# Grammatical Assimilation

Ted Briscoe
Computer Laboratory
University of Cambridge

## 1 Introduction

In this paper, I will review arguments for and against the emergence and maintenance of an innate language acquisition device (LAD) via genetic assimilation. By a LAD, I mean nothing more or less than a learning mechanism which incorporates some language-specific inductive bias.[1] I will adopt a coevolutionary model in which natural languages are treated as complex adaptive systems undergoing often conflicting selection pressures, some of which emanate from the LAD, which itself evolved in response to (proto)languages in the environment of adaptation

Of course, the existence of an innate LAD has not gone unquestioned, and it is certainly the case that many arguments that have been proposed in its favour are either questionable or wrong (e.g. Pullum and Scholz, forthcoming; Sampson, 1989, 1999). It is not my intention to review this debate here. However, I will argue that all remotely adequate extant models of grammatical acquisition that have been proposed presuppose a LAD, and that genetic assimilation is the only coherent account of its emergence and maintenance. These arguments, even if flawless, do not constitute a proof either of the existence of an innate LAD, or

[1] The term inductive bias is utilised in the field of machine learning to characterise 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 cost metric or prior probability distribution on hypotheses (e.g. Mitchell, 1997:39f).

that it emerged by genetic assimilation. However, they do suggest that the onus is on non-nativists to demonstrate an adequate detailed account of grammatical acquisition which does not rely on a LAD, and on non-assimilationists to propose a detailed and plausible alternative mechanism for the evolution of the LAD.

I use the more succinct phrase 'grammatical assimilation' as shorthand for the more ponderous 'genetic assimilation of grammatical information into the LAD'. The general concept of genetic assimilation is described and discussed in more detail in section 4, where some arguments for and against grammatical assimilation are also presented. Section 2 reviews work on grammatical acquisition and presents the case for assuming the existence of the LAD. Section 3 outlines an account of languages as complex adaptive systems and spells out several consequences for models of grammatical assimilation. Section 5 describes and evaluates extant simulations of genetic and grammatical assimilation, and section 6 describes my model and reports some new experiments with it. Much of this paper is a re-presentation of my own research and simulations addressing the issue of grammatical assimilation (Briscoe, 1997, 1998, 1999, 2000a,b, 2001). However, I also discuss and compare this to related work by Batali (1994), Kirby and Hurford (1997), Livingstone and Fyfe (2000), and Turkel (2001), and I present some extensions of my earlier work which reexamine the issue of the degree of correlation between genotype and phenotype and its likely impact on grammatical assimilation (Mayley, 1996; Yamauchi, 2000a,b).

## 2 Grammatical Acquisition

Adequate accounts of grammatical acquisition during first language learning must satisfy at least the following desiderata. Firstly, there is the desideratum of *coverage*: models should support acquisition of any attested grammatical system and adequately characterise the range of possible mappings from meaning to form in attested systems. This rules out much work which purports to address the issue of grammatical acquisition, for example, all work based on (recurrent) neural networks as models of such systems. A reasonable requirement, given current knowledge, is that the model be capable of acquiring a proper subset of indexed languages, including those exhibiting cross-serial dependencies (e.g. Joshi *et al* 1991). Secondly, models must work with *realistic input*: grammatical acquisition is based on finite positive but noisy input; that is, learners are exposed to a finite sequence of utterances drawn from mixed and non-stationary sources, so speech communities are never totally homogeneous nor static (e.g. Milroy, 1992). Many models instead assume a single non-noisy stationary source, or equivalently a finite sequence of 'triggers' drawn from the target grammar to be acquired (e.g. Gibson and Wexler, 1994). Thirdly, models should work with *realistic input enrichment*: many assume that each 'trigger' is paired reliably with its correct meaning (logical form) and that the learner never hypothesises an incorrect pairing. Such assumptions may facilitate formal learnability results for inadequate algorithms, but they presuppose implausibly that the context of utterance during learning is always highly determinate and redundant – or at least when it isn't, it is reliably recognisable as such, so that the learner knows when to ignore input (e.g. Osherson *et al* 1986:100). Fourthly, there is *selectivity* in acquisition:

learners do not acquire (maximum likelihood style) 'covering' grammars of the input, but rather reject noise and other random or very infrequent data in favour of a single consistent grammar (e.g. Lightfoot, 1999) – modulo sociolinguistic variation, which is neither random nor infrequent when conditioned on the appropriate sociologuistically defined context. Fifthly, *accuracy* is critical: learners do not 'hallucinate' or invent grammatical properties regardless of the input, though they do (over)generalise and, in this limited sense, 'go beyond the data'. Accuracy can be operationalised in terms of a formal learnability requirement, at least under the simplifying assumption that the learner is only exposed to input from the target grammar (e.g. Wexler and Culicover, 1980; Niyogi, 1999)

If accuracy is defined in terms of a learnability requirement from realistic finite positive but noisy sentence-meaning pairs or 'triggers' over a hypothesis space with adequate coverage, even when drawn from a single stationary target grammar, then some form of inductive bias in the acquisition model is essential. In most current work on grammatical acquisition, inductive bias takes the form of a restricted finite hypothesis space of grammars within which individual grammars are selected by setting (finite-valued) parameters. There may also be additional bias in terms of default initial settings for a subset of parameters, creating a preference ordering on grammars in the hypothesis space (e.g. Chomsky, 1981). Models of this form, which do not incorporate a statistical or quantitive component, are not able to deal adequately with noisy input (e.g. Briscoe, 1999, 2001). There is a well-known formulation of inductive bias in terms of Bayesian statistical learning theory (e.g. Mitchell, 1997:154f). Bayes theorem provides a general formula and justification for the integration of prior bias with experience and it has been demonstrated that an accurate Bayesian prior supports learn-

ability from finite noisy data over infinite hypothesis spaces (e.g. Horning, 1969; Muggleton, 1996).

Bayesian learning theory is a general domain-independent formulation of learning. The LAD, as defined in section 1, is a language-specific endowment. So it is at least questionable whether Bayesian models of grammatical acquisition presuppose a LAD. 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 Vityani, 1993). 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. Horning's (1969) work is based on the (infinite) class of context-free grammars, which violates the coverage desideratum introduced above as cross-serial dependencies are not covered. However, Muggleton's (1996) proof is defined over a restricted form of stochastic logic program which does meet the coverage desideratum. Furthermore, both Horning and Muggleton require that the prior distribution over grammars in the hypothesis space is accurate, in the sense that it defines a preference metric over hypotheses that leads the learner to the correct target grammar given realistic input. Gold's (1967) original negative 'in the limit' learnability results are founded on the intuition that any amount of positive data from a target grammar in a class containing grammars capable of generating an infinite set of sentences is always compatible with a hypothesised 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. A prior distribution

or cost metric encoding a preference for smaller, more compressed grammars will, in general, select ones that predict the grammaticality of supersets 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. 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.

This last argument is sufficiently important that a concrete example is warranted. Consider a potential class of languages consisting of clauses constructed from a verb (V) and optionally a subject (S) and object (O), where S and O are always realised 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 sentences 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 goes with 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 $S_i V_i O_i S_j V_j O_j$
(e.g. cats like dogs$_i$ who$_i$ like cats)
b $S_i V_i O_i S_j V_j O_j$
(e.g. who$_i$ like dogs cats$_i$ like cats)
c $S_i V_j O_j S_j V_i V_k O_k S_k O_i$
(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.

This class of languages is a proper subset of the context-free languages (CFLs), as intersecting but not nested dependencies are prevented by the stipulations above. 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 learning sample, 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).[2]

(2) a Sent $\rightarrow$ NP$^S$ V NP$^O$
b NP $\rightarrow$ NP Sent
c NP $\rightarrow$ 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 rel-

[2]Once again, I use superscripted S and O and subscripted indices to show the mapping to predicate argument structure and leave that required to characterise the predicate-argument structure of sentences containing relative pronouns implicit. The details of how this mapping is actually realised 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.

ative clauses. (Given this cost metric, the learner could equally well learn a 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 $\rightarrow$ $N^S$ V $N^O$
b Sent $\rightarrow$ $N_i^S$ $V_i$ $N_i^O$ $N_j^S$ $V_j$ $N_j^O$

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 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).[3]

If the learning sample also includes (1b), containing a prenominal subject-modifying relative clause, then a learner utilising grammar (2) might acquire a further right-recursive rule analogous to (2b), predicting complementary distribution of pre- and post-modifying relative clauses. Whilst one utilising (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 post-nominal 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 above cost metric exposed to SVO unembedded sequences and (1c) would acquire grammar (4) with a total cost of

[3]The point is not, new of course. Chomsky (1965:38) recognises the need for an evaluation measure based on simplicity to choose between grammars during language acquisition, and others criticised the arbitrariness of such measures. Kolmogorov Complexity (e.g. Li and Vitanyi, 1993) and the related Minimum Description Length 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. Osborne and Briscoe, 1997; Ristad and Rissanen, 1994), but with restricted representation languages. These complexities are ignored here to keep the example simple as they do not alter the fundamental point about the domain-dependence of cost metrics or prior distributions defined over restricted hypothesis represenation languages.

16.

(4) a Sent $\rightarrow$ NP$^S$ V NP$^O$
b RC $\rightarrow$ V NP$^O$ NP$^S$
c NP $\rightarrow$ NP RC
d NP $\rightarrow$ N

Thus, the 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 so-called IDLP notation instead of standard CFG, acquiring immediate dominance (ID) rules independently of linear precedence (LP) rules (Gazdar *et al* 1985), but utilising a similar cost metric which also assigns a cost of one to each LP rule, then the preference ordering on specific IDLP grammars predicts that order-free variants of the above grammars with no LP rules will be hypothesised and that the inclusion of examples like (2b) or (2c) in the learning sample will not alter the learner's hypothesis.

