Grammatical Acquisition: Inductive Bias and Coevolution of Language and the Language Acquisition Device

Ted Briscoe
ejb@cl.cam.ac.uk
http://www.cl.cam.ac.uk/users/ejb
Computer Laboratory
University of Cambridge
Pembroke Street
Cambridge CB2 3QG, UK

January 31, 2000

Abstract

An account of grammatical acquisition is developed within the parameter-setting framework applied to a generalized categorial grammar (GCG). The GCG is embedded in a default inheritance network yielding a natural partial ordering (reflecting generality) of parameters which determines a partial order for parameter setting. Computational simulation shows that several resulting acquisition procedures are effective on a parameter set expressing major typological distinctions based on constituent order, and defining 70 distinct full languages and over 200 subset languages. The effects on acquisition of inductive bias, that is, of differing initial parameter settings, are explored via computational simulation.

Computational simulation of *populations* of language learners and users instantiating the acquisition model show: 1) that variant acquisition procedures, with differing inductive biases, exert differing selective pressures on the evolution of language(s); 2) acquisition procedures will evolve towards more efficient variants in the environment of adaptation. The reciprocal evolution of language acquisition procedures and of languages creates a genuinely coevolutionary dynamic, despite the relative speed of linguistic selection for language variants compared to natural selection for variant language acquisition procedures.[1]

[1]I would like to thank seminar audiences at the Universities of Edinburgh, Sussex and Cambridge, and Oregon Graduate Institute, as well as the organizers and participants of the 1st and 2nd International Conferences. on Language and Evolution, University of Edinburgh, 1996, and University of East London, 1998, for helpful feedback on aspects of the work described in this paper. In particular, Bob Berwick, Miriam Eckert and Jim Hurford made suggestions which influenced the subsequent development and presentation of this work. In addition, Ann Copestake, Gerald Gazdar, Hans van Halteren, Simon Kirby, David Lightfoot, David Milward, Geoff Pullum and four anonymous reviewers for *Language* kindly gave comments on earlier drafts which helped considerably improve the present one.

1 Theoretical Background

It is widely accepted that language acquisition is guided by an innate language acquisition procedure and a partial innate specification of the form of language. Language acquisition by children is a near-universal feat, where (partial) failure appears to correlate more with genetic deficits (e.g. Gopnik, 1994) or with an almost complete lack of linguistic input during the critical period (e.g. Curtiss, 1988), than with measures of general intelligence (e.g. Smith and Tsimpli, 1991) or the quality or informativeness of the learning environment (e.g. Bickerton, 1981; Kegl and Iwata, 1989; Ochs and Sheiffelin, 1995).[2] There is considerable psycholinguistic evidence that children have strong inductive biases in language acquisition which shape their linguistic development, the nature of their errors, and the kind of languages they are predisposed to learn. Often these biases are partially incorrect as generalizations about the nature of human languages. For example, 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. Similarly, Clark (1993) argues for a principle of contrast in lexical acquisition, suggesting that children hypothesize novel meanings for novel words, ignoring, at least initially, the hypothesis that a new word may be synonymous with a known one.

This paper presents a model of the grammatical acquisition procedure in order to address the related questions of how such partially inaccurate inductive biases arise and how the language acquisition device evolved. To develop a precise and coherent answer to these questions it is not enough to model the acquisition device, we also need to characterize the environment in which such a device could emerge and evolve (e.g. Pinker and Bloom, 1990). This environment must have included (proto)language(s) capable of creating (natural) selection pressure in favour of more efficient and reliable language acquisition (e.g. Kirby, 1998). In turn, the evolving language acquisition device created (linguistic) selection pressure in favour of more learnable languages (e.g. Kirby and Hurford, 1997a). Inductive bias in language acquisition emerges as a direct consequence of this coevolutionary process.

Section 1.1 briefly surveys the current state of research on grammatical acquisition and provides the theoretical background which motivates the acquisition model presented in section 2. Section 1.2 considers the utility of evolutionary theory in the characterisation of language development, and section 1.3 argues for a consequent coevolutionary account of the development of language and of the language acquisition device. These sections motivate the development of the model and associated computational simulations of populations of language learners and language users in sections 3, 5 and 6.

1.1 Grammatical Acquisition

Grammatical acquisition proceeds on the basis of a partial genotypic specification of (universal) grammar (UG) complemented by a procedure enabling the child to complete this specification appropriately on exposure to finite positive samples from other speakers. The parameter setting framework of Chomsky (1981) claims that learning involves fixing the values of a finite set of finite-valued parameters to select a single fully-specified grammar from within the space defined by the genotypic specification of UG. Formal models of parameter setting have been developed for small sets of grammars, but even the search spaces defined by these models contain local maxima and subset-superset relations which may cause a learner to converge to

[2] See, e.g., Pinker (1994) or Aitchison (1996) for recent positive summaries and discussion of this evidence. See Sampson (1989) for a dissenting view.

an incorrect grammar (Clark, 1992; Gibson and Wexler, 1994; Niyogi and Berwick, 1996; Wexler and Manzini, 1987).

Gibson and Wexler (1994) formalize the concept of a trigger (e.g. Lightfoot, 1992:13f) as a simple (unembedded or degree-0) sentence of primary linguistic data which signals the value of some parameter and can serve to guide the learner to the target grammar. The notion of a trigger is a refinement of that of primary linguistic data, which, through context of use, unambiguously signals a particular surface form (SF) to logical form (LF) pairing (e.g. Wexler and Culicover, 1980). Thus the task of the learner faced with a trigger, or SF-LF pairing, not expressible given the current grammar, is to update a parameter such that the trigger can be parsed appropriately. Frank and Kapur (1996) demonstrate that the existence of locking sequences of such triggers, guaranteeing convergence to a target grammar, depends on the nature of the parameters, on the specific acquisition procedure, and on the initial configuration of the parameter set.

Chomsky (1981:7f) argued that at least some parameters probably have an initially unmarked or default value which will be retained by the learner unless incompatible with input. That is, that the learner is biased towards certain settings of some parameters. Unmarked, default values have been proposed as a mechanism for avoiding premature acquisition of a superset grammar (Hyams, 1986; Wexler and Manzini, 1987; Lightfoot, 1992). The effect of unmarked, default parameter settings is to introduce one type of inductive bias into the acquisition procedure by ordering grammars within the hypothesis space. That is, it introduces preferences which both serve to guide the acquisition process and act as a form of soft constraint guiding learning.[3] Such soft constraints on initial parameter settings do not preclude languages with certain properties, as do hard, innate constraints on parameters of variation. Nevertheless, soft constraints may play a very similar role to innate constraints, in practice, as we will see in section 1.3.

By contrast, formal work on parameter setting has tended to assume arbitrary initial configurations of parameters in evaluating learnability, perhaps because initial unmarked settings have only been proposed and justified for a few putative parameters. Nevertheless, formal learnability results can be affected by this starting point (e.g. Gibson and Wexler, 1994). In addition, there have been few proposals concerning the grammatical representation and formalization of the distinction between initially unset and initially default parameters. Chomsky (1981:8) also proposed that the same mechanism might well be responsible for acquisition of the periphery of marked idiosyncratic constructions for which positive evidence was provided by a given speech community. Thus, such a mechanism might account for some of the partially inaccurate inductive biases, for example of the type noted by Wanner and Gleitman (1982), within the parameter setting framework. Bickerton (1984) also argues that the abrupt transition from pidgin to creole suggests that children are endowed genetically with initial unmarked parameter settings specifying the stereotypical core creole grammar.

Pullum (1983) criticizes the parameter setting framework because it predicts that the space of possible grammars, and thus languages, is vast, though finite (20 independent binary parameters yields 2^{20} or 1048,576 grammars, while 30 such parameters yields 1,073,741,824 distinct grammars), and because few if any psychologically feasible, as opposed to merely computationally tractable, acquisition procedures have been proposed within this framework. For example, brute force

[3] In the literature on machine learning, the term (inductive) bias is often used to characterize both the hard constraints and softer preferences which together constitute the background knowledge utilized by a learning procedure (e.g. Mitchell, 1990). Here I reserve the term (inductive) bias for the soft preferences defined both in terms of the initial settings of parameters and the language acquisition algorithm which can be overridden by data. Hard (innate) constraints defined by the parameterization of UG which cannot be overridden by data are referred to as (innate) constraints.

search through the space of distinct grammars will require time proportional to their number (e.g. Clark, 1992), while the number of positive samples of the language, and hence amount of time required for convergence to a target grammar, can be arbitrarily long depending on the distribution of trigger types in the language (e.g. Niyogi and Berwick, 1996). Brent (1996) argues that Markovian 'memoryless' procedures of the type introduced by Gibson and Wexler (1994) and further investigated by Niyogi and Berwick (1996) are psychologically implausible because they predict that a child may repeatedly revisit the same hypothesis and/or 'jump' randomly around the hypothesis space.[4]

The model presented in section 2 addresses these issues via a modified parameter setting procedure, which can acquire more complex grammars, utilizes limited memory and is, therefore, more directed and less psychologically implausible. The modified procedure is based on a partial ordering on the updating of parameter settings, defining the category set and rule schemata available in a generalized categorial grammar. The partial ordering is obtained by use of a default inheritance network as the grammatical representation language. The generalized categorial grammar (GCG) framework supports a more articulated account of UG than is typically deployed in recent formal work on parameter setting, enabling a wider space of grammars to be explored, and a richer notion of parameter updating to emerge. Parameters are deterministically set in (partial) order of their generality in defining the space of possible grammatical categories, ensuring that grammatical hypotheses remain maximally specific, though further modifiable due to the defeasibility of their consequences. The grammatical representation language provides a formal means for distinguishing unset, default and absolute specification, and thus for distinguishing unset or (un)marked parameters from principles. Variants of this acquisition procedure can be defined based on the criterion adopted for retaining a new parameter setting, on the number of new parameter settings per trigger, on the existence of (maturational) working memory limits during learning, and on the initial configuration of the parameter set.

1.2 Linguistic Evolution

An important insight of diachronic generative linguistics is that language acquisition is the primary engine of language change (e.g. Lightfoot, 1979). In recent generative work on diachronic syntax, language change is primarily located in parameter resetting (reanalysis) during language acquisition (e.g. Lightfoot, 1992, 1999; Clark and Roberts, 1993; Kroch and Taylor, 1997). Differential learnability of grammatical systems, on the basis of learners' exposure to triggering data from varying grammatical sources, causes language change. To a first approximation once the critical period for language acquisition is complete, the grammars internalized by learners do not change further. However, although generative linguistics offers a precise formal characterization of idiolects as the stringset generated by an ideal speaker-hearer's internalized grammar, a similar characterization of language, the (changing) aggregate output of the (changing) membership of a speech community, has been lacking until recently.[5]

To study and characterize language, it is necessary to move from the study of individual (idealized) language learners and users, endowed with a LAD and acquiring an idiolect, to the study of *populations* of such generative language learners and

[4]The term *random* is used throughout in the sense of randomly sampling from an underlying (uniform) distribution. In the computatonal work discussed, this is simulated deterministically using standard techniques.

[5]Chomsky (1986) makes a related distinction between I-language and E-language. I will use *idiolect* and *language* to distinguish the output of a single internalized grammar and the aggregate output of a speech community (or set of internalized grammars) respectively.

users, parsing, learning and generating a set of idiolects constituting the language of the community. Once this step is taken, then the dynamic nature of language emerges more or less inevitably. Occasional misconvergences on the part of language learners can introduce variation into a previously homogeneous linguistic environment, fluctuations in the proportion of learners to adults in the population can skew the distribution of primary linguistic data significantly enough to affect grammatical acquisition, and so forth. Once such variation is introduced, then properties of the acquisition procedure become critical in determining which grammatical forms will be differentially selected for and maintained in the language, with language acquisition across the generations of users as the the primary form of linguistic inheritance.

Niyogi and Berwick (1997a,b; Niyogi, 2000) model language as a dynamical system in which each each state of the system is defined by a population of possibly differing individual adults' grammars. The language of this population is defined in terms of the aggregate output of these grammars and characterizes the triggering data for a new non-overlapping generation of learners. Niyogi and Berwick show that such a dynamical system can be fully characterized in terms of a space of possible grammars, G, a probability distribution, P, on sentences generated by a (mixed) population of grammars, and a parameter setting algorithm, A, which selects a grammar, $g \in G$, given a sequence of triggers. Each state of the dynamical system is represented by P, which defines the data sampled by the next generation of learners. Niyogi and Berwick, thus, formalize language, in a way which takes full account of variation and change, in terms of a changing population of generative grammars. The update rule which maps the dynamical system from one state to the next is defined in terms of G, A and P. In the instantiations of the model they consider, they show language change occurring through misconvergence of learners when P is initially not mixed, and through frequency-driven convergence to one variant grammar when P is mixed. They derive some results concerning the speed of change and stable configurations of the dynamical system.

Niyogi and Berwick's account of language change can be viewed technically as an evolutionary process. Although we typically think of evolution applying to biological organisms, it can be defined more abstractly as a process which occurs whenever we are dealing with a dynamical system in which there is a source of (random) variation, (differential) inheritance and (possibly) selection (e.g. Dawkins, 1983; Cziko, 1995). Niyogi and Berwick's model is either seeded with variation in the initial population of grammars or such variation emerges through misconvergence of some learners to different grammars. Successive generations of learners 'inherit' grammars on the basis of A, but there is no selection preference for particular grammars *inherent* in A. That is, the models of A that they consider either choose $g \in G$ based on the data seen or make an unbiased random selection if the data is not decisive. Thus, A does not incorporate inductive bias. As such, the model they develop is based on differential 'inheritance' of linguistic variants via language acquisition, but it does not involve any form of selection between variants on any basis other than their relative frequency of exemplification in triggering data. There is no sense in their model that one variant is more adaptive than any other. As such, their model is closest to the so-called neutral theory of evolution (Kimura, 1983), in which evolution in finite populations is argued to be primarily a consequence of random genetic drift.

It is simple to modify the Niyogi and Berwick model to incorporate a linguistic analogue of natural selection and thus to model languages as adaptive systems. We need only construct a variant account of acquisition, A, which incorporates inductive bias in some form. For example, if we posit a version of A in which triggers are used to select $g \in G$ but when the sequence of triggers is not decisive, A selects the most preferred $g \in G$ compatible with the triggering data, perhaps on the basis of default, unmarked initial settings of (some) parameters, then we immediately have a model in which it makes sense to think of languages as *adaptive* to their ecological niche; in this case the possibly arbitrary preferences built into the LAD. In this model, the inductive bias of A embodied in the LAD creates a selection pressure for certain variants over others, so that variants more compatible with the bias will stand a better chance of being passed on through successive generations of language learners. When we say that these variants are more *adaptive* we say nothing more than that they are adaptive with respect to this inductive bias. There is no entailment that this makes the variants better in any wider sense. The inductive bias in A may be nothing more than the chance fixation of a random and arbitrary mutation in an ancestral population of protolanguage speakers, though it may also have a deeper functional basis in cognitive capacities and limitations.[6]

One conceptual problem which emerges in a model which characterizes languages as adaptive to A in this way is that, assuming that A is fixed and common to the entire population, it becomes hard to find a plausible source of linguistic variation (at least with respect to these adaptive aspects of languages). So if languages adapt solely to inductive bias during language acquisition, we would expect the history of languages to show nearly inexorable development towards an optimal or most adaptive grammar (e.g. Lightfoot, 1999). Chance factors might temporarily move a language away from such a grammar but, over time, constant and universal selection pressure should (re)assert itself. But this is not what has happened; the development of (proto)language(s) seems to have led rather to increasing linguistic diversity (e.g. Nettle, 1999:2f). However, evolution is *not* a process of steady improvement along a single trajectory leading to a single optimal solution. Sewall Wright (1931) introduced into evolutionary theory the idea of adaptive or fitness landscapes with multiple local optima or peaks, and this idea has been considerably refined since (e.g. Kauffman, 1993:33f). The modern picture of (co)evolution is of a process of local search or hill climbing towards a local optimum or peak in an adaptive landscape which itself inevitably changes. Conflicting selection pressures will cause this landscape to contain many locally optimal solutions, and thus the evolutionary pathways will be more complex and the space of near optimal solutions more varied (e.g. Kauffman, 1993:44f). A simple and well-attested example of conflicting selection pressures from biology is the case of 'runaway' sexual selection for a non-functional marker such as the peacock's tail, counterbalanced by natural selection for efficient movement (e.g. Dawkins, 1989:158f).

The idea that there are competing motivations or conflicting pressures deriving from both the social and functional exigencies of language production, comprehension and acquisition has been developed by linguists working from many different perspectives (e.g. Langacker, 1977; Fodor, 1981; Croft, 1990:192f; Lindblom, 1998; Nettle, 1999:32f). For example, pressure for articulatory economy in production may conflict with pressure for ease of decoding in comprehension, while social pressure for intergroup variation will tend to diffuse variants regardless of their functional adaptiveness. In general, little progress has been made in quantifying these pressures or exploring their interaction. Nevertheless, once we recognize that there may be such conflicting selection pressures, it is easier to see why language change does not move inexorably (and unidirectionally) towards a unique global optimum. No such optimum may exist, and in any case, change will always be relative to and local with respect to the current 'position' in the current adaptive landscape. Languages are complex systems with many interacting levels of structure. Therefore, at any given point many alternative changes in one of these subsystems may be adaptive and these may create widely different further pressures throughout the system.

[6] The term *fixation* is used throughout to refer to the spread of a variant through the members of a population to the point where all members share that variant.

Which, if any, of these alternatives is adopted by a speech community may depend on historical or chance factors. Thus, conflicting linguistic selection pressures will naturally lead to diversity if speech communities are linguistically isolated.[7]

As Niyogi and Berwick (1997a,b) argue, the behaviour of all but the simplest dynamical systems is often unintuitive; while analytic proofs of the behaviour of classes of such systems are only possible when the number of variables involved is severely limited. For these reasons, a computational simulation methodology is utilized here, which allows more complex models to be studied experimentally. It is important that simulations strike the right balance between idealization and ecological validity, ignoring irrelevant complexities, but modelling potentially relevant factors, and making critical assumptions explicit. A simulation of linguistic evolution, at a minimum, needs to provide a source of linguistic variants on which selection can work and a realistic model of language acquisition which will form the basis of both the inheritance of and possibly selection amongst those variants. But before, developing such a model we need to consider the relationship between linguistic evolution and the biological evolution of the LAD.

1.3 Coevolution and Genetic Assimilation

Pinker and Bloom (1990) argue for an adaptationist account of the *biological* evolution of the language acquisition device (LAD) suggesting that the domain-specific linguistic knowledge required to support reliable language learning was genetically assimilated via natural selection for more successful language learners since the emergence of structured language.[8] 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 appropriate trait develops reliably on

[7] Müller, Schleicher and other 19th century linguists speculated that languages evolved according to Darwinian theory, and Darwin (1871) endorsed the idea, quoting with approval from Müller: 'A struggle for life is constantly going on amongst the words and grammatical forms in each language. The better, the shorter, the easier forms are constantly gaining the upper hand, and they owe their success to their own inherent virtue.' See Harris and Taylor (1997:ch14) and McMahon (1994:ch12) for more discussion of the checkered relationship between Darwinian and linguistic theory, and Keller (1994:46f) for a critical discussion of Müller and Schleicher's theories of language. Keller (1994) goes on to develop an account of language very similar to that taken here. The earliest post-generative proposals that languages be viewed as adaptive systems of which I am aware are Lindblom (1986) and Hurford (1987), though this position is articulated far more clearly in more recent work such as Hurford (1999), Kirby (1999), Lindblom (1998), and Nettle (1999). Though the view that languages evolve has a generally bad press in linguistic theory, it is worth reconsidering as the neo-Darwinian synthesis and subsequent analytic and algorithmic thinking about evolution and dynamical systems makes available a panoply of new perspectives and techniques that were not available to nineteenth century or even nineteen sixties linguists. For accessible introductions to this new intellectual landscape see e.g. Dennett (1995), Kauffman (1993), Peak and Frame (1994), Sigmund (1993).

