Chapter 1

Grammatical Assimilation

Ted Briscoe

1.1 Introduction

In this paper, I 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 learning bias in favour of some proper subset of the space of possible languages. Genetic assimilation is a neo-Darwininan mechanism by which organisms can appear to inherit acquired characteristics. Genetic assimilation of grammatical generalisations exemplified in the environment of adaptation of the LAD facilitates more rapid and robust grammatical acquisition by first language learners. I will develop a coevolutionary account of this process 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

The existence of an innate LAD has not gone unquestioned, and it is probably the case that some arguments that have been proposed in its favour are either questionable or wrong (e.g. Pullum and Scholz, 2002; Sampson, 1989, 1999). I will argue that all remotely adequate extant models of grammatical acquisition do presuppose a LAD (in the above weak sense), and that genetic assimilation is the most plausible account of its emergence and maintenance. These arguments do not constitute a proof either of the existence of an innate LAD or of its evolution 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, 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 1.4, where some arguments for and against grammatical assimilation are also presented. Section 1.2 reviews work on grammatical acquisition and presents the case for the existence of the LAD. Section 1.3 outlines an account of languages as complex adaptive systems and spells out several consequences for models of grammatical assimilation. Section 1.5 describes and evaluates extant simulations of genetic and grammatical assimilation.

1.2 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 characterize the range of possible mappings from meaning to form in attested systems. A reasonable requirement, given current knowledge, is that the model be capable of learning the mappings for a proper subset of indexed languages, including those exhibiting cross-serial dependencies (e.g. Joshi *et al* 1991). This rules out much work which purports to address the issue of grammatical acquisition, for example, extant work based on (recurrent) neural networks, as these have only been shown to make (graded) grammaticality judgements for small language fragments (e.g. Lawrence *et al* 1996) and not to recover mappings from meaning to form. 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, as 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 hypothesizes 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, equivalently, that the learner knows when to ignore input (e.g. Osherson *et al* 1986:100). Fourthly, models should account for *selectivity* in acquisition: learners do not acquire '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). Fifthly, models should display *accuracy*: learners do not 'hallucinate' or invent grammatical properties regardless of the input, though they do (over)generalize and, in this sense, 'go beyond the data'.

If accuracy is defined in terms of formal learnability (e.g. Bertolo, 2001; Niyogi, 1999) from realistic, finite, positive but noisy sentence-meaning pairs over a hypothesis space with adequate coverage, even when drawn from a single stationary target grammar, then some form of inductive bias in the acquisition model is essential (see also Nowak *et al* 2001).[1] In much current work on grammatical acquisition within the Principles and Parameters (P&P) framework (e.g. Chomsky, 1981; Gibson and Wexler, 1994), 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, 1981:8f). P&P models, which do not incorporate a statistical or quantitive component, are not able to deal adequately with noisy input (e.g. Briscoe, 1999, 2002). There is a well-known formulation of inductive bias in terms of Bayesian probabilistic learning theory (see e.g. Mitchell, 1997:154f for an introduction). Bayes theorem provides a general formula and justification for the integration of prior bias with experience and it has been demonstrated that an accurate prior supports learnability from finite noisy data over infinite (though restricted) hypothesis spaces (e.g. Horning, 1969; Muggleton, 1996).

Bayesian learning theory is a general domain-independent formulation of learning. The most general formulation of learning in this framework (Kolmogorov Complexity) posits a learner able to learn any generalisation with a domain-independent bias (the so-called 'universal prior') in favour of the smallest, most compressed hypothesis (e.g. Li and Vitanyi, 1997). However, nobody has demonstrated that this general formulation could, even in principle, result in a learning algorithm capable of accurately acquiring a specific grammar of a human language from realistic input. Horning's (1969) work is restricted to the (infinite) class of stochastic 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 (i.e. generalisation due to inductive bias from the input is correct).

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 input. 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 grammars generating a superset of the input is acquired 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 1.1, is unavoidable in any adequate account of grammatical acquisition.

