

Evolutionary Perspectives on Diachronic Syntax

Ted Briscoe

University of Cambridge

(Final Draft for Proc. of DIGS5)

The main purpose of this article is to argue the merits of ‘population thinking’ in gaining insight into linguistic and, in particular, syntactic change. Population-level thinking and modelling can shed new light on many issues in the study of language acquisition and language change, and leads directly to a precise and useful characterisation of E-language. Something which is lacking in current generative linguistics. Moreover, this way of thinking is fully compatible with the major insights of the latter, and integrates them into a framework in which language variation and change are inherent and inevitable, rather than peripheral and/or accidental, properties of language. I will argue that (E-)languages are best modelled as particular kinds of *dynamical systems*; namely, *complex adaptive systems* (where these terms are used in technical senses made precise below).

The article both introduces some relevant ideas and techniques from modern evolutionary theory, and from the mathematical and computational study of dynamical systems, and also offers a critique and review of some recent work on syntactic change in this emerging framework, arguing that a useful population model needs to support overlapping generations of language users and learners and to allow quite detailed modelling of differing demographic scenarios. I utilise simple linguistic scenarios based on constituent order changes to illustrate the ideas and techniques clearly. I abstract away from the sociolinguistic detail of the actuation and diffusion of changes, and also from much of the linguistic detail of attested changes. However, the article also contains extensive references both to further background material and to more specific and detailed work within the general framework exemplified here.¹

1 The Basic Model

A formal model achieves greatest generality, and thus validity, by making as few assumptions as necessary and by omitting as much extraneous detail as possible. It is a matter of judgement, of course, deciding what is extraneous and what essential. Chomsky (1965) defined grammatical competence in terms of the language of (i.e. stringset generated by) an ideal speaker-hearer at a single instant in time, abstracting away from working memory limitations, errors of performance, and so forth. The generative research programme has been very successful, but, one legacy of the idealisation to a single speaker at a single instant has been the sidelining of variation and change. My model will embody similar assumptions: language users will, in essence, be Chomskyan ideal speaker-hearers (and learners), their linguistic interactions will be very simplified, and so forth. An assumption I will make here, in common with most work in diachronic generative syntax, is that language acquisition is the fundamental factor in significant syntactic change, and that there is a critical period for, at least, syntactic acquisition, after which no major change to the acquired adult grammar takes place (e.g. Lightfoot, 1979, 1999).

1.1 A Language Agent

We want to model language users and learners in a manner which abstracts away from many details potentially relevant to their behaviour, but which preserves what we think is essential for a model of language learning and use. A language agent can learn, produce and interpret a language, defined as a well-formed set

¹ McMahon (1994:ch12) is a brief account of the chequered history of evolutionary ideas and terminology in linguistic theory. Keller (1994) provides a detailed critique of Müller and Schleicher’s theories of language and goes on to argue for a similar view of (E-)language to that taken here. My fundamental argument is that, despite this chequered history, it is worth a second try 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. Cziko (1995), Dennett (1995), Kauffman (1993), Peak and Frame (1994), Sigmund (1993).

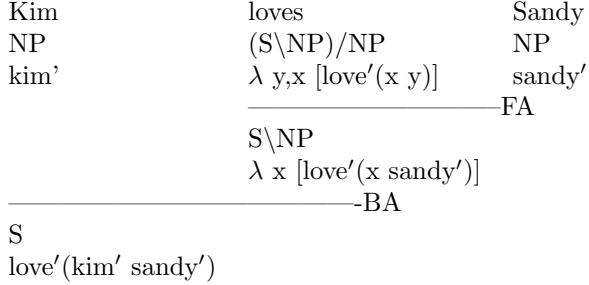


Figure 1: CG Derivation for *Kim loves Sandy*

of strings with associated logical forms, by acquiring and using a generative grammar according to precisely specified procedures.² We can think of language agents as embodying a model of the language acquisition device (i.e. a universal grammar, UG , a parsing procedure, P , and a grammar learning procedure, LP) with the addition of a simple generating procedure, G .

I will define UG to be the space of classical Categorial Grammar (i.e. AB CG e.g. Wood, 1993:7f) with atomic categories S, N and NP, other (complex) categories formed by left-associative combination of categories with slash or backslash (e.g. $S\backslash NP$, ‘English intransitive verb’; $(S\backslash NP)/NP$, ‘English transitive verb’), and the two directional variants of function-argument application defined below (where X and Y are variables over distinct tokens of (sub)categories and X' and Y' denote the semantic values of these (sub)categories):

<table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="text-align: center; padding: 5px;">Forward Application (FA):</td></tr> <tr> <td style="text-align: center; padding: 5px;">$X/Y Y \Rightarrow X \quad \lambda Y' [X'(Y')] (Y') \Rightarrow X'(Y')$</td></tr> <tr> <td style="text-align: center; padding: 5px;">Backward Application (BA):</td></tr> <tr> <td style="text-align: center; padding: 5px;">$Y X\backslash Y \Rightarrow X \quad \lambda Y' [X'(Y')] (Y') \Rightarrow X'(Y')$</td></tr> </table>	Forward Application (FA):	$X/Y Y \Rightarrow X \quad \lambda Y' [X'(Y')] (Y') \Rightarrow X'(Y')$	Backward Application (BA):	$Y X\backslash Y \Rightarrow X \quad \lambda Y' [X'(Y')] (Y') \Rightarrow X'(Y')$
Forward Application (FA):				
$X/Y Y \Rightarrow X \quad \lambda Y' [X'(Y')] (Y') \Rightarrow X'(Y')$				
Backward Application (BA):				
$Y X\backslash Y \Rightarrow X \quad \lambda Y' [X'(Y')] (Y') \Rightarrow X'(Y')$				

Application combines a complex functor category with an argument category to form a derived category (with one less (back)slashed argument category). So, for example, the category associated with an English intransitive verb encodes the fact that it can combine with a subject NP to its left to form a sentence. Grammatical constraints of order and agreement are captured by only allowing directed application to adjacent matching categories. Application is paired with a corresponding determinate semantic operation, shown here in terms of the lambda calculus, which compositionally builds a logical form from the basic meanings associated with lexical items.³ A simple derivation is shown in Figure 1. A trigger sentence is defined as a surface form (SF), a string of words, with an associated logical form (LF), a possibly underspecified formula of the lambda calculus: $t_i = \{< w_1, w_2, \dots, w_n >, LF_i \}$. In the case of the example in Figure 1 (1a) and (1b), respectively.

- (1) a Kim loves Sandy
- b love'($\text{kim}' \text{sandy}'$)
- c Kim:NP loves:S\NP_s/NP_o Sandy:NP

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 , $< w_1 : c_1, w_2 : c_2, \dots, w_n : c_n >$ such that the parse derivation, d_i for this sequence

²I borrow the term ‘agent(s)’ from computer science for this idealisation to emphasise both their autonomy and their artificiality, and because they are going to form part of a decentralised, distributed system.

³Thus, I adopt a deterministic syntax-driven compositional LF construction framework in common with Montague Grammar and most work with CGs (see e.g. Dowty *et al.*, 1981; Wood, 1993:29f). However, the details are not essential to an understanding of what follows, beyond the fact that the LF is determined by the derivation and, given a derivation, a LF is recovered.

of categories yields LF_i , as in (1c).⁴ Given $VCA(t)$, the parse derivation in CG (and thus the LF) will be unique, so the parsing algorithm, P , can be defined precisely and deterministically in terms of a simple shift-reduce parser that orders reduction via forward/backward application before shifting new words onto the analysis stack.⁵

The generation procedure, G , selects a trigger (i.e. an SF:LF pairing) given the agent's grammar and lexicon. A trigger is defined as any one of the finite proper subset of strings generated by a grammar and lexicon which involves at most one level of recursive application, in terms of a single recursive category or a chain of categories involving recursive application.⁶ I will assume a highly-skewed distribution over such degree-0 and degree-1 triggers so that a learning agent has a very high probability of being exposed to a fair and effective sample of them for any given target grammar and lexicon. Given a space of possible grammars, UG , and a learning procedure, LP , a fair and effective sample of triggers, $t_1 \dots t_n$, from the language defined by a target grammar, $L(g^t)$, will allow a learner to converge on g^t with high probability (i.e. $p > 1 - \epsilon$ where ϵ denotes a small error probability), because it will contain enough information to select the target grammar uniquely from others in UG using LP :⁷

$$p[LP(UG, t_1 \dots t_n \in L(g^t)) = g^t] > 1 - \epsilon$$

The acquisition procedure, LP , is error-driven and incremental. A learning agent begins with an empty lexicon and empty category set and incrementally hypothesises new category types and/or word:category associations. On each trigger presentation, LP finds a $VCA(t)$ yielding the appropriate LF. If this can be done using word:category associations available in the current lexicon, then no change takes place. Otherwise, new word:category associations are hypothesised and retained if these lead to successful recovery of the appropriate LF. For concreteness, I will assume LP only hypothesises one new word:category association per trigger, tries existing category types before creating new ones, and never deletes an association.⁸ For example, a learning agent presented with the trigger in (2a,b), with the word:category associations in (2c) as well as, say: *talk*:N/N might hypothesise and retain *red*:N/N.

