What can formal or computational models tell us about how (much) language shaped the brain?

Ted Briscoe
Computer Laboratory
University of Cambridge
ejb@cl.cam.ac.uk

Acknowledgements
Thanks to Simon Kirby for modifying and allowing me to (re)use Figure 2 from Kirby *et al.* (this vol.)

1 Introduction

In this paper, I expand on and update the arguments concerning the evolutionary emergence and maintenance of an innate language acquisition device (LAD) discussed in Briscoe (2003). By a LAD, I mean nothing more or less than a learning mechanism which incorporates some *language-specific* inductive learning bias in favour of some proper subset of the space of possible grammars.[1] The existence of an innate LAD has remained controversial, and it is certainly the case that many arguments that have been proposed in its favour are questionable or wrong (e.g. Pullum and Scholz, 2002; Sampson, 1989, 1999; Lappin and Shieber, 2007). However, I will still argue that all adequate extant models of language acquisition do presuppose a LAD in the sense above. These arguments put the onus on non-nativists to demonstrate an adequate, detailed and precise account of the acquisition of grammar which does not rely on a LAD.

Chomsky has consistently downplayed the role of evolution in the emergence of the LAD, emphasized the discontinuities between human language and animal communication systems, and speculated that the LAD arose as a result of a macromutation or saltationist jump, even in his most recent work (e.g. Hauser *et al.* 2002). Pinker and Bloom (1990) developed an account of the gradual evolutionary emergence of the LAD via genetic assimilation (or in their terms, the Baldwin Effect). More recently, Briscoe (1997), Deacon (1997) and others have argued that languages themselves are adaptive systems and that the universal constraints on grammar that underpin much argumentation for the LAD can be explained as a consequence of convergent evolution under similar linguistic selection pressure. However, I will argue that this important insight does not undermine the existence of the LAD, though it certainly undermines arguments for the LAD based solely on the existence of linguistic universals.

Genetic assimilation is a neo-Darwininan mechanism (e.g. Waddington, 1942) by which organisms can appear to inherit acquired characteristics though, in fact, it is changes in their behaviour or more generally their environment (e.g. niche construction) which create novel selection pressures and thus cause information to be assimiliated into the genome[2]. Genetic assimilation of grammatical information exemplified in the environment of adaptation of the LAD potentially would facilitate more rapid and robust acquisition of grammar by first language learners. Thus, if mastery of language increases fitness, we might expect natural

[1] The term, LAD, is taken from Chomsky (1965). In more recent work, it has been dropped in favour of universal grammar (UG) (e.g. Chomsky, 1981), reflecting the increasing focus on constraints on the space of learnable grammars. Here I stick to the older term as I believe that it is only possible to evaluate empirically claims about UG when they are embedded within a precise account of the acquisition of grammar.

[2] I have reviewed the evidence for genetic assimilation in areas other than language evolution elsewhere (e.g Briscoe, 2003). See also Pigliucci *et al.* (2006), for a recent more extensive review and discussion.

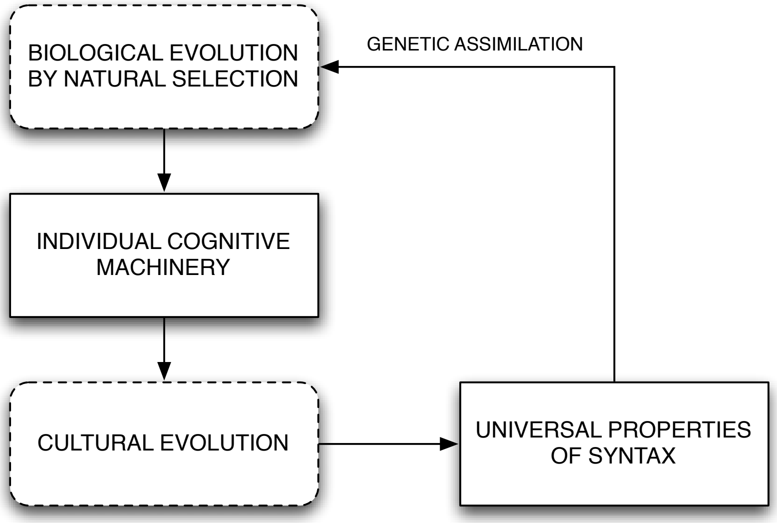
BIOLOGICAL EVOLUTION BY NATURAL SELECTION
GENETIC ASSIMILATION
INDIVIDUAL COGNITIVE MACHINERY
CULTURAL EVOLUTION
UNIVERSAL PROPERTIES OF SYNTAX

Figure 1: The emergent linguistic environment potentially creates new (natural) selection pressures on our cognitive machinery

selection to improve language learning. I have argued (e.g. Briscoe, 2005) for a coevolutionary account incorporating this process in which natural languages are treated as complex adaptive systems undergoing often conflicting selection pressures, only some of which emanate from the LAD or indeed more general cognitive mechanisms, and where the LAD itself evolved via genetic assimilation in response to (proto)languages in the environment of adaptation.