Cost metrics applied to restricted representation languages mean that learners will 'go beyond the evidence' in different ways and, thus, will have different linguistic biases. However, learners that do not utilise cost metrics, or equivalently prior distributions, cannot learn target grammars for human languages accurately, as Gold's (1967) work demonstrated. There are no detailed models of grammatical acquisition utilising an unrestricted representation language with a domain-independent cost metric or 'universal prior'. Extant models assume a LAD, in the (weak) sense of section 1, because they utilise prior distributions or cost metrics defined over restricted representations chosen to facilitate encoding of grammars for human languages. The onus is on non-nativists to develop an ac-

count of grammatical acquisition which meets the above desiderata and does not utilise a LAD. The Bayesian learning framework provides a general and natural way to understand and model how further grammatical bias can be integrated with the language acquisition procedure 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 (e.g. Staddon, 1988; Cosmides and Tooby, 1996). Finally, there is independent psycholinguistic evidence that human language learners are biased in lingustically-specific ways; 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 errors where languages, for example, mark negation morphologically.

# 3 Linguistic Evolution

Learnability frameworks typically assume that the learning sample is generated by a fixed target grammar and accuracy or success is defined in terms of acquisition of this grammar. However, as we noted, in section 2, first language learners are not typically exposed to homogeneous data from an unchanging speech community. Though major and rapid grammatical change is relatively rare, learners typically hear utterances produced by members of other speech communities, and the learning period is sufficiently extended that they may be exposed to ongoing linguistic change within a single community. A major tenet of generative diachronic linguistics is that first language acquisition is the main engine of grammatical change because, faced with such mixed data, learners do acquire grammars that are distinct from those of the previous generation (e.g. Lightfoot,

1999).

We can model the development of the (E-)language of a speech community as a dynamical system in which states encode the distribution of grammars (and lexicons) within the community and transitions between states are defined in terms of changes in this distribution (Briscoe, 2000b; Niyogi and Berwick, 1997). If there is inductive bias in first language acquisition (regardless of its provenance), then E-languages are best characterised as adaptive systems, because learners will preferentially select linguistic variants which are easier to learn and thus more adaptive with respect to the acquisition procedure (Briscoe, 2000b; Kirby, 1998). However, linguistic selection of this kind does not come exclusively from language acquisition. There are other often conflicting selection pressures created by the exigencies of production and comprehension which mean that the fitness (or adaptive) landscape for languages is complex and dynamic with no fixed points or stable attractors (Briscoe, 2000b). For example, a functional pressure for more parsable linguistic variants (Briscoe, 2000a; Kirby, 1999) may be counterbalanced by a social pressure to produce innovative variants (Nettle, 1999) or a functional pressure to produce shorter utterances (Lindblom, 1998). Thus individual languages are complex adaptive systems on rugged and multipeaked fitness landscapes, in the sense of, for example, Kaufmann (1993).

Even restricting consideration to major grammatical change, which is only 'fixed' through first language acquisition, it seems clear that linguistic evolution proceeds via cultural transmission (i.e. first language acquisition) at a faster rate than biological evolution. The populations involved are smaller (speech communities rather than entire species), and language acquisition appears to be 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 clearly delineated or constrained than the biological mechanisms of genetic evolution.

Several consequences emerge from the evolutionary account of (E-)languages as (complex) adaptive systems which must be taken into consideration by any plausible account of grammatical assimilation. 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 signal-meaning correspondences and the (subsequent) emergence of protolanguage(s) with decomposable (minimally grammatical) sentence-meaning correspondences (e.g. Oliphant, 2001; Kirby, 2001a). They conclude that the language acquisition procedure must incorporate inductive bias causing generalisation, and consequent regularisation of the input, in order that repeated rounds of cultural transmission of language regularise random variations into consistent and coherent communication systems. Newport (1999) reports the results of experiments on sign language acquisition from poor and inconsistent signers which clearly exhibits exactly this bias to *impose* regularity where there is variation unconditioned by social context or other factors. Secondly, the account of languages as adaptive systems entails that linguistic universals no longer constitute strong evidence for a LAD. Deacon (1997), Kirby and Hurford (1997) and others make the point that universals may equally be the result of convergent evolution in different languages as a consequence of similar evolutionary pathways and linguistic selection pressures. For example, to return to the first point, the fact that in attested languages irregularity is associated

with hih frequency forms is unlikely to be a consequence of a nativised constraint and much more likely to be a universal consequence of

# 4 Grammatical Assimilation

If there is a LAD, then it is legitimate to ask how this unique biological trait emerged. There are only two clearly distinct possibilities compatible with modern evolutionary theory: some degree of exaptation of preexisting traits combined with saltation and/or genetic assimilation (e.g. Bickerton, 2000).

Genetic assimilation is a neo-Darwinian (and not Lamarckian) mechanism supporting apparent 'inheritance of acquired characteristics' (e.g. Waddington, 1942, 1975). The fundamental insights are that: 1) plasticity in the relationship between phenotype and genotype is under genetic control, 2) novel environments create selection pressures which favour organisms with the plasticity to allow within-lifetime developmental adaptations to the new environment, 3) natural selection will function to 'canalize' these developmental adaptations by favouring genotypic variants in which the relevant trait develops reliably on the basis of minimal environmental stimulus, providing that the environment, and consequent selection pressure, remains constant over enough generations.[4] Waddington, himself, suggested that genetic assimilation provided a possible mechanism for the gradual evolution of a LAD: 'If there were selection for the ability to use language,

[4]Waddington's work on genetic assimilation is a neo-Darwinian refinement of an idea independently proposed by Baldwin, Lloyd Morgan and Osborne in 1896, and often referred to as the Baldwin Effect (see Richards, 1987 for a detailed history). Waddington refined the idea by emphasizing the role of canalization and the importance of genetic control of ontogenetic development – his 'epigenetic theory of evolution'. He also undertook experiments with *Drosphila subobscura* which directly demonstrated modification of genomes via artificial environmental changes (see Jablonka and Lamb, 1995:31f for a detailed and accessible description of these experiments).

then there would be selection for the capacity to acquire the use of language, in an interaction with a language-using environment; and the result of selection for epigenetic responses can be, as we have seen, a gradual accumulation of so many genes with effects tending in this direction that the character gradually becomes genetically assimilated.' (1975:305f). Pinker and Bloom (1990) briefly make essentially the same suggestion, citing the 'Baldwin Effect' and Hinton and Nowlan's (1987) computational simulation showing genetic assimilation of initial node settings facilitating learning in a population of neural networks.

The account proposed in Briscoe (1999, 2000a) is that an initial acquisition procedure emerged via recruitment (exaptation) of preexisting (preadapted) general-purpose (Bayesian-like) learning mechanisms to a specifically-linguistic mental representation capable of expressing mappings from conceptual representations to realisable, essentially linearised, encodings of such representations (e.g. Bickerton, 1998, 2000; Worden, 1998). The selective pressure favouring such a development, and its subsequent maintenance and refinement, is only possible given a coevolutionary scenario in which some protolanguage(s) supporting successful communication and capable of cultural transmission (that is, learnable without a LAD) within a hominid population had already evolved via cultural transmission (e.g. Kirby, 1998; Deacon, 1997). Protolanguage(s) may have been initially similar to those proposed by Wray (2000) in which complete propositional messages are conveyed by undecomposable signals. However, to create selection pressure for the emergence of grammar, and thus a LAD incorporating language-specific grammatical inductive bias, protolanguage(s) must have evolved at some point into decomposable utterances, broadly of the kind envisaged by Bickerton (1998). Several accounts of the emergence of syntax have been developed that predict the

emergence of syntax (e.g. Kirby, 2001a,b; Nowak *et al* 2000). At the point when the environment contains language(s) with minimal syntax, grammatical assimilation becomes adaptive, under the assumption that language confers a fitness advantage on its users, since assimilation makes grammatical acquisition faster and more reliable.

Saltations or macromutations are compatible with evolutionary theory if a single highly-adaptive change in genotype creates a large change in phenotype. Evolutionary theory predicts that macromutations are extremely unlikely to be adaptive (Dennett (1995:282f). Saltationist accounts have been proposed by Chomsky (1988), Gould (1991), Bickerton (1998), Berwick (1998), Lightfoot (2000) and others who variously speculate that the LAD emerged rapidly, in essentially its modern form, as a side-effect of the development of large general-purpose brains (possibly in small heads) and/or sophisticated conceptual representations. These accounts not only speculate that the LAD emerged in a single and extremely unlikely evolutionary step (e.g. Pinker and Bloom, 1990), but also neglect the fact that selection pressure is required to *maintain* a biological trait (e.g. Ridley, 1990). Without such selection pressure, we would expect a trait to be whittled away by accumulated random mutations in the population (i.e. genetic drift, e.g. Maynard-Smith, 1998:24f). However, with such selection pressure, a newly emerged trait will probably continue to adapt, especially if the environmental factors creating the selection pressure are themselves changing, as languages do. A saltationist account, then, requires the assumption that language, and consequently the ability to learn one fast and reliably with a LAD, confers an adaptive advantage just as much as a gradualist account requires the same assumption. Therefore, even if the first LAD emerged by macromutation, evolutionary theory

predicts it may have been further refined by genetic assimilation.

Nevertheless, a number of potential problems have been raised for accounts of either the emergence or subsequent evolution of the LAD via grammatical assimilation, even though some such account appears to be the only possibility which is both plausible and consistent given general evolutionary considerations.

Newmeyer (2000) goes one stage further than other saltationists, arguing that, given the assumptions that: 1) the LAD incorporates a universal grammar based on government-binding theory (Chomsky, 1981); 2) the language(s) extant in the environment of adaptation were exclusively SOV rigid order languages with grammatical properties similar to their attested counterparts; and 3) such attested languages do not manifest most of the universal linguistic constraints posited in government-binding theory, then the LAD, if it exists, could not have emerged as a result of grammatical assimilation and must be the result of saltation. All three of these assumptions are highly controversial. However, the definition of the LAD in terms of inductive bias, developed in section 2 is in no way dependent on any specific linguistic constraints, or even on the existence of any *absolute* linguistic universals, and none of the argumentation in this paper rests on anything like such strong and speculative assumptions about the prehistoric environment of adaptation.