- (2) a give me the red sock
- b give'(me' x) \wedge red'(sock'(x))
- c give:(S/NP)/NP me:NP the:NP/N red:? sock:N

On the other hand, an agent without any instance of N/N would need to hypothesise this new category type in the grammar before being able to add the appropriate word:category association to the lexicon. Thus, LP incrementally expands the lexicon and, if necessary, the category set to find the smallest grammar and lexicon (i.e. that containing the least number of category types and word:category associations) compatible with the analysable subset of the trigger sequence seen so far. The current $g \in UG$ hypothesised by a learning agent is entirely defined by the category set (and lexicon) since UG specifies forward/backward application as the only means of grammatical combination.

Though learning agents may be exposed to different triggers in different orders, given LP a learning agent

⁴I will write SVO for S(subject), O(object), etc to informally indicate relevant aspects of both SF and LF. I will also use the same abbreviations to indicate the LF associated with particular CG categories where it is convenient to suppress details of the semantic framework employed.

⁵Briscoe (1997, 1998) gives a more detailed and precise definition of such a parsing algorithm. In fact, the correct fully-specified and determinate LF cannot always be recovered from $VCA(t)$. For example, if this includes a sequence like: *almost*:(S\NP)/(S\NP) *smiled*:S\NP *deliberately*:(S\NP)\(S\NP), the parsing algorithm outlined yields the left-branching derivation and interpretation where the second adverb has wide scope. Such complications are not relevant here.

⁶Thus, triggers will not involve derivations such as N/N N/N N \Rightarrow N or NP/S NP/S S\NP \Rightarrow NP. This restriction is in accord with psycholinguistic evidence that children are exposed to a preponderance of unembedded (degree-0) triggers with about 16% of child-directed utterances involving one level of (degree-1) embedding (e.g. Newport, 1977). Restricting possible triggers to a finite proper subset of sentences generated by a given grammar also facilitates modelling.

⁷This formal setting and associated assumptions concerning models of language acquisition is based on recent work on formal learnability – see, for example, Niyogi (1996) for a particularly thorough treatment – though its antecedents go back at least to Wexler and Culicover (1980). The assumption that the learner has access to a fair and effective sample circumvents most of the substantive issues of language acquisition. However, our focus here is not on these issues but rather on how the process of acquisition influences language change.

⁸The restriction to a single word:category association per trigger entails that the learner must acquire a lexicon of names, N(P), before acquiring some one-place predicates, S\N(P).

must be exposed to a (not necessarily continuous) sequence of triggers in which the addition of one new word:category association per trigger leads to g^t . This implies that such incremental trigger sequences must be fairly common in the more likely subsets of triggers which will be sampled for $g \in UG$ otherwise LP will not be able to converge with high probability on exposure to a finite specific sequence, t_n . This requirement follows directly from the twin requirements that *one* hypothesised word:category association must yield the *complete* correct LF for a trigger.⁹

LP , as outlined, has no way of guaranteeing that the correct $VCA(t)$ will be recovered for each trigger. Even given the correct LF, there is still ambiguity in the assignment of $VCA(t)$ given arbitrary sequences of triggers. A well-known example (e.g. Clark, 1992) is the ambiguity of SVO triggers when UG includes V2 grammars. LP could ‘prematurely’ learn categories like $S \setminus NP_s / NP_o$ for such sequences, although g_t might be a V2 grammar and also generate OV triggers. With CG there will always be indeterminacy concerning the form of the new category that should be hypothesised. For example, a learner could infer a type-raised category like $(NP/N) \setminus (NP/N)$ with semantics $\lambda P \lambda x NP/N'[(NP/N) \setminus (NP/N)'[P(x)]]$ for *red* in (2a,b) yielding the same LF (e.g. Wood, 1993:42f). Again the assumption that the learner infers the simplest category (i.e. the one requiring the least number of atomic categories and (back)slash operators) would often suffice for this type of indeterminacy. Since we are not currently concerned with how the learner resolves such indeterminacies, I will assume (unrealistically) that a trigger comes labelled with the correct (partial) $VCA(t)$, as in (2c) above, and that the learner always determines the correct new category.

LP , as defined, is a lexically conservative learner because variant category assignments to a word token will not be generalised to other words of the same type. For example, if a learning agent has acquired the word:category associations: *nicely*: $(S \setminus NP) | (S \setminus NP)$ and *quickly*: $(S \setminus NP) / (S \setminus NP)$ (where $|$ is a variable which can be instantiated as slash or backslash), then LP will not generalise the association for *quickly*. I will assume that LP includes lexical (redundancy) rules of the general form: $Cat \Rightarrow Cat$ which get hypothesised appropriately. This assumption allows us to abstract away from lexical issues and the lexical content of triggers, which are also not our primary concern.¹⁰

This completes the linguistic part of our definition of a language agent. I extend this definition with an age, between 1 and 10, and a communicative success ratio (CSR), between 0 and 1, for reasons which will become clear in the next section. So, to summarise, a language agent has the following components:

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

Of course, there is much that is questionable about this simple model, the choice of UG , LP , and so forth. However, my intention is to explore what happens when such relatively simple and comprehensible models of language learners and users form speech communities, and then show how these predictions can be used to refine the model of a language agent.

1.2 Populations of Language Agents

A population is simply a set of language agents. However, we want this set to change over time as new agents are added ('born') and old agents are removed ('die'). One way of achieving this is to not allow generations of learning and adult agents to overlap; that is, to remove all adult agents from the population

⁹Gibson and Wexler (1994), following earlier work on formal learnability, incorporate both these requirements into the Trigger Learning Algorithm. However, Niyogi and Berwick (1996), Frank and Kapur (1996), Dresher (1999) and Fodor (1998) have all questioned one or both and made alternative proposals.

¹⁰Of course, the issue of when to hypothesise a (potentially semi-productive) lexical rule and how to determine its range of application is complex (e.g. Pinker, 1989; Schütze, 1997; Briscoe and Copstake, 1999). They are also relevant to diachronic syntax, since most accounts of lexical rule induction will predict initially slow lexical diffusion, followed by more rapid syntactic change when triggering data licenses induction of the rule, rather than just induction of specific word:category associations.

State	Pop-size	Learners	Non-learners	Ratio
0	4	4	0	∞
1	6	2	4	0.5
2	8	4	4	1.0
3	10	6	4	1.5
4	12	8	4	2.0
5	15	9	6	1.5
6	19	11	8	1.375
7	18	12	6	2.0
8	22	14	8	1.75
9	27	16	11	1.454
10	34	19	15	1.266
11	40	24	16	1.5
12	47	29	18	1.611
13	55	34	21	1.619
14	66	40	26	1.538
15	78	47	31	1.516
16	92	56	36	1.555
17	110	67	43	1.55
18	132	80	52	1.53
19	158	96	62	1.54
20	187	114	73	1.56

Table 1: Growth and Composition of a Population of Agents

once the learning period is finished, declare the learners the adults, and add a new batch of learners. This makes the population dynamics simple but unrealistic. My model supports overlapping generations because demographic dynamics, for example altering the proportion of learners to adults, may well be a factor in some types of language change (see e.g. section 4 below).

The constitution of the population changes over time according to some prespecified rules; agents are usually removed at age 10, and 2 distinct adult ($age \geq 4$) agents ‘reproduce’ one and only one new (age 1) learning agent during each of their adult ages (so agents can help create up to 7 offspring). We have already effectively defined a dynamical system; that is, a system which changes over time in a manner determined by the previous state. The system has a set of states, corresponding to distinct agent ages, and an update rule which determines the state of the system at time $t + 1$ in terms of its state at t :

$$s^{t+1} = Update(s^t)$$

Update will update the age of the agents at each time step of the system, remove any over age 10, and reproduce new agents from the set of adults. To see how the system behaves, suppose that we start in s^0 with 4 agents all aged 4. Table 1 shows the behaviour of the system through the first 20 states. Each row shows the state number, the total population size, the number of learners (aged 1–4), the number of non-learners (aged 5–10) and the ratio of learners over non-learners.