In terms of Kirby *et al.* (this vol., Fig. 2), the question we are considering is whether it is appropriate to add a further arrow to their diagram depicting the interaction of natural selection and linguistic selection which 'closes the loop' between natural selection for cognitive machinery and the linguistic environment created via cultural transmission, as illustrated in Figure 1. Thus I am not arguing that cultural evolution and linguistic selection have not had a profound effect on the nature of natural languages, nor that many if not most linguistic universals have emerged as a consequence of cultural or linguistic selection for more learnable or otherwise cognitively or socially advantageous linguistic forms or constructions. The only question I will consider is whether genetic assimilation of language-specific grammatical information into the LAD is also plausible given the coevolutionary scenario entailed by Figure 1. In particular, I will consider the extent to which formal or computational models could and do contribute to an answer to this question.

I do not intend to revisit all the arguments for and against genetic assimilation reviewed in Briscoe (2003), nor to rereview the various models discussed there. Instead I will focus on some recent arguments, sometimes supported by models and simulations, against genetic assimilation of linguistic information. However, before addressing these arguments the next section defines grammatical acquisition and presents a Bayesian account of the task.

2 Grammatical Acquisition

In Briscoe (2003), I discuss five desiderata that adequate accounts of grammatical acquisition during first language learning must satisfy: 1) *coverage* of attested grammatical constructions; 2) *realistic input* to the learner consisting of a finite, positive, but partly noisy sample from the target language; 3) *realistic contextual enrichment* of this sample with only partial, noisy representations of the form-meaning mapping; 4) *selectivity* in which a consistent grammar is acquired and random noise is rejected; and 5) *accuracy* in which the acquired grammar captures the form-meaning mappings of the target grammar – learners do not 'hallucinate' or invent grammatical properties regardless of the input, though they do (over)generalize and, in this sense, 'go beyond the data'. If accuracy is defined in terms of formal learnability from realistic, finite, positive but noisy sentence-meaning pairs over a hypothesis space with adequate coverage, even when drawn from a single stationary target grammar, then inductive bias in the acquisition model is essential.[3]

The term inductive bias is utilized in learning theory to characterize both hard constraints on the hypothesis space considered by a learner, usually imposed by a restricted representation language for hypotheses, and soft constraints which create preferences within the hypothesis space, usually encoded in terms of a cost metric or prior probability distribution on hypotheses (e.g. Mitchell, 1997:39f). Bayesian probabilistic learning theory is a general domain-independent formulation of learning (see e.g. Mitchell, 1997:154f for an introduction) which relies on statistical inference and can thus cope with noise. Bayes' theorem provides a general formula and justification for the integration of prior bias with experience:

$$P(H \mid D) = \frac{P(H)P(D \mid H)}{P(D)} \tag{1}$$

We compute the posterior probability of a hypothesis, H, given some data, D, by multiplying the prior probability of the hypothesis by its likelihood given the data, and normalize to obtain a probability by dividing the result by the overall probability of the data. We typically choose the hypothesis with the highest posterior probability. If we don't need to know this exact probability we can skip the normalisation step and simply choose the highest value hypothesis after

[3] See also Lappin and Shieber, 2007; Nowak *et al.* 2002 for related discussions of learning theory drawing similar conclusions.

multiplying prior and likelihood:

$$H = argmax\ P(H)P(D \mid H) \tag{2}$$

The most general formulation of learning in this framework (Kolmogorov Complexity) posits a learner able to learn any generalisation with a domain-independent bias (the so-called 'universal prior') in favour of the smallest, most compressed hypothesis (e.g. Li and Vitanyi, 1997). However, nobody has demonstrated that this general formulation could, even in principle, result in a learning algorithm capable of accurately acquiring a specific grammar of a human language from realistic input. However, there have been many demonstrations that grammars from more restrictive though infinite hypothesis spaces, such as the class of context-free grammars, can be acquired given a general bias in favour of the smallest or most probable hypothesis (e.g. Horning, 1969). However, when such a *general* bias is applied to a domain-specific and restrictive representation, then it will create bias in favour of certain form-meaning mappings. This is where domain-specific inductive bias appears to be unavoidable if the desideratum of learning accuracy is to be met. And thus, this is the basis on which a LAD, in the sense of section 1, is unavoidable in any adequate account of grammatical acquisition. To relate, this back to the equations above, if the space of possible hypotheses, H, is that of unrestricted rewrite rules or Turing machines, then we might argue reasonably that we have a domain-independent inductive bias. On the other hand, if this space is defined as the (infinite) class of context-free or indexed grammars, which cannot express some types of possible dependencies within sequential strings and thus some possible mappings between meaning and form, then we are positing a LAD, possibly with additional soft bias deriving from the prior.