Consider a potential class of languages consisting of clauses constructed from a verb (V), a subject (S) and object (O), where S and O are always realized as single (pro)nouns (N) or as noun phrases consisting of a noun and a (relative) clause – the S and O labels are a shorthand for the mapping from 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 is an argument of which V, when there is more than one V in a sentence. Thus, without further stipulation, any clausal ordering of S, O and V is possible, as well as any arrangement of root and relative clauses like those in (1).

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

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

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

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

Example (1c) provides evidence for a root SVO language containing 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 cost metric exposed to SVO unembedded sequences and (1c) would acquire grammar (4) with a total cost of 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, this learning model predicts that mixed root and embedded constituent orders is a dispreferred or more marked option that will only be adopted when the learner is forced to do so by positive evidence.

By contrast, if the learner represents the class of CFLs in IDLP notation instead of standard CFG, acquiring immediate dominance (ID) rules independently of linear precedence (LP) rules (e.g. Gazdar *et al* 1985), but utilizing a similar cost metric which also assigns a cost of one to each LP rule, then the preference ordering on specific IDLP grammars predicts that order-free variants of the above grammars with no LP rules will be preferred and that the inclusion of examples like (1b) or (1c) in the input will not alter the learner's hypothesis.

Cost metrics applied to such restricted hypothesis representation languages entail that learners will 'go beyond the evidence' in different ways and, thus, will have different specifically-linguistic inductive biases. However, learners without cost metrics, or equivalently prior distributions, cannot acquire target grammars accurately, as Gold's (1967) work demonstrated. Extant models assume a LAD, in the (weak) sense of section 1.1, because they utilize prior distributions or cost metrics defined over restricted hypothesis representation languages selected to facilitate encoding of grammars for human languages. The onus is on non-nativists to develop an account of grammatical acquisition which meets the above desiderata and does not utilize a LAD in this sense.[4]

Independently of these theoretical arguments, there is psycholinguistic evidence that human language learners are biased in linguistically-specific ways.

There are learning stages in which overgeneralisation of regular morphology is common, tense is assigned to auxilaries and main verbs with auxiliariy inversion, and so forth. Whilst, the exact interpretation of such phenomena is a matter of analysis, psycholinguists most often describe them as linguistically-specific biases, for instance, Wanner and Gleitman (1982:12f) argue that children are predisposed to learn lexical compositional systems in which 'atomic' elements of meaning, such as negation, are mapped to individual words. This leads to transient production errors where languages, for example, mark negation morphologically.

1.3 Linguistic Evolution

First language learners are not typically exposed to homogeneous data from a static 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 can acquire grammars that are distinct from those of the previous generation (e.g. Lightfoot, 1999:77f).

We can model the development of the 'external' (E-)language of a speech community as a dynamical system in which states encode the distribution of 'internalized' individual grammars (and lexicons) (i.e. I-languages or idiolects) within the community. In such dynamical systems, transitions between states are defined in terms of changes in the distribution of internalized grammars or I-languages (Briscoe, 1997, 2000b; Niyogi and Berwick, 1997). If there is inductive bias in first language acquisition (regardless of its provenance), then languages are best characterized 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, 1997, 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 language 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, dynamic and multipeaked fitness landscapes, in the sense of, for example, Kaufmann (1993).

Linguistic evolution proceeds via cultural transmission (i.e. first language acquisition) at a faster rate than biological evolution. The populations involved are generally smaller (speech communities, rather than entire species), and language acquisition is a more flexible and efficient method of information transfer than genetic mutation. Clearly, vocabulary learning and, at least, peripheral grammatical development are ongoing processes that last beyond childhood, so that linguistic inheritance is less clearly delineated or constrained than the biological mechanisms of genetic evolution.