Lightfoot (1999, 2000) argues that the LAD is not fully adaptive and, therefore, could not have evolved by gradual genetic assimilation since, by definition, this is an adaptive process. He uses the example of the putative universal constraint against some forms of subject extraction from tensed embedded clauses, which prevents the asking of questions like: **Who do you wonder whether/how solved the problem?*. Lightfoot argues that such phenomena show that aspects of the

LAD are dysfunctional since the constraint reduces the expressiveness of human languages and provides evidence that the constraint is circumvented by various ad hoc strategies in different languages – in English, such questions become grammatical if the normally optional complementiser *that* is obligatorily dropped: *Who do you think (*that) solved the problem?*. He argues that the presence of such a maladaptive constraint in the LAD entails that the LAD could not have evolved gradually, even though this constraint is a by-product of an adaptive more general condition on extraction. However, evolutionary theory does not predict that traits will be or will remain optimal in all environments. Although evolution is an optimisation process, complex and dynamic fitness landscapes typically contain many local optima which are far more likely to be discovered than the global optimum, should it exist (e.g. Kauffman, 1993). It may simply be that any genetically encodable extraction constraint aiding parsability and/or learnability also has unwanted side-effects for expressiveness. Furthermore, a dynamic fitness landscape entails that even an optimal solution at one time can become suboptimal. For instance, one might speculate that this constraint on extraction was assimilated into the LAD in an environment in which protolanguage(s) did not exhibit embedded clauses at all (e.g. Carstairs-McCarthy, 2000). Subsequently, when embedded clauses evolved through linguistic selection for greater expressiveness, aspects of the previously fully adaptive constraint became maladaptive, but by this stage this component of the LAD had gone to fixation in the population.

Waddington, himself, (1975:307) noted that if there is an adaptive advantage to attenuating the process of grammatical acquisition, then we might expect assimilation to continue to the point where no learning (plasticity) would be re-

quired, because a fully specified grammar had been genetically assimilated. In this case acquisition would be instantaneous and fitness would be maximised in a language-using population. Clearly, this hasn't happened, as there are around 6000 attested languages each with (varyingly) different grammatical systems, as well as distinct vocabulary. Given a coevolutionary scenario, in which languages themselves are complex adaptive systems, a likely explanation for continuing grammatical diversity is that social factors favouring innovation and diversity create linguistic selection pressure (e.g. Nettle, 1999). Genetic transmission, and thus assimilation, will be much slower than cultural transmission, therefore, continued plasticity in grammatical acquisition is probable, because assimilation will not be able to 'keep up with' all grammatical change, and too much assimilation will reduce individuals' fitness, if linguistic change subsequently makes it hard or impossible for them to acquire an innovative grammatical (sub)system.

Deacon (1997) and Worden (2001) also assume a coevolutionary scenario, but argue that genetic assimilation of specifically linguistic, grammatical information is unlikely precisely because languages evolve far faster than brains. If languages evolve one or more orders of magnitude faster than brains, since attested languages have shifted major grammatical system within 1000 years which is a mere 50 or so generations, then, they argue, it is far more likely that grammatical systems have evolved to be learnable by a preexisting general-purpose learning mechanism than that this mechanism adapted to language. One might question this argument on the grounds that the relevant hominid population was in all likelihood small, and therefore genetic evolution would have been faster, whilst linguistic evolution might well have been slower, particularly if there were close contact between most members of this population in the environment of adap-

tation. However, the main weakness of the argument is that it fails to take any account of the potential space of possible grammatical systems. Most linguists would probably argue that this space was potentially infinite prior to the emergence of a LAD, whilst the claim that the LAD restricts the class of possible grammars of human languages to be finite, even if vast (on the order of 30 million grammars), remains controversial (e.g. Pullum, 1983). In this case, no amount of rapid change between attested grammatical systems can count as evidence against grammatical assimilation of linguistic constraints that rule out the (infinitely) many unattested grammars that could not have been sampled in the period of evolutionary adaptation (Briscoe, 2000a).

Given that grammatical assimilation only makes sense under a coevolutionary scenario in which (proto)language(s) create selection pressure and given that languages change, Waddington's notion of genetic assimilation should probably be replaced with the more general one of coevolution (e.g. Futuyma and Slatkin, 1983; Kauffman, 1993:242f) which is taken to include the possibility of gene-culture interactions (e.g. Durham, 1991). Even without considering putative gene-culture interactions which may, for example, underlie the higher incidence of lactose intolerance in humans from cultures with little history of animal husbandry and milk consumption, it is clear that general evolutionary considerations do not rule out the possibility of cross-species gentic interactions, and thus genetic assimilation, in situations where there is considerable asymmetry in the speed of evolution. Host-parasite coevolution is well-documented and in one influential theory is the explanation for the evolution of sexual reproduction, and thus potentially faster recombination of favourable mutations, to enable hosts to evolve parasitic defences more quickly (e.g. Maynard-Smith, 1998:234). From

this perspective, genetic assimilation of environmental information is not precluded by the fact that aspects of the the environment, in this case the parasite, are changing faster than the host organism can evolve. However, the coevolutionary perspective does imply that biological evolution is not guaranteed to succeed: an assimilative response to the environment may well become maladaptive if the environment subsequently changes. This in turn has led to the hypothesis that sexual reproduction also preserves genetic heterogeneity, which helps preserve the plasticity of immune systems and thus a species' ability to respond to parasites. The logic of this argument is analogous to the argument given above for preserving plasticity in the LAD in the face of grammatical change. However, the putative mechanism is reduced genetic specification rather than more heterogeneous specification as is the case with the immune system.

The remaining problems for grammatical assimilation accounts concern the assumed relationship between genotype and phenotype. Biological pathways from genome to phenotype are not well understood for most traits, and certainly not for the putative LAD. This raises two issues. Firstly, has there been enough time since hominid divergence from chimps for the gradual evolution of the LAD? And secondly, is there enough correlation between the genetic encoding of the LAD and its (variant) behaviour to support grammatical assimilation?

There have probably been about 400,000 generations since hominids diverged from chimps. There is an upper bound to the rate at which evolution can alter the phenotype of a given species. The rate of evolution of any trait is dependent on the strength of the selection pressure for that trait, but too much selection pressure and a species will die out. Estimates of the the upper bound vary between 1 and 400 bits of new information per generation (Worden, 1995; Mackay,

1999) creating an upper bound of between 4Kbits and 160Mbits of new genetic information expressed in the species' phenotype. If the correct answer is close to the lower estimate then this places severe demands on any account of the emergence of a species-specific LAD, and means that exaptation of preexisting cognitive apparatus will play a critical part of any plausible gradualist scenario. On the other hand, if the higher estimate is closer to the truth, then it appears that there has been time for the *de novo* evolution of quite complex cognitive traits. Finally, the logic of the speed-limit argument collapses, given a saltationist account based on macromutation. Under this scenario, a single genetic mutation event brings about a complex of extremely unlikely but broadly adaptive phenotypic changes which spreads rapidly through the population.

The second and related argument is based on the observation that the relationship between genes and traits is rarely one-to-one and that epistasis (or 'linkage') and pleiotropy are the norm. In general, the effect of epistasis and pleiotropy will be to make the pathways more indirect from selection pressure acting on phenotypic traits to genetic modifications increasing the adaptiveness of those traits. Therefore, in general terms, we would expect a more indirect and less correlated genetic encoding of a trait to impede or perhaps even prevent genetic assimilation. Mayley (1996) presents a general exploration of the effects of manipulating the correlation between genotype (operations) and phenotype (operations) on genetic assimilation. In his model, individuals are able to acquire better phenotypes through 'learning' (or another form of within lifetime plasticity), thus increasing their fitness. However, the degree to which the learnt phenotype can be assimilated into the genotype of future generations, thus attenuating learning and/or increasing its success and further increasing fitness, depends on the cor-

relation between possible 'moves' in learning 'space' between variant phenotypes and mutations in genotype 'space' encoding these phenotypes.

# 5 Simulations of Genetic and Grammatical Assimilation

One way to explore the arguments and counter arguments outlined in the last section is to build a simulation and/or a mathematical model. The latter is, in principle, preferable as analytic models of dynamical systems yield absolute results, whilst those generated by stochastic computational simulation are statistical (e.g. Renshaw, 1991). However, to date, no analytic model of grammatical assimilation has been developed.

There are a number of commonalities between these simulation models that I will introduce before briefly describing general simulations of genetic assimilation and then describing and evaluating extant simulations of grammatical assimilation. Each model consists of an evolving population of individuals. Individuals are endowed with the ability to acquire a trait by learning. However, the starting point for learning, and thus individuals' consequent success is determined to an extent by an inherited genotype. Furthermore, the fitness of an individual, that is the likelihood with which individuals will produce offspring, is determined by their successful acquisition of the trait. Offspring inherit starting points for learning (genotypes) which are based on those of their parents. Inheritance of *starting* points for learning prevents any form of Lamarckian inheritance of acquired characteristics, but allows for genetic assimilation, in principle. Inheritance either takes the form of crossover of the genotypes of the parents, resulting in a shared mixed inheritance from each parent, and overall loss of variation in

genotypes over generations, and/or random mutation of the inherited genotype, introducing new variation.