Gold's (1967) original negative 'in the limit' learnability results are founded on the intuition that any amount of finite, positive data from a target grammar in a class containing grammars capable of generating an infinite set of sentences is always compatible with a hypothesized grammar generating all and only the data seen so far and also with any one of a potentially infinite set of other grammars from the candidate class which generate some superset of the learning sample. Notwithstanding more recent developments in learnability theory and machine learning (e.g. Nowak *et al.* 2002; Lappin and Shieber, 2007), this basic point still holds. A prior distribution or cost metric encoding a preference, for example, for smaller, more compressed grammars will, in general, select a single grammar which predicts the grammaticality of of a specific superset of the learning sample. The exact form of the representation language in which candidate grammars are couched and/or the addition of factors other than just size to the prior distribution or cost metric will determine which of the potentially infinitely many grammars generating a superset of the learning sample is selected by the learner.

Consider a potential class of languages consisting of clauses constructed from a verb (V), a subject (S) and object (O), where S and O are always realized as single (pro)nouns (N) or as noun phrases consisting of a noun and a (relative) clause – the S and O labels are a shorthand for the mapping from forms to meanings (in this instance just predicate-argument structure). By stipulation, there is one root clause per sentence and all relative clauses modify the immediately preceding or following noun. Potentially, grammatical sentences in this class of languages can consist of any infinite sequence of Ss, Vs and/or Os, where we will use subscripts to indicate which S or O is an argument of which V, when there is more than one V in a sentence. Thus, without further stipulation, any clausal ordering of S, O and V is possible, as well as any arrangement of root and relative clauses like those in (1).

(1) a $S_iV_iO_iS_jV_jO_j$
(e.g. cats like dogs$_i$ who$_i$ like cats)

b $S_iV_iO_iS_jV_jO_j$
(e.g. who$_i$ like dogs cats$_i$ like cats)

c $S_iV_jO_jS_jV_iV_kO_kS_kO_i$
(e.g. cats$_i$ like dogs who$_i$ like eat mice who$_j$ cats$_j$)

These examples illustrate that post- and pre-nominal relative clauses with clause-initial and -final relative pronouns are all potentially grammatical sequences.

A learner over context-free grammars (CFGs) with preterminals N and V will be capable, in principle, of acquiring any target grammar in this space. Suppose that the learner prefers, a priori, the smallest grammar compatible with the input, defined as the grammar with the least number of nonterminals and the least number of rules with the least number of daughters (where each nonterminal and rule costs one and each daughter of each rule costs one). Then a learner exposed to a sample of unembedded SVO sequences and (1a) might learn the grammar (2).[4]

(2) a Sent → NP^S V NP^O

b NP → NP Sent

c NP → N

This grammar has a cost of 2 for nonterminals, 3 for rules and 6 for daughters (making 11), and predicts the grammaticality of postnominal subject-modifying relative clauses and of centre-embedded and right-branching sequences of relative clauses. (Given this cost metric, the learner could equally well learn a

[4] Once again, I use superscripted S and O and subscripted indices to show the mapping to predicate argument structure and leave implicit that required to characterize the predicate-argument structure of sentences containing relative pronouns. The details of how this mapping is actually realized formally are not important to the argument, but either a rule-to-rule semantics based on the typed lambda calculus or a unification-based analogue would suffice.

non-recursive variant of (2b) with N substituted for NP as leftmost daughter.) Without the preference for smaller grammars, defined as above, a learner might have acquired the less predictive (3).

(3) a Sent → N^S V N^O
b Sent → N_i^S V_i N_i^O N_j^S V_j N_j^O

This grammar has a cost of 1 for nonterminals, 2 for rules and 10 for daughters (making 13), and it does not predict the grammaticality of subject-modifying relative clauses or multiply-embedded relative clauses. Moreover, a cost metric which assigned a cost of 2 to each rule would also select (3) in preference to (2).[5]

If the input also includes (1b), containing a prenominal subject-modifying relative clause, then a learner utilizing grammar (2) might acquire a further right-recursive rule analogous to (2b), predicting complementary distribution of pre- and post-modifying relative clauses. A learner utilizing (3) might acquire a further rule analogous to (3b) predicting only subject-modifying prenominal relative clauses.

Example (1c) provides evidence for a root SVO language containing postnominal VOS relative clauses. A learner with no cost metric might well acquire a grammar with a rule analogous to (3b) with 9 daughters predicting this and only this exact sequence. A learner with the cost metric exposed to SVO unembedded sequences and (1c) would acquire grammar (4) with a total cost of 16.

(4) a Sent → NP^S V NP^O
b RC → V NP^O NP^S
c NP → NP RC
d NP → N

Thus, this learning model predicts that mixed root and embedded constituent orders is a dispreferred or more marked option that will only be adopted when the learner is forced to do so by positive evidence.