From s^1 to s^4 , *Update* simply adds 2 new agents reproduced by the 4 original adults. At s^5 and s^6 , 3 and then 4 new agents are added. At s^7 the total number of agents drops because the original 4 agents have reached age 10 and are removed. At this point learners outnumber non-learners 2:1. The population continues to expand exponentially. However, the ratio of learners over non-learners lies between 1.51 and 1.57 from s^{14} onwards. Thus, these few apparently simple linear rules have some quite surprising and unintuitive properties, highlighting both the need for simulation and/or mathematical analysis in order to understand the behaviour of even very simple dynamical systems, and the need to keep the model as simple as is consonant with our modelling purposes.

The *Update* rule above is unrealistic because there is no bound on population growth. Real populations

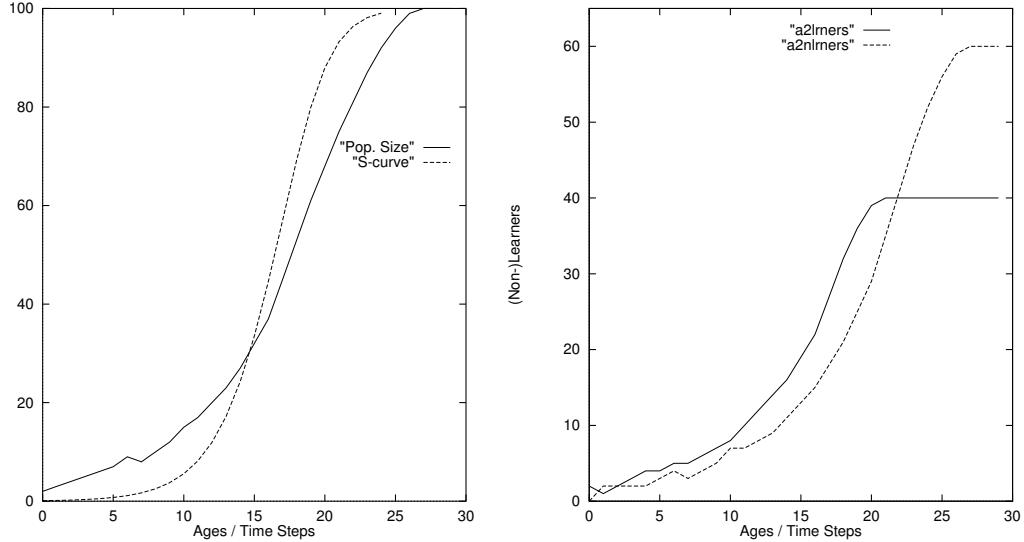


Figure 2: Population Dynamics

compete for finite resources and, therefore, typically show approximately logistic or S-shaped growth (e.g. Maynard Smith, 1998:15f; Peak and Frame, 1994:132f). We can impose a limit (crudely) by specifying an upper bound to the total number of agents that can be reproduced in a single time step. If we do this then we place an upper bound on the size of the population defined by the age of death multiplied by this limit. Furthermore the maximum number of learners is defined by this limit multiplied by the age at which learning ceases (i.e. the end of the critical period). From these two upper bounds we can derive the ratio of learners to non-learners once the population has converged to these upper bounds. For example, if we set an upper bound of 10 new agents per time step, then the example above will converge to a population of 100, 40 of whom are learners, giving a learner/non-learner ratio of 0.66. These two linear components of *Update* now define a nonlinear dynamical system.

The first plot in Figure 2 shows the composition of a population constructed as described, starting from 2 age 4 agents. Superimposed on this plot of total population size across time is a S-shaped curve constructed using the logistic map. This represents an approximation of the actual population growth; for example, not modelling the drop in population which occurs when the original agents die. The number of learners and non-learners is shown in the second plot. As can be seen there is fluctuation in the ratio until the population settles into its steady state.

The choices I have made here are different to those taken by Niyogi and Berwick (1997) in which a speech community is modelled using non-overlapping generations in a population of fixed size. These and other abstractions enable them to simulate the behaviour of the model in terms of updates rule which reduce to quadratic or higher polynomial maps, and thus prove analytically properties of the resulting dynamics. In effect, they simulate the *average* behaviour of a learner at each time step of the system given the probability distribution of triggers predicted from the adult population. If we simply took the logistic map as a correct model of population growth, then we could abstract away from the specific rules of birth, death and reproduction presented in this section and model the population dynamics abstractly. However, we would lose the flexibility and low-level variation inherent in the ‘microscopic’ rules introduced above. Below we will see that this flexibility is important for modelling some types of language change.¹¹

¹¹Maynard Smith (1998:15f) discusses the difference between modelling at the microscopic and the macroscopic level with respect to population dynamics. As modelling at the macroscopic level inevitably builds in more (perhaps unwarranted) assumptions and abstraction, I prefer the microscopic approach. In the long run derivation of mathematical laws of language change may, though, warrant macroscopic modelling. In the short run, we need microscopic modelling to discover the appropriate (if any) macroscopic model(s). In fact, the dynamics of the *Update* rule I have adopted here are probably better modelled macroscopically by the tent map (see Peak and Frame (1994:137f) for discussion and a general accessible introduction to nonlinear dynamical systems).

In all versions of the model, agents in the population attempt to interact with each other a prespecified number of times during each time step of the system; therefore, the time steps will be called interaction cycles. The *Update* rule then maps the population to the next state, and another interaction cycle commences. Note that a generation, defined as the time between birth and reproductive age, is 4 interaction cycles. The number of times agents interact during an interaction cycle is set so that, with high probability, each learning agent will be exposed to enough triggers to acquire any $g \in UG$ during the critical period. Thus if it takes 100 triggers to guarantee convergence to the ‘hardest’ grammar with $p > 0.99$, then the number of interactions might be set to a mean of 75 per agent per cycle, giving a mean of 300 interactions per learner during the critical period. In a mean 150 of these interactions, a learning agent will be the listening / parsing agent. Therefore, with high probability, learners will be given adequate triggering data to converge accurately to an adult grammar.¹²

The precise number of interactions per cycle depends on population size, so is calculated after each time step of the model. For each interaction, two distinct agents are randomly chosen from the population and one is randomly selected to be the speaking agent, the other the listening agent.¹³ The speaking agent generates a trigger given its current grammar (if any) and the listening agent attempts to parse it. If the listening agent is able to recover the same LF from the trigger that the speaking agent associates with it, the interaction is successful. If not, and the listening agent is still learning, then it attempts to modify its grammar and reparse the trigger. If this results in recovery of the right LF, then the interaction is successful, otherwise it fails. A population of agents constitutes a speech community if >90% of interactions between adult agents are successful. This is a fairly arbitrary cutoff, but nevertheless some such definition is needed to distinguish a population with no language or not much (shared) language from a speech community. The communicative success of an agent is just the ratio of successful to all interactions in which it has participated, regardless of whether it was the parsing or generating agent. The CSR is recalculated after each interaction cycle for each agent. For the moment, the only role that this ratio will play is to track whether a population constitutes a speech community.

For now, I will assume that reproduction simply creates a new agent with identical properties to all the other agents in the population. That is, the population of agents changes, and the ratio of learners to non-learners can vary, along with factors such as age of reproduction, death, and so forth. However, the population does not *evolve* at the genetic level because all agents have the same UG , LP , G and P . In this case, it is irrelevant which agents reproduce new agents. Nevertheless, I define reproduction in terms of two parent agents to facilitate extension to a model in which agents’ language faculties do evolve. Figure 3 summarises the main properties of the model introduced in sections 1.1 and 1.2.

1.3 A Simple Example of Language Change

Before we begin to study language change, it is important to consider under what conditions a speech community will be maintained, and under what conditions a language will *not* change. In general, speech communities do not seem to undergo such fundamental language change that communication breaks down significantly, even if during periods of major change there may be a small degree of intergenerational miscommunication (e.g. Lightfoot, 1999:7f). An intuitive requirement for maintenance of a speech community is that if there is no initial linguistic variation in that community and learners are presented with enough triggers, then they should converge to g^t . However, this is not necessary to maintain a speech community, nor sufficient to preclude language change in our model. It is possible for most learning agents in several overlapping generations to ‘misconverge’ to g^t and still maintain >90% successful interactions when adults.

¹²We could try to characterise this probability more precisely via a probabilistic analysis in terms of the (normal) distribution and standard deviation on triggers. However, it is also straightforward to find values which yield a stable model empirically.