Several consequences emerge from the evolutionary account of languages as adaptive systems which must be taken into consideration by any plausible account of 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 form-meaning correspondences and the (subsequent) emergence of protolanguage(s) with decomposable (minimally grammatical) sentence-meaning correspondences (e.g. Oliphant, 2002; Kirby, 2002, Brighton, 2002). They conclude that the language acquisition procedure must incorporate inductive bias resulting in generalisation, and consequent regularisation of the input, in order that repeated rounds of cultural transmission of language regularize random variations into consistent and coherent communication systems.[5]

Moreover, 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, the fact that in attested languages irregularity is associated with high frequency forms is unlikely to be a consequence of a nativized constraint and much more likely to be a universal consequence of the fact that low frequency irregular forms are less likely to be reliably learned by successive generations of first language learners (Kirby, 2001).

1.4 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).

1.4.1 Genetic assimilation

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.[6]

A simple putative example of genetic assimilation is the propensity to develop hard skin on certain regions of the body on the basis of quite limited environmental stimulation. Selection for individuals who developed hard skin more rapidly, and subsequent canalization of this trait prevented infection, aided mobility, and so forth. A more complex putative case is Durham's (1991) example of gene-culture interaction resulting in extended lactose tolerance in human populations in which animal husbandry is well-established. The ability to consume milk in maturity was selected for in an environment in which it was one of the most reliable and beneficial sources of nutrition.

One form of plasticity in primates is the ability to learn from the environment. The Bayesian learning framework provides a general and natural way to understand and model how more and more accurate prior distributions over hypothesis spaces with better and better 'fit' with the environment can evolve. Staddon (1988) and Cosmides and Tooby (1996) both argue at length that Bayesian learning theory is an appropriate framework for modelling learning in animals and humans and that evolution can be understood within this framework as a mechanism for optimizing priors to ‘fit’ the environment and thus increase fitness.

1.4.2 Genetic assimilation of grammatical information

Pinker and Bloom (1990) develop a gradual assimilationist account of the evolution of the LAD. However, they rely heavily on linguistic universals as their evidence. Waddington, himself, suggested earlier that genetic assimilation provided a possible mechanism for the evolution of a LAD:

> ‘If there were selection for the ability to use language, 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)

Briscoe (1999, 2000a) speculates that an initial acquisition procedure emerged via recruitment (exaptation) of preexisting (preadapted) general-purpose (Bayesian-like) learning mechanisms to a specifically-linguistic cognitive representation capable of expressing mappings from decomposable meaning representations to realisable, essentially linearized, encodings of such representations (see also Bickerton, 1998, 2000; Worden, 1998). The selective pressure favouring such a development, and its subsequent maintenance and refinement, is only possible if some protolanguage(s), supporting successful communication and capable of cultural transmission (that is, learnable without a LAD) within a hominid population, had already emerged (e.g. Deacon, 1997; Kirby and Hurford, 1997). Protolanguage(s) may have been initially similar to those advocated 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 for 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 models of the emergence of syntax have been developed (e.g. Kirby, 2001, 2002; 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 will make grammatical acquisition more rapid and reliable.

Saltations or macromutations are compatible with evolutionary theory if a single change in genotype creates a large highly-adaptive change in phenotype, though general considerations predict that such genetic macromutations are extremely unlikely to be adaptive (e.g. 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 entail that the LAD emerged in a single and extremely unlikely evolutionary step (see e.g. Pinker and Bloom (1990) for detailed counterarguments), 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 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.

1.4.3 Counterarguments to assimilation

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 (GB) 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 GB theory, then the LAD, if it exists, could not have emerged as a result of grammatical assimilation and must be the result of saltation. Newmeyer (2002) develops related arguments, for instance arguing that grammatical subordination would have rarely been manifest in preliterate speech communities and therefore that universal constraints relating to such constructions could not have been assimilated. Deacon (1997:307f) argues that, since attested grammatical systems display a trade-off between syntactic and morphological encoding of predicate-argument structure and since these distinct linguistic devices are also neurally distinct, the changing linguistic environment could not have created consistent selection pressure on either neural mechanism. These arguments all rest on specific assumptions about what precisely is assimilated. However, the account of the LAD in terms of inductive bias, developed in section 1.2 is in no way dependent on any specific linguistic constraints and does not rest on (speculative) assumptions about linguistic phenomena manifest in the prehistoric environment of adaptation. Assimilation of linguistic constraints or preferences into the LAD only requires that *some* neurally encodable generalisations were manifest in the environment of adaptation for the LAD.