By contrast, if the learner represents the class of CFLs in ID/LP notation instead of standard CFG, acquiring immediate dominance (ID) rules independently of linear precedence (LP) rules (e.g. Gazdar *et al.* 1985), but utilizing a similar cost metric which also assigns a cost of one to each LP rule, then the preference or-

[5] The point is not new, of course. Chomsky (1965:38) recognized the need for an evaluation measure based on simplicity to choose between grammars during language acquisition, and others criticized the arbitrariness of such measures. Kolmogorov Complexity (e.g. Li and Vitanyi, 1997) and the related Minimum Description Length (MDL) Principle (e.g. Rissanen, 1989) provide a less arbitrary metric based on the cost of compressing a hypothesis. The MDL principle can and has been applied to grammatical acquisition (e.g. Ristad and Rissanen, 1994), but once again coupled with restricted hypothesis representation languages. These complexities are ignored here to keep the example simple as they do not alter the fundamental point about the domain-specificity of cost metrics or prior distributions defined over restricted hypothesis representation languages.

dering on specific ID/LP grammars predicts that order-free variants of the above grammars with no LP rules will be preferred and that the inclusion of examples like (1b) or (1c) in the input will not alter the learner's hypothesis. Thus by changing the hypothesis representation language but keeping the cost metric the same, we create inductive bias in favour of different grammars which generalize in different ways from the evidence. Similarly, by keeping the representation language the same but modifying the cost metric, we can also create differing inductive biases.

The Bayesian learning framework also provides a general and natural way to understand and model how stronger grammar-specific inductive biases might have come to be integrated with the LAD, in terms of the evolution of more and more accurate prior distributions over the hypothesis space with better and better 'fit' with languages in the environment of adaptation. Cosmides and Tooby (1996), Geisler and Diehl (2003) and Staddon (1988) argue in detail that Bayesian learning theory is an appropriate framework for modelling learning in animals and humans and that evolution can be understood within this framework as a mechanism for optimizing priors to 'fit' the environment, and thus increase fitness. Thus, it provides a framework for making precise the effects of genetic assimilation, as section 4 details. Cost metrics applied to such restricted hypothesis representation languages entail that learners will 'go beyond the evidence' in different ways and, thus, will have different specifically-linguistic inductive biases (i.e. different mappings between form and meaning). However, learners without cost metrics, or equivalently prior distributions, cannot acquire target grammars accurately, as Gold's (1967) and Horning's (1969) work demonstrated.

All extant models which learn form-meaning mappings assume a LAD, in the sense of section 1, because they utilize prior distributions or cost metrics defined over restricted hypothesis representation languages selected to facilitate encoding of grammars for human languages. The onus is on non-nativists to develop a precise account of grammatical acquisition which meets the above desiderata and does not utilize a LAD in this sense. Work utilizing simple recurrent neural networks or other forms of statistical classification purporting to address issues of grammar learning is largely irrelevant as such models can at most learn to classify segmemnts of the input and/or predict the class of the next unit of input. They do not learn a form-meaning mapping which requires the ability to construct a relational encoding using two-place predicates over constituents or lexical heads, such as 'subject-of', 'object-of', and so forth

Independently of these logical and theoretical arguments, there is psycholinguistic evidence that human language learners are biased in linguistically-specific ways. There are learning stages in which overgeneralisation of regular morphology is common, tense is assigned to auxilaries and main verbs in subject-auxiliary inverted constructions, and so forth. Whilst, the exact interpretation of such phenomena is a matter of complex analysis within a theoretical framework, psycholinguists most often describe them as linguistically-specific biases, for instance, Wanner and Gleitman (1982:12f) argue that children are predisposed to learn lexical compositional systems in which 'atomic' elements of meaning, such as negation, are mapped to individual words. This leads to transient production errors, for example, where languages mark negation morphologically.

In summary then, non-nativists must develop an effective detailed grammar learning procedure which meets the desiderata outlined in the opening paragraph of this section which doesn't utilize some form of cost metric applied to a restricted representation language. Until this is done, we must continue to assume that grammar learning requires at least a weak inductive bias able to choose betwen different form-meaning mapping rule sets (grammars) which predict the grammaticality of different supersets of the learning data.

3 Linguistic Evolution

Linguistic evolution proceeds via cultural transmission (primarily, first language acquisition) at a faster rate than biological evolution. The populations involved are generally smaller (speech communities, rather than entire species), and language acquisition is a more flexible and efficient method of information transfer than genetic mutation. Clearly, vocabulary learning and, at least, peripheral grammatical development are ongoing processes that last beyond childhood, so that linguistic inheritance is less delineated or constrained than the biological mechanisms of genetic evolution. Several consequences emerge from the evolutionary account of languages as adaptive systems which must be taken into consideration by any plausible account of grammar learning. Firstly, several researchers have considered what type of language acquisition procedure could not only underlie accurate learning of modern human languages but also predict the emergence of protolanguage(s) with undecomposable form-meaning correspondences and the (subsequent) emergence of protolanguage(s) with decomposable (minimally grammatical) sentence-meaning correspondences (e.g. Oliphant, 2002; Kirby, 2002, Brighton, 2002). They conclude that the language acquisition procedure must incorporate inductive bias resulting in generalisation, and consequent regularisation of the input, in order that repeated rounds of cultural transmission of language regularize random variations into consistent and coherent communication systems.[6] Secondly, the account of languages as adaptive systems entails that linguistic universals no longer constitute strong evidence for a LAD. Deacon (1997), Briscoe (1997) and others make the point that universals may equally be the result of convergent evolution in different languages as a