¹³By random here and below, I mean randomly sampled with uniform probability, so that any two distinct agents have an equal chance of interacting during every interaction and each interacting agent has an equal chance of being the speaker or listener. Randomly sampling with respect to a uniform distribution is tantamount to making the weakest assumptions; in this case, about who talks to whom. We could, for example, assume that learners talk more to their parents and siblings or that populations are structured into subgroups with greater interactions within rather than across such groups. However, though these moves would make the model more concrete, they would also potentially make it less general if such assumptions do not hold in every situation. Niyogi (2000) contrasts the dynamics which result if learners are exposed to triggers just from their parents, from the entire adult community, or according to a spatial distribution of adults.

LAGt:	
$LP(UG, t)$	$= g$
$P(g, t)$	$= LF$
$G(g, LF)$	$= t$
Age :	$[1 - 10]$
CSR :	$[0 - 1]$
Pop:	
POP_n :	$\{LAGt_1, LAGt_2, \dots, LAGt_n\}$
INT:	$\text{Gen}(LAGt_i, LF_k), \text{Parse}(LAGt_j, t_k), i \neq j,$
SUCC-INT:	$\text{Gen}(LAGt_i, LF_k) = t_k \wedge \text{Parse}(LAGt_j, t_k) = LF_k$
REPRO:	$\text{Create-LAgt}(LAGt_i, LAGt_j), i \neq j$

Figure 3: The Basic Model

For example, a community in which learning agents outnumber adults may fixate on a proper subset grammar of g^t , because a high proportion of the triggers the learners are exposed to come from the ‘intermediate’ subset grammars of other learners.¹⁴ Thus, linguistic stasis requires not only a high probability of convergence to g^t during learning – which, in turn, requires little or no variation in triggers – but also population stability too. We will begin by considering stable models, in which the ratio of learning agents to adults will not significantly skew the distribution of triggers, and where the size of the population means that very occasional misconvergence by a single learner, typically to a subset language, will not skew the subsequent distribution of triggers to further learners enough to cause a chain reaction. In this situation, both the speech community and a homogeneous language will be maintained if there is little or no initial linguistic variation.¹⁵

As a first example, consider the following scenario. Initially there are 60 adult agents, of whom half have acquired grammar, g^1 , and half g^2 , in equal proportions. These grammars are identical except that g^1 generates nominal postmodifiers by associating them with category $N \setminus N$, and verbal postmodifiers $(S \setminus NP) \setminus (S \setminus NP)$, while g^2 generates nominal and verbal premodifiers by replacing the (highest-level) backslash operator with slash for these categories. The model is set up to be stable so that a learning agent exposed to triggers exclusively from g^1 (or g^2) would converge to that grammar with very high probability; so, in the absence of variation, linguistic homogeneity would be maintained within the speech community. However, once the variation is introduced (perhaps through contact between previously isolated communities) the situation is very different. A learner will, on average, be exposed to equal numbers of variant triggers, such as (3a) and (3b) with the same logical form (3c) but variant VCAs (3d) or (3e).

- (3) a Daddy gave you the sock red nicely
- b Daddy nicely gave you the red sock
- c nicely'(give'(daddy' me' x) \wedge red'(sock'(x)))
- d daddy:NP gave:((S\NP)/NP)/NP you:NP the:NP/N sock:N red:N\N
 nicely:(S\NP)\(S\NP)
- e daddy:NP nicely:(S\NP)/(S\NP) gave:((S\NP)/NP)/NP you:NP the:NP/N
 red:N/N sock:N

For example, if we assume that 1/12 of triggers generated from g^1 (or g^2) exemplify the variation, then a

¹⁴Briscoe (1998) discusses change via contraction of this type in more detail and argues that a realistic model of a language agent must incorporate a counteracting pressure for expressivity to prevent communities fixating on easily learnable but more restrictive subset languages.

¹⁵Briscoe (1998, 2000b) discusses the conditions required for linguistic stasis in more detail within a very similar model. Taking a stable model of this form as the starting point for the study of change represents a rather different philosophy from that of Niyogi and Berwick (1997) in which, given the learning procedure and conditions they model, a population initially speaking a -V2 language always fixates on +V2 grammars. See also Robert Clark (1996) for further discussion.

learning agent has, on average, a $1/24$ chance of seeing data distinguishing g^1 from g^2 every time it is exposed to a trigger (since, on average, half the triggers will be generated by agents who have acquired g^1). If the mean number of triggers the learning agent is exposed to is 100, then it will be exposed to both variations one or more times with probability $p = 0.971$.¹⁶

LP , as defined in section 1.1, will acquire both variants if it sees each at least once during the learning period, so learning agents will almost certainly converge to a grammar, g^3 , capable of generating triggers like (4) as well as all the original variant triggers in (3).¹⁷

- (4) a Daddy nicely gave you the sock red
- b nicely'(give'(daddy' me' x) \wedge red'(sock'(x)))
- c daddy:NP nicely:(S\NP)/(S\NP) gave:((S\NP)/NP)/NP you:NP the:NP/N
 sock:N red:N\N

The introduction of g^3 into the community represents language change in two ways. Firstly, (4a) is a new type of potential trigger; previously it was not possible for a single speaker to combine pre/postmodifiers. Secondly and consequently, the distribution of trigger types will change because the distribution of grammars amongst agents in the community has changed. Once grammar, g^3 , has entered this community, the overall proportion of triggers exhibiting pre/postmodifier variation will rise. Subsequent learners will be exposed to triggers like (3) from the original adult agents and to triggers like (3) and (4) from new adults who have acquired g^3 . This creates a chain reaction, that is, positive feedback, so that, unless further variation is introduced, the community will soon fixate on g^3 (in the sense that all adult agents will have acquired g^3). The first generation of learners will each acquire g^3 with $p = 0.971$. The joint probability that all will acquire g^3 , assuming 10 learners, is $p = 0.745$. The probability that more than half will is $p = 0.999$. Subsequent generations of learners, exposed to g^1 , g^2 , and increasingly g^3 , will converge to g^3 with increasingly higher probability. Then the original adult population of g^1 and g^2 agents will begin to die, so the proportion of g^3 agents will increase further. If no learner acquires a subset grammar, or g^1 or g^2 , then the community can fixate on g^3 within 6 interaction cycles.¹⁸

Now consider what happens if we start with 59 adult agents with g^1 and 1 agent with g^2 . The initial generation of learning agents each has a $1/60$ times $1/12$ chance of being exposed to the variant triggers in g^2 so the probability of this happening during the learning period is $p = 0.130$. Assuming 10 new learners again, the probability of half or more of them being exposed to the g^2 variant is $p = 0.005$, so the probability of subsequent generations of learners acquiring the g^2 variant is lower not higher, as the relative frequency of the variant trigger(s) will decline as long as more than half of the new agents acquire g^1 . Thus, it is very unlikely that a variant seen this infrequently during acquisition would ever spread through the community. Another way of saying this is that the community (as defined by the model) is unlikely to acquire variations based on individual innovations. However, in this community if $1/12$ of the adult agents (ie. 5) generate a variant trigger with a $1/12$ chance, a new grammar is quite likely to be acquired by subsequent generations of learners; that is, the probability that more than half of the first generation of learners will acquire a

¹⁶We think of the triggering data as 100 Bernoulli trials in which success, i.e. a variant trigger, has probability $1/24$ for each trial. Then by the binomial theorem, the probability that the learner will see 1 or more variants from g^1 (equivalently g^2):

$$\begin{aligned} P(X > 0) &= (1 - P(X)) = 0 \\ &= 1 - \binom{100}{0} \left(\frac{1}{24}\right)^0 \left(1 - \frac{1}{24}\right)^{100} \\ &= 1 - \left(\frac{23}{24}\right)^{100} \\ &= 0.985 \end{aligned}$$

So the probability of seeing both variants 1 or more times is 0.985 multiplied by itself. (See e.g. McColl (1995) for a straightforward introduction to probability theory, Bernoulli experiments, and the binomial theorem.)

¹⁷Notice that an assumption that either appropriate lexical rules are induced or the triggering data is sufficiently lexically uniform is critical to maintenance of the speech community in this case. Otherwise different learning agents exposed to different triggers may acquire incompatible word:category associations for the same nominal and verbal modifiers. That is, one agent may only be exposed to *nicely* as a postmodifier and another as a premodifier during the learning period. As adults these agents would not be able to successfully interact using any sentence containing *nicely*.

¹⁸Analytically characterising the exact behaviour of the model given such scenarios is very complex because of the use of overlapping generations and ‘horizontal’ (learner \rightarrow learner) as well as ‘vertical’ (adult \rightarrow learner) interactions. However, in this fairly clearcut case the lower bound on fixation within 6 cycles is $p \approx 0.994$.

grammar covering the variant is $p > 0.6$, so the positive feedback dynamics have a reasonable chance of getting started. This type of analysis allows us to characterise more precisely the degree of variation that a stable model, maintaining linguistic stasis, will tolerate, given LP and the population dynamics.

