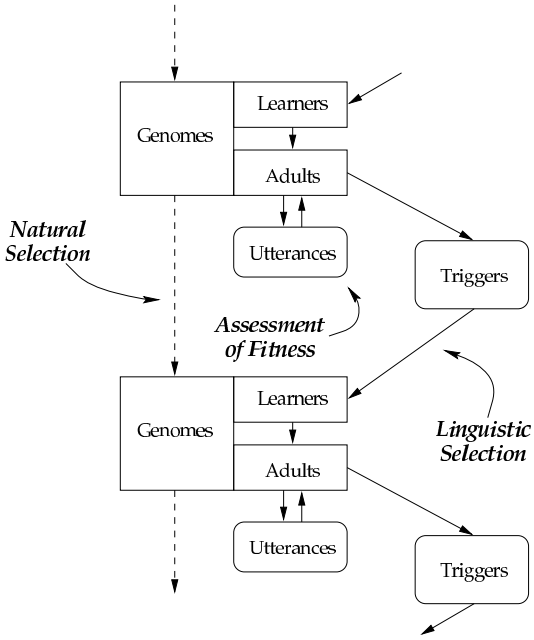

Figure 1: An overview of the simulation model.

On the right hand side of this diagram, we have the part of the model that deals with cultural transmission. Each generation of adults produces a set of triggers randomly in line with their grammars, and this acts as training data for the next generation of learners. In this way, languages survive across generations, although there is not perfect transmission. Language change can occur through: failure to learn, linguistic selection in the modified TLA, and (as we shall see below) language contact.

The fitness of the adults is assessed as described above after learning has finished. This fitness assessment is used to select which individuals will mate to produce the next generation of learners. Rank selection is used — for the results presented here, the top 90 percent of the population have an equal chance of reproducing (the bottom 10 percent have no chance of reproducing). The new population of genomes is formed using one point crossover with a mutation probability of 0.001 per allele. At every generation, the entire population is replaced. Notice crucially the utterances that the adults produce are kept separate from the triggers that are given to the next generation. In a sense there are two "games" taking place in this model: an adult-to-child game which results in cultural transmission, and an adult-to-adult game that results in natural selection.

In order to model language change, it is important for the arena of use to be organised spatially [24, 29]. This is achieved in the simulation by organising the individuals in the population in a one-dimensional loop, as in figure 2. For the results reported, breeding is not spatially organised. It was found that organising *both* linguistic and genetic interaction spatially lead quickly to extreme genetic heterogeneity in the population. An investigation of this phenomenon is beyond the scope of this paper.

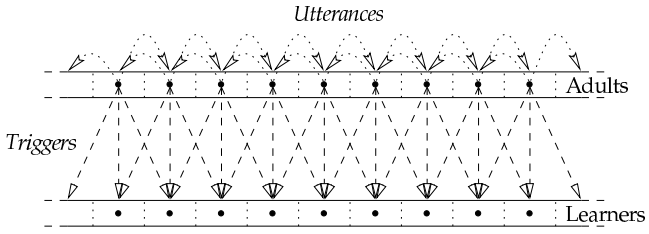

Figure 2: How the arena of use is spatially organised in the simulation.

6 Results

For the results in this section, the parsability scoring function is chosen arbitrarily to prefer 1s in the first 4 bits of the grammar. The score ranges from 0.0 to 1.0 proportional to the number of 1s. There are many ways of modelling the parsability of languages in the hypothetical 8 bit grammar space; this one is chosen here simply so that there are some specific parameter settings that lead to more parsable utterances and some that have no impact on parsability. Clearly, if the simulation is responding to functional pressures we should find the languages distribution at the end of the simulation to reflect the preference for grammars that start $< 1, 1, 1, 1, \ldots >$.

6.1 Natural selection for functional LADs

Firstly we wish to see if phylogenetic functionalism works. To do this we "turn on" natural selection, but "turn off" linguistic selection. In terms of the simulation, this means that the proportion of *triggers* that have their parsability scored in the TLA is zero, on the

other hand parsability is measured for 10 percent of *utterances* in assessing fitness.[8] There are 100 individuals in the population, all of which start with a genome that is fully "plastic", in other words with no innate principles: $<?,?,?,?,?,?,?,?>$. The critical period is set at 200 triggers.[9] The initial arena of use (the triggers fed to the first generation of learners) is completely random.