[6] Newport (1999) reports the results of experiments on sign language acquisition from poor and inconsistent signers which clearly exhibit exactly this bias to *impose* regularity where there is variation unconditioned by social context or other factors.

consequence of similar evolutionary pathways and linguistic selection pressures. For example, the fact that in attested languages irregularity is associated with high frequency forms is unlikely to be a consequence of a nativized constraint and much more likely to be a universal consequence of the fact that low frequency irregular forms are less likely to be reliably learned by successive generations of first language learners (see Kirby, 2001, for an elegant simulation).

Zuidema (2003) has argued, following Deacon (2007), that if languages have evolved to be learnable this undermines the learnability arguments of Nowak *et al.* (2002) that for speech communities to evolve, the probability of children being able to learn a target grammar must be higher than a 'coherence threshold', below which no single communal grammar, and thus language, can be maintained. He presents a simulation of an iterated learning model in which early generations of learners do not acquire the target language, but a compression-based prior bias for small context-free grammars leads to the evolution of languages which can be acquired accurately by this learning procedure. Thus over generations, the population of learners evolves languages which meet the coherence threshold even though the starting conditions do not. However, to achieve this result, Zuidema must assume that the learners in his population come equipped with an invariant learning algorithm equvalent to that of Horning (1969), as a prior bias for small stochastic context-free grammars is equivalent to a compression-based learner of context-free grammars (see e.g. Rissanen, 1989). Thus, contrary to his claims, the model does not really address Gold's 'in the limit' negative results, because of the assumed inductive bias for smaller grammars. However, the model does show very elegantly how the fit between languages and prior bias is predicted to become very close in many if not all such models (e.g. Griffiths and Kalish, 2007; Kirby *et al.* 2007)[7] The question, I wish to address here is where might this grammar-specific bias have come from, given that it is evolutionarily implausible to assume that it simply emerged *de novo* before the emergence of (proto)language.

4 Genetic Assimilation

Although Pinker and Bloom (1990) and many others use the term 'Baldwin Effect', I prefer Waddington's (1942, 1975) notion of genetic assimilation to describe the process by which changes in the behaviour of a population, i.e. niche construction, can cause changes to the environment of adaptation, and thus create

[7] Kirby *et al.*(2007) argue, contra Griffiths and Kalish, that cultural transmission in the form of an 'information bottleneck' (i.e. exposure to a finite positive sample of a language which doesn't completely determine the target grammar) can overcome prior bias for learners who select the most probable grammar rather than selecting a grammar with a bias determined by the posterior probability distribution over grammars. However, this result is questionable given the need for noise in lingustic production, which essentially reincorporates the effect of posterior biased selection of a grammar in the original simulation.

novel selection pressures on that population. Unlike Baldwin, and others writing before the modern synthesis, Waddington was able to demonstrate experimentally with fruit flies that environmental changes combined with artificial selection for flies that responded in a phenotypically specific way to such changes results in "canalization" of the phenotypic response, i.e. phenotypic plasticity is supplanted by a genetically encoded invariant response in the evolved population, which no longer requires the original environmental stimulus[8] .

Deacon (this vol.) argues that in addition to the unmasking of genes to novel selection pressure demonstrated by Waddington, niche construction may also mask selection for other genes. He gives the example of the loss of the ability in the primate lineage to internally synthesize asorbic acid, as a consequence of masking of selection for a gene which coded for a protein essential to this process caused by adoption of a diet containing fruit and thus an external supply of asorbic acid. Deacon characterizes the process of genetic assimilation as the unmasking of selection pressure on genes coding for cognitive neural mechanisms (e.g. the LAD) as a consequence of niche construction (e.g. the emergence of (proto)language). However, he argues that masking of selection on the genes coding for neural mechanisms and their consequent "relaxation" is a more plausible explanation for our linguistic abilities because "highly distributed synergistic organisation emerges from this type of process" and because "epigenetic parsimony" entails that the genes should only encode what cannot be offloaded to "self-organizing developmental processes" in interaction with the environment (Deacon, this vol.).