To close this section, let us consider what happens if we introduce a sequence of variations into such a speech community. Suppose that every 16 interaction cycles the current adult population is increased by around 1/3 new adult agents who have acquired a grammar with variant trigger(s) exemplified in their output 1/12 of the time. I will call such events migrations and introduce ongoing variation into the model via such ‘population movements’. Each of these successive variant triggers has a good chance of being incorporated into a new grammar acquired by subsequent learners. If the variant grammars are otherwise close, it is very likely that the speech community will be maintained; such migrations model contact between dialects, rather than mutually incomprehensible languages. Nevertheless, language change will be characterised by the acquisition of increasingly large ‘covering grammars’ which successively incorporate more and more variation and allow successively greater interactions between variants originally localised in earlier smaller grammars. This is a profoundly unrealistic dynamic for grammatical change, and probably much lexical change too. It implies that as languages change, grammars will tend to incorporate more and more of the possibilities defined by UG . There is no evidence that grammars grow monotonically in this fashion, incorporating ever increasing variation and optionality (e.g. Lightfoot, 1999:77f). Yet there is abundant evidence that linguistic heterogeneity rather than homogeneity is the norm during language acquisition.

2 Data-Selective Learning Procedures

A fundamental principle of change and variation is that most variation is mutually exclusive; that is, grammars or linguistic variants tend to compete rather than coexist as optional variants (e.g. Kroch, 1989; Lightfoot, 1999:92f). A grammar learning procedure which selects between variants rather than acquiring them all (at least by default) will model this aspect of language change better. There are several ways in which agents could be modified to yield data-selective learners. Firstly, UG could be parameterised so that the sets of categories defining full grammars was disjoint. For example, at the moment there is nothing to prevent a learning agent acquiring both categories, $(S/NP_s)/NP_o$ and $(S\backslash NP_s)/NP_o$ either for overlapping or disjoint sets of transitive verbs, thus acquiring I-languages with mixed VOS or SVO clause orders. However, if the set of categories available is parameterised so that once the order of subjects is determined for one (verbal) category all other (verbal) categories must conform to this ordering, then UG will not include grammars with mixed subject ordering in canonical clauses, and thus LP will not be able to acquire them. Secondly, LP could be modified so that there is data-driven competition between variant categories. For example, LP might count the instances in which variant categories (i.e. ones generating the same LF up to lexical variation) occur, and G might select the current most frequent one for trigger generation.

These alternatives would lead to different outcomes assuming once again the scenario of section 1.3. Given a UG which parameterised cross-categorially for head-initial/final modification, then a learning agent could acquire g^1 or g^2 , but not g^3 .¹⁹ If we assume that LP sets such parameters deterministically never altering a parameter once it has been set via triggering data (e.g. Briscoe, 1997), then the first variant trigger a learning agent is exposed to will determine the relevant part of the category set available thereafter. Thus, learning agents will each select g^1 or g^2 solely on the basis of the specific sequence of triggers they are exposed to. The initial generation each have an equal chance of seeing a postmodifier or premodifier first, so *on average* learners have an equal chance of selecting g^1 or g^2 . However, the actual probability of exactly half the learning agents acquiring g^1 is only $p = 0.246$, so the probability that the first generation of learners will alter the subsequent distribution of triggers in favour of either g^1 or g^2 is $p = 0.754$ (rising to $p = 0.999$ by the 5th generation). Similarly, if, following Gibson and Wexler (1994) and others, we assume that parameters are continuously reset on parse failure (i.e. that the learning procedure is ‘memoryless’ and free to revisit

¹⁹Once we posit parameterisation of UG or, more generally, a data-selective learner, then it may no longer be possible to characterise the E-language of the community in terms of any I-language grammar which can be acquired by an individual learning agent. Lightfoot (1999:81f), for example, argues that ‘social grammars’ are, at best, fictitious linguistic constructs and will probably have to embody very different properties from ‘biological grammars’ generating I-languages. This section amplifies this point and demonstrates how either properties of UG and its parameterisation *or* of LP may make it true. However, ‘social grammars’ might still turn out to be useful constructs for modelling triggering data succinctly.

previous hypotheses), then the dynamics remains identical but the grammar selected is now dependent on the *last* variant trigger seen. We could also posit a different parameterisation in conjunction with either of these models of *LP*, such as one that licensed, say, verbal postmodifiers and nominal premodifiers but not both types of modifier for one type of head, then learning agents could acquire 1 of 4 possible grammars. However, the dynamics of which grammar the community fixated on would remain essentially the same. The model of language change developed by Niyogi and Berwick (1997) and Niyogi (2000) is of the type just outlined. Learners select between variants (i.e. set a parameter) on the basis of the last relevant trigger they see before the learning period ends. This accounts for the (unrealistic) preference for +V2 over -V2 languages in their simulations, because in the *UG* fragment developed by Wexler and Gibson (1994) +V2 languages have more determinate +V2 triggers than -V2 languages have determinate -V2 triggers. Therefore, a learner is more likely to see +V2 triggers, on average, if +V2 grammars are present in the adult population and trigger generation is random.

Learners of this type predict that a speech community will drift randomly between variant grammars until one or other variant reaches fixation. Even though there will be positive feedback favouring any variant which is slightly better exemplified in the learner's data, random effects, such as the particular sequences of triggers seen by individual learners, may override an incipient trend towards one variant. However, at some point, one or other variant will become dominant enough to fixate with virtual certainty. This pattern is well-known and well-studied in population genetics and evolutionary theory as (random) *genetic drift* (e.g. Maynard Smith, 1998:24f). Eventual fixation on one variant is inevitable in finite populations, so we would typically expect to see a longish period of random fluctuations followed by an exponential increase in the frequency of one variant, leading to fixation. Notice that this pattern is very different from S-shaped logistic growth. As far as I know, this pattern of 'grammatical drift' has not been attested.²⁰

The alternative approach of modifying *LP* to track the frequency with which variant categories are exemplified in the learning data will yield different dynamics again. Consider first a non-parameterised learner as defined in section 1.1 with a modified version of *LP* which records the number of times each word:category association hypothesised is used in a successful parse of a trigger. At any given point, the learning agent generates using only the most frequent word:category association when there are variant alternatives generating the same LF. At the end of the learning period, the agent stops counting, so the most frequent word:category associations at that point are the ones which define the acquired grammar. This 'statistical' learning procedure will acquire the 'majority grammar' which incorporates the most frequent variants from potentially multiple source grammars. For the pre/postmodification example, 4 grammars would be available to the learning agent representing the possible combinations of pre/postmodification with nominal/verbal heads. Thus, superficially this approach resembles the weaker parameterisation discussed above. However, the dynamics of language change will now depend much more closely on the frequency with which variants are exemplified across the whole learning period, and thus on the true frequency of the variants in the community. If initially variants are equally represented in the community, then the grammar acquired by a learner will depend on random fluctuations in triggering data. However, as soon as one variant is incorporated into more than half of the grammars, learners will tend more consistently to acquire that variant, because their final hypothesis will be much less sensitive to the particular sequence of triggers they were exposed to. Thus the amount of random drift will decrease and patterns of change should, other things being equal, show less fluctuation and proceed more consistently towards fixation on one variant.

The differences between these dynamics are illustrated graphically in Figure 4 which shows the spread of a single variant grammar through a stable population of 100 agents, 40 of whom are learners (see section 1.2). The y-axis of the left hand graph indicates the proportion of all agents who have acquired the variant; that of the right hand graph, the proportion of learning agents acquiring the variant in the same simulation run. The S-Lrner curve shows the spread of the variant when learning agents use the statistical version of *LP* described above. R-Lrner1 models this spread when *LP* acquires the variant determined by the first (or last) relevant trigger, and R-Lrner2 represents the effect of *LP* selecting a grammar as a probabilistic function of the proportion of agents who have acquired each variant grammar at that point. The simulations were initialised with equal proportions of adult agents with g^1 and g^2 . In each of the three runs selected,

²⁰The notion of drift being employed here is, of course, very different from that introduced into (historical) linguistics by Sapir (1921:150), who used this term to designate stereotypical and thus apparently 'directed' processes, such as loss of case marking leading to increased periphrasis.