Lightfoot (1999, 2000) argues that the LAD is not fully adaptive and, therefore, could not have evolved by 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 subordinate clauses, which prevents the asking of questions like (5).

(5) *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 he 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 complementizer *that* is obligatorily dropped, as in (6).

(6) Who do you think (*that) solved the problem?

He argues that the presence of such a maladaptive constraint 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. It may be that *any* genetically encodable extraction constraint aiding parsability and/or learnability also has unwanted side-effects for expressiveness. Complex fitness landscapes typically contain many local optima which are far more likely to be discovered than any global optimum, should it exist (e.g. Kauffman, 1993). Furthermore, a dynamic fitness landscape entails that a once optimal solution can become suboptimal.

Given that grammatical assimilation only makes sense in a scenario in which evolving (proto)languages create selection pressure, Waddington's notion of genetic assimilation should be embedded in the more general one of coevolution (e.g. Kauffman, 1993:242f). Waddington, himself, (1975:307) noted that if there is an adaptive advantage to attenuating grammatical acquisition, then we might expect assimilation to continue to the point where no learning would be needed because a fully specified grammar had been encoded. In this case acquisition would be instantaneous and fitness would be maximized in a language-using population. Given a coevolutionary scenario, in which languages themselves are complex adaptive systems, a plausible explanation for continuing grammatical diversity is that social factors favouring innovation and diversity create conflicting linguistic selection pressures (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. Furthermore, 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 (2002) 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. Attested languages have shifted major grammatical system within 1000 years (or a mere 50 or so generations), so 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. The main weakness of this argument is that it fails to take any account of the potential size of the hypothesis space of grammatical systems. The standard classes of languages familiar from formal language theory are infinite, so the hypothesis space for even the regular languages contains an infinite class of regular grammars. Even if we assume, as does the P&P framework, that the *evolved* LAD restricts the class of possible grammars of human languages to be finite, most linguists (implicitly) agree the hypothesis space remains vast (on the order of 30 million grammars) as they typically posit around 30 binary-valued independent parameters (e.g. Roberts, 2001).[7] No amount of rapid change between attested grammatical systems can count as evidence against grammatical assimilation of linguistic constraints that ruled out the many unattested grammars that could *not* have been sampled in the period of evolutionary adaptation (e.g. Briscoe, 2000a). If, for example, arbitrarily intersecting dependencies of the kind exhibited by the MIX family of context-sensitive languages (e.g. Joshi, *et al* 1991) were unattested, but a proper subset of grammars in the hypothesis space generated constructions with such dependencies, then assimilation of a hard constraint or preference against such grammars might be possible. Rapid change within the proper subset of grammars *not* generating MIX languages would not alter the adaptiveness (i.e. learnability advantages) of ruling out or dispreferring the unattested constructions.

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 causes a species to 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 neural mechanisms 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 traits. The logic of the speed-limit argument collapses, given a saltationist account based on macromutation – a single genetic change brings about a complex of extremely unlikely but broadly adaptive phenotypic changes which spreads rapidly through the population. A 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 acquired phenotype can be assimilated into the genotype of future generations, thus attenuating learning and/or increasing its success, and further increasing fitness, depends critically on this correlation.

1.5 Computational Simulations of Assimilation

One way to explore the arguments and counter arguments outlined in section 1.4.3 is to build a simulation and/or a mathematical model. The latter is, in principle, preferable as analytic models of dynamical systems yield more reliable conclusions (given the assumptions underlying the model), whilst those generated by stochastic computational simulation are statistical (e.g. Renshaw, 1991). However, to date, no detailed analytic model of grammatical assimilation has been developed.[8]

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.