If Deacon is right (and the analogy with the development of more complex song in the Bengalese Finch is certainly compelling), then masking in the linguistic niche means that we evolved into a genetically "degenerated ape" rather than a more finely-adapted one. However, even under this scenario, "stabilizing selection" for the suite of epigenetic responses to linguistic stimuli is still required for maintenance of our language learning abilities. So under this scenario, genetic assimilation still plays a role, but a reduced one in which the emergent complexity of language and its acquisition is more a consequence of serendipitous synergies amongst various less-constrained epigenetic developmental processes, rather than of active selection for a genetically-encoded LAD. One possible problem with this account is that it relies on synergies whose probability may not turn out to be much higher than those required by saltationist accounts of the emergence of the

[8] Longa (2006) argues, rather incoherently, that I resort to Waddington's mechanism of genetic assimilation in order to motivate my account of the emergence of a LAD via the Baldwin Effect. He claims that I conflate the two processes and that somehow my arguments and simulation model rest on the parity of the two processes. In fact, I only refer to the Baldwin Effect at all because of its widespread use by others to mean something like genetic assimilation where phenotypic plasticity is supplied by a within-lifetime learning mechanism. I am not particularly concerned with the pre-modern synthesis speculations of Baldwin and others nor with the various (re)interpretations of these speculations, and the coevolutionary model and account of the emergence and maintenance of the LAD in no way rests on them.

LAD (see Pinker and Bloom, 1990). Whether one places the emphasis more on masking or unmasking, it is hard to see why this would impact on the issue of language-specificity given the arguments in 2 and 3 above. Deacon endorses the simulation and modelling work of Yamauchi (e.g. 2001), which Yamauchi argues undermine the plausibility of genetic assimilation. However, in my own modelling work (Briscoe, 2005) replicating Yamauchi's decorrelation of genotypic and phenotypic space within a coevolutionary model of the evolution of the LAD and of languages themselves, I showed that these results rest more on the simplifying assumptions of his model than on any substantive extension of Mayley's (1996) original work on decorrelation and genetic assimilation. I believe the distinction between an unmasking and a masking account reduces to one of causation in an evolutionary (pre)history that we as yet have only very indirect access to. Either way, there is a critical role for genetic assimilation and on balance I believe current evolutionary theory suggests unmasking (i.e. Waddingtonian genetic assimilation) would play the larger causative role in the development of novel traits.

5 Models and Simulations

The value of formal modelling and computational simulation of linguistic evolution and of associated cognitive neural evolution is that it can lend greater precision to argumentation concerning interactions between at least two complex and only partially understood domains. However, if a model supports a particular argument, this does not mean the argument is correct. Rather the required precision and detail needed to make a particular prediction exposes the assumptions, some perhaps implicit, whose plausibility can then be more directly evaluated.

For instance, Deacon (1997) argues, along with others, that genetic assimilation could not have been a significant factor in the development of the LAD because the speed of linguistic evolution so outpaces biological evolution that genes tracking grammatical regularities would not have time to go to fixation in the population before these changed and the associated selection pressure they entailed disappeared. This sounds plausible, but when tested by modelling and simulation turns out to require an unstated assumption that the full range of grammatical possibilities available in the hypothesis space of grammars be manifest during the period of adaptation. No matter how much apparent linguistic change is manifested, if this only covers a proper subset of the hypothesis space of grammars, then there will be selection pressure for genes which constrain the hypothesis space to just this proper subset under the assumptions that this makes learning more robust and efficient and that mastery of language confers a fitness benefit. Of course, this doesn't prove that genetic assimilation of this kind occurred, but it does suggest Deacon's argument is flawed in this form[9].

In Briscoe (2003, 2005) as well as in earlier work referenced there, I review, evaluate and model a number of arguments and models both for and against genetic assimilation of grammatical information, concluding that this remains a coherent and evolutionarily plausible account of the emergence and maintenance of the LAD. One theme that is often implicit but always present in this work is that designing a useful model and deriving results from it is a non-trivial business which, although apparently largely a mathematical and computational exercise, is in fact replete with complex judgements about the appropriate level of abstraction to adopt and what simplifying assumptions it is legitimate to make.

For instance, Christiansen *et al.* (2006), summarized in Kirby *et al.* (this vol.), revisit the relative speed of change argument, albeit without considering either Deacon's arguments or my own work, and present a series of simulations which they argue demonstrate that only functionally motivated features of language can become genetically encoded because of the rapidity of linguistic change compared to biological evolution. They take this as a refutation of Pinker and Bloom's (1990) claim that arbitrary features of language might become encoded in universal grammar (the LAD) to make language learning more robust. The model of learning is based on that of Hinton and Nowlan (1987) and language change is simulated by introducing a new language at each time step of the model. They do not measure the communicative success of the evolving learning agents after each time step and they do not investigate the proportion of the original hypothesis space explored during an average simulation run. The description of the simulation isn't detailed enough to infer either, however, it seems likely that the former will be low, contrary to attested language change, and the latter high when change is not closely correlated with the genetic make-up of the population at the previous time step. It is, therefore, not surprising nor particularly interesting that genetic assimilation does not occur under these conditions. They perform a second simulation run in which agents are selected on the basis of their communicative success where they observe genetic assimilation but take this to mean that only functionally motivated traits can be assimilated. They appear to miss the point that arbitrary features of grammar if assimilated become functional in this sense if they make learning more efficient and thus increase communicative success – reinforcing Pinker and Bloom's original point that "parity" is functional in its own right without any additional "functional" requirements for ease of acquisition, product or comprehension.