[8] This aspect of their argument, at least, is distinct from the question of whether the LAD emerged via a biological saltation or gradually. Berwick (1997) argues that the Merge operation of the Minimalist Program (e.g. Chomsky, 1991) might have been exapted via genetic drift. This specific proposal is quite compatible with the framework presented here, in which function-argument application plays a similarly central role to Merge (see section 2.1). Indeed, as Steedman (1996:14f) notes, the deterministic mapping via categorial rules of application, composition, and so forth from SF to LF strengthens the case for an evolutionary pathway in terms of the development of such rules of 'realization' for pre-existing conceptual structures (see e.g. Bickerton, 1998; Worden, 1998). However, the question of the origin of the LAD, as opposed to its subsequent evolution and maintenance, is not addressed further in this paper.

the basis of minimal environmental stimulus, providing that the environment, and consequent selection pressure, remains constant over enough generations.[9]

As an example of genetic assimilation, Durham (1991) discusses in detail the case of widespread, though by no means universal, lactose tolerance in adult humans. Many of us, uniquely amongst mammals, continue to be able to easily digest milk after weaning. In many parts of the world the growth of animal husbandry created a new and reliable source of nutrition – milk. Creating the potential for selection for individuals more able to exploit this resource for longer periods of their lifetime. Lactose tolerance has been genetically assimilated by the great majority in populations where milk has been reliably available over many generations. Although it is not possible to relate lactose tolerance directly to specific genetic differences (yet), Durham demonstrates convincingly that the incidence of intolerance correlates, in a manner compatible with a genetic explanation, with a fairly recent introduction of diary products and with warm climates, where lack of Vitamin D is less potentially problematic.[10]

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, 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 the same suggestion, citing a computational simulation by Hinton and Nowlan (1987) showing genetic assimilation of initial 'gene' settings facilitating learning.

In Hinton and Nowlan's (highly abstract) simulation, a population of 1000 individuals with 20 'genes' or switches which could be unset (?), on (1), or off (0) was evolved using a genetic algorithm (see e.g. Holland, 1995). A 'successful' individual had all 20 switches on, but individuals were initialized randomly with switch probabilities of 0.5 for ? and 0.25 for 1 or 0 at each position. Each individual was able to set ? switches through learning on the basis of 1000 trials during its lifetime. The fitness of an individual was defined as $1 + 19(1000 - n)/1000$ where n is the number of trials after it has learnt the correct settings, making an individual 'born' with all settings correct (i.e. 1) 20 times fitter than one which never learnt them. Reproduction of offspring individuals was by crossover of the *initial* switch settings from two parents whose selection was proportional to their fitness. In the early generations most individuals had the same minimum fitness as a consequence of being born with one or more 0 settings. However, this soon gave way to exponential increases in individuals with more 1 settings, less ? settings and no 0 settings. Then, in the later stages, the increase of 1s and decrease of ?s tailed off, once all

[9]Waddington's work on genetic assimilation is a neo-Darwinian refinement of an idea independently discovered 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).

[10]Evolutionary biologists accept the possibility of genetic assimilation (e.g. Maynard Smith, 1987, 1993:319f; Rose, 1997:217f), however, some (e.g. Dawkins, 1982:284) regard it as a 'hypothetical' mechanism because, though it has been demonstrated experimentally, it has not been conclusively shown to occur naturally. It is extremely difficult to prove a case of natural, adaptive genetic assimilation. Nevertheless, the developmental view of evolution, which Waddington pioneered, is gaining ground as more is understood about the relationship between genes and environment in morphogenesis (e.g. Jablonka and Lamb, 1995). Cognitive learning theorists, such as Cosmides and Tooby (1996) and Staddon (1988) argue that evolution has equipped organisms with specialized domain-specific learning mechanisms incorporating inductive bias assimilated from the environment of adaptation for that learning mechanism.

the individuals in the population were capable of learning successfully.

Hinton and Nowlan point out that the search space defined by the model – its adaptive landscape – is like a needle in a haystack, only one overall setting of all 20 switches confers any selective advantage whatsoever. Therefore, evolution unguided by learning would be expected to take on the order of 2^{20} trials (i.e. individuals) to find a solution. If increased fitness required evolution of two such individuals in the same generation, as would be the case for coordinated behaviour, evolution would be expected to take around 2^{400} trials to find a solution (without considering further restrictions created by spatial distribution of the population). However, with learning (modelled as random flipping of ? settings) the model always converges within 10-15 generations on a successful individual (i.e. after generating less than 150,000 individuals). Furthermore, the model shows clearly that, once successful individuals appear, their superior performance rapidly leads to the spread of genotypes (i.e. initial switch settings) which support successful learning and which move their genotypes close to the optimal solution.[11]

One potential complication for an account of the evolution of the LAD in terms of genetic assimilation is that it does not explain why the process should not have continued until the point where a fully-specified grammar had been assimilated, and grammatical acquisition became redundant. Waddington (1975:307) remarks: 'Evolution is quite capable of performing such a feat... But in the case of language, there is certainly little reason to see why it would have been advantageous to press the matter further. If a child which had never met a language-user developed the ability to talk, who after all would it talk to?' Nevertheless, the propensity to use a fully-specified grammar, given minimal triggering input, would simplify the language acquisition problem to one of vocabulary acquisition. Pinker and Bloom (1990), following Hinton and Nowlan (1987), argue that selection pressure to set the remaining initial unset genes in Hinton and Nowlan's individuals is weak once they have evolved to learn reliably. Harvey (1993) demonstrates that this is an artifact of Hinton and Nowlan's simulation. Without mutation as well as crossover of the initial settings, later more effective individuals almost invariably evolve from a single ancestor, causing 'premature' (and artifactual) fixation on some unset switches, and thus preventing the population from evolving further. However, with mutation, Hinton and Nowlan's fitness function is too weak to overcome the effects of random genetic drift. As long as there is selection pressure for a fully-developed capacity, we would expect no learning, and thus no delay in acquisition of the trait, to be the optimal solution.[12]

Deacon (1997:102f,327f) rejects any account of the evolution of a LAD via genetic assimilation, on the basis that genetic assimilation requires an unchanging environment to create sustained and constant selection pressure over the many generations required for genotypic adaptation. Pinker and Bloom (1990) simply assume that linguistic universals are evidence of enough constancy in the environment to allow genetic assimilation. However, the usual explanation for the existence of (absolute) universals is that they are embodied in the LAD, so there is a danger of circularity here. Deacon (1997:116f) instead argues for the contrary position that all linguistic 'universal[s]... emerged spontaneously and independently in each evolving language, in response to universal biases in the selection processes affecting language transmission. They are *convergent* features of language evolution in the same ways as dorsal fins of sharks, ichthyosaurs, and dolphins are independent convergent adaptations

[11] Turkel (2000) discusses the Hinton and Nowlan model in greater detail from a linguistic perspective, and develops an account of the evolution of the LAD using a slight variant.

[12] Ackley and Littman (1991) and Cecconi *et al.* (1996) describe unrelated simulations which, unlike Hinton and Nowlan, distinguish phenotype and genotype, do not make use of a fixed externally-defined fitness function, and do model learning cost – in these simulations learning is eventually entirely displaced, given a constant environment, as expected.

of acquatic species.' He suggests, in particular, that languages have evolved to be easily learnable by a procedure which 'starts small', following Newport (1990) and Elman (1993), with a limited working memory only capable of 'seeing' local grammatical dependencies. Furthermore, Deacon (1997:328f) argues that the surface grammatical organization of languages changes with such speed relative to genetic evolution that there could not have been consistent enough selection pressure for genetic assimilation. Thus, Deacon argues that universals are a consequence of convergent evolution of languages in response to inductive bias causes by working memory limitations and perhaps other general cognitive limitations.

Deacon's position can be criticized on three levels. Firstly, it is unclear that he recognizes the import of linguistic learnability arguments and the relevance or existence of abstract universals (without clear 'surface' effects). For example, the language acquisition procedure presented in section 2 can parse and acquire grammatical constructions involving cross-serial grammatical dependencies, such as those exemplified in the formal language $a^n b^n c^n$ (where $n \geq 1$), Swiss German syntax and Bambara morphology (e.g. Shieber, 1985; Gazdar 1988), but not constructions involving the MIX or Bach language variant in which any ordering of equal numbers of the *a*s, *b*s and *c*s is grammatical, creating arbitrarily intersecting dependencies. Whether a language exhibits cross-serial or arbitrarily intersecting dependencies is an apparently rather abstract feature which does not fit well into traditional more 'surfacy' characterizations of languages as, say, inflecting, agglutinating or isolating, or head-initial /-final, and so forth. Nevertheless, this has profound consequences for the kind of rule system capable of expressing the mapping from SF to LF. Not least, that a formal proof of learnability has been found for grammatical frameworks capable of expressing cross-serial dependencies (Joshi *et al.*, 1991), but not for those able to express arbitrarily intersecting dependencies. The genetic assimilation of a language-specific rule *system* (the UG component of the LAD) remains a theoretical possibility, even if the emergence of such abstract universals can be traced to non-domain-specific factors, such as working memory limitations, and even if languages change their surface characteristics, such as constituent order, to fast for genetic assimilation.

Secondly, Deacon relies heavily on the 'starting small' hypothesis and Elman's (1993) experiments training recurrent neural networks to approximate recognition of context-free languages. While these experiments demonstrate a clear requirement for initially training on short sequences containing local grammatical dependencies, it is unclear what consequences this has for grammatical acquisition by human learners. Elman's networks do not have the expressive power to encode SF-LF mappings and, therefore, to underpin a model of language production and interpretation. It is not, *a priori* obvious that when we move to consider models with this capacity, and their associated acquisition procedures, that a similar effect will be observed. In fact, though the experiments reported in sections 4.1 and 7 do show that the assumption of maturational memory limitations during language acquisition does affect predictions concerning the differential learnability of languages in the framework developed here, they do not show any effect on the learnability of languages *per se*.

Thirdly, in recent years, the increased use of mathematical tools and computational simulation has demonstrated the probability of extensive coevolutionary interactions across species, such as predator-prey interactions, competitive and benign host-parasite interactions, plant-insect interactions, and so forth (e.g. Futuyma and Slatkin, 1983; Kauffman, 1993:242f; Maynard Smith, 1998:285f). Most of these interactions involve species evolving at different rates, as the lifespan of the parasite is usually far shorter than that of the host. Though Waddington's neo-Darwinian mechanism of genetic assimilation remains the basis for (co)evolution in response to environmental change, this work suggests that relative speed alone cannot conclusively be used to reject the possibility of genetic assimilation in response to pressure from an evolving linguistic environment. Interestingly, though Deacon (1997:112-13) draws the analogy between language and symbiotic bacteria (for example, those found in the human gut which aid digestion) and subtitles his book 'co-evolution of language and brain', he does not explicitly discuss the recent literature on coevolution, or whether this might warrant reconsideration of how environmental changes affect genetic assimilation. The speed at which linguistic changes can diffuse through a population will be far faster than that at which genetic change can do so. However, there is clearly a speed limit to linguistic change within a successfully communicating population, and that speed limit means that only a small part of the space of possible grammars may be sampled over the period required for biological evolution. The experiments reported in section 7 suggest this can lead to a constant enough selection pressure capable of supporting genetic assimilation of a LAD. However, equally crucially the fact of linguistic change provides a natural barrier to total genetic assimilation of a fully-specified grammar even in the face of an adaptive advantage for faster language acquisition.

The simulation presented in section 3 models both natural selection for variant language acquisition procedures and linguistic selection for languages. Therefore, it is possible to both explore what kind of acquisition procedure might evolve and what effects different acquisition procedures might have on the grammatical systems which evolve, given varying assumptions about the role of maturational working memory limitations, the adaptive advantage of language to language users, and so forth. The simulation can be set up to model either a neutral relationship or benign, symbiotic relationship between languages and their potential users. That is, one in which the ability to communicate via language either confers no selective advantage (or disadvantage) or one which confers some (unspecified) selective advantage to its users. Additionally, in some experiments, the ability to communicate using a more learnable, expressive or interpretable variant language can confer greater relative advantage. Roughgarden (1983) argues that mutualistic coevolution between 'host' (language users) and 'guest' (language idiolects) organisms will only occur when the host benefits (and the experiments reported in section 7 bear out this prediction).

On the assumption that language confers selective advantage, linguistic variants will compete for language users on the basis of their relative learnability, and, possibly their interpretability and/or expressiveness. Language users will also evolve language faculties which improve their capacity to acquire and use language. Given this scenario, a language can be viewed as a parasitic coevolving species. Under the alternative assumption that language confers no selective advantage, linguistic variants will compete for language users solely on the basis of their learnability with respect to whatever acquisition procedure is in place. However, there will be no pressure for this acquisition procedure to evolve to favour any particular linguistic variants. Thus, a language can still be seen as a dynamical system adapting to the requirements of learnability, but language will have no influence on biological evolution. Nevertheless, given the implausibility of assuming that language confers no selective advantage, whatever form this might take, the coevolutionary scenario seems highly likely.

There are several ways in which linguistic evolution and biological evolution might be argued to be qualitatively different, in addition to such quantitative differences as relative speed of change. Linguistic variants may compete for language users, but it can be argued they do not have a fitness, in the technical sense of expected number or proportion of offspring (e.g. Maynard Smith, 1998:36f). Rather the primary mechanism of linguistic inheritance is through a child language learner *actively* learning their idiolect, rather than the gene replicating via the medium of DNA (e.g. Keller, 1994). The degree to which this distinction can be upheld depends on the extent to which a gene is defined as a biochemical object, as opposed to a unit of information.[13] In the simulation model, language users may have (relative) fitness as a consequence, primarily, of their communicative success, while languages have (relative) cost to users depending, primarily, on their fit with their acquisition procedures. A different but related question concerns the units of linguistic selection, and whether there can be a corresponding distinction between phenotype and genotype in linguistic evolution. Linguistic variation is defined in terms of *competing* constructions which form part of the linguistic environment (or phenotype). Such variants compete by virtue of being in parametric variation or, perhaps more generally, because they are variant means of expressing the same meaning. In terms of the model of the LAD developed in section 2 and simulation model of section 3, the parameters of UG which define the range of specific grammars form the ultimate units of linguistic selection.

2 The Language Acquisition Device

A realistic model of the LAD must incorporate a UG with associated parameters, a parser, and an algorithm for updating initial parameter settings on parse failure during acquisition (e.g. Clark, 1992). The following sections present such a model, which builds on and extends previous work reviewed in section 1.1.

2.1 The Parameter Set

Classical (AB) categorial grammar (e.g. Wood, 1993:7f) employs a category set consisting of a small number of atomic categories (e.g. S,N,NP) and derived complex functor categories (e.g. N/N 'prenominal adjective', S\ NP, 'intransitive verb'), and uses one rule of application which combines a complex functor category, containing a (back)slash, with an argument category to form a derived category (with one less slashed argument category). Grammatical constraints of order and agreement are captured by only allowing directed application to adjacent matching categories. Generalized categorial grammars (GCGs) extend the AB system with further rule schemata (e.g. Wood, 1993:34f). The rules of forward application (FA), backward application (BA), generalized weak permutation (P) and forward and backward composition (FC, BC) are given in Figure 1 (where X, Y and Z are category variables, | is a variable over slash and backslash, and ... denotes zero or more further functor arguments). Generalized weak permutation enables cyclical permutation of argument categories, but not modification of their directionality. Each rule has an associated semantic operation represented here in terms of the (Typed) Lambda Calculus. Once permutation is included, several semantically equivalent derivations for *Kim loves Sandy* become available, Figure 2 shows the non-conventional left-branching one.[14] Composition also makes alternative non-conventional semantically-equivalent (left-branching) derivations available, as Figure 3 illustrates. Steedman (1988, 1996) presents arguments for the linguistic utility of composition rules.

GCG as presented is inadequate as an account of UG or of any individual grammar. In particular, the definition of atomic categories needs extending to deal with featural variation, further unary/lexical rules will be needed (e.g. Bouma and van

[13] Dawkins (1982:109f) and Dennett (1991:341f) make similar points discussing the differences between genes and memes (minimal ideational units of cultural inheritance putatively subject to cultural selection).

[14] Generalized weak permutation (P) is more powerful than the rule sometimes called associativity (e.g. Wood, 1993:37f) which licenses $(X/Y)\backslash Z \Rightarrow (X\backslash Z)/Y$ but not $(X/Y)/Z \Rightarrow (X/Z)/Y$, since the latter is also licensed by P. However, P is less powerful than permutation in the extended Lambek calculus **LP** (e.g. Wood 1993:64f; Moortgat, 1988:45f) in which directional constraints are no longer maintained.

Forward Application (FA):	
$X/Y \ Y \Rightarrow X$	$\lambda \ y \ [X(y)] \ (y) \Rightarrow X(y)$
Backward Application (BA):	
$Y \ X\backslash Y \Rightarrow X$	$\lambda \ y \ [X(y)] \ (y) \Rightarrow X(y)$
Forward Composition (FC):	
$X/Y \ Y/Z \Rightarrow X/Z$	$\lambda \ y \ [X(y)] \ \lambda \ z \ [Y(z)] \Rightarrow \lambda \ z \ [X(Y(z))]$
Backward Composition (BC):	
$Y\backslash Z \ X\backslash Y \Rightarrow X\backslash Z$	$\lambda \ z \ [Y(z)] \ \lambda \ y \ [X(y)] \Rightarrow \lambda \ z \ [X(Y(z))]$
(Generalized Weak) Permutation (P):	
$(X\|Y_1)\ldots\|Y_n \Rightarrow (X\|Y_n)\|Y_1\ldots$	$\lambda \ y_n \ldots, y_1 \ [X(y_1 \ldots, y_n)] \Rightarrow \lambda \ldots y_1, y_n \ [X(y_1 \ldots, y_n)]$

Figure 1: GCG Rule Schemata

Kim	loves	Sandy
NP	$(S\backslash NP)/NP$	NP
kim'	λ y,x [love′(x y)]	sandy′
	————–P	
	$(S/NP)\backslash NP$	
	λ x,y [love′(x y)]	

————————————BA

S/NP

λ y [love′(kim′ y)]

————————————————————FA

S

love′(kim′ sandy′)

Figure 2: GCG Derivation for *Kim loves Sandy*

Noord, 1994), and the rule schemata, especially FC, BC and P, must be restricted in various parametric ways so that overgeneration is prevented for specific languages (e.g. Morrill, 1994). Nevertheless, GCG does represent a plausible kernel of UG; Hoffman (1995, 1996) explores the descriptive power of a very similar system, in which P is not required because functor arguments are interpreted as multisets. She demonstrates that this system can handle (long-distance) scrambling elegantly and generate some mildly context-sensitive, though not some MIX, languages (Joshi *et al*, 1991).

The relationship between GCG as a theory of UG (GCUG) and as a specification of a particular grammar is captured by embedding the theory in a default inheritance network.[15] Figure 4 illustrates schematically and informally a fragment of a such a network. The network defines intensionally the set of possible categories and

[15]This can be formalized as a semi-lattice of typed default feature structures (TDFSs) representing subsumption and default inheritance relationships (Lascarides *et al*, 1996; Lascarides and Copestake, 1999) supporting multiple orthogonal (default) inheritance. The TDFS formalism allows absolute specification, default specification, or unset values in feature structures. These possibilities correspond here to inherited principles and default-valued or unset parameters, respectively.