1.5.1 Genetic assimilation

Hinton and Nowlan (1987) describe the first computational simulation of genetic assimilation. In their (very abstract) simulation of a population of 1000 neural networks with 20 potential connections, which can be unset $\boxed{?}$, on $\boxed{1}$, or off $\boxed{0}$, was evolved using a genetic algorithm. The target was a network with all 20 connections set to $\boxed{1}$, but networks were initialized randomly with connection ('gene') frequencies of 0.5 for $\boxed{?}$ and 0.25 for $\boxed{1}$ or $\boxed{0}$ at each position. Each network was able to set $\boxed{?}$ connections through learning (modelled as random search of connection settings) 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 acquired the correct settings, making a network with all $\boxed{1}$ connections initially 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 $\boxed{0}$ settings, however this soon gave way to exponential increases in networks with more $\boxed{1}$ settings, less $\boxed{?}$ settings and no $\boxed{0}$ settings. In the later stages, the increase of $\boxed{1}$ settings and decrease of $\boxed{?}$ settings asymptotes, 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. However, with learning, the simulation always converges within 10-15 generations on a viable genotype (i.e. after generating 100-150K networks). Once successful networks appear, their superior performance rapidly leads to the spread of genotypes which support successful learning. However, networks with $\boxed{?}$ settings persist despite the pressure exerted by the fitness function to minimize the number of learning trials required to find the solution. 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 initial settings. When a single successful genotype evolves and dominates subsequent generations, it is possible for a $\boxed{?}$ setting 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 predefined 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.

1.5.2 Grammatical assimilation

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 make grammaticality judgements for sentences generated by a restricted class of unambiguous CFGs, can be improved by genetic assimilation. An evolving population of RNNs with randomly initialized weights was exposed to languages from this class and the networks best able to judge sentences 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 error rates for sentences from any of this class of languages. 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 the desiderata identified in section 1.2 above, because the RNNs do not model the mapping between form and meaning.

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-organized 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 organized 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 (2002) adapts Hinton and Nowlan's (1987) simulation more directly by adopting a P&P 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 ($\boxed{1}$), off ($\boxed{0}$) or unset ($\boxed{?}$). $\boxed{?}$ settings represent parameters which are set during lifetime learning, $\boxed{0}$/$\boxed{1}$ settings represent nativized principles of the LAD. Learned settings of parameters define variant phenotypes of a given genotype interpreted as different grammars learnable from the inherited variant of the LAD. The fitness of a genotype is determined by the speed with which individuals acquire compatible settings for unset parameters. A population of randomly initialized individuals each with 10 parameters attempts to set them in order to communicate with another random individual via 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 form-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 a set of sentences in terms of the principle/parameter settings required to accurately parse them and by utilizing a modified version of 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 $\boxed{1}$ 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 only parameters who are exposed to enough sentences to be able to learn some grammar. As the population evolves, fitness increases through grammatical assimilation of $\boxed{1}$/$\boxed{0}$ 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 (i.e. ones yielding grammars generating less parsable sentences). However, in conjunction with linguistic selection, the population converges 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 nativized. 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.

Yamauchi (2000, 2001) 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 initial setting 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 $\boxed{1}$/$\boxed{0}$/$\boxed{?}$ setting (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 setting of the first locus of the phenotype. To compute, the setting 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 operation, altering the value of single phenotypic locus, and a genetic operation, 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 $N/2$ genetic assimilation is considerably slowed, and for very high values ($K = N - 1$) 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 match phenotypes and thus 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 theoretical 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). But despite these caveats, Yamauchi's simulation suggests 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, as Mayley (1996) argued.