Similarly, Reali and Christiansen (submitted) argue that there is evidence of considerable overlap in the cognitive machanisms used in sequential learning and

[9] I'd like to make clear that I focus here on Terry Deacon's work not because I think it is generally flawed but, on the contrary because I find it very stimulating and often very convincing, and this provokes me to evaluate it carefully, even to the extent of building and modifying quite complex computational models

language learning. They evolve the initial weights of a population of simple recurrent neural networks (SRNs) to perform optimally on a sequential learning task. They then used these evolved SRNs to "learn" (in fact, predict string sequences of) languages in an iterated learning model, also allowing the SRN weights to further evolve, subject to the proviso that they maintained the same performance on the original sequential learning task. The result is that languages emerge with consistent head ordering, but the networks themselves do not evolve further. Reali and Christiansen interpret this to mean that a sequential learning mechanism exapted for language learning predicts that languages will evolve in typologically plausible ways without any specific linguistic biases being genetically assimilated. I think this is a potentially interesting claim and line of research which is unfortunately undermined by the use of SRNs, which are incapable, in principle, of grammar learning (see 2 and by the fact that the specific class of SRNs deployed may be unable to even reliably predict the sequences of many languages in the space explored given any possible weight settings. Rather than using the iterated learning paradigm with a population of learners it would have been more informative to demonstrate the predictability of plausible and implausible word order sequences by networks with various weight settings, and then demonstrate that a network optimized for non-linguistic sequential learning incorporates a bias against certain word order sequences.

6 Conclusions

Modelling and simulation are potentially very valuable in such a complex domain of enquiry where the constraints on theory are weak given the available evidence. However, such modelling has a largely negative impact, mostly exposing the flaws and implicit implausible assumptions in arguments. Even to achieve this much, models must meet certain criteria before they become relevant. They must model the acquisition task realistically, track communicative success in many contexts, and make realistic assumptions about rates and types of language change.

To date, I believe that the evolutionarily most plausible account of the emergence and maintenance of the LAD is that a representation language evolved out of the (compositional) 'language of thought' capable of mapping meaning to sequential or spatial realizations which disambiguate argument relations to predicates. Most likely the simplest such mappings, requiring the least additional apparatus, embody substantive constraints on such mappings and thus are low or intermediate on the Chomsky hierarchy of language classes and associated automata. In this sense, the LAD already incorporated grammar/language-specific bias. However, the linguistic niche created new selection pressures for robust and efficient language acquisition, and genetic assimilation provided the mechanism by which adaptations encoding ever more informative prior biases could evolve.

These would most likely be weak biases rather than hard constraints, in the face of continuing linguistic evolution and then subsequent change within the space of modern human languages, and would asymptote at the point where, given such variation in the linguistic environment of adaptation, no further gains were possible or all relevant genetic variation had gone to fixation.

References

Brighton, H. (2002) 'Compositional syntax through cultural transmission', *Artificial Life, vol.8.1,* 25–54.

Briscoe, E.J. (1997) 'Co-evolution of language and of the language acquisition device', *Proceedings of the 35th Assoc. for Comp. Ling.,* Morgan Kaufmann, pp. 418–427.

Briscoe, E. (2003) 'Grammatical Assimilation' in (eds.) M. Christiansen and S. Kirby (ed.), *293–316,* Oxford University Press, pp. Language evolution: the states of the art.

Briscoe, E.J. (2005) 'Coevolution of the language faculty and language(s) with decorrelated encodings' in (ed.) Tallerman, M. (ed.), *Language Origins: Perspectives on Evolution,* Oxford University Press, pp. 310–333.

Chomsky, N. (1965) *Aspects of the theory of syntax,* MIT Press.

Chomsky, N. (1981) *Government and binding,* Foris, Dordrecht.

Christiansen, M. and Reali, F. (2006) 'The Baldwin effect works for functional, but not arbitrary, features of language' in World Scientific Publishing, London (ed.), *Cangelosi, A., Smith, A. and Smith, K.,* 27–34, pp. Proc. of 6th Conf. on Evolution of Language.

Cosmides, L. and Tooby, J. (1996) 'Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgement under uncertainty', *Cognition, vol.58,* 1–73.

Deacon, T. (1997) *The symbolic species: coevolution of language and brain,* MIT Press, Cambridge MA.

Deacon, T. (this vol.) *Relaxed selection and the role of epigenesis in the evolution of language,*.

Gazdar, G., Klein, E., Pullum, G., and Sag, I. (1985) *Generalized Phrase Structure Grammar,* Blackwell, Oxford, UK.

Geisler, W. and Diehl, R. (2003) 'A Beyesian approach to the evolution of perceptual and cognitive systems', *Cognitive Science, vol.27,* 379–402.