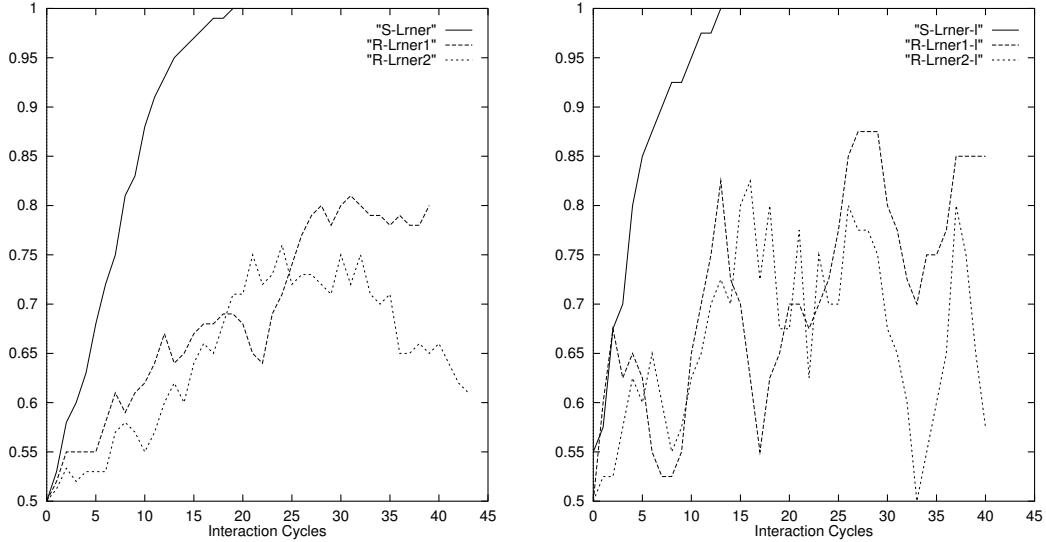


Figure 4: Spread of a Variant

representing the three different variants of *LP*, g^1 spreads at the expense of g^2 and the y-axis starts at 0.5. The statistical version of *LP* shows steep, though still logistic spread of g^1 until all learners are acquiring this variant (by interaction cycle 13 on the right hand graph). Fixation in the adult population is slower (as indicated in the left hand graph) because once all new learners are reliably acquiring g^1 , the remaining g^2 adult population must die out. This slope can be made more gradual by increasing the age of death relative to the critical period or by increasing the proportion of adults to learners. The dynamics of the spread of g^1 with the other versions of *LP* are more random, with downward as well as upward fluctuations and no fixation within the time scale plotted, as predicted by the analysis above. Niyogi and Berwick (1997) argue that logistic spread follows from but is not ‘built into’ the model of language change they develop. Robert Clark (1996) shows that in many situations their model actually produces exponential rather than logistic growth. These microscopic simulations suggest that their framework, and in particular the Trigger Learning Algorithm, will also frequently predict random grammatical drift. The particular derivation of logistic spread proposed here follows from a more realistic modelling of the composition of the population (with overlapping generations and variant proportions of learners to non-learners).²¹

The statistical version of *LP* is quite compatible with (any) parameterisation of *UG*. For example, if we, as before, posit a head-initial/final parameter which selects for pre/postmodifiers cross-categorically, then the learning agent will simply select between 2 rather than 4 grammars on the basis of the frequency with which pre/postmodification is exemplified cross-categorically in the triggering data. In general, parameterisation will yield a more efficient *LP* which explores a smaller hypothesis space and thus more robustly selects a final grammar on the basis of a smaller number of associated cues in triggering data (e.g. Dresher, 1999). Therefore, there are independent reasons for expecting *UG* to be parameterised, apart from data-selectivity. Nevertheless, parameterised or not, a *LP* which selects between grammars using single trigger instances predicts very random dynamics in language change. These have not been observed to my knowledge, and certainly would not lead us to generally expect logistic patterns of change.²²

²¹Whether this is a correct account of any particular case of syntactic change remains to be seen. Kroch (1989) demonstrates that logistic change via competition between grammatical subsystems occurs, but is not explicit about whether this competition is between grammars internalised by single speakers, across speakers, or a mixture of both. This account does not address the bilingual or diglossic scenario. However, diglossia is much easier to account for if learners track parameters of variation rather than utilising a ‘memoryless’ *LP* because it is straightforward to augment the statistical version of *LP* so that the evidence for alternative settings of each parameter is recorded. Simon Kirby (pc) points out that logistic change would also be expected if learners acquired both variants but produced one in preference to the other. Hurford (2000) discusses such ‘production bottlenecks’ in the context of a detailed comparison of the various ways in which language changes have been modelled computationally.

²²A statistical learning procedure of the type outlined can also account for robust acquisition in the face of indeterminacies

3 Learning Procedures with Inductive Bias

So far we have explored, the interaction of population dynamics and LP in determining the dynamics of selection or absorption of linguistic variants exemplified in a speech community, assuming that LP has no preferences concerning the grammars it learns. That is, that within the space defined by UG , there is no inductive bias favouring some grammars over others. Formal learnability work on language acquisition has tended to, at least implicitly, take the position that there is no such bias, by assuming that the starting point for learning is arbitrary or random (e.g. Gold, 1967; Gibson and Wexler, 1994) and by defining learnability in terms of reachability of any $g \in UG$ (given triggers from g) from any such starting point.²³ More substantively-oriented work on parameters has tended to assume that some, perhaps most, parameters will have unmarked or default values (e.g. Chomsky, 1981:7f; Hyams, 1986; Wexler and Manzini, 1987). Such an assumption is a form of inductive bias, or soft constraint, over and above the hard constraints determined by UG , as it creates a preference ordering on the acquisition of $g \in UG$ by LP . For example, if we assume that there is a V2 parameter which has an initial unmarked or default -V2 value, we are saying that -V2 grammars will be ordered higher and considered before +V2 grammars within the space of possible grammars defined by UG . So a learner must be exposed to robust positive evidence to reset this parameter to the non-default or marked value. The unmarked default value, however, will dictate the shape of the acquired grammar in the absence of (robust enough) evidence.

Inductive bias can also be introduced by assuming that parameters of UG are set in a particular order and that the effects of parameters are partly dependent on the settings of others (e.g. Briscoe, 1997, 1998; Dresher, 1999). For example, if the effects of a head-initial/final parameter, having been (re)set, can be partly overridden by the setting of a more specific parameter which determines the order of heads and arguments for a specific category, then LP will first consider ‘harmonic’ head-initial/final grammars, but adopt ‘mixed’ grammars on the basis of positive evidence of specific categorial exceptions.²⁴ Similarly, a statistical approach to the setting of parameters (or choice between variants) can naturally be extended to incorporate inductive bias by positing a Bayesian approach to learning in which the final (posterior) setting (or choice) rests on the interplay between a prior probability of a setting (or choice) and its probability given the triggering data (e.g. Briscoe, 1999). For example, if a head-initial/final parameter has a prior probability of 2/3 for the value head-initial (and thus a prior of 1/3, i.e. 1 - 2/3, for head-final), then we can now quantify precisely the degree of evidence required for LP to reset the parameter from its default value using Bayes theorem. In this case, the probability derived from the triggering data must be >2/3 for the posterior probability to favour the marked setting.²⁵

of parameter expression (e.g. Clark, 1992), such as the ambiguity of SVO triggers given V2 grammars discussed in section 1.1 (see Briscoe, 1999). Deterministic and/or local parameter setting procedures of the type described by Gibson and Wexler (1994), Niyogi and Berwick (1996), Dresher (1999), and Briscoe (1997, 1998) suffer from possible ‘premature’ and irrecoverable convergence to an incorrect grammar in the face of such indeterminacies.

²³In fact, an assumption of inductive bias is the norm rather than the exception in learning theory; for example Cosmides and Tooby (1996) and Staddon (1988) argue that evolution equips organisms with specialised domain-specific learning mechanisms incorporating inductive bias assimilated from the environment of adaptation for the learning mechanism. More statistically-oriented approaches to language learnability also often rely on an assumption of ‘closeness’ or informativeness of prior knowledge in proofs of learnability or high probability convergence (e.g. Muggleton, 1996).

²⁴Briscoe (1997, 1998) details how UG can be modelled using a default inheritance hierarchy describing possible categories of a CG. In this model, the setting of more general cross-categorial parameters, by default, sets more specific parameters, so that the learner predicts harmonic grammars and only revises such predictions if subsequent triggers force this.