I have developed a coevolutionary model and associated simulation (Briscoe, 1997, 1998, 1999, 2000a, 2002, forthcoming) which supports linguistic selection for grammatical variants, based on learnability, parsability and/or expressiveness, and natural selection for variant LADs based on communicative success. It incorporates a detailed account of grammatical acquisition, meeting the desiderata of section 1.2, which, in turn, supports much more detailed modelling of the grammars acquired. Language agents (LAgts) learn and deploy Generalized Categorial Grammars (GCGs) using a Bayesian learning procedure which acquires the most probable grammar capable of representing the form-meaning mapping manifested by a noisy, finite, unordered sequence of form-meaning pairs generated by other random members of the current population of LAgts.

The starting point for learning is represented by a prior probability distribution over 20 binary-valued principles/parameters defining around 300 viable distinct GCGs. An unset parameter is represented by an unbiased prior (i.e. a uniform distribution over the two possible values), a parameter with a default initial setting by a biased prior capable of being reversed during the learning period, a principle by a strongly-biased prior that cannot be reversed given the amount of data that can be observed during the learning period. Mutation and one-point crossover operators can alter this prior probability distribution converting unset parameters to default parameters, parameters to principles, and so forth randomly so not to bias evolution towards any LAD within the space available. The acquired grammar utilizes just those parameters which are consistently expressed in the data so LAgts can acquire grammars of subset languages. LAgts who communicate successfully with others because they have acquired (partly) compatible grammars reproduce in proportion to their overall relative success. LAgts who have acquired subset grammars or grammars incompatible with that dominant in the population will tend to have lower communicative success. Linguistic variation can be introduced by seeding initial populations with different grammars or by introducing successive migrations of new adult LAgts deploying a grammar different from that currently dominant in order to simulate language contact.

A number of results relevant to grammatical assimilation emerge from this model. Firstly, assimilation occurs when and only when LAgts reproduce according to communicative fitness. This creates selection pressure for attenuating the learning period and making it more robust to noise, so the population assimilates default parameter settings and principles at the expense of unset parameters (Briscoe, 1999). Secondly, populations converge on LADs that further restrict the class of learnable grammars to ones generating subset languages, unless there is an additional conflicting selection pressure on LAgts to acquire more expressive grammars which counteracts the pressure for learnability (Briscoe, 2000). Thirdly, as in the work of Kirby and Hurford (1997), natural selection for communicative success does not guarantee assimilation of functional constraints. However, if parsability inhibits learning or biases the distribution of form-meaning pairs manifest during learning, then assimilated LADs become biased towards more parsable languages (Briscoe, 2000). Fourthly, if language change is as rapid as is consistent with maintenance of a speech community (defined as a mean 90% or better communicative success), assimilation still occurs but asymptotes well before the LAD defines a single grammar. In addition, default initial parameter settings (i.e. preferences) are selected over principles (i.e. hard constraints) as subsequent changes can render principles acquired by a proper subset of the population highly maladaptive (Briscoe, 1999, 2000). Fifthly, in a population with a fixed LAD exposed to homogeneous linguistic input manifesting dispreferred parameter settings, successive generations of learners reliably acquire the correct grammar. However, if their input is heterogeneous and manifests conflicting values, the prior distribution assimilated into the evolved LAD will tend to predominate. Briscoe (2002) suggests this can provide the basis for an account of creolisation and perhaps other attested major historical change resulting from contact. Finally, Briscoe (forthcoming) progressively decorrelates the effects of the mutation operator from the updating of parameter settings during the learning process. Major decorrelation prevents assimilation and most mutations which spread are preemptive 'side effects' rather than assimilative, causing rapid concomitant linguistic change. Consequently, populations eventually evolve LADs which predefine simple subset languages in which learning is redundant despite natural selection for expressiveness. Intermediate levels of decorrelation slow assimilation and increase the proportion of preemptive mutations which spread, but populations are not forced towards subset languages. Low levels of decorrelation have no significant effects as preemptive mutations fail to spread through communities with consistently stable and accurate cultural transmission of language.