Hinton and Nowlan (1987) describe the first computational simulation of genetic assimilation. In their (very abstract) simulation a population of 1000 neural networks with 20 potential connections which can be unset (?), on (1), or off (0) was evolved using a genetic algorithm. A 'successful' neural net had all 20 connections on, but networks were initialised randomly with connection ('gene') frequencies of 0.5 for ? and 0.25 for 1 or 0 at each position. Each network was able to set unset connections through learning on the basis of 1000 trials during its lifetime. The fitness of a network was defined as $1 + 19n/1000$ where $n$ is the number of trials after it has learnt the correct settings, making a network with all connections initially set to on 20 times fitter than a network which never learnt to set them correctly. Reproduction of offspring was by crossover of *initial* connections from two parents whose selection was proportional to their fitness. In the early generations most networks had the same minimum fitness through being born with one or more off settings, however this soon gave way to exponential increases in networks with more on settings, less unset settings and no off settings, then in the later stages the increase of on settings and decrease of unsets tailed off once the population had evolved to genotypes enabling successful learning.

Hinton and Nowlan point out that the fitness landscape for this model is like a needle in a haystack: only one final setting of all 20 connections confers any fitness advantage whatsoever. Therefore, evolution unguided by learning would be expected to take on the order of $2^{20}$ trials (i.e. genotypes) to find a solution. If increased fitness required evolution of two such networks in the same generation, as would be the case for coordinated communicative behaviour, evolution would

be expected to take around $2^{400}$ trials to find a solution (without even considering further restrictions created by limits on population size). However, with learning (modelled as random search of connection settings) the model always converges within 10-15 generations on a viable genotype (i.e. after generating 100-150K networks). Furthermore, the model shows clearly that once successful networks appear, their superior performance rapidly leads to the spread of genotypes which support successful learning. However, networks with unset connections persist for over 500 generations despite the pressure exerted by the fitness function to minimise the number of learning trials required to find the solution (Harvey, 1993). Hinton and Nowlan suggest that this is a result of weak selection pressure once every network is capable of successful learning. Harvey (1993) analyses the model using the tools of population genetics and argues that, since many settings in genotypes of successful networks derive from the genotype of the first such successful network to emerge, there is a significant chance factor in the distribution of on and unset initial connections within the population at each connection site. Given this 'bottleneck' factor when a single successful genotypes evolves and dominates subsequent generations, it is possible for an unset connection to become 'prematurely' fixated, despite the selective pressure exerted by the fitness function in favour of shorter learning periods. The use of a mutation operator would presumably allow populations to converge to the optimum genotype, provided that selection pressure was strong enough to curtail the effects of subsequent random mutation and genetic drift.

This initial result has been extended by Ackley and Littman (1991), Cecconi *et al*(1995) and French and Messenger (1994), variously demonstrating genetic assimilation can occur without a fixed externally-defined fitness criterion, can result

in complete assimilation of a trait where learning has a significant cost and the environment remains constant, and, when this occurs, can result in loss of the now redundant learning component through deleterious genetic drift. An important caveat on these positive results is that Mayley (1996) demonstrates that assimilation can be slowed and even stopped if the degree of neighbourhood correlation between genotype and phenotype is reduced. In Mayley's model individuals have separate encodings of genotype and corresponding phenotype. Learning alters the latter, whilst the directness of the encoding of phenotypes in genotypes and the relationship between learning rules and genetic operators determines the degree of genetic assimilation possible, in interaction with the shape of the fitness landscape and the cost of learning.

The first computational simulation of grammatical assimilation is that of Batali (1994), who demonstrates that the initial weight settings in a recurrent neural network (RNN), able to learn by backpropagation to recognise strings generated by a restricted class of deterministic context-free grammars, can be improved by genetic assimilation. An evolving population of RNNs with randomly initialised weights was exposed to languages from this class and the networks best able to recognise strings from these languages were kept and also used to create offspring with minor variations in their initial settings. RNNs evolved able to learn final weights which yielded much lower recognition error rates for strings from any of this class of languages. Batali also discusses how his approach might yield an account of the critical period for language acquisition. However, from the perspective of this paper, this work is chiefly relevant for its demonstration of the potential for genetic assimilation in a precise computational setting on a non-trivial learning task. The RNN model of grammatical acquisition fails to meet

most of the desiderata identified in section 2 above, because the RNNs do not model the mapping between structure and meaning and cannot (approximately) recognise strings from languages exhibiting cross-serial dependencies.

In a related simulation, Livingstone and Fyfe (2000) start with a population of networks able to represent the mapping between undecomposable finite signal-meaning correspondences and demonstrate that spatially-organised networks will genetically assimilate an increased production capacity by switching on further hidden nodes in their networks, given selection for interpretative ability and exposure to a larger vocabulary. They argue that in a spatially organised setting this amounts to a form of kin selection since networks receive no direct benefit from an increased production ability. They suggest that their approach might be extended to grammatical competence. However, it is difficult to see how, as the network architecture is only able to represent *finite* signal-meaning correspondences.

Turkel (2001) adapts Hinton and Nowlan's (1987) simulation more directly by adopting a principles and parameters model of grammatical acquisition. Individuals in the evolving population are represented by a genotype of 20 binary-valued principles/parameters which can be set to on (1), off (0) or unset (?) again. Unset values represent parameters which are set during lifetime learning, and values represent nativised principles of universal grammar. Learned settings of unset parameters define variant phenotypes of a given genotype interpreted as different grammars learnable from the inherited variant of universal grammar. The fitness of a genotype is determined by the speed with which individuals acquire compatible settings for unset parameters. A population of randomly initialised individuals each with 10 unset parameters attempts to set them in order to com-

municate with another random individual using the same grammar. Individuals able to communicate are more likely to produce offspring with new genotypes derived from their own by crossover with those of another individual. Populations evolved genotypes which increased the speed and robustness of learning. However, despite the cost of learning, they did not converge on genotypes with no remaining parameters, probably for similar reasons to those identified by Harvey in his analysis of Hinton and Nowlan's original work. Turkel's approach does not suffer from the weaknesses of neural network based models, because he does not specify how genotypes encode grammars capable of generating sentence-meaning correspondences. Turkel, like Hinton and Nowlan, sees the simulation more as an abstract demonstration of how genetic assimilation provides a mechanism for canalizing a trait, and thus, as a demonstration of how a LAD might have arisen on the basis of natural selection for communicative success. However, because of the unspecified relationship between genotypes and actual grammars, the only really substantive difference from Hinton and Nowlan's model is the use of a frequency-dependent rather than fixed fitness function which creates an overall lower degree of selection pressure.

Kirby and Hurford (1997) extend Turkel's model by encoding sentences in terms of the principle/parameter settings required to accurately parse them and by modifying Gibson and Wexler's (1994) Trigger Learning Algorithm. Appropriate parameter settings are learnt by individuals as a function 1) of the parsability of individual sentences, where more parsable sentences are generated by grammars defined by on settings at the first 4 loci, and 2) of their distance from the individual's current parameter settings. This introduces linguistic selection into the model as grammars which generate more parsable sentences can be learnt more

easily. The initial population consists of individuals with all settings unset (i.e. no LAD) exposed to enough sentences to be able to learn some grammar. As the population evolves, fitness increases through grammatical assimilation of settings which shorten the learning period and therefore increase communicative success. Kirby and Hurford demonstrate that grammatical assimilation without linguistic selection results in attenuation of the acquisition period, but also often results in assimilation of linguistically non-optimal settings in the genotype. However, in conjunction with linguistic selection, the population fixates on a genotype that is compatible with the optimal grammars, because linguistic selection guarantees that the population converges on optimally parsable languages, via the inductive bias built into the learning algorithm, before genetic assimilation has time to fixate individual loci in the genotype. They conclude that functional constraints on variation will only evolve in the LAD if prior linguistic selection means that the constraints are assimilated from an optimal linguistic environment, and thus, that natural selection for communicative success is not in itself enough to explain why *functional* constraints could become nativised. This work is important because it develops a coevolutionary model of the interaction between linguistic selection for variant grammars via cultural transmission with natural selection for variant LADs via genetic assimilation. However, the model remains underspecified in terms of the connection between genotypes and actual grammars, the grammatical acquisition procedure does not meet the desiderata of section 2 even when this mapping is fully specified (e.g. Brent 1996; Briscoe, 1999, 2000a).

Yamauchi (2000a,b) replicates Turkel's simulation but manipulates the degree of correlation in the encoding of genotype and phenotype. He continues to represent a grammar as a sequence of $N$ principles or parameters but determines the

inherited value at each locus from a look-up table which uses $K$ 0/1s (where $K$ can range from 1 to $N$-1) to encode each on/off/unset value (and presumably ensure that all possible genotypes can be encoded). A genotype is represented as a sequence of $N$ 0/1s. A translator reads the first $K$ genes from the genotype and uses the look-up table to compute the value of the first locus of the phenotype. To compute, the value of the second locus of the phenotype, the $K$ genes starting at the second locus of the genotype are read and looked up in the table, and so on. The translator 'wraps around' the genotype and continues with the first locus when $K$ exceeds the remaining bits of the genotype sequence. Yamauchi claims, following Kauffman (1993), that increases in $K$ model increases in pleiotropy and epistasis. Increased $K$ means that a change to one locus in the genotype will have potentially more widespread and less predictable effects on the resulting phenotype. It also means that there is less correspondence between a learning 'move', altering the value of single phenotypic locus, and a genetic 'move', potentially altering many in differing ways or none depending on the look-up table. For low values of $K$, genetic assimilation occurs, as in Turkel's model, for values of $K$ around half $N$ genetic assimilation is considerably slowed, and for very high values it is stopped.