The	big	bad	wolf
NP/N	N/N	N/N	N
$\lambda P\ [\mathrm{the}'(x) \wedge (P(x))]$	$\lambda P\ [\mathrm{big}'(P)]$	$\lambda P\ [\mathrm{bad}'(P)]$	$\lambda x\ [\mathrm{wolf}'(x)]$

———————————————-FC

$\lambda P\ [\mathrm{the}'(x) \wedge \mathrm{big}'(P(x))]$

————————————————————-FC

$\lambda P\ [\mathrm{the}'(x) \wedge \mathrm{big}'(\mathrm{bad}'(P(x)))]$

——————————————————————————-FC

$\mathrm{the}'(x) \wedge (\mathrm{big}'(\mathrm{bad}'(\mathrm{wolf}'(x))))$

Figure 3: GCG Derivation for *The big bad wolf*

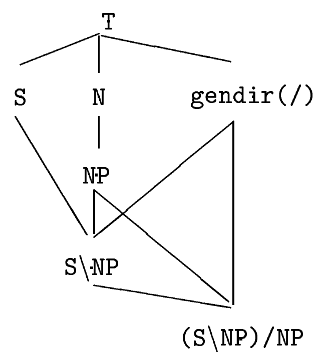

Figure 4: Network fragment for a category set

rule schemata via type declarations on nodes. For instance, an intransitive verb is treated as a subtype of verb, inheriting subject directionality by default from a type **gendir** (for general direction). For English, **gendir** is default **right** (/) but the node of the (intransitive) functor category, where the directionality of subject arguments is specified (**subjdir**), overrides this to **left** (\), reflecting the fact that English is predominantly right-branching, though subjects appear to the left of the verb. A transitive verb inherits its structure from the type for intransitive verbs and an extra NP argument with default directionality specified by **gendir**, and so forth. A full specification of English will also declare English verbs, such as *smile* and *love*, to be instances of the appropriate categories (types). Wood (1993) further discusses techniques for embedding categorial grammars, including rule schemata, in constraint / unification-based representation languages.

For the purposes of the computational simulation described in section 3, GC(U)Gs are represented as a sequence of *p-setting*s (where *p* denotes principles or parameters) based on a flat (ternary) sequential encoding of such default inheritance networks. The inheritance network provides a partial ordering on parameters, which is exploited in the acquisition procedure. For example, the atomic categories, **N** and **S** are each represented by a parameter encoding the presence, absence or lack of specification (T(rue)/F(alse)/?, respectively) of the category in the (U)G. Since they are unordered in the semi-lattice, their ordering in the sequential coding is arbitrary. However, the ordering of the directional types **gendir** and **subjdir** (with values L(eft)/R(ight)) is significant as the latter is a more specific type. The distinctions between absolute, default or unset specifications also form part of the encoding (A/D/?, respectively). Figure 5 shows several equivalent and equally correct sequential encodings of the fragment of the English type system described above.

N	**NP**	**S**	**gendir**	**subjdir**	**applic**
A T	A T	A T	D R	D L	D T
N	**gendir**	**applic**	**S**	**NP**	**subjdir**
A T	D R	D T	A T	A T	D L
applic	**N**	**NP**	**gendir**	**subjdir**	**S**
D T	A T	A T	D R	D L	A T
...					

Figure 5: Sequential encodings of the network fragment

	NP	**N**	**S**	**P**	**COMP**	**argorder**	**Applic**	**Perm**	**Comp**
SVO	T	T	T	T	T	T	T	T	T
VSO	T	T	T	T	T	F	T	F	T
SOV	T	T	T	T	T	T	T	T	?

Figure 6: The Category Parameters

A set of grammars, $g \in GCUG$, based on typological distinctions defined by constituent order (e.g. Greenberg, 1966; Hawkins, 1994), was defined with binary-valued parameters encoding the availability of specific atomic and functor categories, the ordering of arguments in functor categories, and several others encoding, for example, the availability of P 'post-lexically' during a derivation, and the mapping from SF to LF in functor categories. The eight basic language families are defined in terms of the unmarked, canonical order of verb (V), subject (S) and objects (O). Languages within families further specify the order of modifiers and specifiers, the order of adpositions, and further phrasal-level ordering parameters.

Figure 6 lists the parameters controlling the availability of specific atomic and functor categories, of specific rule schemata, and the mapping from functor categories to LF. Examples of settings based on familiar languages such as SVO, "English", VSO, "Tagalog", SOVv2, "German", and SOV, "Japanese" are given.[16] For full languages most if not all category parameters will be T(rue) – here, for example, "Japanese" is modelled as lacking only (relative) complementizer (?, unset). **Argorder** determines the mapping from functor argument to LF, it is T(rue) for languages in which the innermost argument of the unpermuted functor is subject and F(alse) otherwise. Hence, for "Tagalog" it is F because the subject argument is outermost canonically. Figure 7 lists the language-specific ordering parameters used to further define the functor category set and thus the full set of grammars. "English" is an SVO language with prepositions, in which specifiers, complementizers and some modifiers precede heads of phrases. The L(eft)/R(ight) specifications refer to the directionality encoding of arguments in functor categories; for example, in "English" specifiers are functor categories looking for (nominal) arguments to their R(ight), while relative clauses are treated as arguments of categories like NP/Rc and thus, this category (**relcl**) is also R. There are other grammars in the SVO family in which all modifiers follow heads, there are postpositions, and so

[16]Throughout double quotes are used around language names, as convenient mnemonics for familiar combinations of parameters. Since not all aspects of these actual languages are represented in the grammars, conclusions about actual languages must be made with care. As well as SVO etc., I use the similar abbreviations, subscripted to functor arguments, to indicate the LF associated with particular CG categories where it is convenient to suppress details of the semantic framework employed; e.g. $S/NP_s/NP_o$ (VOS) etc.

	gen	v1	n	subj	obj	v2	mod	spec	relcl	adpos	compl
SVO	R	F	R	L	R	F	R	R	R	R	R
SOVv2	R	F	R	L	L	T	R	R	R	R	R
SOV	L	F	L	L	L	F	L	L	L	L	?

Figure 7: The Ordering Parameters

forth. Not all combinations of parameter settings correspond to attested languages and one entire language family (OSV) is either unattested or extremely rare (see Pullum, 1981). "Japanese" is an SOV language with postpositions in which specifiers and modifiers precede heads. There are other languages in the SOV family with less consistent left-branching syntax in which specifiers and/or modifiers precede phrasal heads, some of which are attested. "German" is a more complex SOV language in which the verb-second (v2) parameter ensures that the surface order in main clauses is usually SVO.[17]

There are 20 p-settings which determine the rule schemata available, the atomic category set, the 'shape' of functor categories, and the mapping to LF. The parameter set licenses just under 300 grammars, $g \in GCUG$. Not all of the associated languages, $L(g)$, are (stringset) distinct and some are proper subsets of other languages. "English" without P results in a weakly-equivalent stringset-identical language: the grammar assigns different derivations to some strings, though the associated LFs are identical. "English" without composition results in a proper subset language of full "English". Some combinations of p-settings result in 'impossible' grammars – for example, ones without any rule schemata. Others yield equivalent grammars, for example, different combinations of default settings (for types and their subtypes) can define an identical category set.

The grammars, $g \in GCUG$, generate (usually infinite) stringsets of lexical syntactic categories. These strings are sentence types since each defines a finite set of grammatical sentences (tokens), formed by selecting a lexical item consistent with each lexical syntactic category. Such sequences of lexical syntactic categories can be viewed as potential triggers (determinate SF-LF pairings) because in this framework knowing the lexical syntactic category of each word in a sentence is enough to deterministically recover a LF. Languages, $L(g)$, are represented as a finite subset of sentence types generated by the associated grammar. These are a proper subset of the degree-0/1 unembedded or singly embedded triggers for the language (e.g. Lightfoot, 1992:22f). Subset languages are exemplified by between 3 and 9 such sentence types and full languages by 12 sentence types. The constructions exemplified by each sentence type and their length are equivalent across all $L(g)$ for $g \in GCUG$, but the sequences of lexical categories can differ. For example, two SOV language renditions of a sentence type corresponding to *The man who Bill likes gave Fred a present*, one with premodifying and the other postmodifying relative clauses, both with a relative pronoun at the right boundary of the relative clause, are shown below with the differing category highlighted.

[17] In the GCG framework, v1 corresponds to verbs being assigned two categories allowing both initial and medial position, as in "Welsh", SOVv1, in conjunction with a relaxation of default ordering of the argument interpreted as subject being 'outermost' (*argorder*), as in "Tagalog", canonical VSO. v2 is encoded by requiring auxiliary verbs to take an underspecified NP argument to their left and a (S\NP) argument to their right with features and interpretation of this missing NP in the main verb's argument list bound to the leftward argument of the auxiliary, as in "German", SOVv2.

Bill	likes	who	the-man	a-present	Fred	gave
NP_s	$(\mathrm{S}\backslash\mathrm{NP}_s)\backslash\mathrm{NP}_o$	$\mathrm{Rc}\backslash(\mathrm{S}\backslash\mathrm{NP}_o)$	$\mathbf{NP}_s\boldsymbol{\backslash}\mathbf{Rc}$	NP_{o2}	NP_{o1}	$((\mathrm{S}\backslash\mathrm{NP}_s)\backslash\mathrm{NP}_{o2})\backslash\mathrm{NP}_{o1}$

The-man	Bill	likes	who	a-present	Fred	gave
$\mathbf{NP}_s/\mathbf{Rc}$	NP_s	$(\mathrm{S}\backslash\mathrm{NP}_s)\backslash\mathrm{NP}_o$	$\mathrm{Rc}\backslash(\mathrm{S}\backslash\mathrm{NP}_o)$	NP_{o2}	NP_{o1}	$((\mathrm{S}\backslash\mathrm{NP}_s)\backslash\mathrm{NP}_{o2})\backslash\mathrm{NP}_{o1}$

The finite subsets of triggers for each language are prespecified so that they discriminate that language uniquely. It is necessary for a learner to be exposed to such a fair sample of triggers for any $g \in GCUG$ so that convergence to the correct grammar is possible, as discussed in section 2.3. There is psycholinguistic evidence that children are exposed primarily to simple unembedded data and to a few singly embedded triggers (e.g. Newport, 1977). The expressiveness of a grammar and associated language is modelled (crudely) in terms of the proportion of sentence types which can be generated and parsed from the finite subset for the associated full language; for example, the expressiveness of a full language is 1 and of a proper subset language exemplified by 9 sentence types is 3/4 (i.e. 9/12).

2.2 The Parser

The parser uses a deterministic, bounded-context shift-reduce algorithm (see Briscoe, 1987 for further details and justification). It represents a simple and natural approach to parsing with GCGs which involves no grammar transformation or precompilation operations, and which directly applies the rule schemata to the categories defined by a GCG. The parser operates with two data structures, an input buffer or queue, and a stack or push down store. Lexical categories are shifted from the input buffer to the analysis stack where reductions are carried out on the categories in the top two cells of the stack, if possible. When no reductions are possible, a further lexical item is shifted onto the stack. When all possible shift and reduce operations have been tried, the parser terminates either with an atomic 'S' category in the top cell, or with one or more non-sentential or complex categories indicating parse failure. The algorithm for the parser working with a grammar which includes the rules of application, composition and permutation is given in Figure 8.

A parse history analysing *Kim loves Sandy* is shown in Figure 9. The first two columns show the state of the stack and buffer after each step. The third column names the operation which has applied to produce the state shown at this step. The final column gives the step number. This algorithm finds the most left-branching derivation for a sentence type because Reduce is ordered before Shift. In Figure 9 this results in *Kim loves* being reduced to a functor from NPs to Ss by permutation on the category for *loves*, and then backward application. The algorithm also finds the derivation involving the least number of parsing operations because only one round of permutation occurs each time application and composition fail.[18] The category sequences representing the sentence types in the data for all $L(g), g \in GCUG$ are designed to be unambiguous relative to this 'greedy, least effort' algorithm, so it will always assign the appropriate LF to each sentence type. However, there are frequently alternative less left-branching or more 'expensive' derivations of the same LF, and in some cases a distinct LF could be recovered by generating all permutations of functors before attempting application or composition. For example, if permutation is not available to the parser at step 3 in Figure 9, the parser will fail to reduce, and instead shift *Sandy* onto the stack, reducing *loves Sandy* first.

The parser is augmented with an algorithm which computes working memory load during an analysis. This algorithm is based on three uncontroversial features

[18]The preference for left-branching derivations and those involving the least number of parsing operations can be seen as a precise and computationally-tractable instantiation of an analogue of the Economy Principle of the Minimalist Program (e.g. Chomsky, 1991:447f) within this framework.

1. THE REDUCE STEP: if the top 2 cells of the stack are occupied, then try
   a) Application, if match, then apply and goto 1), else b),
   b) Composition, if match then apply and goto 1), else c),
   c) Permutation, if match then apply and goto 1), else goto 2)
2. THE SHIFT STEP: if the first cell of the Input Buffer is occupied, then pop it and move it onto the Stack together with its associated lexical syntactic category and goto 1),
   else goto 3)
3. THE HALT STEP: if only the top cell of the Stack is occupied by a constituent of category S,
   then return Success,
   else return Fail

THE MATCH AND APPLY OPERATION: if a binary rule schema matches the categories of the top 2 cells of the Stack, then they are popped from the Stack and the new category formed by applying the rule schema is pushed onto the Stack.

THE PERMUTATION OPERATION: each time step 1c) is visited during the Reduce step, permutation is applied to one of the categories in the top 2 cells of the Stack (until all possible permutations of the 2 categories have been tried in conjunction with the binary rules). The number of possible permutation operations is finite and bounded by the maximum number of arguments of any functor category in the grammar.

Figure 8: The Parsing Algorithm

of human working memory. Firstly, working memory is limited, as evidenced, for example, by people's inability to remember sequences of more than a few unrelated digits. Secondly, there is a strong recency effect on working memory which ensures that recent or recently-revisited elements of a sequence are better recalled. And thirdly, the greater the degree of analysis or depth of processing of elements, the greater the chance of recall (see e.g. Baddeley, 1992; Gathercole and Baddeley, 1993; King and Just, 1991).

Limitations of working memory are modelled in the parser by associating a cost with each stack cell occupied during each step of a derivation, and recency and depth of processing effects are modelled by resetting this cost each time a reduction occurs: the working memory load (WML) algorithm is given in Figure 10. Figure 11 gives the right-branching derivation for *Kim loves Sandy*, found by the parser utilizing a grammar without permutation. The WML at each step is shown for this derivation. The overall WML (16), found by summing the WML at each step, is higher than for the left-branching derivation (9).

The WML algorithm ranks sentence types, and thus indirectly languages, by parsing each sentence type from the data exemplifying each language with the associated grammar and then taking the mean of the WML obtained for all exemplifying sentence types. "English" with permutation has a lower mean WML than "English" without permutation, though they are stringset-identical. A hypothetical mixture of SOV clausal order with "English" phrasal syntax has a mean WML which is 25% worse than that for "English" with permulation.

The parser and WML algorithm are broadly in accord with existing psycholinguistically and typologically motivated theories of parsing complexity (e.g. Briscoe,

Stack	Input Buffer	Operation	Step
	Kim loves Sandy		0
Kim:NP:kim′	loves Sandy	Shift	1
loves:(S\NP)/NP:λ y,x(love′ x, y) Kim:NP:kim′	Sandy	Shift	2
Kim loves:S/NP:λ y(love′ kim′, y)	Sandy	Reduce (P,BA)	3
Sandy:NP:sandy′ Kim loves:S\NP:λ y(love′ kim′, y)		Shift	4
Kim loves Sandy:S:(love′ kim′, sandy′)		Reduce (FA)	5

Figure 9: Parsing *Kim loves Sandy*

After each parse step (Shift, Reduce, Halt (see Fig 8):

1. Assign any new Stack entry in the top cell (introduced by Shift or Reduce) a WML value of 0
2. Increment every Stack cell's WML value by 1
3. Push the sum of the WML values of each Stack cell onto the WML-record

When the parser halts, return the sum of the WML-record which gives the total WML for a derivation.

Figure 10: The WML Algorithm

1987; Gibson, 1991, 1998; Hawkins, 1994; Rambow and Joshi, 1994).[19] For example, the predictions on Rambow and Joshi's (1994) German center-embedded subordinate clause data with and without extraposition or (long-distance) scrambling mirror their acceptability judgements, as the WML values, parsing with the associated categories using composition and permutation post-lexically, illustrate for the following pairs:

a)	WML: 505					
daB	Peter	dem Kunden	den Kuhlschrank	zu reparieren	zu helfen	versucht
that	Peter	the client	the fridge	to repair	to help	tries
Su/S	NPn	NPd/N N	NPa/N N	VP\NPa	(VP\NPd)|VP	(S\NPn)|VP
b)	WML: 117					
daB	Peter	versucht	dem Kunden	den Kuhlschrank	zu reparieren	zu helfen
Su/S	NPn	(S\NPn)|VP	NPd/N N	NPa/N N	VP\NPa	(VP\NPd)|VP

[19]The combination of GCG and shift-reduce bounded-context parsing allows fully incremental left-to-right interpretation (e.g. Milward, 1995) and, although the model as presented here, is deterministic, it could be straightforwardly extended to a nearly-deterministic interactive parser (Briscoe, 1987) or a bounded parallel parser (Gibson, 1991) in order to model the resolution of ambiguity.

Stack	Input Buffer	Operation	Step	WML
	Kim loves Sandy		0	0
Kim:NP:kim′	loves Sandy	Shift	1	1
loves:(S\NP)/NP:λ y,x(love′ x, y) Kim:NP:kim′	Sandy	Shift	2	1 2 (=3)
Sandy:NP:sandy′ loves:(S\NP)/NP:λ y,x(love′ x, y) Kim:NP:kim′		Shift	3	1 2 3 (=6)
loves Sandy:S/NP:λ x(love′ x, sandy′) Kim:NP:kim′		Reduce (A)	4	1 4 (=5)
Kim loves Sandy:S:(love′ kim′, sandy′)		Reduce (A)	5	1

Figure 11: WML for *Kim loves Sandy*

c)	WML: 180					
daB	niemand	den Kuhlschrank	zu reparieiren	zu versuchen	versprochen	hat
that	noone	the fridge	to repair	to try	promised	has
Su/S	NPn	NPa/N N	VP\NPa	VP/(S\NPn)	(S\NPn)|VP	VP|VP

d)	WML: 103				
daB	niemand	versprochen hat	zu versuchen	den Kuhlschrank	zu reparieiren
Su/S	NPn	(S\NPn)|VP	VP/(S\NPn)	NPa/N N	VP\NPa

2.3 The Parameter Setting Algorithm

The parameter setting algorithm is an extension and modification of Gibson and Wexler's (1994) Trigger Learning Algorithm (TLA) to take account of the inheritance-based partial ordering, the role of memory in parameter setting, variant criteria for retaining new parameter settings, and so forth. The TLA is error-driven – parameter settings are altered in constrained ways when a learner cannot parse trigger input and when the alteration results in a successful parse. Trigger input is defined as primary linguistic data which, because of its structure or context of use, is determinately unparsable with the correct interpretation.

The TLA is memoryless in the sense that a history of parameter updates is not maintained, in principle, allowing the learner to revisit previous hypotheses. This is what allows Niyogi and Berwick (1996) to formalize parameter setting as a Markov process. However, as Brent (1996) argues, the psychological plausibility of this algorithm is doubtful – there is no evidence that children blindly move between neighbouring grammars along paths that revisit previous hypotheses. Therefore, in the modified algorithm each parameter can only be updated once during the acquisition process, resulting in a procedure with (limited) memory. As Brent points out, this results in a *consistent* algorithm which utilizes triggers in the most efficient manner possible to traverse the search space. However, because of the use of default specification in the grammatical representation language, this does not lead to strictly monotonic refinement of grammatical hypotheses. Thus, despite the ordered acquisition procedure, the sequence of hypothesized grammars can involve overriding or retraction of decisions, because parameters encode a *default* inheritance network. For example, a learner of German might incorrectly hypothesize a SVO grammar by updating **gendir** to R(ight) and **subjdir** to L(eft), but subsequently override this and hypothesize underlying SOV by resetting **objdir** to L(eft).

Now the default effects of **gendir** will only apply (correctly) within NP and PP – see e.g. Clark (1992) for discussion of the indeterminacy of parameter expression and its consequences for learnability. Meisel (1995) argues that a limited memory which prevents resetting of parameters already updated by the learner is essential for any account of the acquisition of core grammar given the presence of a marked periphery of constructions.