Figure 3 shows the average fitness of the population over time where each speaker produces 100 utterances in fitness testing, and is scored 1 for each successfully received utterance, and 1 for each successfully transmitted utterance. Figure 4 shows the average proportions of parameters and the two types of principle during the same run. This graph is typical for runs of the simulation with the initial conditions described. The final graph in figure 5 shows the proportions of the different alleles in the genomes at the end of the run — essentially the make up of the average LAD after evolution. This varied wildly from run to run.

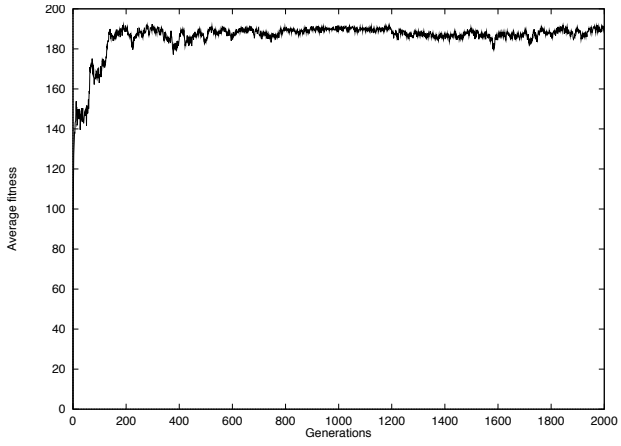

Figure 3: Average fitness against generations for typical run.

It should be clear from these results that evolution has failed to respond to the functional pressure to have only 1s in the first four positions of the grammars. In fact, the first grammatical parameter has been almost completely nativised as a '0' principle. This means that the individuals in the simulation simply cannot produce or parse optimal utterances. This explains why their fitness, although it increases initially never reaches the maximum possible 200. Similar results where also forthcoming when fitness was based solely on reception behaviour or transmission behaviour.

[8]Various degrees of parsability testing were tried. Different results only arise at the extremes of the range.

[9]Again, different values were tested, each giving a different degree of nativisation (see below). This value gives results in the middle of the range.

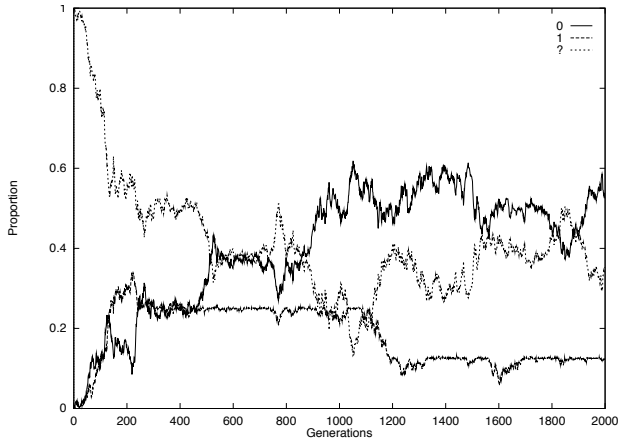

Figure 4: Average proportions of 0s, 1s and ?s in the LADs of the population.

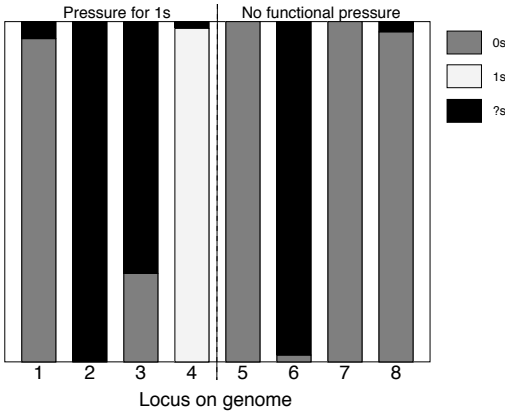

Figure 5: Average proportion of different alleles at the end of run.

6.2 Linguistic selection for functional languages

For the next run of the simulation we enable linguistic selection in the TLA. As with the adult-to-adult utterances, we score the parsability of the adult-to-child triggers 10 percent of the time. Apart from the inclusion of linguistic selection during transmission of triggers, the set up is identical to that in the previous section.