Yamauchi does not consider how the progressive decorrelation of phenotype from genotype affects the degree of communicative success achieved or how linguistic systems might be affected. In part, the problem here is that the abstract nature of Turkel's simulation model does not support any inference from configurations of the phenotype to concrete linguistic systems. Yamauchi, however, simply does not report whether decorrelation affects the ability of the evolving population to match phenotypes via learning. The implication, though, is that, for high values

of $K$, unless the population starts in a state where genotypes are sufficiently converged to make learning effective, then they cannot evolve to a state better able to support communication. Kauffman's original work with the $NK$ model was undertaken to find optimal values of $K$ for given $N$ to quantify the degree of epistasis and pleiotropy likely to be found in systems able to evolve most effectively. Both theretical predictions and experiments which allow $K$ itself to evolve suggest intermediate values of $K$ are optimal (where the exact value can depend on $N$ and other experimental factors). Despite these caveats, Yamauchi's simulation demonstrates that (lack of) correlation of genotype and phenotype with respect to the LAD is just as important an issue for accounts of grammatical assimilation as it is for accounts of genetic assimilation generally.

# 6 My Simulation Model

The model that I have developed (Briscoe, 1997, 1998, 1999, 2000a,b, 2001) is similar to that of Hurford and Kirby (1997), in that it supports both linguistic selection for grammatical variants and natural selection for variant LADs, however, it incorporates a considerably more detailed and adequate account of grammatical acquisition which, in turn, supports a much more precise account of the range of grammatical systems that can potentially be adopted by a speech community. Thus, the model makes more concrete linguistic predictions based on the presence or absence of grammatical assimilation.

## 6.1 Language Agents

A language agent (LAgt) is a model of a language learner and user consisting of a 1) learning procedure, $LP$, which takes a definition of a universal grammar, $UG$,

and a sentence-meaning ('logical form') pair or 'trigger', $t$ and returns a specific grammar, $g$; 2) a parser, $P$ which takes a grammar and a trigger, $t$, and returns a logical form, $LF$, for $t$ if $t$ is parsable with $g$ and otherwise reports failure; and 3) a generator, $G$, which given a grammar, $g$, and a randomly selected $LF$ produces a trigger compatible with this $LF$.

I have developed several accounts of $LP$ based on a theory of $UG$ known as Generalized Categorial Grammar and an associated parsing algorithm $P$. In what follows, I assume the Bayesian account of parametric learning developed in Briscoe (1999) with minor modifications. Briscoe (1999, 2001) outlines how the general purpose Bayesian learning mechanism might have been integrated with the grammatical representation, itself an exaptation and minor modification of a preadapted conceptual representation system. Grammatical acquisition consists of incrementally adopting the most probable, and thus most compact, grammar defined by $UG$ compatible with the $n$th trigger in the sequence seen so far:

$$g = argmax_{g \in UG}\ p(g)\ p(t_n \mid g)$$

Briscoe (1999) shows how this formula can be derived from Bayes theorem and how prior probability distributions can be placed on $g \in UG$ in terms of the number and type of parameters required to define $g$, broadly favouring regularity and compactness. The probability of $t$ given $g$ is defined in terms of the posterior probabilities of the grammatical categories required to parse $t$ and recover the correct $LF$. These posterior probabilities are updated according to Bayes theorem after each new trigger is parsed and $LP$ searches a local space, defined parametrically, around $g$, to find a parse for $t$ if necessary. This account of $LP$ meets the desiderata described in section 2 (Briscoe, 1999, 2000a).

In the experiments reported below, $LP$ does not vary, however, the starting point for learning and the hypothesis space can be varied. $UG$ is defined by a *P-setting* consisting of 20 binary principles/parameters which define possible grammars and the exact prior probability distribution on them. Each individual p-setting is represented by a fraction: $\frac{1}{2}$ represents an unset parameter with no prior bias on its value; $\frac{1}{5}$ and $\frac{4}{5}$ represent default parameters with a prior bias in favour of specific settings. However, this bias is low enough that consistent evidence for the alternative setting during learning will allow $LP$ to move the posterior probability of this parameter through the $\frac{1}{2}$ (unset) point to take on its other setting. Principles, which have been nativised, have prior probabilities so close to 1 or 0, typically $\frac{1}{50}$ or $\frac{49}{50}$, that $LP$ will not see enough evidence during learning to alter their settings. How a *P-setting* defining $UG$ is initialised for specific LAgts determines their exact inductive bias and hypothesis space. The 'weakest' minimal or no LAD variant is one in which all p-settings are unset parameters. If all have absolute values, either a single grammar is available to a LAgt or no grammar is available (as some 'off' settings preclude any form of message decomposition or are mutually incompatible). Mutation and one-point crossover operators are defined over *P-setting*s and designed to not bias evolution towards adoption of any one of the three types of p-setting. However, if default settings or principles are acquired this clearly constitutes grammatical assimilation because it creates either soft or absolute inductive bias in favour of subclasses of grammars with specific linguistic properties. This bias is additional too the general and domain-independent bias in favour of generalisation or compactness inherited through exaptation of the general purpose learning mechanism.

The space of possible grammars in $UG$ is defined in terms of canonical constituent

order, possible non-canonical variations and categorial complexity. These are adopted from the typological literature on attested variation (e.g. Croft, 1990) but model most as independent. There are 70 full languages and a further 200 subset languages of these full languages, generated by grammars which have some parameters unset or off. Further details of these grammars and language fragments are given in Briscoe (1997, 1998, 2000a). Assimilated default or absolute settings for any of these parameters, therefore create clear and concrete forms of specifically grammatical inductive bias in favour of specific constituent orders, and so forth.

In addition, each LAgt has an age, between 1 and 10, and a fitness, between 0 and 1. LAgts can learn until they exceed age 4 and interact (i.e. parse or generate) with whatever grammar they have internalised between 1 and 10. The simplest version of fitness measures LAgts' communicative success as a ratio of successful to all interactions, but other factors can be included in fitness such as the degree of expressiveness of the grammar acquired. A successful interaction occurs when the trigger generated by a 'speaking' LAgt can be parsed by the 'hearing' LAgt to yield the same $LF$ which does not require that the LAgts share identical grammars. So, to summarise, a language agent has the following components:

LAgt:
$LP(UG, t) = g$
$P(g, t) = LF$
$G(g, LF) = t$
$Age: [1 - 10]$
$Fitness: [0 - 1]$

## 6.2 Populations and Speech Communities

A population is a changing set of LAgts. Timesteps of the simulation consist of interaction cycles during which each LAgt participates in a prespecified number of interactions. On average each LAgt generates for half of these interactions. LAgt pairs participating in interactions are drawn randomly without bias from the entire population. After each time step, the age of each LAgt is incremented and those over age 10 are removed, the fitness of each LAgt over that interaction cycle is computed, and LAgts aged 4 or more who have greater than mean fitness can reproduce a new LAgt by single-point crossover of their *P-setting*s with another such LAgt with whom they have successfully interacted. The resulting *P-setting* can also optionally undergo random unbiased single-point mutation creating new p-setting values at specific loci. The number of new LAgts per timestep is capped to prevent the proportion of learning LAgts to exceed one-third of the overall population.

The mean number of interactions is set so that accurate grammatical acquisition is possible from many initialisations of $UG$, including 'no LAD' for which all p-settings are unset parameters. Therefore, if a simulation run is initialised with no mutation and a mixed age 5 or over population of LAgts endowed with the same p-settings who have internalised the same full grammar, then grammatical acquisition will be 99% accurate or better, and communicative success will be around 98%, the 2% accounting for learners who have temporarily internalised a subset grammar. In this case, the population constitutes a stable homogeneous speech community, in which no significant grammatical variation is present and no grammatical change takes place. If grammatical variation is introduced into

such a population, then linguistic drift, analogous to genetic drift, means that the population will converge on one variant within around 2N timesteps (where N is population size) due to sampling effects on learning LAgts' input (Briscoe, 2000b). Grammatical variation can be introduced by initialising the simulation with LAgts who have internalised different grammars or by periodic 'migrations' of groups of such LAgts. In this case, it makes sense to think in terms of contact between speech communities. However, the dynamic of the simulation is always to recreate a single such community with a high overall communicative success because all LAgts in the current population interact with each other with equal probability regardless of the grammar they have internalised, their provenance, or their age.

Linguistic selection, as opposed to drift, occurs whenever any factor other than the frequency of a grammatical variant plays a role in its ability to be passed on to successive generations of learning LAgts. Such factors might be the relative parsability of variants and their consequent learnability, the probability with which they are generated, or the degree to which inductive bias in the LAD militates for or against them, their expressiveness, social prestige, and so forth. In simple cases of linguistic selection (Briscoe, 1998, 2000a), the population typically converges to the more adaptive variant within N timesteps (where N is population size). In this simulation model, once linguistic variation is present there is a tendency for populations to converge on subset grammars and associated languages. These grammars require fewer parameters to be set and thus can be learnt faster. However, if all LAgts utilise the same subset language then communicative success will remain high. This tendency can be countered by introducing a further factor into LAgt fitness which adds an extra cost for utilising

a subset grammar each time a LAgt generates a sentence. This creates selection for grammars able to express the widest range of $LF$s. In the experiments below the only form of linguistic selection considered will be that created by natural selection for communicative success and expressiveness.

## 6.3 Coevolution and Grammatical Assimilation

Linguistic selection can occur without natural selection for, or any genetic evolution of, LAgts (i.e. *P-settings*), if they are initialised with p-settings creating inductive bias. However, if mutation is enabled on *P-setting*s and reproduction is random, then simulation runs inevitably end with populations losing the ability to communicate because accumulated genetic drift in p-settings eventually prevents learning LAgts acquiring any grammar. If reproduction is fitness-guided, then there is modest selection pressure for p-settings which attenuate the learning process and increase fitness at age 4, and more severe pressure for p-settings which allow reliable accurate grammatical acquisition by the end of the learning period at age 5.