The TLA is local in the sense that only one parameter can be updated on parse failure. In the modified algorithm, this requirement can be relaxed to n parameters per parse failure. Bertolo (1995) argues that this relaxation of the TLA does not alter fundamental results concerning local maxima and learnability. The motivation for relaxing the single-value constraint and adopting a n-local variant of the TLA is twofold: firstly, the selection of an unbiased sample of sentence types with respect to WML creates trigger paths requiring differing numbers of parameter updates per trigger to successfully acquire different $g \in GCUG$; secondly, the parameter n can be varied in the simulation, creating a wider range of acquisition procedures to select from.[20]

The TLA is unordered in the sense that on parse failure a parameter is chosen at random to be updated. In the modified algorithm, parameters are updated starting with the most general, in terms of the partial order defined by the inheritance network. Once updated they are not revisited because the procedure utilizes limited memory. The TLA is greedy in the sense that a parameter updated on parse failure is retained if that setting allows the current trigger to be reparsed successfully. The acquisition procedure can be made more incremental, as well as greedy, by relaxing the requirement that parameter updates must result in a completely successful parse for the new setting(s) to be retained. Retaining a parameter update if it results in an improved parse, defined as the recovery of more of the target LF, results in a model closer to Dresher and Kaye's (1990) cue-based approach, as it places more emphasis on the degree of evidence provided by a trigger for an individual parameter setting, rather than on obtaining a successful parse. The WML measure for a sentence type can be used to determine whether it can function as a trigger at a particular stage in learning, thus filtering random presentation of triggers and ensuring that triggers are presented in (partial) order of decreasing parsability. This corresponds to incorporating an analogue of Newport's (1990) 'starting small' hypothesis into the model (see section 1.3).

The algorithm is summarized in Figure 12. A valid category assignment to a trigger ($VCA(t_i)$) is defined as a pairing of a lexical syntactic category with each word in the SF of t_i, $< w_1 : c_1, w_2 : c_2, ...w_n : c_n >$ such that the parse derivation, d_i for this sequence of categories yields LF$_i$ A p-setting defines a grammar which in turn defines a parser (where the subscripts indicate the output of each function given the previous trigger). A parameter is updated on parse failure and, if this results in a (better) parse, the new setting is retained. The core of the algorithm is the update function, which is applied to a sequential p-setting encoding as described in section 2.3. A default parameter can be reset to its opposite value and the 'D' encoding changed to a 'R" to record that this default parameter has been reset. Unset parameters can be set to the correct value required to parse the trigger, according to $VCA(t)$.[21]

[20] To determine whether each $g \in GCUG$ can be acquired, in principle, by a non-incremental acquisition procedure with, say, n set to 1 would require an exhaustive specification of the set of potential triggers for each language – see Gibson and Wexler, 1994; Niyogi and Berwick, 1996. Because of the larger parameter set explored here this would be a non-trivial undertaking which is beyond the scope of the current paper.

[21] Random updating of parameters, as in the TLA and related proposals, would simply require more exposure to appropriate triggers, since possible settings are finite and only retained if they result in a successful parse. However, the deterministic approach is used here in order not to bias the simulation model presented in section 3 towards a preference for default initial settings.

Data: $\{S_1, S_2, \ldots S_n\}$

unless
  PARSER$_i$(GRAMMAR$_i$(p-setting$_i$))($VCA(t_j)$) = Success
then
  p-setting$_j$ = UPDATE(p-setting$_i$)($VCA(t_j)$)
  if
  PARSER$_j$(GRAMMAR$_j$(p-setting$_j$))($VCA(t_j)$) = Success
  then
    RETURN p-setting$_j$
  else
    RETURN p-setting$_i$

UPDATE(p-setting)($VCA(t)$):
Reset the first n default parameter(s) or set the first n unset parameter(s) in a 'left-to-right' search of the p-settings (consistent with the partial order encoding their generality) according to the following table:

Input:	D 1	D 0	? ?
Output:	R 0	R 1	? 1/0

(where 1 = T/L and 0 = F/R – see Figures 6 and 7)

Figure 12: The Learning Algorithm

In summary, this account of the parameter setting procedure is consistent, error-driven, greedy, possibly incremental, n-local, partially-ordered, utilizes limited memory, and can incorporate maturationally-developing working memory limitations. Finally, the initial configuration of the parameters in the TLA is usually taken to be any arbitrary grammar, though as Gibson and Wexler (1994) point out, assuming (some) specific unmarked initial settings can remove local maxima. In this model, parameters can be initially unset (?) or have a default (D) value (see section 2.3). The precise choice of parameters, of their initial settings, of the n updatable parameters per trigger, and of the update success criterion, defines a space of variant acquisition procedures for the experimenter (or the simulation) to select from.[22]

The relative learnability of languages in the model is ranked in terms of the number of parameters that must be updated to converge to the target grammar, and also in terms of the maximum number of parameters which must be updated for a single trigger given an optimal presentation sequence of triggers to a non-incremental procedure. This ranking is calculated by assuming a learner with all parameters unset initially (see section 4 below). However, the ranking can also be made more dynamic by recalculating it for different potential initial p-settings and acquisition procedures.

[22] In the simulation,, sentence types used as triggers are represented by p-setting schemata with precomputed associated memory loads for each $g \in GCUG$ to avoid the need for continuous on-line parsing of triggers. Thus, the model circumvents issues of indeterminacy in parameter expression, the need to deal with 'noise' in the input, and any consequent errors by the learner (see Clark, 1992). Nevertheless, the current model can be extended to deal with noise and indeterminacy by embedding it in a statistical learning framework (Kapur and Clark, 1994; Briscoe, 1999, 2000a,b), though this involves abandoning the strictly ordered, consistent acquisition procedure presented here.

3 The Computational Simulation

The simulation models the behaviour of a changing population of language agents. A language agent (LAgt) is defined as a language learner and/or user endowed with a LAD, as described in section 2, and equipped with a simple sentence generator based on (usually random) generation of a sentence type on the basis of the LAgt's current grammar (if any). In addition, LAgts have an age which is used to determine the critical period for language acquisition, their reproductive potential and their 'death'.[23]

A linguistic interaction consists of a randomly selected LAgt emitting a sentence type and another randomly selected LAgt attempting to parse it. LAgts can generate and parse sentence types from the prespecified finite subset of $L(g)$ selected by their current p-settings. Linguistic interactions are successful if LAgts' p-settings are compatible. Compatibility is defined in terms of the ability to map from a given SF to the same LF, rather than in terms of sharing of an identical grammar.[24] LAgts aged 1–4 are learners, still potentially updating parameters. LAgts aged 5–10 have completed acquisition and converged to a fixed grammar. LAgts aged 4–10 are capable of reproducing new LAgts.[25]

Simulation states are defined in terms of interaction cycles. An interaction cycle consists of a specified number of interactions between LAgts. This number is computed dynamically so that, regardless of population size, a learning LAgt will be exposed to enough triggers to acquire any $g \in GCUG$ with $p > 0.99$ during the critical period, assuming that the proportion of adult LAgts in the population is >60% and that the adult population is linguistically homogeneous; that is, that all adults have internalized $g_t \in GCUG$, and thus that g_t is the target grammar for the learning LAgts. In a model with overlapping generations of adult and learning LAgts and a range of variant acquisition procedures, specifying this number is complex, and is discussed further in section 4. At the end of each interaction cycle, the age of each LAgt is incremented and any who are over age 10 are removed from the population. In addition, some agents reproduce new agents, up to a prespecified maximum per interaction cycle which is used to ensure that the number of learning LAgts never exceeds 40% of the population. Reproduction is either random or according to the relative fitness of LAgts, depending on whether the simulation is modelling linguistic selection alone or the interaction of natural selection for LAgts and linguistic selection for grammars. One full generation of Lagts corresponds to 4 interaction cycles as LAgts can only reproduce from age 4.

In experiments which utilize natural (biological) selection for LAgts, the relative fitness of a LAgt is a function of the proportion of its linguistic interactions which have been successful, and optionally of the learnability, expressiveness and/or parsability of the grammar(s) used by that LAgt during a cycle of interactions. Thus, fitness is dependent on an agents' linguistic compatibility with other agents, creating a form of frequency-dependent selection (e.g. Maynard Smith, 1998:69f). It is also potentially dependent on properties of the current grammar internalized.

[23] The term *agent* is used to emphasize both that a LAgt is a model of a language learner / user and also that LAgts are elements of a decentralized distributed system. It should be clear from the definition of a LAgt that the simulation does not attempt to characterize the emergence or origin of the LAD. It assumes a prior population of 'Saussurean' LAgts, to use Hurford's (1989) term, with at least the capacity to represent and learn word-meaning associations, and with the basic architecture of a LAD.

[24] P-setting compatibility implements a weak notion of communicative success. Thus, there is no Gricean entailment of successful transmission of speaker intentions, or of a shared interpretation. Consequently, the model builds in no strong assumptions about the function(s) of language, whether this be to influence others, communicate (mis)information, or whatever (see e.g. Pinker and Bloom, 1990; Keller, 1994:84f for insightful discussion).

[25] The critical period is simply stipulated – Hurford (1991) and Kirby and Hurford (1997b) describe simulations in which such a criticial period emerges given certain assumptions.

1. Generate cost: 1 (GC)
2. Generate subset language cost: GSC(g)
3. Parse cost: 1 (PC)
4. Parse failure cost: 1 (PF)
5. Parse memory cost: WML(st|g)
6. Parse/Generate success benefit: 1 (SI)
7. Parameter update cost: 1 (PU)
8. Parameter update success benefit: 1 (PSU)
9. Maximum updatableable parameters: n (MUP)

Figure 13: Cost/Benefit per Interaction

Communicative Success:	$\frac{SI}{GC+PC}$
Expressiveness:	$\frac{GC}{GC+GSC}$
Learnability:	$\frac{1}{MUP} \times \frac{PSU}{PU}$
Parsability:	$\frac{PC-PF}{WML(st\|g_1 \ldots st\|g_n)}$
Full Fitness Function:	$w1(CS) \times w2(Exp.) \times w3(Lrn.) \times w4(Pars.)$

Figure 14: LAgt Fitness

Learnability is measured in terms of the agent's success rate at correctly setting parameters and the maximum number of parameters which can be updated per trigger. Parsability is measured as the mean WML created by an agent's parsing interactions (see section 2.2). Lack of expressiveness is measured as an additional graded cost for generating a sentence type with a proper subset grammar, defined as the reciprocal of expressiveness (see section 2.1). The costs and benefits which a LAgt accrues with each interaction in a given interaction cycle and which make up the components of the various possible fitness functions are summarized in Figure 13 and the components of the fitness functions and consequent full fitness function are given in Figure 14. The weights in the full fitness function are used to ensure that the different components of fitness contribute equally to overall fitness. For the calculation of parsability only successfully parsed sentence types are utilized, hence parse failures (PF) are subtracted from the total number of parse interactions for a LAgt.

During learning, a LAgt can update genuine parameters which either were unset or had default settings 'at birth'. However, p-settings with an absolute value (principles) cannot be altered during the lifetime of a LAgt. This is the manner in which the distinction between principles, or universal grammar (the genetic endowment), and parameters, to be updated during the acquisition of a particular grammar, is modelled. Successful LAgts reproduce at the end of interaction cycles by one-point crossover of (and, optionally, single point mutation of) their *initial* 'at birth' p-settings – ensuring neo-Darwinian rather than Lamarckian inheritance; that is, LAgts inherit (a composite of) their parents' genetic endowment and not their acquired (learnt) characteristics.[26] In reproduction there is a high chance of the reproducing LAgts' p-settings being mixed by crossover, where the p-settings are cut and cross-spliced at a randomly chosen point. There is also, in some experiments, a low chance of a single element in the resulting p-setting being mutated to an alternative value.

Fitness-based reproduction ensures that successful and somewhat compatible p-settings are preserved in the population and continually resampled in the search for better versions of the LAD. Thus, although the parameter setting algorithm *per se* is fixed, a range of alternative acquisition procedures can be explored based on the definition of the set of parameters, their initial settings, and optionally mutation of the n updatable parameters per trigger. Crossover and mutation can, without bias, turn an absolute (inherited) principle into a default or unset parameter and vice versa, change values of either, and so forth.[27]

In experiments investigating linguistic selection, there is a need to provide a source of linguistic variation. In reality, variation is generated by language contact and borrowing, linguistic innovation, reanalysis during language acquisition, and so forth (see e.g. Harris and Campbell, 1995; Lightfoot, 1999; Milroy, 1992). In the simulation, this is modelled by introducing additional adult LAgts with a different full language at regular intervals, or by initializing the simulation with two adult groups speaking different full languages. That is, variation is a consequence of population make-up or movement and no attempt is made to model the actuation of linguistic change. The simulation can be configured to introduce 'migrations' of prespecified proportions of adult LAgts speaking a language which is a prespecified 'distance' in terms of parameter settings from the current dominant language. Typically migrations are used to increase linguistic variation when communicative success is high and linguistic variation is low, and the migrating adult LAgt initial p-settings are set identically to the majority settings of the current population. Thus, they act to reduce rather than increase genetic / p-setting variation in the population.

Figure 15 summarizes the model graphically. There are two interacting evolving domains of LAgts and of languages. Selection for languages operates on sentence types (`st1,st2...stN`) some of which act as triggers during language learning by LAgts. Linguistic selection is either simply in terms of the learnability of triggers or more generally in terms of the parsability and expressiveness of sentence types for all LAgts, depending on the fitness function utilized, if any. Thus, the pool of sentence types in the linguistic arena of use (e.g. Hurford, 1987) changes over time as LAgts select from language variants. The ultimate units of linguistic evolution are the principles and parameters encoded in LAgts' *initial* p-settings, but selection operates directly on sentence types. This is the analogue of the distinction between phenotype and genotype in linguistic evolution. Natural selection operates on LAgts and is driven by their communicative success, but can also take account of the working memory resources used in parsing, learnability and expressiveness, depending on the fitness function utilized. As LAgts' p-settings evolve, this can affect the relative learnability of languages.

[26] The encoding of p-settings allows the deterministic recovery of the initial setting because reset parameters are those preceded by 'R', or '?' followed by a determinate value. '?' parameters are reset to unset values and default 'R' parameters are reset again to the opposite value.

[27] The use of crossover and mutation operators with the p-setting code is based on genetic algorithms (see e.g. Holland, 1993, 1995). However, the simulation is not technically a genetic algorithm as fitness is internal to each LAgt and generations overlap. Also this use of genetic algorithm techniques should not be confused with Clark (1992) and Clark and Roberts (1993) model of the parameter setting *procedure* as a genetic algorithm. LAgts are, in fact, more similar to Holland's Echo Agents as the fitness of an agent is relative to the properties of other agents in the population.

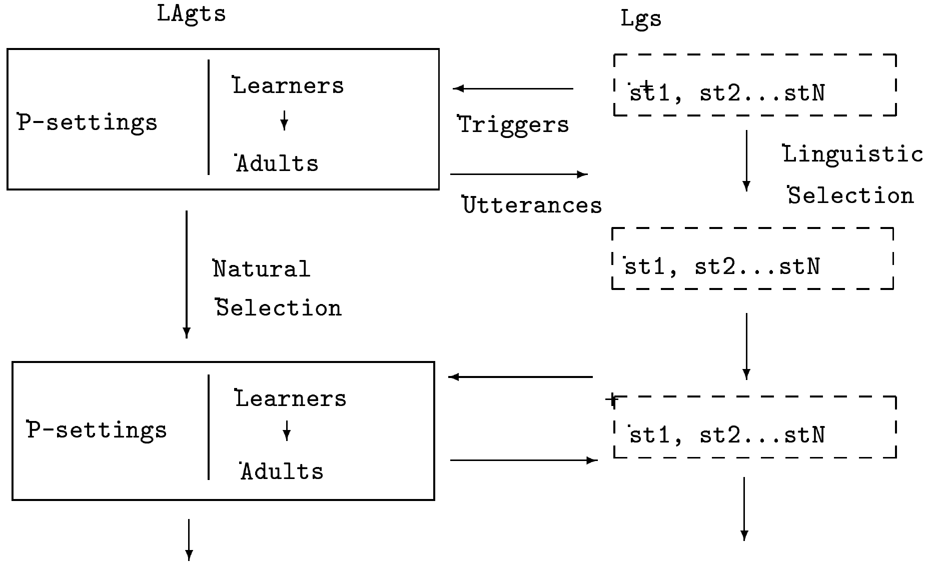

Figure 15: The Simulation Model

4 Preliminary Experiments

The computational model must have several properties to qualify as a useful simulation of the grammatical acquisition process and of the (co)evolution of language and of the LAD. Firstly, it must be clear that, for $g \in GCUG$, at least some acquisition procedures, in the space of possibilities definable in terms of LAgts' p-settings, are able to acquire these grammars given finite and feasible (positive) input. Secondly, learning LAgts should converge reliably on the homogeneous language of a population of adult LAgts, to model language maintenance and the continuity of speech communities. Thirdly, it should be clear that coordinated grammars can evolve at some point during simulation runs quasi-randomly initialized with a population of non-communicating LAgts. Otherwise the model does not provide an environment in which the emergence and maintenance of structured language and its learners is likely (see e.g. Turkel, 2000).

4.1 Effectiveness of Acquisition Procedures

Four acquisition procedures were predefined on the basis of two initial p-settings, unset and default, and incremental and non-incremental variants of the parameter update procedure. Unset learners were initialized with p-settings consistent with a minimal inherited GCUG consisting of application with the **N** and **S** atomic categories already present. All the remaining p-settings were genuine parameters for both learners. The unset learner was initialized with all these unset, while the default learner had default settings for the parameters **argorder**, **gendir**, **subjdir**, **v1** and **v2** which specify a minimal SVO right-branching grammar. The unset learner represents a 'pure' principles-and-parameters learner with innate knowledge of the noun-verb distinction and their (predicate-argument) mode of combination. The default learner is loosely modelled on Bickerton's (1984) bioprogram hypothesis, representing, additionally, a language learner with a preference for SVO canonical order and predominantly right-branching syntax.[28] These two initial p-settings

[28] Briscoe (2000b) applies the simulation presented here to the Hawaiian pidgin-creole transition and argues that a statistical version of the acquisition procedure can account for rapid creolization,

Learner	Language							
	SVO	SVOv1	VOS	VSO	SOV	SOVv2	OVS	OSV
Unset (n4)	26 (32)	26 (31)	18 (26)	18 (25)	18 (25)	27 (33)	27 (30)	17 (20)
Default (n4)	14 (26)	17 (24)	16 (24)	17 (25)	15 (23)	15 (25)	18 (26)	26 (27)
Unset (i1)	32 (98)	30 (82)	30 (89)	31 (84)	31 (78)	31 (97)	31 (84)	31 (45)
Default (i1)	29 (93)	28 (72)	30 (74)	30 (73)	30 (70)	28 (89)	30 (73)	31 (49)

Table 1: Effectiveness of Four Acquisition Procedures

were combined with two update procedures. One, n4, in which 4 parameters were updatable per trigger but updates were only retained if they resulted in a complete LF, and a second, i1, in which only one parameter could be updated per trigger but the updated value was retained if it resulted in recovery of more of the LF.

Each variant learner was tested against an adult LAgt initialized to generate one of seven full languages in the set which are close to an attested language; namely, "English" (SVO, predominantly right-branching), "Welsh" (SVOv1, mixed order), "Malagasy" (VOS, right-branching), "Tagalog" (VSO, right-branching), "Japanese" (SOV, left-branching), "German" (SOVv2, predominantly right-branching), "Hixkaryana" (OVS, mixed order), and a hypothetical OSV language with left-branching phrasal syntax. In these tests, a single learner LAgt interacted with a single adult LAgt. The adult always randomly generated a sentence type and the learner always attempted to parse and learn from it. The first figure in Table 1 shows the mean number of triggers (i.e. number of input sentences) required by the four learners to converge on each of the eight languages. The figure in brackets shows the mean number of triggers required for convergence when the WML metric was used to filter the learners' access to triggers in accordance with the 'starting small' hypothesis (see sections 1.1 and 2.2). These figures are each calculated from 1000 trials and rounded down to the nearest integer. When no memory constraints were imposed, each learner converged with less than a 1% error rate, to the target grammar on the basis of 100 random presentations of trigger sentences. 200 trigger sentences were required to achieve convergence with this reliability when memory constraints were imposed. Thus, we can conclude with reasonable confidence that all these learners will converge for the languages tested, given this distribution and amount of data, with $p \geq 0.99$. (See Niyogi and Berwick, 1996 for detailed discussion of high-probability convergence from finite data.)