1.6 Conclusions

In summary, extant models predict that grammatical assimilation would have occurred given three crucial assumptions. Firstly, communicative success via expressive languages with compositional syntax conferred a fitness benefit on their users. Secondly, the linguistic environment for adaptation of the LAD consistently manifested grammatical generalisations to be assimilated – rapid linguistic change does not preclude generalisations ruling out or dispreferring areas of the hypothesis space generating unattested constructions. Thirdly, some of these generalisations were neurally and genetically encodable with sufficient correlation to support assimilation. None of the counterarguments reviewed in section 1.4.3 or simulations discussed in section 1.5 undermines these assumptions. Thus, the case for grammatical assimilation as the primary mechanism of the evolution of the LAD remains, in my opinion, strong.

Nevertheless, the coevolutionary perspective on grammatical assimilation raises two important caveats. Firstly, as languages themselves are adapted to be learnable (as well as parsable and expressive) and as languages change on a historical timescale, some of the grammatical properties of human languages were probably shaped by the process of cultural transmission of (proto)language via more general-purpose learning (e.g. Kirby, 1998). Secondly, whether the subsequent evolution of the LAD was assimilative, encoding generalisations manifest in the linguistic environment, or preemptive, with mutations creating side-effects causing linguistic selection for new features, the fit between the inductive bias of the

LAD and extant languages is predicted to be very close.

Finally, it is important to emphasize that modelling and simulation, however careful and sophisticated, is not enough to establish the truth of what remains a partly speculative inference about prehistoric events. The value of the simulations, and related mathematical modelling and analysis, lies in uncovering the precise set of assumptions required to predict that grammatical assimilation will or will not occur. Some of these assumptions relate to cognitive abilities or biases which remain manifest today, these predictions are testable. 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 of the linguistic evolution of modern languages from protolanguage(s). Other assumptions, such as the correlation between genetic and neural encoding are theretically plausible but empirically untestable using extant techniques.

Key Further Readings

Nowak *et al*2002) provide a brief synopsis of formal language theory and learnability theory, and develop evolutionary models of language change and of the emergence of the LAD ('universal grammar' in their terms, though they make the point that it is neither universal or a grammar). Mitchell (1997) provides a more detailed and introductory treatment of learning theory. Joshi *et al*1991) summarises extant knowledge concerning the expressive power of human languages in terms of formal language theory. Jablonka and Lamb (1995) describe Waddington's work and the concept of genetic assimilation. Durham's (1991) theory of gene-culture interactions provides the basis for a coevolutionary account of grammatical assimilation. Bertolo (2001) is a good collection of recent work in the P&P framework. Cosmides and Tooby (1996) makes the case for integrating the Bayesian learning framework with evolutionary theory as a general model of human learning.

Acknowledgements

I am grateful to the editors for providing helpful feedback, both from them and from their students on the first draft, which helped me improve this one. Remaining errors or infelicities are entirely my responsibility.

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.

Bertolo, S. (2001) 'A brief overview of learnability' in S. Bertolo (ed.), *Language Acquisition and Learnability,* Cambridge University Press, Cambridge, pp. 1–14.

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.

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

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. (2002) 'Grammatical acquisition and linguistic selection' in E. Briscoe (ed.), *Language acquisition and linguistic evolution: formal and computational approaches,* Cambridge: Cambridge University Press.

Briscoe, E. (2002, forthcoming) *Coevolution of the language faculty and language(s) with decorrelated encodings,* Paper presented at Evolang02, Boston.

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.

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.

Durham, W. (1991) *Coevolution, Genes, Culture and Human Diversity,* Stanford University Press, Pala Alto, Ca..

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

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

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

Kirby, S. 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.

Lawrence, S., Giles, C.L. and Fong, S. (1996) 'Can recurrent neural networks learn natural language grammars?', *Proceedings of the Int. Conf. on Neural Networks (ICNN96),* Washington, DC, pp. 1853–1858.

Li, M. and Vitanyi, P. (1997) *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 formation 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..

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. 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.