Previous experiments with a variety of different *P-settings* and several variants of $LP$ have demonstrated that grammatical assimilation occurs with natural selection for communicative success in this simulation model and that populations continue to utilise full grammars and associated languages if there is also natural selection for expressiveness (Briscoe, 1998, 1999, 2000a, 2001). Reproduction is typically by high probability crossover of two above mean fitness LAgts' *P-setting*s with a low probability of subsequent unbiased mutation of one p-setting value to another, representing a different p-setting type. Natural selection prefers p-setting variants which aid grammatical acquisition in the current linguistic en-

vironment. Inducing rapid linguistic change in the environment does not prevent grammatical assimilation, though it does cause it to asymptote rather than continue to the point where the population fixates on a single nativised grammar. Rapid linguistic change also creates a preference for the assimilation of default p-settings over principles, since the latter are potentially more damaging when subsequent linguistic change renders a principle maladaptive for learners. If the population were exposed to the entire space of grammatical variation within the time taken for a variant p-setting to go to fixation, then assimilation would not occur. However, for this to happen, the rate of linguistic change would be so great that communication would breakdown and the population would not consitute a speech community in which the majority of interactions are successful. Below I describe one such experiment using the $LP$ and simulation model outlined above, taken from Briscoe (1999).

Populations of LAgts were initialised with LADs consisting of 3 principles and 17 unset parameters all speaking one of seven attested full grammars. 10 runs were performed under each condition. Simulation runs lasted for 2000 interaction cycles (about 500 generations of LAgts). Reproduction was proportional to communicative success and expressiveness, and was by crossover and mutation of the initial p-settings of the 'parent' LAgts. Constant linguistic heterogeneity was ensured by migrations of adult LAgts speaking a distinct full language with 1-3 different parameter settings at any point where the dominant (full) language utilised by the population accounted for over 90% of interactions in the preceding interaction cycle. Migrating adults accounted for approximately one-third of the adult population and were initialised to have initial p-settings consistent with the dominant settings already extant in the population; that is, migrations were

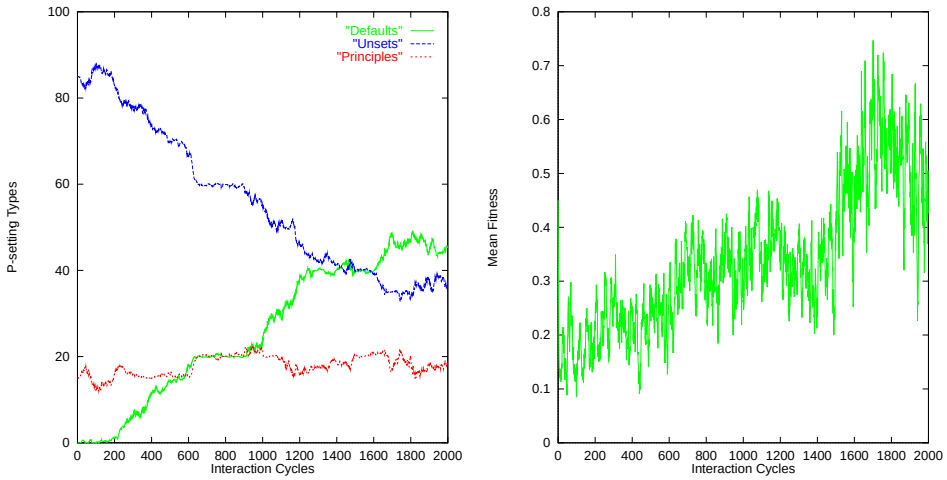


Figure 1: Proportions of P-setting Types and Mean fitness

designed to increase linguistic, not genetic, variation.

Over all these runs, the mean increase in the proportion of default parameters in all such runs was 46.7%. The mean increase in principles was 3.8%. These accounted for an overall decrease of 50.6% in the proportion of unset parameters in the initial p-settings of LAgts. Qualitative behaviour in all runs showed increases in default parameters and either maintenance or increase in principles. Figure 1 shows the relative proportions of default parameters, unset parameters and principles in the overall population and also mean fitness for one such run. Overall fitness increases as the learning period is truncated, though there are fluctuations caused by migrations and/or by increased proportions of learners.

In these experiments, linguistic change (defined as the number of interaction cycles taken for a new parameter setting to go to fixation in the population) is about an order of magnitude faster than the speed with which a genetic change (new initial p-setting) can go to fixation. Typically, 2-3 grammatical changes occur during the time taken for a principle or default parameter setting to go to

fixation. Grammatical assimilation remains likely, however, because the space of grammatical variation (even in the simulation) is great enough that typically the population is only sampling about 5% of possible variations in the time taken for a single p-setting variant to go to fixation (or in other words, 95% of the selection pressure is constant during this period). Though many contingent details of the simulation are arbitrary and unverifiable, such as the size of the evolving population, size of the grammar set, and relative speed at which both can change, it seems likely that the simulation model massively *underestimates* the size of the potential space of grammatical possibilities. Thus, there would very probably have been more opportunity to restrict the hypothesis space by grammatical assimilation than is predicted by the simulation model. Nevertheless, there is a limit to grammatical assimilation in the face of ongoing linguistic change, in simulation runs with LAgts initialised with all default parameters, populations evolve away from such fully-assimilated LADs when linguistic variation is maintained.

## 6.4 Neighbourhood Correlation

The above experiments did not take account of the potential effect of (lack of) neighbourhood correlation. However, we know from Mayley's (1996) and now from Yamauchi's (2000a,b) work that this may undermine the results. In Briscoe (2000a), I pointed out that my model assumes full correlation between genotype and phenotype and operations defined on them in genetic and learning 'space'. In the absence of any great understanding of the biological pathways from genes to neural mechanisms, or of the neural mechanisms underlying the putative LAD, it is impossible to draw definitive conclusions about this assumption. However, Kauffman's (1993) work on *NK* models does suggest that such pathways are likely

to evolve towards a degree of epistasis and pleiotropy so that systems evolve optimally 'at the edge of chaos'. It is fairly straightforward to explore the experimental effects of introducing progressive decorrelation in several different ways into the simulation model. As is often the case, the results of these experiments were surprising (at least to me) and much more subtle in terms of their interpretation than expected, underlining once again the need for careful simulation of the interaction of variant assumptions in such complex dynamical models.

The modified model does not distinguish genotype and phenotype, instead utilising a single *P-setting* which encodes both the initial state ($UG$) and subsequent states during learning, as parameter probabilities, and consequently their settings, change. *P-settings* are defined by a sequence of fractions which define the prior probability of each of three possible settings: unset, default parameter and absolute principle. Arbitrary manipulation of denominators and numerators is very likely to result in values outside the range 0-1. A *NK*-like scheme based on a binary sequential 'genetic' encoding of these fractions with single-point mutation by bit flipping will nearly always produce new absolute principles (under the fairly natural assumption that values outside the range 0-1 are interpreted this way). So instead the mutation operator was modified so that it created unbiased movement of parameters between default and unset settings at multiple random points in a *P-setting*. The maximum number of p-settings that could be modified in a single mutational event was the main parameter varied in these simulation runs, but the exact number modified, the points in the p-setting modified and the resultant settings, were all independent stochastic variables of each such event. The fractional values defining prior probabilities remained fixed, as defined in section 6.1 above.

New simulation runs were performed identical to those reported in section 6.3, except that the parameter controlling the degree of decorrelation between mutation and parameter (re)setting was varied, and half the runs did not include migrations. The general effect of migrations is to add greater linguistic variation and therefore to increase the potential for linguistic selection. However, linguistic selection will also occur in a population in which the LAD is evolving because different variants of the LAD can force even a homogeneous speech community to shift to a new language. For example, if a new LAgt inherits a mutated p-setting which alters a directional default, that learner may acquire a variant grammar compatible with this default if the input sample does not exemplify the non-default setting reliably enough to override the default. If that LAgt and some of its descendents achieve better than mean fitness, because this default only affects a proper subset of sentences in the language or because it is reset effectively, then the default initial setting may spread through the population. The likelihood of such LAgts achieving better than mean fitness is lower in an environment where the remaining population are all learning accurately, but is increased in one in which some other new LAgts are inheriting mutated *P-settings* which disadvantage them more seriously.

The main effect of progressively decorrelating the mutation operator from the learning procedure is to increase the rate of linguistic selection and, despite natural selection on the basis of expressiveness, to cause populations to tend to reconverge to successively less expressive subset languages. Often, linguistic change is coextensive with a few new LAgts appearing who fail to learn any language. However, swift shifts to variant (often less expressive) languages mean that other genetically similar LAgts do acquire the variant language. Thus, although decor-

relation modestly increases the number of subset learning, mislearning and non-learning LAgts, this, in turn, creates linguistic selection for other more learnable languages given the evolving genetic make up of the learning population. When the decorrelation rate is very high, potentially affecting all of the *P-setting* during one mutational event, then the number of non-learners appears to go through a phase transition increasing about a thousand-fold over the previous increment. In these runs, populations always converge to a minimal subset language, which is learnable by setting three p-settings, and in most cases several of these settings have evolved as default parameters across the population.

Tracking the rate of evolution of default parameters over these runs reveals that this rate *increases* by about 5% over runs without decorrelation, as measured by the number of default parameters in the population at the end of each run. This increase is broadly constant across all the runs regardless of the level of decorrelation and the presence or absence of migrations. However, as the decorrelation rate increases the standard deviation of the mean also increases reflecting the size of the potential changes induced by the increasingly dramatic mutational events. That is, for higher rates of decorrelation, distinct runs diverge more as the stochastic factors in the mutational operator affect the exact behaviour of individual runs to a greater extent.