Figure 6 shows a space-time diagram of the languages that are present in the arena of use. The two languages that predominate after only 200 generations are $< 1,1,1,1,0,0,1,1 >$ and $< 1,1,1,1,0,1,1,1 >$. Both of these languages are optimally parsable. What seems to be happening here is that languages are very rapidly evolving historically to become easier to parse. Even after only 57 generations, the main languages

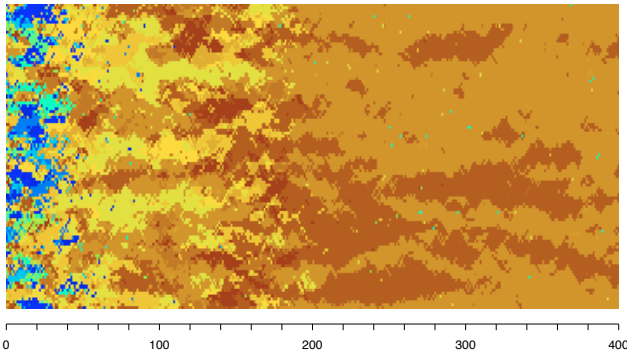

Figure 6: A space-time graph for the first 400 generations of the run. Each colour is assigned to one language type. Time runs horizontally left to right.

are $< 1,1,1,1,0,0,1,1 >$, $< 1,1,1,1,0,1,1,1 >$, $< 1,1,1,1,0,0,0,1 >$ and $< 1,1,1,1,1,0,0,1 >$. This looks like glossogenetic functionalism.

If we look at what is happening to the LADs in the same simulation run we see a different picture (figure 7). After 57 generations, only one of the parameters has been nativised as a principle. This makes it clear that it is linguistic, not natural, selection that is improving parsability. However, eventually the LADs do evolve, as more of the parameters become principles. The shape of

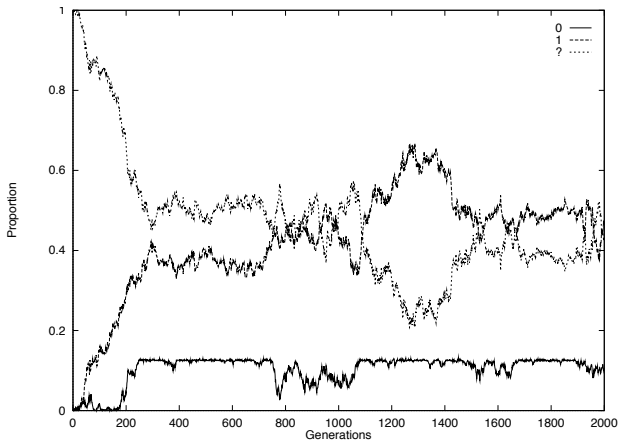

Figure 7: Average proportions of 0s, 1s and ?s in the LADs of the population with linguistic selection.

the evolved LADs is shown in figure 8. The interesting feature of these results is that the LADs appear to have evolved to at least partially constrain learners to learn languages that are functional. This is exactly what is predicted by phylogenetic functionalism, but this result does not emerge without a prior *glossogenetic* evolution.

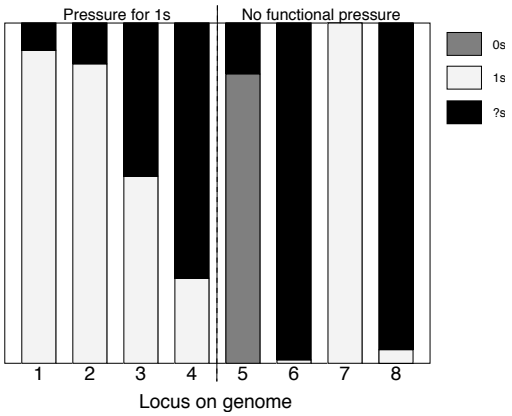

Figure 8: Average proportion of different alleles at the end of run in a population with linguistic selection.

A natural selection pressure to communicate efficiently (either as a speaker or hearer) is not on its own enough to account for functional constraints on variation. In fact, it appears to lead to positively dysfunctional constraints on variation being built into the genome. The linguistic selection of triggers in the cultural transmission of language, on the other hand, accounts for the appearance of functional constraints very rapidly. It also, in combination with a natural selection pressure for communication, leads to a partial nativisation of these constraints in the LAD.