Newmeyer, F. (2002) 'Uniformitarian assumptions and language evolution research' in A. Wray (ed.), *The Transition to Language,* Cambridge University Press, Cambridge, pp. 359–375.

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.

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

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

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, vol.19.1-2,* 1–50.

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.

Roberts, I. (2001) 'Language change and learnability' in S. Bertolo (ed.), *Language Acquisition and Learnability,* Cambridge University Press, Cambridge, pp. 81–125.

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. (2002) '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.

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. (2002) '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. (2000) *Evolution of the LAD and the Baldwin Effect,* MA Dissertation, University of Edinburgh, Dept. of Linguistics.

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

Footnotes

1. The term inductive bias is utilized in the field of machine learning to characterize both hard constraints on the hypothesis space considered by a learner, usually imposed by a restricted representation language for hypotheses, and soft constraints which create preferences within the hypothesis space, usually encoded in terms of cost metric or prior probability distribution on hypotheses (e.g. Mitchell, 1997:39f).

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

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

4. Nowak *et al*(2001, 2002) make the stronger claim that the LAD (in my sense) is a logical necessity given the theoretical results of learnability theory and formal language theory.

5. 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.

6. 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).

7. Even this degree of finiteness remains controversial (e.g. Pullum, 1983). For instance, it would be falsified if a language with a parametrically-specified maximum of four syntactically-realized arguments developed a predicate, analogous to English *bet* requiring five such arguments: *(np Kim) bet (np Sandy) (np £ 10) (scomp that she would win) leadsto (np Kim) bet (np Sandy) (np £ 10) (pp for Red Rum) (vpinf to win).*

8. Nowak *et al* 2002 briefly describe the general form of a model capable, in principle, of incorporating grammatical assimilation / coevolution. However, the simplifying assumptions required to yield deterministic dynamical update equations make it very difficult to address many of the arguments in section 1.4.3. For instance, no counteracting (socio)linguistic selection for diversity/variation is modelled, so the equilibrium point for many instantiations of their model may be a LAD encoding a single grammar/language.

Abstract

Genetic assimilation is a possible neo-Darwininan mechanism for the emergence and subsequent refinement of the putative innate human language acquisition device (LAD). The LAD, in the weak sense of language-specific inductive bias, is a part of all extant models of grammatical acquisition. A survey of arguments and counterarguments for the assimilation of such bias during the period of adaptation of the LAD and a review of relevant modelling and simulation work suggests that genetic assimilation is the most plausible extant account of the evolution of the LAD.

Keywords: Genetic Assimilation, Language Learning, Grammatical Acquisition, Linguistic Evolution, Evolution of the Language Acquisition Device

Biography

Ted Briscoe is Reader in Computational Linguistics at the Computer Laboratory, University of Cambridge where he has been a member of staff since 1989. He works on statistical and constraint-based approaches to natural language processing as well as evolutionary modelling and simulation of language development and change.

Index list

genetic assimilation, grammatical assimilation, language acquisition, grammatical acquisition, language acquisition device (LAD), coevolution, natural / linguistic selection — variation, exaptation saltation, macromutation, mutation, crossover, computational simulation, (language as) complex adaptive system, cross-serial dependencies, triggers, trigger learning algorithm, inductive bias, hypothesis space, principles and parameters (P&P) / default — unset — binary parameters, Bayesian learning, prior distribution, cost metric, learnability, (stochastic) context-free/context-sensitive grammars/languages, immediate dominance / linear precedence (ID, LP, IDLP), MIX languages, E-/I-language, protolanguage / (proto)language, gene-culture interaction, cultural / genetic transmission, parsability, expressiveness, fitess landscapes – dynamic — rugged — multipeaked, (recurrent) neural networks (RNNs), communicative success, fitness – criterion — benefit, epistasis, pleiotropy, correlation of phenotype and genotype, gene-culture interaction, Baldwin Effect, Kolmogorov Complexity, Minimum Decsription Length Principle, genetic algorithm