Gold, E. (1967) 'Language identification in the limit', *Information and Control, vol.10,* 447–474.

Griffiths, T. and Kalish, M. (2007) 'Language evolution by iterated learning with Bayesian agents', *Cognitive Science, vol.31,* 441–480.

Hauser, M., Chomsky, N. and Fitch, W. (2002) 'The Faculty of Language: What Is It, Who Has It, and How Did It Evolve?', *Science, vol.298,* 1569–1579.

Hinton, G. and Nowlan, S. (1987) 'How learning can guide evolution', *Complex Systems, vol.1,* 495–502.

Horning, J. (PhD, Computer Science Dept., Stanford University) *A study of grammatical inference,.* 1969

Kirby, S. (2001) 'Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity', *IEEE Transactions on Evolutionary Computation, vol.5.2,* 102–110.

Kirby, S. (2002) 'Learning, bottlenecks and the evolution of recursive syntax' in E. Briscoe (ed.), *Language Acquisition and Linguistic Evolution: Formal and Computational Approaches,* Cambridge University Press, Cambridge, pp. 173–204.

Kirby, S., Dowman, M. and Griffiths, T. (2007) 'Innateness and culture in the evolution of language', *Proc. of the Nat. Academy of Sciences, vol.104,* 5241–5245.

Kirby, S., Christiansen, M. and Chater, N. (this vol.) *Syntax as an adaptation to the learner,.*

Li, M. and Vitanyi, P. (1997) *An Introduction to Kolmogorov Complexity and its Applications,* Springer-Verlag, Berlin.

Longa, V. (2006) 'A misconception about the Baldwin effect', *Folia Linguistica, vol.40,* 305–318.

Mayley, G. (1996) 'Landscapes, learning costs and genetic assimilation' in Peter Turney, Whitley, D., and Anderson, R. (ed.), *Evolution, learning and instinct: 100 years of the Baldwin effect,* MIT Press, Cambridge MA.

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

Newport, E. (1999) 'Reduced input in the acquisition of signed languages: Contributions to the study of creolization' in M. DeGraff (ed.), *Language Creation and Language Change: Creolization, Diachrony, and Development,* MIT Press, Cambridge MA.

Nowak, M., Komarova, N., and Niyogi, P. (2002) 'Computational and evolutionary aspects of language', *Nature, vol.417,* 611–617.

Oliphant, M. (2002) 'Learned systems of arbitrary reference: the foundation of human linguistic uniqueness' in E. Briscoe (ed.), *Language Acquisition and Linguistic Evolution: Formal and Computational Approaches,* Cambridge University Press, Cambridge, pp. 23–52.

Pigliucci, M., Murren, C. and Schlichting, C. (2006) 'Phenotypic plasticity and evolution by genetic assimilation', *J. Exp[erimental Biology, vol.209,* 2362–2367.

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

Pullum, G. and Scholz, B. (2002) 'Empirical assessment of stimulus poverty arguments', *The Linguistic Review, vol.19.1-2,* 1–50.

Reali, F. and Christiansen, M. (submitted) ‘Sequential learning and the interaction of biological and lingustic adaptation in language evolution’, *Interaction Studies, vol.*.

Rissanen, J. (1989) *Stochastic Complexity in Statistical Inquiry,* World Scientific, Singapore.

Sampson, G. (1989) ‘Language acquisition: growth or learning?’, *Philosophical Papers, vol.XVIII.3,* 203–240.

Sampson, G. (1999) *Educating Eve: The Language Instinct Debate,* Continuum International, London.

Shieber, S. and Lappin, S. (2007) ‘Machine learning theory and practice as a source of insight into universal grammar’, *J.Linguistics, vol.43,* 1–34.

Staddon, J. (1988) ‘Learning as inference’ in Evolution and Learning (ed.), *Bolles, R. and Beecher, M.,* Lawrence Erlbaum, Hillside NJ..

Waddington, C. (1942) ‘Canalization of development and the inheritance of acquired characters’, *Nature, vol.150,* 563–565.

Waddington, C. (1975) *The evolution of an evolutionist,* Edinburgh: Edinburgh University Press.

Wanner, E. and Gleitman, L. (1982) ‘Introduction’ in E. Wanner and L. Gleitman (ed.), *Language acquisition: the state of the art,* MIT Press, Cambridge MA, pp. 3–48.

Yamauchi, H. (2001) ‘The difficulty of the Baldwinian account of linguistic innateness’ in J. Keleman and P. Sosik (ed.), *Advances in Artificial Life: Proceedings of the 6th European Conference on Artificial Life,* Springer-Verlag, Heidelberg.

Zuidema, W. (2002) ‘How the poverty of the stimulus solves the poverty of the stimulus’ in Advances in Neural Information Processing Systems 15 (Proceedings of NIPS’02) (ed.), *(eds.) S. Becker, S. Thrun and K. Obermayer,* Cambridge, MA: MIT Press.