These results suggest that, in general, the default learners are more effective than the unset learners, though the difference is small and probably insignificant for the incremental learner. For the OVS language (OVS represents 1.24% of the world's languages; Tomlin, 1986), for the unattested or very rare OSV language, and for SOVv2, the default (SVO) n4 learner appears less effective. In memory-constrained learning, learners pass through 4 maturational stages at each of which the allowable memory load during parsing is increased, and 25% of the triggers are presented at each stage. Unsurprisingly, given that the acquisition procedures are deterministic and consistent, memory-based filtering of triggers does not affect convergence. However, it does slow it down since more triggers are required at each stage to ensure the learner has a high chance of being exposed to a convergent set of triggers overall. The variable performance of the different learners on the various languages suggests that many, perhaps intuitively insignificant, aspects of an acquisition procedure can affect its performance on a specific language, and

provided the demographic situation is modelled and it is assumed that learners utilize limited data from super/substratum languages.

consequent predictions concerning the relative learnability of languages.

Many other variant procedures result in effective learners for some or all of the eight languages tested, given varying amounts of triggering data. Testing the above learners on randomly-generated full languages suggests that these learners are capable of converging on any grammar defined by the parameter set. However, stronger conclusions would require either exhaustive experimentation or development of a formal proof of convergence (see Gibson and Wexler, 1994; Niyogi and Berwick, 1996; Osherson *et al.*, 1986). Nevertheless, these results demonstrate that effective limited-memory parameter updating algorithms can be developed within the CGUG framework for quite complex parameter sets, given inductive bias in the form of partial ordering of parameter updating and possibly some initial default settings for parameters.

4.2 Language Maintenance

The simulation employs random interactions within a population, some of whom will be learners. Thus, a proportion will involve learning LAgts interacting with each other or generating input for adult LAgts, before they have converged on the target grammar. Even in an initially homogeneous linguistic environment with a critical period for learning, if, for example, the proportion of learners to adults in the population becomes too high, then learners may not converge to the target grammar, since the distribution of sentence types may become too skewed towards those of subset languages. Initial experiments were undertaken to discover levels of reproduction, death and average interactions for LAgts which supported languge stasis, given particular acquisition procedures.

A series of 50 interaction cycle simulations were run each initialized with either 32 adults endowed with the unset n4 acquisition procedure or 32 default n4 adults, all speaking one of the eight languages described in section 4.1. LAgts reproduced randomly (without mutation) and died as described in section 3. Given the p-settings of the initial population, LAgts were only able to reproduce further LAgts endowed with the unset or default n4 acquisition procedure. Both acquisition procedures were also run with and without maturational memory limitations during the learning period. All conditions were run at least 10 times.[29]

With the levels of reproduction and death set so that the proportion of adult LAgts never fell below 60% and with the mean number of interactions per cycle per LAgt set to 65 for the memory-constrained procedures and 25 otherwise, the population continued to speak the original language and all learners reliably converged to that language by adulthood in all simulation runs. Thus, any subset language speakers in the population at the end of an interaction cycle were, without exception, learners. At first sight, the property of language maintenance or stasis may seem somewhat contradictory for an evolutionary model of language change. However, it is critical that the model possess the *potential* to be stable (once all/most genetic and/or linguistic variation is removed) if it is to represent a plausible model of language acquisition and change. Though very occasional misconvergence during learning in ideal (i.e. homogeneous) conditions is probably possible, few would argue that this *alone* is the source of language change. Rather most linguistic change is probably a consequence of variation introduced through invention or contact between speech communities, and the consequent linguistically heterogeneous data supplied to the learner. Briscoe (2000a) argues that this property of the simulation is critical for insightful modelling of attested changes.

[29]In these and subsequent experiments reported in this paper, the relevant qualitative results were observed in all runs (and many others not discussed), so no statistical analysis beyond means is reported. The use of standard deviations, error bars and/or tests of significance would only be informative if results were less clearcut or a detailed quantitative evaluation was relevant.

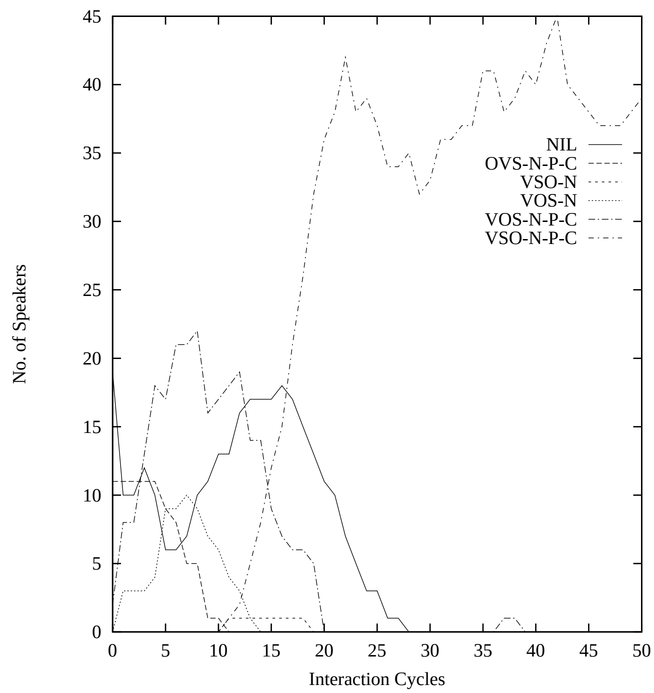

Figure 16: Emergence of language(s)

4.3 Emergence of Structured Language and its Learners

To explore the emergence and persistence of structured language (and consequently the emergence of effective acquisition procedures) in the simulation, a series of simulation runs of 500 cycles (approximately 125 full generations of LAgts) were performed. 32 LAgts' p-settings were randomly initialized for any combination of p-setting values, with a probability of 0.25 that a setting would be an absolute principle, and 0.75 a parameter with unbiased allocation for default or unset parameters and for values of all settings. All LAgts were initialized to be age 1, memory-limited n4 learners with a critical period of 4 interaction cycles, a maximum age of 10, and the ability to reproduce by crossover (0.9 probability) and mutation (0.05 probability) from 4–10. The full fitness function defined in section 3 was utilized. In around 5% of the runs, language(s) emerged and persisted to the end of the run; that is, the population discovered a shared language before it became extinct through a total lack of successful communication.

When languages emerged, those with close to optimal parsability typically came to dominate the population quite rapidly. However, sometimes less parsable languages were initially selected, and occasionally these persisted despite the later appearance of a more optimal language (but with few speakers). Typically, a minimal subset language dominated – although full and intermediate languages did appear briefly, they did not survive against less expressive but more easily learnable and parsable subset languages. Figure 16 is a typical plot of the emergence (and extinction) of language variants in one of these runs. In this run, 10 LAgts converged on a minimal OVS language and 3 others on a VOS language. The latter is more parsable and learnable and both are of equal expressiveness so, as expected, the VOS language acquired more LAgts over the next few cycles. A few LAgts also converged on VOS-N, a more expressive but less easily parsable extension of

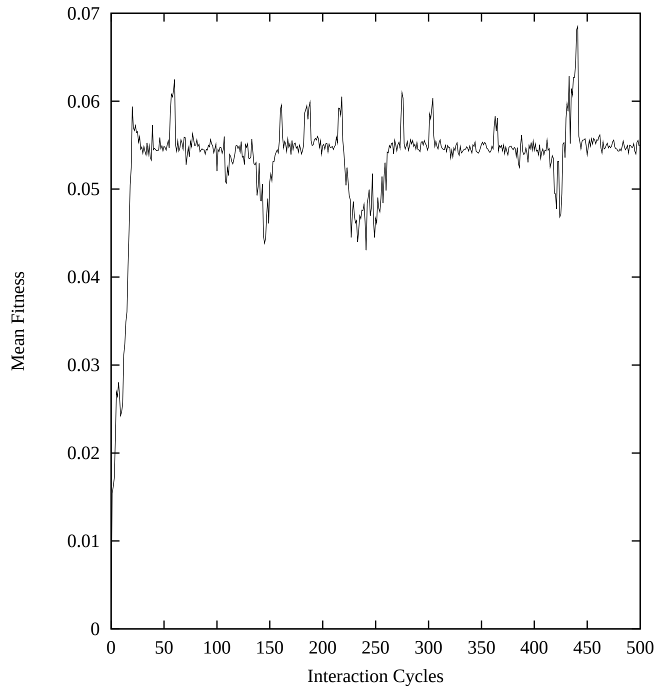

Figure 17: Mean fitness with language emergence

VSO-N-P-C.[30] However, neither this nor the OVS language survived beyond cycle 14. Instead a VSO language emerged at cycle 10, which has the same minimal expressiveness of the VOS language but is more parsable, and this language dominated rapidly and eclipsed all others by cycle 40. Figure 17 is a plot of the mean fitness of all LAgts at the end of each interaction cycle through this entire run. As can be seen, mean fitness improves rapidly early in the run, once a single dominant language emerges. Subsequent downward fluctuations are mostly caused by the occasional re-emergence of a few non-speaking LAgts who fail to learn the language, and upward fluctuations by a lower proportion of learners in the population, or by the increased use of a more learnable and/or parsable language.

As full languages did not emerge in these runs, a second identical series of runs was undertaken, except that the initial population now contained 2 speakers of one of the full languages defined in section 4.1. This resulted in the evolution of acquisition procedures capable of acquiring such full languages across the population. For example, in 10 runs initialized with two SOVv2 "German" speaking adult LAgts, the population converged on a full SOVv2 language 7 times, on the intermediate subset language SOVv2-N 2 times, and once on the minimal subset language SOVv2-N-P-C. These experiments suggest that if a full language defines the environment of adaptation then a population of randomly initialized LAgts is more likely to converge on a (related) full language, evolving a LAD capable of acquiring such a full language. Thus, although the simulation does not model the development of expressiveness well, it does appear that it can model the emergence of effective acquisition procedures for (some) full languages.

[30] The names for languages are intended to be mnemonic: the first element indicates basic constituent order, the remaining elements delimited by '−/+', say what (un)marked features were present/absent from the associated grammar. For example '−N' indicates no complex multiword NPs, −P indicates no permutation operation, −C no composition operations, and so forth. NIL represents speakers of no language.

The pattern of language emergence and extinction followed that of the previous series of runs: more parsable languages were selected from those that emerged during the run. However, often the initial locally optimal SOVv2 itself was lost before enough LAgts evolved capable of learning this language. The mean fitness and communicative success measures show very similar patterns to that of the previous runs. However, learning rates are worse, reflecting the more complex linguistic environment. There are clear changes in the percentages of absolute, default or unset p-settings within the population during the runs: the mean number of absolute principles declined by 6.1% and unset parameters by 17.8%, so the number of default parameters rose by 23.9% on average between the beginning and end of the 10 runs sampled. This contrasts with the previous series of runs in which there was a greater increase in absolute principles than increase in default parameters. This may also reflect the more complex linguistic environment, in which (incorrect) absolute settings are more likely to handicap, rather than simply be irrelevant to, the performance of the LAgt.

The experiments reported in this section demonstrate linguistic selection, chiefly for more learnable and/or parsable languages, and natural selection for LAgts with initial p-settings supporting effective grammatical acquisition. They also suggest coevolution is occurring: there is a clear preference for absolute principles or default parameters over unset parameters. Unset parameters represent the least informative p-setting, while both default parameters and absolute principles provide more information about the linguistic environment. The preference for default parameters over absolute principles in the environment of more complex (full) languages may reflect the fact that in these simulations a dominant language is emerging as acquisition procedures are evolving, so flexibility is selected over further attenuation of the acquisition procedure, in the absence of a constant and homogeneous linguistic environment. Nevertheless, as the acquisition procedure evolves this exerts selection pressure in favour of the type of languages which the population has sampled. The experiments reported in subsequent sections investigate each of these factors independently and in more depth.

5 Linguistic Selection

Selection for grammars and thus languages might occur as a consequence both of acquisition procedures and of the conflicting preferences for more parsable and more expressive languages. In the simulations discussed in the previous section, selection for more learnable and/or parsable languages occurred (although in some cases a more optimal language did not survive because it was spoken by too few speakers). But, it is not possible to definitely say whether it is parsability, properties of the acquisition procedure, or the proportion of speakers which is the causal factor in any given case. It is not even clear what precise form of the acquisition procedure is being deployed at any point in the randomly-initialized populations.

5.1 Linguistic Selection between Language Pairs

In more circumscribed experiments, linguistic selection for more parsable or more learnable languages, and the interplay between these two pressures as well as the 'robustness' of critical triggers, can be demonstrated directly. A series of 300 cycle simulations was run in which a population of 32 LAgts was initialized with differing proportions of unset n4 adult LAgts speaking two different full languages which contrasted in learnability and/or parsability. There were no differences in the initial p-settings in the population and no mutation. All conditions were run at least 10 times.

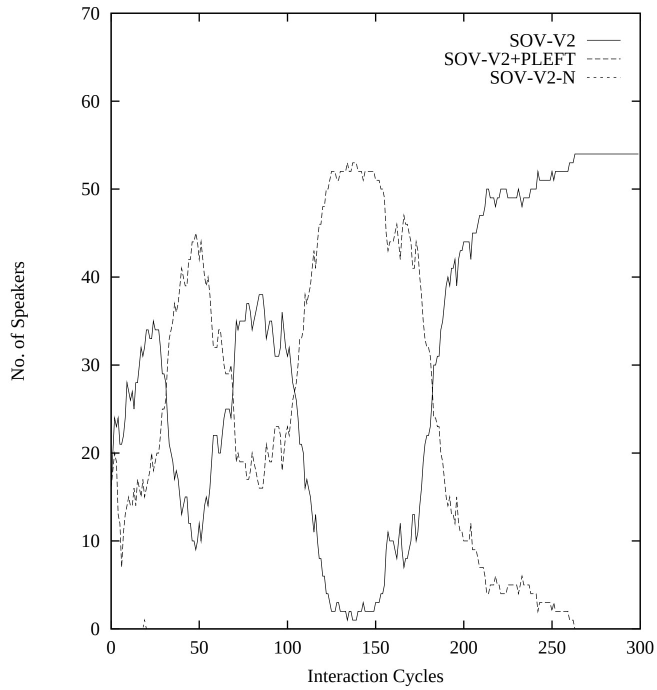

Figure 18: Random drift

In one such series of experiments, the population was initialized with speakers of "German", SOVv2, and "German with postpositions", SOVv2+Pleft. These languages differ only in one directional parameter setting. There is no inherent learnability advantage, in terms of numbers of parameters to be set, for either variant given the unset acquisition procedure. However, there is a small difference in the parsability of the two languages according to the WML metric (as Hawkins (1994) also predicts) with a preference for the more consistently right-branching phrasal syntax of "German". Utilizing the full fitness function but no memory limitations during learning, and initializing with equal numbers of speakers of each language, the population converged on one or other variant within a mean 143 interaction cycles. Which variant was selected was dominated by the random factors in the simulation – principally, how many learning LAgts happened to be exposed to a critical prepositional (or postpositional) trigger first.[31] Figure 18 shows a typical run in which SOVv2 happens to emerge as the dominant language around cycle 260. When the proportion of postpositional speakers was reduced to one third of the initial population, then in two-thirds of the runs "German" emerged as the dominant language. When this proportion was further reduced to one fifth of

[31] In an evolutionary model, random drift alone with no selection will (eventually) lead to the eradication of variation. This effect is well known in population genetics and can be analyzed probabilistically under certain assumptions about population size and the distribution of offspring within the population (e.g. Maynard Smith, 1998:25f; Roughgarden, 1979). The upshot is that for small finite populations fixation on a linguistic variant by random drift can be expected to occur within $2N$ interaction cycles with a standard error a little greater than N, where N is population size. As the population in these simulations initialized with 32 LAgts typically rises to around 60 quite quickly and then stabilizes, consistent fixation within 50 interaction cycles in 10 runs constitutes reliable evidence of linguistic selection. This approach to modelling language change in finite populations contrasts with Niyogi and Berwick's (1997a,b) use of a macroevolutionary model in which they derive deterministic update rules by abstracting to an infinite population (see Briscoe, 2000a).

the initial population, then "German" became dominant (within about 50 cycles) in all the runs. However, inverting these experiments so that the frequency of prepositional triggers is progressively lower does not produce a symmetric effect. It is only when less than one-tenth of the initial population are producing prepositional triggers that selection of the postpositional variant occurs reliably in all runs.

These runs demonstrate the interaction of linguistic selection for parsability with frequency-dependent sampling effects on triggering data. They show that general parsability factors for both adults and learners can (weakly) select for a linguistic variant by increasing the fitness of LAgts with that variant, creating asymmetry in the frequecy-dependent effects. This general explanation for linguistic selection has been questioned though, since it relates a linguistic factor such as parsability directly to LAgt fitness. Lightfoot (1991) and others have argued that it is unlikely that *specific* properties of a language spoken by speakers would lead directly to increased numbers of offspring. It might be more plausible to argue that such properties might lead indirectly to increased offspring by increasing communicative success. Such an indirect effect could easily be incorporated into the simulation model by positing that the probability of successful interaction is partly a function of the WML of the sentence type chosen.

An alternative assumption is that maturational working memory limitations will decrease the chances of less parsable sentence types functioning as effective triggers (see e.g. Kirby, 1997, 1998, 1999 for a similar position). To simulate this scenario, the same set of runs was done with random reproduction of LAgts but with memory limitations during learning. The results show a very similar pattern to those reported above, though the selection effect is weaker and it is only when the proportion of initial postpositional speakers is less than one sixth that the prepositional language dominates reliably.

A final variant of this experiment is to assume initially equal numbers of speakers of each variant but weight the production of sentence types by their parsability, under the assumption that speakers avoid less parsable sentence types, perhaps to improve their chances of communicative success (e.g. Hawkins, 1994:180f). Altering the LAgts' generation algorithm so that sentences selected with WMLs above 40 have a less than 100% chance of being uttered, falling from 80% for a WML over 40 to 20% for WMLs over 200, ensured that in all runs the population converged on SOVv2.

The simulation model predicts that, given the correctness of any one of these assumptions allowing an effect of parsability on language learning, production and/or interpretation, parsability will cause linguistic selection.[32]

The interplay between parsability and learnability can be seen in simulation runs initialized with equal numbers of "German", SOVv2, and "Japanese", SOV, speakers. SOVv2 has a slightly lower mean WML, and thus parsability, than SOV (largely because the freer constituent ordering options of Japanese relative to German are not modelled effectively in "Japanese" (see e.g. Hawkins, 1994)). Figure 19 shows the languages which are present during one run with the full fitness function. SOVv2 comes to dominate the population after 5 interaction cycles. The other language which persists, SOVv2-N, is a subset language spoken by learners of SOVv2. SOV and various subsets of both languages (not shown in the key) virtually disappear once the population converges to SOVv2 and learning is taking place in a more homogeneous environment. All runs exhibited the same clear effect. However, with

[32]Caution should be used when making inferences from these results concerning the frequency thresholds of the postpositional trigger. A different and probably more plausible acquisition procedure which 'damped' response to an initial trigger and tracked the relative frequencies of conflicting triggers in the input before finally setting a parameter (e.g. Kroch, 1989; Niyogi and Berwick, 1997b; Briscoe, 1999; 2000a,b) would make different predictions, though the basic conclusion concerning the potential for linguistic selection would still hold.