7 Discussion

7.1 Why phylogenetic functionalism cannot work alone

It seems that phylogenetic functionalism, although an apparently obvious explanation for linguistic constraints, does not work alone. Evolution seems unable to optimise the innate constraints on learners in such a way that they are constrained to produce functional languages. Why might this be so?

The answer seems to lie in what selection pressures face a linguistic individual. If the pressure is to produce utterances that are potentially easy to parse, or to be able to analyse easily parsable utterances, then we would expect evolution to select the individual whose LAD reflects parsing pressures. However, an overriding concern for a communicating individual is to *correctly learn the language of her speech community.* Without a grammar that accurately reflects that of her peers, an individual will not be understood, nor be able to understand others, whatever the functionality of her grammar.

Consider the possible situations an individual with a mutated LAD might find himself in:

1. The individual may belong to a speech community that is speaking an optimal language. In this case, as long as the mutation does not constrain the individual in such a way that he cannot learn this language, there is no preferential selection for a more functional LAD over a less functional one.

2. The individual may belong to a speech community that is speaking a sub-optimal language. In this case, if the mutation produces a more functional LAD the individual is actually selected *against* because he will not be able to learn the language of his community.

In this rather simplified characterisation, then, there is no way that a mutation that increases the functionality of the LAD (in the sense that it constrains languages to be parsable) can give a direct fitness advantage to an individual. This is true even though the fittest population would be one where everyone possessed just such an LAD.

7.2 The Baldwin Effect

This result stands in contradiction to the fact that there *do* seem to be at least some functional innate linguistic constraints that humans are born with. The Subjacency Condition, reviewed earlier, appears to be one of them. There are also functional constraints on variation that are harder to account for in terms of innate constraints on learnability — the word order universals expressed in the Branching Direction Theory, for example. The results of the second simulation run seem to capture this *partial* innateness of functional constraints rather well. Why does adding linguistic selection make such a difference to the results?

The results show that even where there is no genetic change (i.e. the whole population is still completely plastic) the languages in the simulation converge on a (sub)set of optimal languages. Wherever there is linguistic variation in the input to a particular learner, for example when there are two languages in contact in the arena of use, there may be differential uptake of competing parameter settings. As Robert Clark [6] has proved, in this simple situation this will inevitably lead to languages becoming adapted to maximise their own transmission potential — in other words, more functional parameter settings survive.

Given a linguistic environment which is adapted glossogenetically to communicative function, the LADs in the population gradually evolve phylogenetically to mirror the existing constraints on variation. This is a clear example of the Baldwin Effect [1, 34, 19, 2] in operation. One of the predicted outcomes of the interaction of learning and evolution is that wherever there is a cost attached to learning (be it risk of making mistakes, or delay in knowledge acquisition), there will be a pressure to make innate those features of the learning task which are *predictable*. If the same parameter settings are consistently expressed in the trigger experience of generation after generation, there is no disadvantage to that parameter setting "becoming" a principle; it means less work for the learner. If, on the other hand, there are parameter settings that are highly variable glossogenetically, then there is a pressure for them to remain learnt by the population.

Notice that this pressure to nativise only exists where there is a disadvantage to learning. In our model this disadvantage is the risk of failing to converge on the correct grammar before the critical period. This is why, when the critical period is changed, the degree of eventual nativisation changed. Only when the critical period is extremely severe do we see a complete nativisation of the functional constraints. (See [26] for a model of how the critical period itself evolves.)

In summary, we have a two stage process:

1. From initially random initial conditions, linguistic selection leads to a glossogenetic adaptation of the languages in the arena of use. This results in observable constraints on variation, although the individuals are still completely plastic and so could potentially learn languages outwith these constraints.

2. This glossogenetic adaptation *enables* the phylogenetic adaptation of the LADs in the population through the Baldwin Effect. Over time, some of the regularities in the linguistic input become nativised. This means that at the end of this process, the constraints on variation "harden up" so that the individuals in the population could not even potentially learn the dysfunctional languages.

8 Extensions

The simulation in this paper is clearly a fairly abstract idea model — a first step towards explaining the complex interaction involved in the evolution of a learning mechanism for a shared culturally transmitted trait like language. In this section we briefly review the directions in which the model might be extended.