An increase in the number of default parameters in the LAD only counts as grammatical assimilation if the evolved defaults are compatible with the language(s) in the environment. Examining the timing of changes in *P-settings* and linguistic changes reveals that decorrelation is often the *cause* of a linguistic change, rather than assimilatory. Grammatical assimilation is defined as the assimilation of environmental constancy to aid grammatical acquisition. The changes to the LAD

which are not assimilative, yet become adaptive and spread, are ones which drive rapid linguistic change, so that these preemptive genetic changes rapidly become indistinguishable from assimilative ones. A default setting which is correct and thus assimilative in the current linguistic environment reduces the number of parameter settings required to learn the language, attenuating grammatical acquisition and making it more robust against sampling variation in learner input. If a default is assimilated, then it is likely to spread through the population, creating added linguistic selection pressure for subsequent linguistic change to remain compatible with the default setting. If a mutated default is incompatible with the current linguistic environment but spreads to other LAgts, either because grammatical acquisition is generally less accurate or because sampling variation allows enough learners to override the default without significant fitness cost, then it will exert increasing linguistic selection pressure, both because more learners will have the default setting and because less LAgts will generate the counterexamples that would cause the default setting to be overridden.

The lefthand plot in Figure 2 shows the rate of increase of default parameter settings within the population for a low and high degree of decorrelation in two runs with no migrations and otherwise identical initialisations. The lefthand plot shows the corresponding decrease in the number of parameters which are (re)set by learners in the same two runs. Although the overall increase in defaults is consistently higher, and the number of (re)sets is mostly correspondingly lower with more decorrelation, in this run resets converge towards the end, because there is a less close a 'fit' between the languages of the speech community and the form of the LAD with higher degrees of decorrelation at the end of the run. This is a tendency in other runs with no migrations. However, the effect is

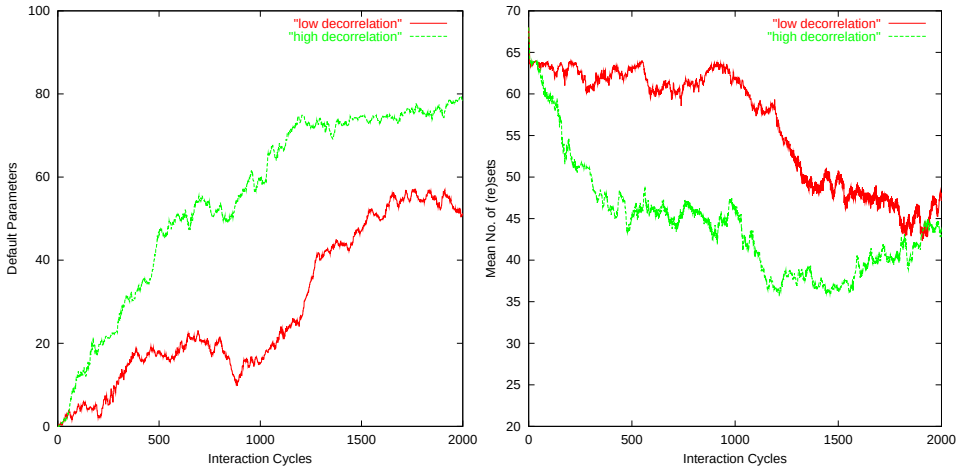


Figure 2: Change in Defaults and Resets with Low/High Decorrelation

removed and to some extent even reversed in runs with migrations, presumably because migrations provide linguistic variation supporting more rapid linguistic selection of variants more compatible with the evolving LAD.

One conclusion that can be drawn from these experiments is that if non-assimilatory random mutations were a factor in the evolution of the LAD, then these mutations would rapidly mesh with linguistic systems, because of the greater speed and responsiveness of linguistic selection. Subsequently such mutations appear to be cases of grammatical assimilation, unless one has access to the precise nature and timing of the mutational event, linguistic environment and any subsequent linguistic change – something one cannot hope to have access to outside the simulation 'laboratory'. The experiments also suggest that high degrees of decorrelation and thus non-assimilatory changes are unlikely, in line with Kauffman's (1993) more general results. The predicted consequence of such changes is that linguistic prehistory would be punctuated by the periodic emergence of mislearners and nonlearners sometimes coupled with bursts of rapid linguistic

change, often in the direction of less expressive languages and despite natural selection for expressiveness. This is contrary to what most non-assimilationists have argued, but, it is not clear whether any evidence beyond the simulation results and general theoretical conclusions bears on the issue.

A simulation model in which greater expressiveness, or acquisition of innovative grammatical variants, outweighed learnability in LAgt fitness might predict natural selection for less restrictive LADs or loss of emerging LADs. A model which integrated some of the social pressures maintaining linguistic diversity discussed, for example, by Nettle (1999), might counterbalance the tendency for learnability to outweigh expressiveness. However, given that there is a strong relationship between the size of the hypothesis space and the amount of data required to reliably acquire a specific grammar (e.g. Nowak *et al* 2001), a model in which expressiveness regularly overcame learnability would predict that the learning period would increase over time or, if a critical period had been nativised, that the reliability of grammatical acquisition would degrade. The existence of a critical period for grammatical acquisition, the accuracy of grammatical acquisition, and its selectivity in the face of variant input (e.g. Lightfoot, 1999) all suggest that this is an implausible evolutionary dynamic. However, integration of a more realistic account of expressiveness into the simulation model would certainly be a worthwhile extension of these experiments.

A further set of similar experiments was undertaken in which the mutational operator was modified so that the fractional values defining initially unset parameters mutated randomly by increasingly large amounts. The increasing bias of this operator is to create absolute parameters as the base of the fractions increases and as they exceed the 0-1 range, so that $LP$ becomes unable to move

them through the $\frac{1}{2}$ threshold which alters a setting. Unsurprisingly, in these experiments, there were many more cases of nonlearners, since principles rather than just default settings were acquired. This mutation operator is exceedingly unlikely to create new unset parameters, and increasingly likely to only create principles with greater degrees of decorrelation. Overall the rate of increase in defaults and principles was slightly higher in these experiments. However, just as in the previous ones, many of the mutational events are at least partly preemptive rather than assimilative, and where the preemption results in principles incompatible with the linguistic environment, a learning LAgt has less chance of reproducing, unless the overall accuracy of grammatical acquisition in the population has degraded significantly. Thus, as in the previous experiments, the trend in linguistic change is towards successive reconvergence on subset languages until the population is speaking a minimal subset language compatible with default parameter settings or principles that have spread through the population.

In general, the greater the degree of decorrelated mutational events involving preemptive non-assimilatory changes, the more the simulation model predicts that the coevolutionary dynamic would bias the hypothesis space until only one grammatical system remained. If the mutation operator is prevented from creating principles or increasingly stronger defaults, as in the first series of experiments, then there is a limit to this effect, but removing this, as in the second series of experiments, strengthens this tendency. A more direct implementation of the *NK* model of neighbourhood decorrelation utilised by Kauffman (1993), Mayley (1996) and Yamauchi (2000a,b) would further increase this effect for increasing values of $K$.

# 7 Conclusions

The simulation case for grammatical assimilation as the primary mechanism of the emergence and subsequent evolution of the LAD remains, in my opinion, strong. However, low degrees of decorrelation of genotype and phenotype predict that some genetic changes might have been partially preemptive rather than fully assimilative, forcing subsequent linguistic change. One way of interpreting this finding would be to argue that it partially vindicates some aspects of the saltationist position, since it implies that some commonalities amongst human languages are a consequence of nonadaptive side effects or 'spandrels' in the evolution of the LAD. Another would be to argue, in the spirit of Deacon (1997), that preemptive mutations are not evidence for nativisation of grammatical knowledge. Rather rapid consequent linguistic evolution has created the close 'fit' between extant languages and the language acquisition procedure.

High degrees of decorrelation of genotype and phenotype are unlikely on general evolutionary grounds (Kauffman, 1993) and predict an implausible coevolutionary dynamic in the case of the coevolution of the LAD and languages, at least according to the current simulation model. There is no evidence that linguistic evolution is punctuated by a series of dramatic mutational events leading to significant breakdown in the cultural transmission of language and followed by subsequent rapid linguistic evolution of new but often less expressive grammatical systems. The natural fixed point of such a dynamic is one expressively restrictive but genetically specified grammatical system. Clearly, this is not what has occurred. Instead the evidence suggests that humans have converged on a genetic specification of a LAD which supports robust acquisition of a wide range

of grammatical systems.

As always, it is important to emphasise that simulation work, however careful and sophisticated, is not enough to establish the truth of what remains a partly speculative inference about prehistoric events. The value of simulations, and related mathematical analysis, lies in uncovering the precise set of assumptions required to predict that grammatical assimilation will or will not occur. Since many of these assumptions relate to cognitive abilities or biases which should remain manifest today, these predictions are not, in principle, untestable. For example, we have seen that inductive bias is at the heart not only of (grammatical) assimilation but also of any satisfactory model of grammatical acquisition and the linguistic evolution of modern languages from protolangue(s). On the other hand, the relative weight of factors relating to learnability and expressiveness in the LAgt fitness function remain largely speculative, though not, in principle, untestable, since they should, for example, be manifest in attested grammatical changes, including those occurring right now (e.g. Kegl *et al* 1999).

# References

Ackley, D. and Littman, M. (1991) 'Interactions between learning and evolution' in C. Langton and C. Taylor (ed.), *Artificial life II,* Addison-Wesley, Menlo Park, CA, pp. 487–509.

Batali, J. (1994) 'Innate biases and critical periods: combining evolution and learning in the acquisition of syntax' in R. Brooks and P. Maes (ed.), *Artificial Life IV,* MIT Press, Cambridge, Ma., pp. 160–171.

Berwick, R. (1998) 'Language evolution and the minimalist program: the origins of syntax' in J. Hurford, M. Studdert-Kennedy and C. Knight (ed.),

*Approaches to the evolution of language,* Cambridge University Press, Cambridge, pp. 320–340.

Bickerton, D. (1998) 'Catastrophic evolution: the case for a single step from protolanguage to full human language' in J. Hurford, M. Studdert-Kennedy and C. Knight (ed.), *Approaches to the evolution of language,* Cambridge University Press, Cambridge, pp. 341–358.