²⁵We assume that ‘relevant’ triggers provide positive evidence for head-final or positive evidence for head-initial. Therefore the so-called likelihood probability (derived from the triggering data) for head-final is given by:

$$P_l(X = final) = \frac{f(t_{final})}{f(t_{final}/initial)}$$

where f denotes frequency counts (i.e. the maximum likelihood estimate over the relevant triggers). The posterior probability is given by:

$$P_{po}(X = final | t_n) = \frac{P_{pr}(X=final) P_l(X=final)}{P(t_{final}/initial)}$$

(i.e. by Bayes theorem). The probability of the relevant triggers, $P(t_{final}/initial)$, is a constant normalising factor which can be ignored. We need only compare the unnormalised posterior for head-final to that for head-initial and choose the highest. As

If LP/UG includes inductive bias, then the dynamics of language change will no longer be determined so directly by properties of the triggering data (within the overall space defined by UG). Instead the relative frequency with which a variant is exemplified in triggering data will interact with the bias inherent in LP , as well as, of course, the population dynamics. Returning to the example of pre/postmodification, if we assume, as is natural in CG, that the same UG parameter also selects between head-final/initial modifiers, that head-initial is the default unmarked value, and that this default has a relatively strong prior of 4/5, then a community fixated on g^1 , a premodifier grammar would need to be swamped by migrating agents generating postmodifiers so that the triggering data to the next generation contained $>4/5$ triggers exemplifying postmodification. Thus, the Bayesian approach is one way of making concrete and precise Kiparsky's (1996) proposal that change results from the interplay of triggering data with internal preferences created by LP/UG .

To address the question of how likely it is that the language faculty incorporates inductive bias, the population model introduced in section 1.2 can be modified to incorporate natural selection for language agents by using the CSR as a measure of fitness and making reproductive success relative to fitness. For this to be meaningful there must be variation in the language faculty with which language agents are initially endowed. In Briscoe (1997, 1998) I describe in detail an encoding of UG which allows the 'starting point' for learning to be varied, in terms of what counts as a principle or parameter and which parameters have default or unset values. Mutation and crossover operators are defined over this encoding so that the creation of new agents can both introduce and spread (the analogue of) genetic variants. Thus, we modify the basic population model with a more evolutionarily realistic version of reproduction (see Figure 3 above for comparison):

REPRO: Create-LAgt(Mutate(Crossover(LAgt_i, LAgt_j))), $i \neq j$, CSR_i \wedge CSR_j > Mean-CSR

The crossover operator combines the encoding of UG randomly from the two higher than average fitness 'parent' agents, while the mutate operator with equal and low probability converts a parameter to a principle, or vice-versa, flips the value of a default parameter, or converts a default to an unset parameter. Once such genetic variation is introduced, it is essential to posit natural selection for language agents on the basis of their CSR, otherwise populations inevitably cease to learn and communicate after a few hundred cycles, as random variation in language faculties builds up. However, with natural selection, the population evolves to a (local) optimum in the predefined space of UG s, determined in part by the range of grammars/languages sampled before they cluster around such an optimum. For example, if the languages sampled during a significant period during adaptation are OV, then the population is likely to genetically assimilate an unmarked OV parameter setting, since this speeds up learning.

This is an example of genetic assimilation (e.g. Waddington, 1942), or the so-called Baldwin Effect (Baldwin, 1896), in which changes in a species' behaviour (the advent of language) create new selection pressures (the need to learn language efficiently).²⁶ When language change is as rapid as is consonant with maintenance of a speech community during adaptation, genetic assimilation results in around twice as many soft constraints (inductive bias) than hard constraints (incorporation of principles into UG); as the former have a less fatal effect on successful learning if subsequent change renders an assimilated principle incorrect. Additionally, genetic assimilation slows and ultimately ceases once a near optimal balance between learning speed and the speed of linguistic change has been found; in this model, when around half of the possible parameters in the predefined space of UG s have been converted to principles or default-valued parameters. Without continuous

the prior, P_{pr} , is multiplied by the likelihood, P_l , the unnormalised posterior for head-final will only exceed that for head-initial (when $P_{pr}(X = final) = 1/3$) if the associated likelihood probability is $>2/3$. For example, if 5 out of 6 relevant triggers are head-final, then the two unnormalised posteriors are computed as follows:

$$P_{po}(X = final \mid t_n) = \frac{1}{3} \times \frac{5}{6} = \frac{10}{36}$$

$$P_{po}(X = initial \mid t_n) = \frac{2}{3} \times \frac{1}{6} = \frac{4}{36}$$

Briscoe (1999, 2000a) develops a Bayesian approach to parameter setting in greater detail. Sivia (1996) provides a good introduction to Bayesian data analysis.

²⁶Waddington (1975) and Pinker and Bloom (1990) both propose that genetic assimilation played a role in the formation of the language faculty. Kirby and Hurford (1997) and Briscoe (1997, 1998, 2000b) develop detailed models of this process and provide a more thorough discussion of the effect and its likely relevance.

linguistic change, there is no reason why the process of genetic assimilation should not proceed until the need for (grammatical) learning is eradicated (*pace* Pinker and Bloom, 1990). However, with constant linguistic change, too much constraint on learning becomes maladaptive.

These simulations cannot, of course, *prove* that the language faculty incorporates inductive bias via genetic assimilation, but they do help clarify the (prehistoric) conditions under which this would have been likely. In particular, the relative speed of language change compared to biological evolution and the size of the relevant population during the period of adaptation are unlikely to have been critical factors (*pace*, Deacon, 1997:328f), and the possibility that the language faculty arose *de novo* via a single macromutation or exaption (e.g. Berwick, 1998; Bickerton, 1998) does not affect the argument because organs which arise via exaptions of spandrels or by macromutations are still susceptible to *subsequent* modification by natural selection (e.g. Ridley, 1990; Lieberman, 1991, Kirby, 1998). Indeed in these simulations, language agents do begin with language faculties (either random or converged) and these are merely refined by subsequent evolution. However, the current model does assume a one:one correlation between changes to the genetic encoding and phenotypic changes to *LP/UG*. Such a close correlation would be unlikely in nature, and has the effect of speeding up genetic assimilation effects (e.g. Mayley, 1996). Until we know more about the genetic encoding and neurological basis of the language faculty, it is impossible to quantify how long it might actually have taken for assimilation to occur.

If we accept that there is inductive bias in language acquisition, as I think we should, given both the evidence from the simulation work and from empirical studies of language acquisition (e.g. Wanner and Gleitman, 1980) and general considerations of learning theory (e.g. Cosmides and Tooby, 1996), how does this affect our view of language change?²⁷

4 Languages as Adaptive Systems

So far I have argued that (E-)languages are dynamical systems, the aggregate output of a set of grammars which change over time as the membership of a speech community changes. However, if grammar learning is data-selective and biased, then languages are better seen as *adaptive* systems which will inevitably evolve to fit their unique ecological niche – the human language faculty and wider cognitive system. Under this view, languages are evolving on a *historical* rather than biological timescale, and the primary source of *linguistic selection* is the language acquisition ‘bottleneck’ through which successful grammatical variants must pass repeatedly with each new generation of language learners. Evolution and selection are not being used here metaphorically, but in their technical ‘universal Darwinist’ sense (e.g. Cziko, 1995; Dawkins, 1983; Dennett, 1995:343) of (random) variation, selection amongst variants, and thus differential inheritance. These terms are potentially applicable to any dynamical system, whether that system is ‘implemented’ in biological organisms, silicon, or cognitive linguistic representations.²⁸

Returning to the example of pre/post-modification and the Bayesian parameter setting learner outlined in the previous section, assume again that *LP/UG* incorporates a cross-categorial head-initial/final parameter with a bias for head-initial and also category-specific parameters which, by default, pick up their value from the more general cross-categorial parameter, but can be reset by positive evidence in triggering data. This creates a bias for harmonic head-initial grammars. Change in (E-)languages can now be seen in terms of selection amongst grammatical variants which are more or less ‘natural’ with respect to *LP/UG*. There will clearly be an asymmetry in terms of the dynamics of change as non-harmonic and head-final grammatical variants will need to be exemplified in the triggering data frequently enough to overcome the bias in *LP* and to force the resetting of more specific parameters. Thus, the competition between grammatical variants is

²⁷Lightfoot (e.g. 1999:165f) criticises some specific arguments for inductive bias as an explanation of grammatical change and also makes the general point that if a bias explains how a variant was eradicated, we also need an account of how the variant can arise. I address the more general point in section 5. The detail of the specific arguments are beyond the scope of this paper, but I also agree with Lightfoot (1999:218f) that if there is bias, this must interact with triggering data in explaining change. The Bayesian model of *LP* allows us to precisely quantify the interplay of data and bias.