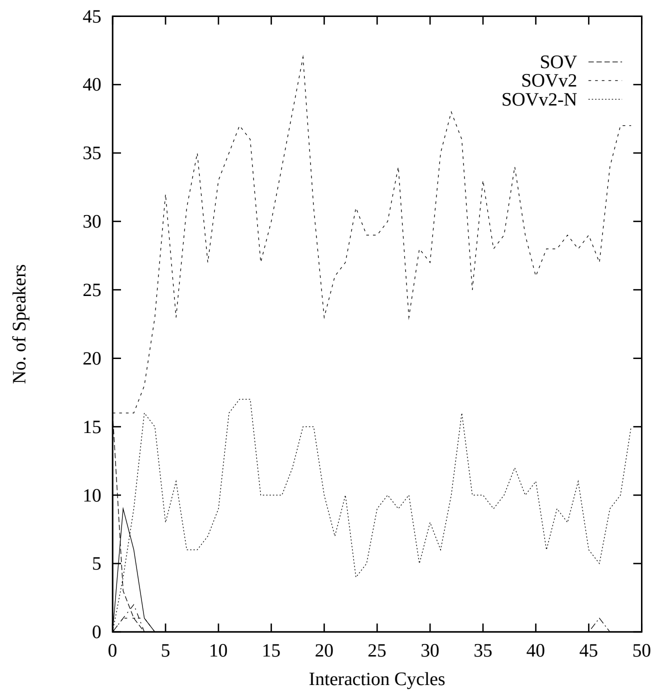

Figure 19: Selection for SOVv2 over SOV

parsability not a factor in LAgt fitness, the opposite result was obtained – in all runs SOV came to dominate with SOV subset languages, again spoken exclusively by learners. SOV is consistently selected in runs when parsability is not a factor because SOVv2 requires the setting of the v2 parameter as well as mixed ordering parameters, reflected in the greater number of triggers required for convergence in Table 1. Thus, in the case of these two languages, ease of parsability, for both learners and users, creates greater overall linguistic selection pressure than that created by the (other) requirements of learnability.

5.2 Linguistic Selection with Migrations

To demonstrate more general linguistic selection for more (locally) optimal grammars, a number of experiments were undertaken with populations of LAgts with identical initial p-settings operating in a continuously heterogeneous linguistic environment, providing the variation on which linguistic selection could work. Migrations of adult LAgts speaking a different language, whenever the population was close to convergence on a single language, ensured heterogeneity. Language change occurs when learners select preferentially one or other grammar, or a mixture, or a subset, while exposed to data from more than one source grammar. There is also an increased possibility of selection of a grammar not directly exemplified in the adult population when the (uniform) distribution of triggers from a single source is skewed by the presence of several sources. This is particularly true when some parameters have default initial settings.

In this series of experiments, approximately one third additional adults were added to the population at regular intervals, all speaking a full language not currently exemplified in the linguistic environment. This proportion of adult LAgts ensured that the new language had a reasonable chance of surviving a number of cycles and thus influencing learners. LAgts added in this fashion had identical initial p-setting configurations as the existing population, so no genetic variation resulted. The maximal 'distance' between an existing dominant language and the new language was three parameters. Migrations of this type occurred every other cycle provided that a clearly dominant language had emerged at the end of the previous cycle. Thus, migrations ensure a constant source of linguistic heterogeneity throughout a simulation run. The amount of variation introduced was tuned to the maximum consistent with the population maintaining a mean communicative success rate of 90% or better. After the first interaction cycle in all runs with migrations there are always at least two and up to ten languages present in the linguistic environment at any one time.

In the first set of experiments, 500 cycle runs were used, and all LAgts utilized the memory-constrained unset n4 procedure, as defined in section 4.1. LAgts reproduced randomly, but because all LAgts were using an effective acquisition procedure, because the simulation was initialized with a single full language, and because the amount of linguistic variation was controlled, in all runs communicative success averaged over 90%. The overall mean costs of the languages adopted by the population were reduced during the course of this and other runs via linguistic selection for learnability, as illustrated in Figure 20. The figure plots an integrated measure for the mean learnability, parsability and expressiveness of the languages present in each interaction cycle, and also breaks this down into the three components, so it can be seen clearly that the population is optimizing learnability at the expense of expressiveness. In this and other runs with random LAgt selection, the population selected subset languages, which are less expressive but more easily learnable as they require fewer parameters updated. As memory load plays a role in learnability via the filtering of triggers, often, but not in every case, parsability was also selected for. Similar results were obtained from all ten runs. These results confirm that linguistic selection can occur without any *natural* selection for LAgts whatsoever. The inductive bias of the acquisition procedure which the LAgts use is enough to create a process of linguistic selection for the most learnable languages.

Kirby (1997, 1998, 1999) explores in detail this form of linguistic selection, as languages, or more accurately triggers, pass repeatedly through the 'bottleneck' of language acquisition. Essentially, triggers compete for learners and those which are more able to pass through the filter of the acquisition procedure will set more parameters in more learners. In this way languages will, over time, adapt to the acquisition procedure. Kirby argues that, on the assumption that parsability is *identical* to learnability, languages will, therefore, evolve to be optimally parsable, and demonstrates that this form of linguistic selection can account for various typological skews in constituent order, and related implicational universals, without the need to posit innate UG constraints. However, Kirby only models differential acquisition of competing variants. Once a more realistic acquisition procedure is defined, the possibility of simply *not* learning arises, and thus the possibility of converging on a subset language. This is exactly what is seen in the runs of the simulation described above – there is no pressure for LAgts to prefer a more expressive, and thus costly, language, so, even if the population is initialized to use such a language, the community soon selects for subset languages. A counteracting pressure for expressiveness is needed to prevent this tendency.

Other runs were performed using communicative success, parsability, expressiveness and/or learnability as components of the fitness function on LAgt reproduction. In the runs where expressiveness was a component of selection, the population did not converge on subset languages despite the linguistic variation in the learning environment created by migrations. When the full fitness function was utilized, LAgts' mean fitness typically did not vary greatly, except where migrations removed them temporarily from a (local) optimum. The mean language costs for parsability, learn-

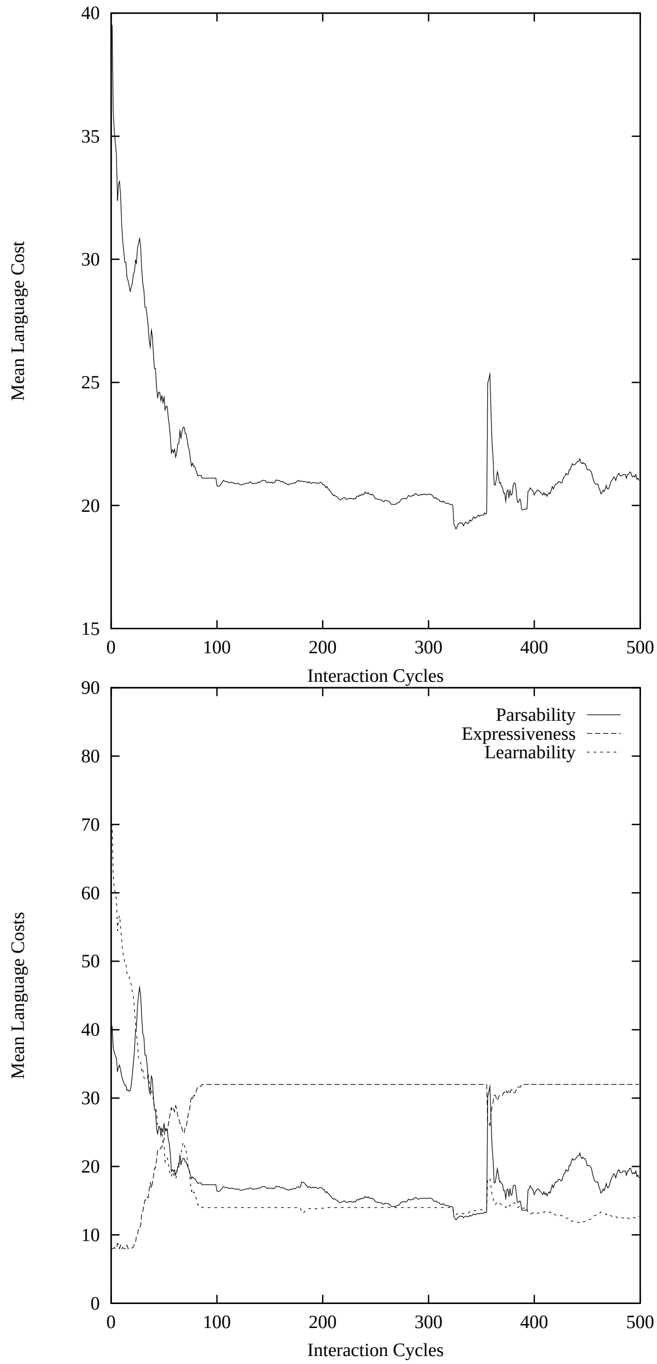

Figure 20: Language costs with random reproduction and migrations

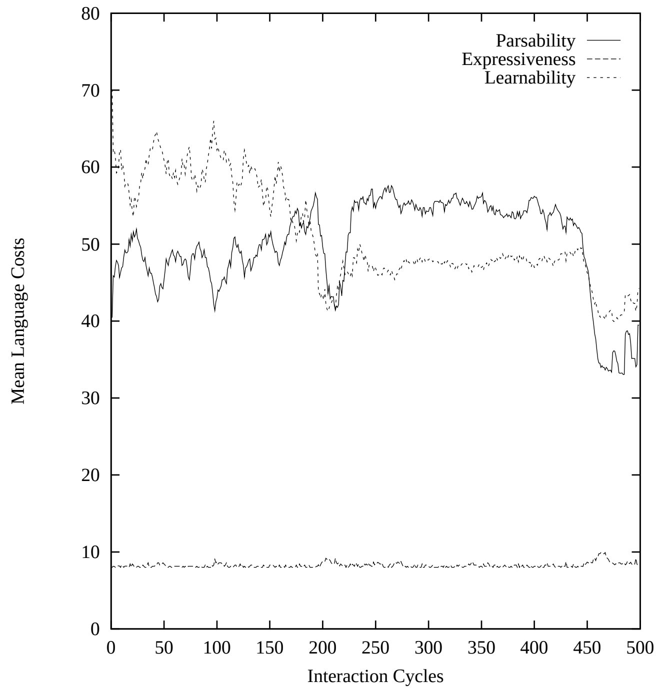

Figure 21: Language costs with natural selection and migrations

ability and expressiveness displayed in Figure 21 demonstrate consistent linguistic selection for more easily learnable and parsable full languages. This is typical of such runs where the full fitness function is utilized.[33] Comparing this with Figure 20 above demonstrates the contrast with linguistic selection without natural selection. Once again, the specific simulation model utilized here builds expressiveness into the LAgt fitness function and it could be objected (along the lines of Lightfoot, 1991) that it is implausible that linguistic expressiveness influences number of offspring; that is, that linguistic expressiveness is a component of *natural* selection and, thus, an indirect influence on linguistic selection. As with the discussion of parsability in section 5.1, it would be straightforward to incorporate expressiveness into the measure of communicative success, by assuming, for example, that an inability to convey some LFs (economically) can lead to failed interactions. However, incorporating expressiveness into learnability and thus making it a purely linguistic selection pressure seems less plausible than in the case of parsability: learning theorists generally argue that there is inductive bias in favour of simpler hypotheses (Ockham's Razor; e.g. Mitchell, 1997) not more complex ones.

The experiments reported in this section demonstrate that linguistic selection occurs whenever there is linguistic variability under a range of different possible assumptions about the precise impact of learnability, parsability, expressiveness,

[33]In some cases, migrations still cause the population to settle on a less optimal language, though this is far less frequent with natural selection for LAgts. The use of random interaction between LAgts idealizes a vast range of sociolinguistic factors which influence selection between linguistic variants, such as the prestige, charisma, economic power or ideology of the speakers of the variants, and so forth. In reality, these factors probably significantly outweigh considerations of selection for parsability or learnability in many situations; for example, where the migrants are conquering invaders. In addition, the simulation does not address differences in death rates between linguistic groups due to disease, genocide, and so forth. Dixon (1997) and Pullum (1981) provide an extended discussion of such factors. Nettle (1999:29f) discusses social as opposed to functional linguistic selection within a similar evolutionary framework to that developed here.

and communicative success. As Kirby's work shows, linguistic selection may form the basis for an explanation of functionally-motivated universals and tendencies in attested languages (as Deacon, 1997:116f) speculates; see section 1.3). However, the more realistic model of acquisition used here suggests that linguistic selection must include a competing pressure for expressiveness (either directly or indirectly via LAgts' fitness), since learnability and parsability alone will predict evolution towards structurally simpler less expressive languages.

6 Natural Selection

The following experiments explore the relative efficiency of several variant acquisition procedures on a range of full languages, by holding the linguistic environment constant but allowing natural selection between variant acquisition procedures to occur. The role of memory limitations in learning is also explored. The experiments provide the background for exploring the coevolutionary dynamic between linguistic selection amongst variant languages and natural selection from variant acquisition procedures.

6.1 Evolution of Acquisition Procedures

A series of 300 interaction cycle runs was performed in which the population was initialized with 16 adult LAgts endowed with the default acquisition procedure and 16 with the unset procedure, as defined in section 3, all speaking one of six full languages. They either all used the incremental acquisition procedure and were able to update one parameter per trigger (i1) or all used the non-incremental procedure able to update 4 parameters per trigger (n4). There was no mutation, so natural selection was only able to select between the five initial parameter settings which distinguish the unset from default learners. Crossover alone cannot change the initial default value of a parameter, only its status as default or unset so the question being explored is under what conditions default parameters, with the default values specified by the default (SVO) procedure of section 3, would be retained in preference to unset versions of these parameters. In all runs, the full fitness function was used, and all conditions were run with and without maturational memory limitations during learning.[34]

The results of these experiments demonstrate marked interacting effects of language type, acquisition procedure and memory limitations on the propensity for specific default initial parameter settings to go to fixation. Table 2 shows the percentage of simulation runs under the varying conditions for which default-valued parameters went to fixation in the population within 300 interaction cycles. For example, with the memory-limited n4 acquisition procedure applied to "English", SVO, the **argorder** parameter was default-valued for every member of the population by the end of 80% of runs. On the other hand, with the memory-limited i1 procedure, **argorder** fixated to a default parameter in only 10% of runs. Conversely, in 20% of runs with n4 and 90% of runs with i1, an unset version of **argorder** went to fixation in the population.[35]. The figure in brackets after each percentage in-

[34] Similar results in most runs would be obtained by running the same experiments with a fitness function based purely on communicative success, or communicative success and learnability, because the effects of expressiveness and parsability are rendered negligible by the linguistically homogeneous environment. However, some form of natural selection for LAgts is required in order to preclude simple random drift amongst variant acquisition procedures (which, given enough variation, invariably results in loss of language in the population).

[35] In about 2% of cases, specific parameters did not go to fixation for either type, and the population retained variation. In these cases, the percentage count assigned to default-valued parameters was halved in Table 2

Learner	Language					
	SVO	SVOv1	VOS	VSO	SOV	SOVv2
n4+ML						
arg0	80% (29)	80% (26)	50% (29)	60% (39)	50% (41)	70% (44)
gendir	80% (30)	90% (23)	50% (30)	60% (61)	50% (36)	80% (54)
subjdir	90% (28)	80% (30)	100% (32)	80% (58)	80% (64)	80% (111)
v1	40% (70)	90% (53)	100% (62)	90% (54)	90% (56)	100% (55)
v2	40% (41)	90% (53)	100% (50)	90% (76)	90% (50)	100% (67)
n4–ML						
arg0	80% (40)	100% (19)	90% (30)	90% (31)	80% (43)	80% (41)
gendir	80% (38)	100% (19)	90% (30)	80% (41)	80% (39)	90% (35)
subjdir	80% (33)	80% (43)	100% (28)	70% (63)	50% (59)	95% (99)
v1	30% (40)	90% (50)	95% (88)	90% (55)	50% (57)	100% (55)
v2	35% (77)	90% (48)	100% (59)	90% (88)	50% (68)	100% (51)
i1+ML						
arg0	10% (25)	50% (23)	20% (24)	20% (35)	30% (30)	10% (23)
gendir	10% (22)	60% (33)	30% (27)	20% (43)	30% (32)	20% (27)
subjdir	10% (22)	50% (32)	60% (27)	20% (43)	60% (63)	0% (27)
v1	10% (22)	75% (33)	20% (27)	20% (43)	10% (102)	0% (27)
v2	0% (37)	60% (150)	30% (63)	10% (49)	0% (49)	0% (36)
i1–ML						
arg0	30% (28)	20% (29)	0% (33)	40% (22)	50% (32)	40% (34)
gendir	40% (48)	60% (32)	0% (40)	20% (42)	50% (38)	40% (35)
subjdir	30% (48)	20% (33)	0% (40)	40% (42)	80% (38)	40% (35)
v1	10% (48)	50% (33)	60% (40)	20% (42)	35% (38)	30% (35)
v2	20% (120)	45% (174)	40% (150)	20% (131)	35% (113)	40% (236)

Table 2: Percentage of Default-Valued Parameters and Mean Fixation Times

dicates the mean number of interaction cycles to fixation for each condition. In similar experiments with random reproduction of LAgts, this mean across all conditions was 139, which gives an estimate of the average time taken to fixation under random drift. Thus, mean fixation times of 50 cycles or less with a strong bias (say, $\geq 80\%$ or $\leq 20\%$) towards either type of parameter constitute reasonable evidence of consistent selection pressure. However, either a strong bias coupled with a higher mean fixation time or lack of a strong bias, even with a lower fixation time, cannot be considered reliable evidence.

If we consider the subject direction parameter (**subjdir**) and the non-incremental memory-constrained learner (n4+ml), we can see that there appears to be quite strong selection pressure in favour of the default initial (leftward) setting when learning SVO, VOS or SVOv1, weaker pressure for VSO and SOV and either very weak or no pressure for SOVv2. Similar interactions with language are in evidence with the other parameters and acquisition procedures. On the other hand, this pressure for a default initial value for **subjdir**, and other parameters, is generally reduced or gone when n4 is not memory-constrained, as can be seen from the generally higher fixation times and lower percentages. With i1 there appears to be no consistent pressure for either type of initial value, especially without memory constraints; or an opposite pressure for an unset initial value, as with **subjdir** and i1+ml learning SVO, for example.

The trend in favour of default values with n4 is what we would predict, given the results summarized in Table 1 of section 4.1, which show that there is a greater efficiency gain with default initial settings for most of the languages tested for this acquisition procedure. The trend against some default settings with i1 is not so predictable, and underlines the need for this second type of experiment if the dynamics of such procedures are to be thoroughly explored. The effect of maturational memory limitations is to decrease fixation times and to (mostly) increase selection for default initial values, though this is far less clearcut with the incremental acquisition procedure.

The conclusions that can be drawn from these experiments are limited because the range of variation available for selection amongst acquisition procedures is very constrained. Nevertheless, they demonstrate that the precise form of acquisition procedure which emerges will be very dependent on the environment of adaptation. While genetic assimilation may occur in a wide variety of scenarios, a SVO default learner is only likely to emerge in the presence of some and not other dominant languages, and may even require additional assumptions, such as utilization of a non-incremental memory-constrained acquisition procedure in the ancestral population.

7 Coevolution

The experiments of section 5 demonstrated evolution of languages on a historical timescale within a genetically-invariant population of LAgts. Those of section 6 demonstrated evolution of acquisition procedures, within circumscribed limits, with maintenance of a single dominant language (see section 4.2). To demonstrate coevolution, it is necessary to allow the LAgt population to evolve and to create a significant degree of linguistic variation in the same run. LAgts' initial p-settings were varied by allowing mutation of a single element of a LAgt p-setting (with probability 0.05) during LAgt reproduction. Successful variant initial settings could then propagate through the population via single-point crossover (with probability 0.9). This allowed much less circumscribed evolution of initial p-settings than in section 6. In addition, the parameter n which determined the number of updatable parameters per trigger could mutate by $+/-1$ with probability 0.05 during LAgt reproduction. The full fitness function was used.