Bickerton, D. (2000) 'How protolanguage became language' in C. Knight, M. Studdert-Kennedy and J. Hurford (ed.), *The Evolutionary Emergence of Language,* Cambridge University Press, Cambridge, pp. 264–284.

Brent, M. (1996) 'Advances in the computational study of language acquisition', *Cognition, vol.61,* 1–38.

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

Briscoe, E. (1998) 'Language as a complex adaptive system: co-evolution of language and of the language acquisition device ' in (eds) Coppen, P., van Halteren, H. and Teunissen, L. (ed.), *8th Meeting of Comp. Linguistics in the Netherlands,* Rodopi, Amsterdam, pp. 3–40.

Briscoe, E. (1999) 'The Acquisition of Grammar in an Evolving Population of Language Agents', *Electronic Trans. of Art. Intelligence (Special Issue: Machine Intelligence, 16. (ed) Muggleton, S., vol.Vol 3(B), www.etaij.org,* 44–77.

Briscoe, E. (2000a) 'Grammatical Acquisition: Inductive Bias and Coevolution of Language and the Language Acquisition Device', *Language, vol.76.2.*

Briscoe, E. (2000b) 'Evolutionary perspectives on diachronic syntax' in Susan Pintzuk, George Tsoulas and Anthony Warner (ed.), *Diachronic Syntax:*

*Models and Mechanisms,* Oxford University Press, Oxford, pp. 75–108.

Briscoe, E. (2001, in press) 'Grammatical acquisition and linguistic selection' in E. Briscoe (ed.), *Language acquisition and linguistic evolution: formal and computational approaches,* Cambridge: Cambridge University Press.

Carstairs-McCarthy, A. (2000) 'The distinction between sentences and noun phrases: an impediment to language evolution?' in C. Knight, M. Studdert-Kennedy and J. Hurford (ed.), *The Evolutionary Emergence of Language,* Cambridge University Press, Cambridge, pp. 248–263.

Cecconi, F., Menczer, F. and Belew, R. (1996) 'Maturation and the evolution of imitative learning in artificial organisms', *Adaptive Behaviour, vol.4,* 29–50.

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

Chomsky, N. (1988) *Language and Problems of Knowledge,* MIT Press, Cambridge, MA.

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.

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

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

Dennett, Daniel (1995) *Darwin's dangerous idea: evolution and the meanings of life,* Simon and Schuster, New York.

Futuyma, D. and Slatkin, M. (1983) *Coevolution,* Sinauer Associates, Sunderland, Ma..

Gazdar, G., Klein, E., Pullum, G., and Sag, I. (1985) *Generalized Phrase Struc-*

*ture Grammar,* Blackwell, Oxford, UK.

Gibson, E. and Wexler, K. (1994) 'Triggers', *Linguistic Inquiry, vol.25.3,* 407–454.

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

Gould, S. (1991) 'Exaptation: a crucial tools for an evolutionary psychology', *Journal of Social Issues, vol.47,* 43–65.

Harvey, I. (1993) 'The puzzle of the persistent question marks: a case study of genetic drift' in S. Forrest (ed.), *Genetic algorithms: proceedings of the 5th International Conference,* Morgan Kaufmann, San Mateo, CA.

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

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

Jablonka, E. and Lamb, M. (1995) *Epigenetic Inheritance and Evolution,* Oxford University Press, Oxford.

Joshi, A., Vijay-Shanker, K. and Weir, D. (1991) 'The convergence of mildly context-sensitive grammar formalisms' in Peter Sells, Stuart Shieber and Tom Wasow (ed.), *Foundational issues in natural language processing,* MIT Press, Cambridge MA, pp. 31–82.

Kauffman, S. (1993) *The Origins of Order: Self-Organization and Selection in Evolution,* Oxford University Press, New York.

Kegl, J., A. Senghas, and Coppola, M. (1999) 'Creation through contact: Sign language emergence and sign language change in Nicaragua' in M. DeGraff (ed.), *Language Creation and Language Change: Creolization, Diachrony, and*

*Development,* MIT Press, Cambridge MA.

Kirby, S. (1998) 'Fitness and the selective adaptation of language' in J. Hurford, M. Studdert-Kennedy and C. Knight (ed.), *Approaches to the evolution of language,* Cambridge University Press, Cambridge, pp. 359–383.

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

Kirby, S. (2001a, in press) '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.

Kirby, S. (2001b, in press) 'Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity', *IEEE Transactions on Evolutionary Computation; special issue on Evolutionary Computation and Cognitive Science,*

Kirby, S. and Hurford, J. (1997) 'Learning, culture and evolution in the origin of linguistic constraints' in Phil Husbands and Imran Harvey (ed.), *4th European Conference on Artificial Life,* MIT Press, Cambridge, MA., pp. 493–502.

Li, M. and Vitanyi, P. (1993) *An Introduction to Kolmogorov Complexity and its Applications,* Springer-Verlag, Berlin.

Lightfoot, D. (1999) *The Development of Language: Acquisition, Change, and Evolution,* Blackwell, Oxford.

Lightfoot, D. (2000) 'The spandrels of the linguistic genotype' in C. Knight, M. Studdert-Kennedy and J. Hurford (ed.), *The Evolutionary Emergence of Language,* Cambridge University Press, Cambridge, pp. 231–247.

Lindblom, B. (1998) 'Systemic constraints and adaptive change in the forma-

tion of sound structure' in J. Hurford, M. Studdert-Kennedy and C. Knight (ed.), *Approaches to the evolution of language,* Cambridge University Press, Cambridge, pp. 242–264.

Livingstone, and Fyfe (2000) 'Modelling language-physiology coevolution' in C. Knight, M. Studdert-Kennedy and J. Hurford (ed.), *The Evolutionary Emergence of Language,* Cambridge University Press, Cambridge, pp. 199–218.

Mackay, D. (1999) *Rate of Information Acquisition by a Species subjected to Natural Selection,* Ms. http://wol.ra.phy.cam.ac.uk/mackay.

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.

Maynard Smith, J. (1998) *Evolutionary Genetics,* Oxford University Press, Oxford, 2nd ed..

McColl, J. (1995) *Probability,* Edward Arnold, London.

Milroy, J. (1992) *Linguistic Variation and Change: on the Historical Sociolinguistics of English,* Basil Blackwell, Oxford.

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

Muggleton, S. (1996) 'Learning from positive data', *Proceedings of the 6th Inductive Logic Programming Workshop,* Stockholm.

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.

Niyogi, P. (1999) *The Informational Complexity of Learning from Examples,* Kluwer, Drodrecht.

Niyogi, P. (2001, in press) 'Theories of Cultural Change and their Application to Language Evolution' in E. Briscoe (ed.), *Language Acquisition and Linguistic Evolution: Formal and Computational Approaches,* Cambridge University Press, Cambridge.

Niyogi, P. and Berwick, R. (1997) 'Evolutionary consequences of language learning', *Linguistics and Philosophy, vol.20,* 697–719.

Newmeyer, F. (2000) 'On the reconstruction of 'proto-world' word order' in C. Knight, M. Studdert-Kennedy and J. Hurford (ed.), *The Evolutionary Emergence of Language,* Cambridge University Press, Cambridge, pp. 372–390.

Nowak, M., Plotkin, J., and Jansen, V. (2000) 'The evolution of syntactic communication', *Nature, vol.404,* 495–498.

Nowak, M., Komarova, N., and Niyogi, P. (2001) 'Evolution of universal grammar', *Science, vol.291,* 114–118.

Oliphant, M. (2001, in press) '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.

Osherson, D., Stob M. and Weinstein, S. (1986) *Systems that learn,* Cambridge University Press, Cambridge.

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

Pullum, G. (1983) 'How many possible human languages are there?', *Linguistic Inquiry, vol.14,* 447-467.

Pullum, G. and Scholz, B. (2002) 'Empirical assessment of stimulus poverty arguments', *The Linguistic Review,*

Renshaw, E. (1991) *Modelling Biological Populations in Space and Time,* Cambridge University Press, Cambridge.

Richards, R. (1987) *Darwin and the Emergence of Evolutionary Theories of Mind and Behaviour,* University of Chicago Press, Chicago.

Ridley, M. (1990) 'Reply to Pinker and Bloom', *Behavioral and Brain Sciences, vol.13,* 756.

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

Ristad, E. and Rissanen, J. (1994) 'Language acquisition in the MDL framework' in E. Ristad (ed.), *Language Computation,* American Mathematical Society, Philedelphia.

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.

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

Turkel, W. (2001, in press) 'The learning guided evolution of natural language' in E. Briscoe (ed.), *Language acquisition and linguistic evolution: formal and computational approaches,* Cambridge University Press, Cambridge.

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.

Wexler, K. and Culicover, P. (1980) *Formal principles of language acquisition,* MIT Press, Cambridge MA.

Worden, R. (1995) 'A speed limit for evolution', *J. Theor. Biology, vol.176,* 137–152.

Worden, R. (1998) 'The evolution of language from social intelligence' in J. Hurford, M. Studdert-Kennedy and C. Knight (ed.), *Approaches to the evolution of language,* Cambridge University Press, Cambridge, pp. 148–168.

Worden, R. (2001, in press) 'Linguistic structure and the evolution of words' in E. Briscoe (ed.), *Language Acquisition and Linguistic Evolution: Formal and Computational Approaches,* Cambridge University Press, Cambridge.

Wray, A. (2000) 'Holistic utterances in protolanguage: the link from primates to humans' in C. Knight, M. Studdert-Kennedy and J. Hurford (ed.), *The Evolutionary Emergence of Language,* Cambridge University Press, Cambridge, pp. 285–302.

Yamauchi, H. (2000a) *Evolution of the LAD and the Baldwin Effect,* MA Dissertation, University of Edinburgh, Dept. of Linguistics.

Yamauchi, H. (2000b) 'The difficulty of the Baldwinian account of linguistic innateness', *Proceedings of the European Conf. on ALife,*