Diversity One of the crucial aspects of the arena of use in the simulation is spatial organisation. One of the effects of this is to increase the sustainable level of diversity in the languages in the population. This is important because the evolving LADs are responding to regularities in the input, and without some variation in the input, there is simply a pressure to nativise a single language. The final degree of diversity is still low in the results shown in this paper, however. Within the model, this can be changed by altering the rate at which languages can spread sideways through the population. It would be interesting to experiment with some of the other features of the arena of use that Nettle [29] argues impact

on the maintenance of diversity, such as social selection and varying competing functional pressures [24].

More complex functional constraints We have seen that only some of the parsing pressures get nativised, but the ones that do or do not seems to be arbitrary. It would be interesting if there was some way in which we could predict what sorts of functional pressure would be left to glossogenetic adaptation and what sorts would become part of the innate LAD. We have applied the model to a situation where the functionality of a particular parameter setting cannot be measured in isolation, but instead depends critically on other parameter settings. For example, the simulation has been tested on a situation where the optimal grammars are those where adjacent pairs of parameter settings are the same. In this regime, some optimal grammars would be: $< 1,1,0,0,0,0,1,1 >$ or $< 0,0,1,1,0,0,1,1 >$ and so on.

It turns out that although glossogenetic adaptation quickly gives us a pattern of variation which captures this functional pressure, there is no way for the Baldwin Effect to nativise it. There is no way of representing this pairing pattern in the LAD. It would require a representation with variables over parameter settings like: $< p_1,p_1,p_2,p_2,p_3,p_3,p_4,p_4 >$.

Interestingly, this kind of cross-parametric interdependence may be just the sort of thing that is going on with the branching direction universals. If we consider the order of head and modifier to be independently specifiable for each phrasal category in a grammar, then it might not be possible to capture in the LAD the parsing preference for consistent ordering across these categories. If this turned out to be true, then we might have an explanation why this word order universal is *statistical* rather than *absolute*. It exists as a constraint that emerges from the glossogenetic process rather than a constraint that is hard-wired into the genome.

The role of the parametric space The model as we have described it makes no direct reference to linguistic features — actual parameters, triggers, or functional pressures. We believe this is the correct first step in understanding the general processes involved, before moving on to more complex models. Briscoe [4] in a fascinating paper shows that it is possible to model the nativisation of more realistic-looking functional pressures. Simplifying somewhat, he models the LAD as a set of 11 parameters which, in combination with a particular syntactic theory (Generalised Categorial Universal Grammar), can generate strings of words which may act as triggers. Briscoe uses a particular theory of working memory to then assess the parsing cost of these triggers.

One of the interesting features of Briscoe's model — and others such as [32, 6] — is that the complexity of mapping from parameter settings to triggers leads to interesting unpredictable dynamics in the glossogenetic evolution of languages due to the ambiguity of some triggers and misconvergence by learners. This means that it is hard to tease apart the effect of the learning model and the functional pressures on the emerging universals. Of course, this is an important insight into the complexities of our object of study.

The problem is that we cannot make any specific predictions until we know more about the parametric space because as Robert Clark [6] shows, a small change in the details of the parameterisation lead to radically different end results. The status of the TLA is far from clear, however. A recent paper by Fodor [12] argues that the TLA is psychologically implausible, and instead suggests a theory of parameter setting that relies on single unambiguous triggers. This is not the place to explore Fodor's theory, suffice to say that some of the complexities introduced by misconvergence in non-trivial parameterisations may be ameliorated with different theories of acquisition.

9 Conclusion

We have shown that phylogenetic functionalism alone cannot work, but this does not mean that functional constraints cannot find their way into the innate Language Acquisition Device. Instead, we show that the introduction of linguistic (as opposed to natural) selection into a model of language acquisition, use, transmission, and evolution has profound effects on the evolutionary trajectory of learners. The very same mechanism (the differential filtering of triggers out of the learners input data due to parsing difficulty) can explain *both* historically emergent universals and innate constraints on variation.

In general, we have shown that for a culturally shared system like language, cultural evolution can bootstrap biological evolution. We are currently exploring the possibility that this kind of mechanism may be involved at an earlier stage of language evolution.

References

[1] J.M. Baldwin. A new factor in evolution. *American Naturalist*, 30:441–451, 1896.

[2] Richard Belew. Evolution, learning, and culture: computational metaphors for adaptive algorithms. *Complex Systems*, 4:11–49, 1990.

[3] R.C. Berwick and A.S. Weinberg. *The Grammatical Basis of Linguistic Performance: Language Use and Acquisition*. MIT Press, 1984.