7.1 Coevolution without Migrations

In the first series of such experiments, the initial population consisted of memory-unconstrained unset n4 adults speaking one of the clearly-attested full languages. Mutations can change a principle to a parameter or vice-versa, alter the type of a parameter, or flip the value of a principle or default parameter. Therefore, they could, in theory, introduce a language variant by altering the value of a principle or default parameter. However the degree of linguistic variation in such runs was typically minimal with populations sampling around 5 closely-related full languages over 500 interaction cycles.

In these runs, the populations always evolved towards initial p-settings which enhanced the learnability of the dominant language in the environment. Figure 22 shows mean fitness for one such population and also the relative proportions of default parameters, unset parameters and principles in the same population. In all such runs, the proportion of default parameters grew at the expense of unset ones, with default values reflecting the language(s) of the environment. In addition, the mean number of updatable parameters per trigger fell until typically the whole population converged on a value of 2 or 3, depending on the dominant language. Consequently, LAgt fitness improved over the course of the run as a result of reduction in learning costs, while mean parsability, expressiveness and communicative

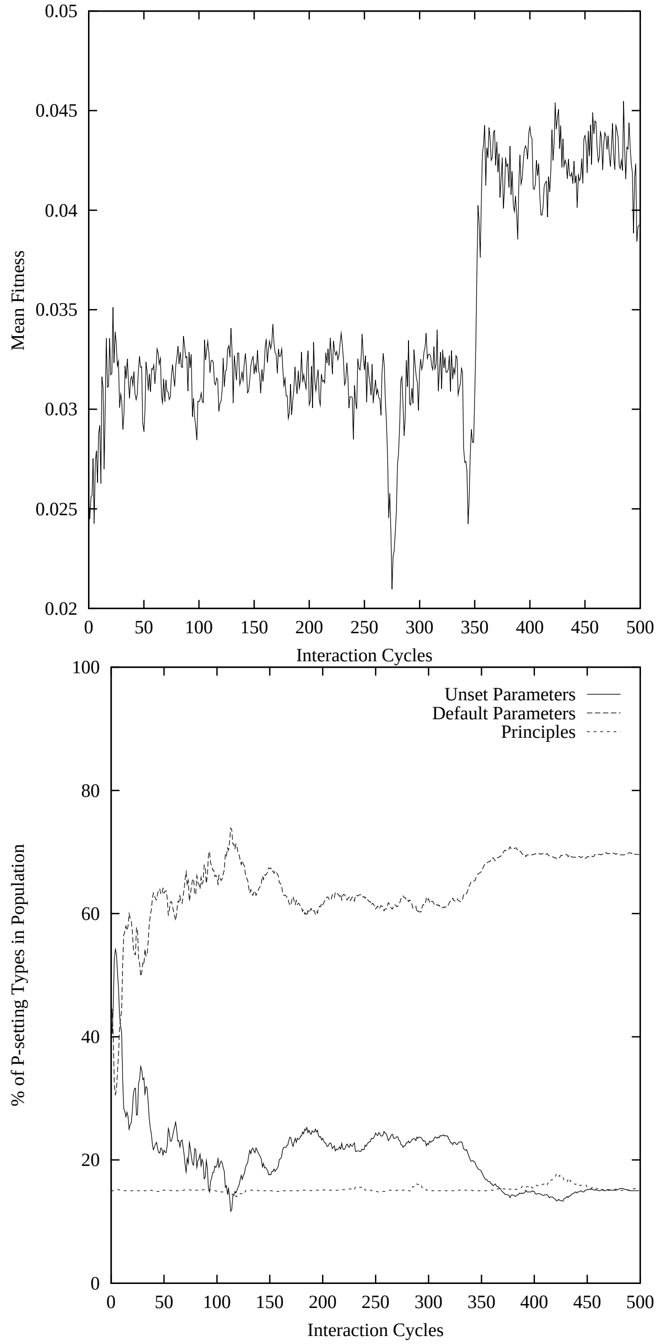

Figure 22: Mean fitness and p-setting types during coevolution without migrations

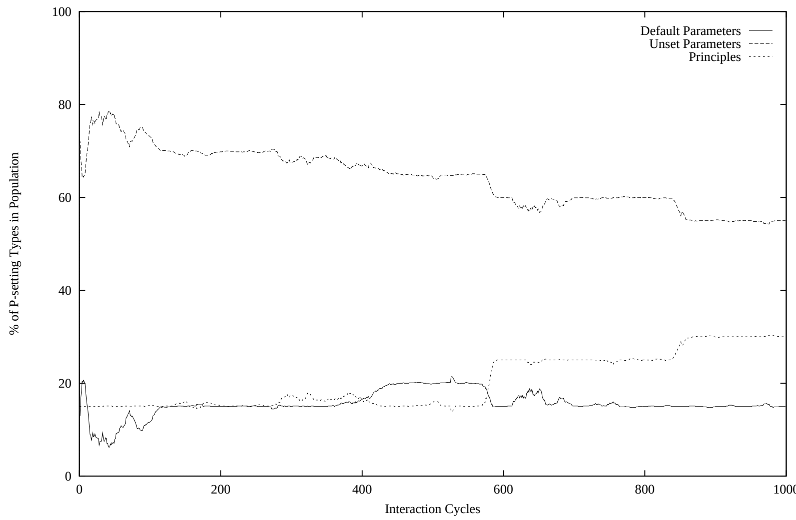

Figure 23: P-setting types during coevolution with migrations

success remained roughly constant. Similar experiments using the i1 procedure with maturational memory limitations showed similar, though weaker effects.

These results are clear evidence of genetic assimilation in which LAgts are evolving to be able to acquire the dominant language(s) more effectively. By replacing unset with default parameters which have initial settings compatible with the dominant language(s), the LAD is evolving an accurate language-specific inductive bias which simplifies grammatical acquisition. At the same time, this bias itself will alter the relative learnability of other languages. However, linguistic variation in these simulations is very limited, caused only by occasional failures of convergence, mutations of default parameter values, or mutations of parameters to principles. Consequently, the rate of linguistic change is very slow, creating a fairly constant selection pressure for genetic assimilation to work on. As discussed in section 1.3, Deacon (1997) has argued that genetic assimilation will not occur because, in practice, languages change faster than genetic mutations can go to fixation in populations.

7.2 Coevolution with Migrations

To see whether genetic assimilation would occur with maximal linguistic variation consonant with communicative success, a second series of experiments was run identical to those described above, except that migrations occurred as often as was compatible with mean 90% communicative success over the entire 1000 cycle run. Figure 23 shows the relative proportions of default parameters, unset parameters and principles for one such run with the population initialized to the unset n4 memory-constrained acquisition procedure.

In these runs, LAgts still evolved LADs which improved learnability despite the fact that typically the dominant language changes about 20 times and approximately 50 languages are sampled by the population. However, as in the run shown, there was a greater tendency to replace unset parameters with principles rather than just with default parameters. Over 10 such runs, the proportion of unset parameters always declined, by a mean 35% leading to around 50% of p-settings being principles or default parameters with roughly an equal number of new principles and default parameters. In other respects, results were identical to the first series of runs with LAgt fitness improving as a consequence of reduced learning cost. However, the greater degree of linguistic variation also allowed greater linguistic selection for more optimal languages.

The replacement of unset parameters by principles is an example of the type of genetic assimilation which Pinker and Bloom (1990) envisage, in which the class of learnable languages is (further) constrained by the LAD in the interests of enhanced learnability. Thus, in these runs we see examples of genetic assimilation of both inductive biases (defaults) and hard constraints (principles), albeit at a slower rate than when the linguistic environment was more constant. To see how long genetic assimilation would continue in a heterogeneous linguistic environment, several such simulations with migrations were run for 10,000 cycles. In these, the mean decline in the proportion of unset parameters was 55% with a mean 65% of p-settings being principles or default parameters at the end of the runs. Once again approximately half of the replaced unset parameters were default parameters. Plots of the proportions of each type of parameter show an asymptotic rate of genetic assimilation for default parameters and principles. Finally, in similar runs with populations initialized to reproduce learners with all default parameters, with values appropriate to the initial language, the population invariably evolved away from such 'total' genetic assimilation towards p-settings containing some unset parameters. Therefore, the model suggests that there is an upper limit to genetic assimilation in the face of such linguistic variability.

7.3 Discussion

Why then is there (partial) genetic assimilation even in the face of great linguistic heterogeneity and rapid linguistic change? And why, when change is rapid, is there a greater tendency for the assimilation of principles as well as default initial parameter values? Firstly, consider the possible mutations which can occur within a p-setting and their expected fitness effects; Table 3 catalogues the possible transitions of individual initial p-settings (which can be created by a single mutation) and their expected fitness cost ($f <$) or benefit ($f >$) in terms of the 'truth/falsity' (T/F) of the resulting p-setting value in the current linguistic environment. The fitness cost or benefit is based on the expected effect on learnability. It is clear that any transition from a false principle (i.e. one which is inconsistent with the current linguistic environment) will incur a fitness benefit, because it will allow a LAgt a chance to learn the dominant language. By contrast, a transition from a true principle to anything other than a true default will have a learning cost because it will either render learning impossible or increase the number of parameters to be updated. Likewise, no transition from a true default creates any benefit and three incur a cost. Three transitions from a false default incur learning benefit, only a transition to a false principle incurs a cost, by making learning impossible. Transitions from unset parameters to true default parameters or true principles are beneficial, while a false principle, as always, incurs a cost. The transition to a false default incurs no cost (or benefit) because during learning it still takes one parameter update to obtain the correct value.

It should be clear from this analysis, that what we would expect to evolve is a population with correct principles, predominantly correct default initial parameter values, and possibly a minority of unset and/or default incorrect parameters. In an unchanging linguistic environment, we would expect the population to eventually fix on all true principles or default parameters. However, in all the experiments reported above the linguistic environment is never entirely homogeneous or static. Therefore, the 'truth/falsity' of a p-setting is an approximation: a value may be predominantly correct in the current environment given the dominant language,

Old		New		
PS-Type	P-value	PS-Type	P-value	Expected Fitness
Absol	F	Def	F	$f >$
Absol	F	Def	T	$f >$
Absol	F	Absol	T	$f >$
Absol	F	Unset	?	$f >$
Absol	T	Def	F	$f <$
Absol	T	Def	T	$f =$
Absol	T	Absol	F	$f <$
Absol	T	Unset	?	$f <$
Def	F	Absol	F	$f <$
Def	F	Absol	T	$f >$
Def	F	Def	T	$f >$
Def	F	Unset	?	$f =$
Def	T	Absol	F	$f <$
Def	T	Absol	T	$f =$
Def	T	Def	F	$f <$
Def	T	Unset	?	$f <$
Unset	?	Absol	T	$f >$
Unset	?	Absol	F	$f <$
Unset	?	Def	T	$f >$
Unset	?	Def	F	$f =$

Table 3: P-setting Transitions and Fitness Effects

but become predominantly or completely incorrect over succeeding cycles (and vice versa). Whether an initially beneficial mutation achieves fixation, or even predominance, within the population will depend not only on the initial benefit it offers the mutated LAgt, but also on the continuing benefit to its descendents. It is here that coevolutionary effects will occur; for example, as a predominantly correct principle spreads through the population, it will create greatly increased linguistic selection for languages which obey this principle. This, in turn, will increase the chance that the principle will go to fixation in the population, rendering languages which do not obey the principle unlearnable. Similar reasoning applies to default-valued parameters.

In a changing environment, we might expect there to be a preference for default parameters over absolute principles, because an initially predominantly correct principle which spread through a proportion of the population would incur a high cost to them if it subsequently became (predominantly) incorrect. By contrast, a default parameter which becomes incorrect, incurs no more cost than an unset parameter, given the acquisition procedure assumed in the current simulation. There does appear to be a bias towards genetic assimilation of default parameters in the experiments reported above with lowish rates of linguistic change (see also section 4.3). However, the migration mechanism, used in the simulation for introducing linguistic variation, tends to reinforce the status of principles which have spread through more than 50% of the population and accelerate their fixation (because it introduces adults with identical initial p-settings to those of the existing majority). So, further experiments are needed to explore the degree of genetic assimilation of principles as opposed to default parameters using different migration models and different language acquisition procedures.

In the experiments reported above with mean 90% communicative success, the fastest observed rate of change from one dominant language variant to another was 4 interaction cycles. The fastest observed rate at which a mutation in a p-setting reached fixation was 43 cycles. This suggests that linguistic evolution of grammatical parameters was only about one order of magnitude faster than 'genetic' evolution of p-settings. Increasing the speed of linguistic change would have resulted in a decrease in communicative success below what is assumed reasonable for a speech community. Increasing the size of the population of LAgts would slow down the fixation of p-settings. Nevertheless, the simulation tells us nothing about the true relative rates of linguistic and biological evolution. There can be no certainty about the size of the ancestral population in which the LAD evolved, though it was probably small (Dunbar, 1993; Rogers and Jorde, 1995). Deacon (1997:329) suggests that linguistic evolution is 'many' orders of magnitude faster than biological evolution, arguing that languages can change their major grammatical properties over thousands of years (historically, 1–2 millenia for the types of constituent order properties modelled here). However, the time taken for a major grammatical change and the time taken for biological evolution will depend critically on population size, geographical dispersal, diffusion rates of genes and of variant grammatical forms, and so forth. In the simulation runs with rapid linguistic change, typically 2-3 major grammatical changes propagate through the population every 50 interaction cycles. Therefore, default parameters and absolute principles are being genetically assimilated and going to fixation in the population typically in the face of several such major linguistic changes.

The key to understanding why genetic assimilation is still likely to occur, almost regardless of the relative speed of change, is that the sample space of possible grammars and associated languages is likely to be vastly larger than the number of grammars which can be sampled by a realistic-sized speech community in the time taken for a principle or default parameter to go to fixation. In the simulation, there are under 300 languages and only 70 distinct full languages. Therefore, in the time taken for a p-setting to go to fixation typically around 5% of the space of grammars might be sampled. This means that 95% of the selection pressure for genetic assimilation of grammatical information remains constant at any one time. In his discussion, Deacon (1997:329f) ignores the issue of the space of grammatical possibilities and the degree to which this can be sampled in the time required for a gene to spread through a population. It is impossible to estimate the real size of this space properly, but few linguists would probably balk at the idea that 30 independent binary grammatical parameters will be required to capture the differences between the world's languages in an account of universal grammar (e.g. Lightfoot, 1999:259f). Given this, there are billions of distinct grammars to explore. This guestimate is based on the existence of an evolved LAD. Prior to the emergence of the LAD, the space of possible grammars would have been infinite. Rapid changes in the tiny subset of potential grammatical systems which the ancestral linguistic population was exposed to could not prevent genetic assimilation on the basis of the many potential systems which were *not* sampled; perhaps, for example, all those potential grammatical systems which would have resulted in arbitrarily intersecting dependencies between constituents (see section 1.3).

8 Conclusions and Further Work

The model of the LAD developed here extends work on grammatical acquisition in the parameter setting framework in several ways. Firstly, the partially-ordered limited-memory parameter setting procedure described integrates a computationally tractable and psychologically feasible algorithm with a more detailed account of UG within the GCG framework. Secondly, this procedure has been shown to be effective experimentally on a more complex parameter set than has been investigated in the parameter setting framework hitherto. Thirdly, the effect of maturational memory limitations ('starting small') has been shown to be largely irrelevant to convergence for this class of consistent procedures. Fourthly, the criterion for retaining a parameter update has been shown to have marked effects on the overall behaviour of the acquisition procedure. Fifthly, the coevolutionary experiments suggest that the starting point of any acquisition procedure will not be arbitrary but will be informed to some extent by the environment of adaptation for the LAD; that is, the LAD will incorporate inductive bias. This latter conclusion is important in any assessment of the significance of learnability arguments which assume arbitrary initial parameter configurations. Briscoe (1999, 2000a,b) extends the acquisition model presented here by developing a Bayesian, statistical approach to parameter updating which can deal with indeterminacy and noise in triggering data.

The evolutionary simulation model demonstrates that embedding a generative model of the LAD in a changing population of LAgts leads naturally to an account in which idiolects are well-defined stringsets, but languages are complex adaptive systems. Linguistic selection is primarily a consequence of properties of the LAD, and slight changes in the model of the LAD may create markedly different selection pressures. The model of working memory load incorporated into the GCG parser predicts that relative ease of parsability will be a factor in linguistic selection under a wide range of possible assumptions about the impact of parsability on language learning or use. The experiments on linguistic selection reported here underline the need for selection for expressiveness to maintain complex language. In general, conflicting linguistic selection pressures will create adaptive landscapes with many local optima capable of accounting for linguistic diversity and complexity, as well as underlying accounts of functionally-motivated universals and tendencies. Briscoe (2000a) demonstrates that linguistic selection on the basis of learnability can be made more sensitive to the relative-frequency of conflicting triggering data by utilizing a statistical approach to parameter updating. The resulting model is potentially better able to account for attested historical changes.

In answer to the question posed in section 1: how do partially inaccurate inductive biases arise and how pervasive are they in language acquisition? The work reported suggests that genetic assimilation of information into the LAD, on the basis of the dominant languages in the environment of adaptation, provides a plausible answer. When LAgts' p-settings can vary, under all experimental conditions genetic assimilation of more 'informative' default parameters or absolute constraints occurs. The general effect of genetic assimilation is to build in as much information concerning the linguistic environment as possible to make learning more efficient and robust. Thus, the idea that the LAD will incorporate both bias, in the form of initial default-valued parameters, and hard constraints, in the form of principles, is broadly supported. If such information is incorporated as principles then, given the coevolutionary scenario developed here, this has the effect of forcing languages to adapt to these evolving constraints. However, if such information takes the form of inductive biases during learning, linguistic variants will be more or less learnable depending on their compatibility with such biases. Therefore, we would expect to see peripheral constructions and typologically rarer grammatical systems which violate some inductive biases, especially if these optimize expressiveness or parsability or have a social motivation. Similarly, the account predicts that assimilation will be partial as a result of linguistic change during the period of adaptation. Bickerton's (e.g. 1984) bioprogram hypothesis, receives qualified support, if it is interpreted as the claim that the LAD incorporates *specific* default parameters, because this will only be the outcome of genetic assimilation if the environment of adaptation for the LAD was dominated by a language, or languages, with grammars (mostly) consistent with such defaults. Briscoe (2000b) argues that a selectionist, parametric account of language acquisition is able to account for rapid creolization if relevant demographic factors are also modelled.

It is important to consider whether *any* simulation model allowing evolution of both LAgts and languages, and conferring selective advantage to communicating LAgts would not show (some) genetic assimilation. Mayley (1996) demonstrates via a model and experiments that, for genetic assimilation to occur, there must be correlation between neighbouring phenotypes, attainable through lifetime adaptations, and their corresponding genotypes. In terms of the current simulation model, there is considerable correlation between steps of the acquisition procedure to converge on a specific grammatical system (i.e. parameter updates) and moves in genotype space representing biological evolution (i.e. changes in LAgts' initial p-settings) which reduce the number of learning steps. If, on the other hand, any small improvement in the acquisition procedure with respect to any target class of grammars required many changes at genotypic level (or vice-versa), then genetic assimilation would be unlikely to occur, even with selective pressure to learn language more efficiently.

Our current lack of knowledge of the neural basis of UG and parameter setting and of its genetic basis does not allow a definitive answer to the question of correlation. However, while the operations involved are no doubt very different from their representation in the simulation model, if we only assume that language acquisition further specifies a partial grammatical representation which itself is specified genetically, it is difficult to see how or why a highly uncorrelated genetic encoding of the neural representation might evolve. Nevertheless, the question does highlight the conditional nature of the conclusions which can be drawn from the results of any such (simulation) model. Not only the assumptions behind the model but also the many contingent, accidental or chance factors in the actual, but prehistoric, evolution of (proto)language(s) and their users may undermine the results. Nevertheless, models of this type have heuristic value in guiding us towards hypotheses which can then be further tested by other means; for example, claims about the effect of working memory on parsing are testable, in principle, via psycholinguistic experimentation or typological investigation, even though claims about the prehistoric development of language are not. Furthermore, such models can be used to evaluate evolutionary theorizing about language which does not utilize a simulation methodology and to expose implicit and, perhaps, incorrect inferences or assumptions in such theorizing; for example, in Deacon's (1997:329f) arguments from rapid linguistic change to the implausibility of genetic assimilation of grammatical knowledge.

References

Ackley, David and Michael Littman (1991) 'Interactions between learning and evolution' in Chris Langton and Charles Taylor (ed.), *Artificial Life II,* Menlo Park, CA: Addison-Wesley, pp. 487–509.

Aitchison, Jean (1996) *The Seeds of Speech,* Cambridge: Cambridge University Press.

Baddeley, Alan (1992) 'Working Memory: the interface between memory and cognition', *J. of Cognitive Neuroscience, vol.4.3,* 281–288.

Bertolo, Stefan (1995) 'Maturation and learnability in parametric systems', *Language Acquisition, vol.4.4,* 277–318.

Berwick, Robert (1985) *The Acquisition of Syntactic Knowledge,* Cambridge MA: MIT Press.

Berwick, Robert (1998) 'Language evolution and the minimalist program: the origins of syntax' in Jim Hurford, Michael Studdert-Kennedy and Chris Knight (ed.), *Approaches to the Evolution of Language,* Cambridge: Cambridge University Press, pp. 320–340.

Bickerton, Derek (1981) *Roots of Language,* Karoma, Ann Arbor.

Bickerton, Derek (1984) 'The language bioprogram hypothesis', *The Behavioral and Brain Sciences, vol.7.2,* 173–222.

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

Bouma, Gosse and Gertjan van Noord (1994) 'Constraint-based categorial grammar', *Proceedings of the 32nd Association for Computational Linguistics,* Las Cruces, NM, pp. 147–154, San Mateo, CA: Morgan Kaufmann.

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

Briscoe, Ted (1987) *Modelling Human Speech Comprehension: A Computational Approach,* New York: John Wiley.

Briscoe, Ted (1997) 'Co-evolution of language and of the language acquisition device', *Proceedings of the 35th Association for Computational Linguistics,* Madrid, pp. 418–427, San Mateo, CA: Morgan Kaufmann.

Briscoe, Ted (1998) 'Language as a complex adaptive system: co-evolution of language and of the language acquisition device', *Proceedings of the 8th Meeting of Computational Linguistics in the Netherlands,* Nijmegan, pp. 3–40, Amsterdam: Rodopi.

Briscoe, Ted (1999) 'The acquisition of grammar in an evolving population of language agents' in Muggleton, Stephen (ed.), *Machine Intelligence, 16: Electronic Transactions in Artificial Intelligence, Special Issue,* http://www.etaij.org/.

Briscoe, Ted (2000a, in press) 'Evolutionary perspectives on diachronic syntax' in Susan Pintzuk, George Tsoulas and Anthony Warner (ed.), *Diachronic Syntax: Models and Mechanisms,* Oxford: Oxford University Press.

Briscoe, Ted (2000b, in press) 'Grammatical acquisition and linguistic selection' in Ted Briscoe (ed.), *Language Acquisition and Linguistic Evolution: Formal and Computational Approaches,* Cambridge: Cambridge University Press.

Cecconi, Federico, Filippo Menczer and Richard Belew (1996) 'Maturation and the evolution of imitative learning in artificial organisms', *Adaptive Behaviour, vol.4.1,* 29–50.

Chomsky, Noam (1981) *Government and Binding,* Dordrecht: Foris.

Chomsky, Noam (1986) *Knowledge of Language: Its Nature, Origin and Use,* New York: Praeger.

Chomsky, Noam (1991) ‘Some notes on economy of derivation and representation’ in Robert Freidin (ed.), *Principles and Parameters in Comparative Grammar,* Cambridge MA: MIT Press, pp. 417–454.

Clark, Eve (1993) *The Lexicon in Acquisition,* Cambridge: Cambridge University Press.

Clark, Robin (1992) ‘The selection of syntactic knowledge’, *Language Acquisition, vol.2.2,* 83–149.

Clark, Robin and Ian Roberts (1993) ‘A computational model of language learnability and language change’, *Linguistic Inquiry, vol.24.2,* 299–345.

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

Curtiss, Susan (1988) ‘Abnormal language acquisition and the modularity of language’ in Frederick Newmeyer, vol. 2 (ed.), *Linguistics: The Cambridge Survey,* Cambridge: Cambridge University Press, pp. 96–116.

Cziko, Gary (1995) *Without Miracles: Universal Selection Theory and the Second Darwinian Revolution,* Cambridge MA: MIT Press.

Darwin, Charles (1871, 1989) *The descent of Man, Selection in Relation to Sex,* London: Pickering and Chatto.

Dawkins, Richard (1982) *The Extended Phenotype,* Oxford: Oxford University Press.

Dawkins, Richard (1983) ‘Universal Darwinism’ in Douglas Bendall (ed.), *Evolution: From Molecules to Men,* Cambridge: Cambridge University Press, pp. 403–425.

Dawkins, Richard (1989) *The Selfish Gene,* Oxford: Oxford University Press, 2nd edition.

Deacon, Terrence (1997) *The Symbolic Species: Coevolution of Language and Brain,* Cambridge MA: MIT Press.

Dennett, Daniel (1991) *Darwin's Dangerous Idea: Evolution and the Meanings of Life,* New York: Simon and Schuster.

Dixon, Robert (1997) *The Rise and Fall of Languages,* Cambridge: Cambridge University Press.

Dresher, Ellen and Paul Kaye (1990) ‘A computational learning model for metrical phonology’, *Cognition, vol.34,* 137–195.

Dunbar, Robin (1993) ‘Co-evolution of neocortex size, group size and language in humans’, *Behavioral and Brain Sciences, vol.16,* 681–734.

Durham, William (1991) *Coevolution, Genes, Culture and Human Diversity,* Pala Alto, CA: Stanford University Press.

Elman, Jeffrey (1993) ‘Learning and development in neural networks: the importance of starting small’, *Cognition, vol.48,* 71–99.

Frank, Robert and Shyam Kipur (1996) ‘On the use of triggers in parameter setting’, *Linguistic Inquiry, vol.27.4,* 623–660.

Futuyma, Douglas and Slatkin, Mark (1983) *Coevolution,* Sunderland, MA: Sinauer Associates.

Gathercole, Susan and Alan Baddeley (1993) *Working Memory and Language,* Hove, UK: Lawrence Erlbaum.

Gazdar, Gerald (1988) ‘Applicability of indexed grammars to natural languages’ in Uwe Reyle and Christian Rohrer (ed.), *Natural Language Parsing and Linguistic Theories,* Reidel, Dordrecht, pp. 69–94.

Gibson, Edward (1991) *A Computational Theory of Human Linguistic Processing: Memory Limitations and Processing Breakdown,* Doctoral dissertation, Carnegie Mellon University.

Gibson, Edward (1998) ‘Linguistic complexity: locality of syntactic dependencies’, *Cognition, vol.68,* 1–76.

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

Gopnik, Myrna (1994) 'Impairments of tense in a familial language disorder', *J. of Neurolinguistics, vol.8,* 109-133.

Greenberg, Joseph (1966) 'Some universals of grammar with particular reference to the order of meaningful elements' in Joseph Greenberg (ed.), *Universals of Grammar,* Cambridge MA: MIT Press, pp. 73–113.

Harris, Alice and Lyle Campbell (1995) *Historical Syntax in Cross-Linguistic Perspective,* Cambridge: Cambridge University Press.

Harris, Roy and Talbot Taylor (1997) *Landmarks in Linguistic Thought,* vol1, Routledge, London.

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

Hawkins, John (1994) *A Performance Theory of Order and Constituency,* Cambridge: Cambridge University Press.

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

Hoffman, Beryl (1995) *The Computational Analysis of the Syntax and Interpretation of 'Free' Word Order in Turkish,* PhD dissertation, University of Pennsylvania.

Hoffman, Beryl (1996) 'The formal properties of synchronous CCGs', *Proceedings of the ESSLLI Formal Grammar Conference,* Prague, pp. 85–98.

Holland, John (1993) *Echoing emergence: objectives, rough definitions and speculations for echo-class models,* Santa Fe Institute, Technical Report 93-04-023.

Holland, John (1995) *Hidden Order: How Adaptation Builds Complexity,* Menlo Park, CA: Addison-Wesley.

Hurford, Jim (1987) *Language and Number,* Oxford: Blackwell.

Hurford, Jim (1989) 'Biological evolution of the Saussurean sign as a component of the language acquisition device', *Lingua, vol.77,* 187–222.

Hurford, Jim (1991) 'The evolution of the critical period for language acquisition', *Cognition, vol.40,* 159–201.

Hurford, Jim (1998) 'Introduction: the emergence of syntax' in Jim Hurford, Michael Studdert-Kennedy and Chris Knight (ed.), *Approaches to the Evolution of Language,* Cambridge: Cambridge University Press, pp. 299–304.

Hyams, Nina (1986) *Language acquisition and the theory of parameters,* Dordrecht: Reidel.

Jablonka, Eva and Marion Lamb (1995) *Epigenetic Inheritance and Evolution,* Oxford: Oxford University Press.

Joshi, Aravind, K. Vijay-Shanker and David Weir (1991) 'The convergence of mildly context-sensitive grammar formalisms' in Peter Sells, Stuart Shieber and Tom Wasow (ed.), *Foundational Issues in Natural Language Processing,* Cambridge MA: MIT Press, pp. 31–82.

Kapur, Shyam and Robin Clark (San Mateo, CA: Morgan Kaufmann, 1994) 'The automatic construction of a symbolic parser via statistical techniques', *Proceedings of the Association for Computational Linguistics Workshop on Integration of Statistical and Symbolic Systems,* Las Cruces, NM.

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

Kegl, Judith and G. Iwata (1989) 'Lenguage de signos nicargüese: a pidgin sheds light on the "creole?" ASL', *Proceedings of the 4th Ann. Meeting of the Pacific Linguistics Conference,* Eugene, OR.

Keller, Rudi (1994) *On Language Change: The Invisible Hand in Language,* London: Routledge.

King, J and Just, M. (1991) 'Individual differences in syntactic processing: the role of working memory', *Journal of Memory and Language, vol.30,* 580–602.
Kimura, Motoo (1983) *The Neutral Theory of Molecular Evolution,* Cambridge: Cambridge University Press.
Kirby, Simon (1997) 'Competing motivations and emergence: explaining implicational hierarchies', *Language Typology, vol.1,* 5–32.
Kirby, Simon (1998) 'Fitness and the selective adaptation of language' in Jim Hurford, Michael Studdert-Kennedy and Chris Knight (ed.), *Approaches to the Evolution of Language,* Cambridge: Cambridge University Press, pp. 359–383.
Kirby, Simon (1999) *Function, Selection and Innateness: The Emergence of Language Universals,* Oxford: Oxford University Press.
Kirby, Simon and Jim Hurford (1997a) '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.
Kirby, Simon and Jim Hurford (1997) *The evolution of incremental learning: language, development and critical periods,* Edinburgh Occasional Papers in Linguistics, 97-2.
Kroch, Anthony (1989) 'Reflexes of grammar in patterns of language change', *Language Variation and Change, vol.1,* 199–244.
Kroch, Anthony and Ann Taylor (1997) 'Verb movement in Old and Middle English: dialect variation and language contact' in van Kemenade, A. and Nigel Vincent (ed.), *Parameters of Morphosyntactic Change,* Cambridge: Cambridge University Press, pp. 297–325.
Lascarides, Alex, Ted Briscoe, Ann Copestake and Nicholas Asher (1995) 'Order-independent and persistent default unification', *Linguistics and Philosophy, vol.19.1,* 1–89.
Lascarides, Alex and Ann Copestake (1999) 'Order-independent typed default unification', *Computational Linguistics, vol.25.1,* 55-106.
Lightfoot, David (1979) *Principles of Diachronic Syntax,* Cambridge: Cambridge University Press.
Lightfoot, David (1991) 'Subjacency and sex', *Language and Communication, vol.11,* 67–69.
Lightfoot, David (1992) *How to Set Parameters: Arguments from language Change,* Cambridge: Cambridge University Press.
Lightfoot, David (1999) *The Development of Language: Acquisition, Change, and Evolution,* Oxford: Blackwell.
Lindblom, Björn (1986) 'Phonetic universals in vowel systems' in John Ohala and Jaeger, J.J. (ed.), *Experimental Phonology,* Academic Press, Orlando, Fl., pp. 13–44.
Lindblom, Björn (1998) 'Systemic constraints and adaptive change in the formation of sound structure' in Jim Hurford, Michael Studdert-Kennedy and Chris Knight (ed.), *Approaches to the Evolution of Language,* Cambridge: Cambridge University Press, pp. 242–264.
Mayley, Giles (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,* Cambridge MA: MIT Press.
Maynard Smith, John (1987) 'When learning guides evolution', *Nature, vol.329,* 762.
Maynard Smith, John (1993) *The Theory of Evolution,* Cambridge: Cambridge University Press, 3rd ed..
Maynard Smith, John (1998) *Evolutionary Genetics,* Oxford: Oxford University Press, 2nd ed..
McMahon, April (1994) *Understanding Language Change,* Cambridge: Cambridge University Press.

Meisel, J. (1995) 'Parameters in acquisition' in Paul Fletcher and Brian MacWhinney (ed.), *The Handbook of Child Language,* Blackwell, Oxford, pp. 10–35.
Milward, David (1995) 'Incremental interpretation of categorial grammar', *Proceedings of the 7th European Association for Computational Linguistics,* Dublin, Ireland, pp. 119–126, San Mateo, CA: Morgan Kaufmann.
Mitchell, Tom (1990) 'The need for biases in learning generalizations' in Jana Shavlik and Tom Dietterich (ed.), *Readings in Machine Learning,* San Mateo, CA: Morgan Kaufmann.
Mitchell, Tom (1997) *Machine Learning,* New York: McGraw Hill.
Moortgat, Michael (1988) *Categorial Investigations,* Dordrecht: Foris.
Morgan, James (1986) *From Simple Input to Complex Grammar,* Cambridge MA: MIT Press.
Morrill, Glyn (1994) *Type Logical Grammar: Categorial Logic of Signs,* Dordrecht: Kluwer.
Nettle, Daniel (1999) *Linguistic Diversity,* Oxford: Oxford University Press.
Newport, Elissa (1977) 'Motherese: the speech of mothers to young children' in Castellan, N., Pisoni, D. and Potts, G. (ed.), *Cognitive Theory, Vol II,* Hillsdale, NJ: Lawrence Erlbaum, pp. 177–217.
Newport, Elissa (1990) 'Maturational constraints on language learning', *Cognitive Science, vol.14,* 11–28.
Niyogi, Partha (2000, in press) 'Theories of Cultural Change and their Application to Language Evolution' in Ted Briscoe (ed.), *Language Acquisition and Linguistic Evolution: Formal and Computational Approaches,* Cambridge: Cambridge University Press.
Niyogi, Partha and Robert Berwick (1996) 'A language learning model for finite parameter spaces', *Cognition, vol.61,* 161–193.
Niyogi, Partha and Robert Berwick (1997a) 'Evolutionary consequences of language learning', *Linguistics and Philosophy, vol.20,* 697–719.
Niyogi, Partha and Robert Berwick (1997b) 'Populations of learners: the case of Portuguese', *Proceedings of the Ann. Meeting of Cognitive Science Society,* Hillsdale, NJ: Lawrence Erlbaum.
Ochs, Eleanor and Sheiffelin, B (1995) 'The impact of language socialization on grammatical development' in Fletcher, P. and MacWhinney, P. (ed.), *The Handbook of Child Language,* Blackwell, Oxford, pp. 73–94.
Oliphant, Michael and John Batali (1996, submitted) *Learning and the emergence of coordinated communication,* University of California, San Diego, ms..
Osherson, Dan, Michael Stob and Scott Weinstein (1986) *Systems That Learn,* Cambridge: Cambridge University Press.
Peak, David and Frame, Michael (1994) *Chaos Under Control: The Art and Science of Complexity,* New York: Freeman.
Pinker, Steven (1994) *The Language Instinct,* New York: Morrow.
Pinker, Steven and Paul Bloom (1990) 'Natural language and natural selection', *Behavioral and Brain Sciences, vol.13,* 707–784.
Pullum, Geoffrey (1981) 'Languages with object before subject: a comment and a catalogue', *Linguistics, vol.19,* 147–155.
Pullum, Geoffrey (1983) 'How many possible human languages are there?', *Linguistic Inquiry, vol.14.3,* 447-467.
Rambow, Owen and Aravind Joshi (1994) 'A processing model of free word order languages' in Chuck Clifton, Lyn Frazier and Keith Rayner (ed.), *Perspectives on Sentence Processing,* Hillsdale, NJ: Lawrence Erlbaum, pp. 267–301.
Richards, Robert (1987) *Darwin and the Emergence of Evolutionary Theories of Mind and Behaviour,* Chicago, IL: University of Chicago Press.
Rogers, R.G. and Jorde, L.B. (1995) 'Genetic evidence on modern human origins', *Human Biology, vol.67,* 1–36.

Rose, Stephen (1997) *Lifelines: Biology, Freedom, Determinism,* London: Penguin Press.

Roughgarden, John (1979) *Theory of Population Genetics and Evolutionary Ecology: An Introduction,* New York: Macmillan.

Roughgarden, John (1983) 'The theory of coevolution' in Douglas Futuyma and Mark Slatkin (ed.), *Coevolution,* Sunderland, MA: Sinauer Associates, pp. 33–64.

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

Sigmund, Karl (1993) *Games of Life: Explorations in Ecology, Evolution and Behaviour,* London: Penguin.

Smith, Neil and Ianthi-Maria Tsimpli (1991) 'Linguistic modularity? A case study of a 'savant' linguist', *Lingua, vol.84,* 315–351.

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

Steedman, Mark (1988) 'Combinators and grammars' in Richard Oehrle, Emmon Bach and Dierdre Wheeler (ed.), *Categorial Grammars and Natural Language Structures,* Dordrecht; Reidel, pp. 417–442.

Steedman, Mark (1996) *Surface Structure and Interpretation,* Cambridge MA: MIT Press.

Steels, Luc (1998) 'Synthesizing the origins of language and meaning using coevolution, self-organization and level formation' in Jim Hurford, Michael Studdert-Kennedy and Chris Knight (ed.), *Approaches to the Evolution of Language,* Cambridge: Cambridge University Press, pp. 384–404.

Tomlin, Richard (1986) *Basic Word Order: Functional Principles,* London: Routledge.

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

Waddington, Conrad (1975) *The Evolution of an Evolutionist,* Edinburgh: Edinburgh University Press.

Wanner, Eric and Lila Gleitman (1982) 'Introduction' in Eric Wanner and Lila Gleitman (ed.), *Language Acquisition: The State of the Art,* Cambridge MA: MIT Press, pp. 3–48.

Wexler, Ken and Peter Culicover (1980) *Formal Principles of Language Acquisition,* Cambridge MA: MIT Press.

Wexler, Ken and Rita Manzini (1987) 'Parameters and learnability in binding theory' in Thomas Roeper and Edwin Williams (ed.), *Parameter Setting,* Dordrecht: Reidel, pp. 41–76.

Wood, Mary (1993) *Categorial Grammars,* London: Routledge.

Worden, Robert (1998) 'The evolution of language from social intelligence' in Jim Hurford, Michael Studdert-Kennedy and Chris Knight (ed.), *Approaches to the Evolution of Language,* Cambridge: Cambridge University Press, pp. 148–168.