[4] Ted Briscoe. Language acquisition: the bioprogram hypothesis and the Baldwin Effect. MS, Computer Laboratory, University of Cambridge, 1997.

[5] Noam Chomsky. *Knowledge of Language*. Praeger, 1986.

[6] Robert Clark. Internal and external factors affecting language change: A computational model. Master's thesis, University of Edinburgh, 1996. In preparation.

[7] Robin Clark. The selection of syntactic knowledge. *Language Acquisition*, 2:85–149, 1992.

[8] Robin Clark and Ian Roberts. A computational model of language learnability and language change. *Linguistic Inquiry*, 24:299–345, 1993.

[9] Bernard Comrie. *Language Universals and Linguistic Typology*. Basil Blackwell, 1981.

[10] William Croft. *Typology and universals*. Cambridge University Press, Cambridge, 1990.

[11] Matthew Dryer. The Greenbergian word order correlations. *Language*, 68:81–138, 1992.

[12] Janet Fodor. Unambiguous triggers. *Linguistic Inquiry*, 1997. To appear.

[13] E. Gibson and K. Wexler. Triggers. *Linguistic Inquiry*, 25:407–454, 1994.

[14] Joseph Greenberg. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press, 1963.

[15] Liliane Haegeman. *Introduction to Government and Binding Theory*. Blackwell, 1991.

[16] John A. Hawkins. Explaining language universals. In John A. Hawkins, editor, *Explaining Language Universals*. Basil Blackwell, 1988.

[17] John A. Hawkins. A parsing theory of word order universals. *Linguistic Inquiry*, 21:223–261, 1990.

[18] John A. Hawkins. *A performance theory of order and constituency*. Cambridge University Press, 1994.

[19] G. Hinton and S. Nowlan. How learning can guide evolution. *Complex Systems*, 1:495–502, 1987.

[20] James Hurford. *Language and Number: the Emergence of a Cognitive System*. Basil Blackwell, Cambridge, MA, 1987.

[21] Simon Kirby. Adaptive explanations for language universals: a model of Hawkins' performance theory. *Sprachtypologie und Universalienforschung*, 47:186–210, 1994.

[22] Simon Kirby. Constraints on constraints, or the limits of functional adaptation. To appear in: *Functionalism and Formalism*, 1996.

[23] Simon Kirby. *The Emergence of Universals: Function, Selection and Innateness*. University of Edinburgh, 1996. PhD thesis.

[24] Simon Kirby. Competing motivations and emergence: explaining implicational hierarchies. *Language Typology*, 1:5–32, 1997.

[25] Simon Kirby. Fitness and the selective adaptation of language. In J. Hurford, C. Knight, and M. Studdert-Kennedy, editors, *Evolution of Language: Social and cognitive bases for the emergence of phonology and syntax*. 1997. To appear.

[26] Simon Kirby and James Hurford. The evolution of incremental learning: Language, development and critical periods. Occasional Paper EOPL-97-2, Department of Linguistics, University of Edinburgh, Edinburgh, 1997.

[27] Eric H. Lenneberg. *Biological Foundations of Language*. Wiley, New York, 1967.

[28] Edith Moravcsik, editor. *Functionalism and formalism in linguistics: proceedings of the 23rd UWM linguistics symposium*. Benjamins, 1997. In preparation.

[29] Daniel Nettle. *The Evolution of Linguistic Diversity*. PhD thesis, University College London, 1996.

[30] Frederick J. Newmeyer. Functional explanation in linguistics and the origins of language. *Language and Communication*, 11:3–28, 1991.

[31] Frederick J. Newmeyer. *Language Form and Language Function*, chapter Langauge Typology and its Difficulties. 1997. In preparation.

[32] Partha Niyogi and Robert Berwick. The logical problem of language change. Technical Report AI Memo 1516 / CBCL Paper 115, MIT AI Laboratory and Center for Biological and Computational Learning, Department of Brain and Cognitive Sciences, 1995.

[33] H. Van Riemsdijk and E. Williams. *Introduction to the Theory of Grammar*. MIT Press, 1986.

[34] William Turkel. The learning-guided evolution of natural language. Manuscript, University of British Columbia, 1994.

[35] William Turkel. Noise-induced enhancement of parameter setting. Submitted to Linguistic Inquiry, 1997.