²⁸As far as I am aware, Hurford (1987) is the earliest expression of this view from a syntactician, though the argument is presented much more clearly in Hurford (1999) and Kirby (1998, 1999). Lindblom (1998) and colleagues have developed a similar approach to the evolution of phonological systems. Others who have adopted similar viewpoints include Batali (1998), Deacon (1997), and Steels (1998).

biased. Assuming, as in section 3, that the prior bias for head-initial is 4/5 and this bias is inherited by all the more specific parameters, yielding harmonic head-initial settings, then a community fixated on a (partially) disharmonic or head-final grammar can begin shifting to a more harmonic head-initial one if variant triggers are exemplified with $>1/5$ likelihood probability to any given learner. That is, there is differential linguistic selection in favour of head-initial harmonic grammars as a consequence of *LP/UG*, so (E-)languages are highly likely to adapt to *LP/UG* over time, given some continuing source of (random) variation.

There is no requirement that the bias, or process of change, is ‘functional’ in any deeper sense. For example, if *LP/UG* incorporates such a bias, this may rest on nothing more fundamental than the fact that the languages sampled during the period of adaptation for the *language faculty* were mostly harmonic and head-initial (see section 3). Alternatively, ‘harmonic’ languages may have been preferred initially because they facilitate language processing (see section 5 below), so the bias assimilated into *LP/UG* would rest ultimately not (entirely) on (pre)historical accident but on wider aspects of human cognitive capacities and limitations. Either way, (E-)languages can be said to have adapted to their niche via linguistic selection.

Superficially, characterising (E-)language changes as an adaptive evolutionary process may not seem to add much to the view that languages change or, more specifically, that languages can be modelled as nonlinear dynamical systems. However, it does commit us to the claim that language change is a process of differential selection amongst variants, so it rules out scenarios in which language acquisition is not data-selective and biased, and also ones which involve a significant amount of ‘invention’; that is, going beyond the data. Perhaps, more importantly, modern population genetics provides a battery of tools to analyse the situations under which a variant with a small selective advantage manifested by a single individual, or small minority of individuals, can spread through a population. These tools can be adapted to the study of linguistic change straightforwardly once we have precise enough theories of language acquisition and processing. In particular, the move from a random drift, ‘neutral’ theory of change to an adaptive account may be a necessary prerequisite to an understanding of how some changes can spread from very small beginnings.²⁹

Creolisation has been characterised by Bickerton (1984:173) as a process of ‘invention’ in terms of the learner’s innate bioprogram; first language learners exposed exclusively to an impoverished pidgin subset language acquire a superset creole grammar. If this view is correct, then it would undermine the claim that significant language change can be modelled as an evolutionary process. However, Lightfoot (1992:178f; 1999:167f), while accepting the abruptness of creolisation, challenges the idea that it requires a special account of language acquisition, arguing that properties of the pidgin triggering data lead to reanalysis and consequent parameter resetting across generations. Roberts (1998) argues that some features of Hawaiian creole took two generations to emerge, which also supports the hypothesis that creolisation is very fast language change, rather than the result of a special process of consistent invention by each first language learner.

I have simulated creolisation, especially the situation in Hawaii, in as much linguistic and demographic detail as is practical, given the limits of the model of *LP/UG* developed in Briscoe (1999) and what is known about the demographic factors. Creoles emerge when first language learners are born to a diverse community of indentured workers or slaves with an impoverished pidgin as their *lingua franca*. In Hawaii the proportion of such learners by the end of the first generation (i.e. 20 years) constituted about 35% of this community and was increasing throughout this period. In other cases of plantation creolisation this proportion may have been lower, but was probably increasing throughout the early (overlapping) generations of first language learners.³⁰

The demographic situation is modelled by introducing 6 new learning agents per interaction cycle into a community of 64 adult agents who are not changed during the simulation run. Agents learn for 4 interaction

²⁹ Kirby (1997, 1999) is one example of a detailed and carefully worked out account of linguistic change based on adaptation to language acquisition.

³⁰The estimate of 35% is based on census figures for 1890, 1900 and 1910 kindly supplied by Derek Bickerton. These are incomplete in some areas but indicate an under 15 population of at least 20% by 1900 and 35% by 1910, assuming a similarly high birthrate amongst Hawaiians as amongst Portuguese immigrants (44% by 1910). (This assumption is in turn supported by school attendance records of 5-14 year olds.) Bickerton suggests that creoles emerge when the proportion of under 12s was between 15% and 25%. The speed of spread of the creole through the learner and total population will both be increased if the proportion of learners is greater and the increase in this proportion is steeper. Therefore, it is possible that (features of) Hawaiian creole spread more rapidly (within two generations, according to Roberts, 1998) because the birthrate was high. Only further demographic and linguistic work on other pidgin-creole transitions will tell, but the model predicts such speed differences.

cycles and can reproduce thereafter. The proportion of learners reaches a maximum of 28% at the 6th interaction cycle and then tails off gradually to a stable 15%. We are interested in the proportion of learners converging to a creole grammar across interaction cycles, bearing in mind that 4 interaction cycles is equivalent to one generation (i.e. 20 years).

The statistical version of *LP*, outlined in section 3, was integrated with a version of *UG* which contains 15 parameters which determine the number and type of CG categories in an acquired grammar. These categories can be divided into those that determine ordering of arguments to functor categories and those which determine the availability of specific functor categories. For example, one parameter controls the availability of a functor category which licences relative clause modifiers of nominal heads. However, whether the wh-element precedes or follows the relative clause and whether this is pre/postnominal depends on potentially interdependent but distinct ordering parameters. The bias assumed in *LP* is based on Bickerton's (1984) account of Saramaccan, as the prototypical creole grammar, and models a preference for simple SVO right-branching grammars.

Such learners, exposed exclusively to pidgin data, modelled as clauses with a verb and single word subjects and objects appearing in random orders, tend to acquire a grammar generating a SVO subset language with similar clauses. That is, they do not generalise the pidgin data and invent a creole superset grammar but they do converge to SVO order. Furthermore, by around the 6th interaction cycle (i.e. within two generations) all learners in these simulation runs are converging to SVO grammars. This result can be understood in terms of the interplay of the triggering data and *LP* in conjunction with the increase in SVO triggers as the population of learners grows. Faced with conflicting and equivocal evidence for basic constituent order, *LP* will tend to set parameters in terms of prior biases. However, trigger sampling variation will mean that some early learners may not acquire a SVO grammar, or may hypothesise non-SVO grammars at intermediate stages in acquisition. Nevertheless, if the majority of learners do (eventually) acquire SVO grammars, then the incidence of SVO triggers will increase causing the familiar positive feedback dynamics to take over. Thus, the inductive bias for SVO is enough to kick the system in the right direction, but it is only as the population gains SVO speakers that the chances of learners acquiring non-SVO grammars declines to negligible levels. Thus, this account is purely selectionist and predicts that the birthrate will be a factor in the speed of creolisation, but as yet does not explain how first language learners can acquire a superset creole grammar.

Bickerton (1981, 1984, 1988) has argued that superstratum and substratum languages play no role in the acquisition of the creole. The evidence for this comes from the lack of a consistent grammatical relationship between the creole and these potential sources, as well as the similarities of unrelated creoles (e.g. Roberts, 1998). Bickerton (e.g. 1984:182f) recognises that in many cases, including Hawaii, learners would be exposed to a small proportion of superstratum and substratum utterances but downplays their role as triggers on the basis of their relative infrequency and, in the case of substratum utterances, mutual inconsistency. Nevertheless, if the triggering data to the statistical *LP/UG* model is enhanced to include a small proportion of more complex superstratum triggers, with or without a further small proportion of random substratum language triggers, exemplifying multiword phrases and subordination, learners converge to and fixate on a SVO right-branching superset grammar with essentially the same dynamic and timecourse as that discussed above.

Briscoe (2000a) discusses these simulations and their interpretation in further detail. While they are by no means conclusive and rest on rather sketchy demographic information and consequent assumptions about the nature of the triggering input, they nevertheless add weight to the argument that creolisation should be seen as a case of rapid language change caused by the interaction of ordinary language acquisition with radical demographic changes. Though creolisation poses the most obvious challenge to a selectionist account of language change in terms of inductive bias in language acquisition, there are others difficulties which require that we adopt a more sophisticated model of the evolutionary process which incorporates conflicting and interacting selection pressures on language change stemming from wider cognitive capacities and limitations.

5 Languages as Complex Adaptive Systems

If languages adapt solely to innate preferences during language acquisition, we would expect the history of languages to show nearly inexorable development towards an optimal or most natural grammar. Presumably, we would expect all speech communities to fixate ultimately on creole-like grammars if these represent the most natural optimal or default solutions with respect to *LP*. Chance factors might temporarily move a language away from such a solution but, over time, constant and universal selection pressure should (re)assert itself. While supporters of grammaticalisation have argued for a (prototypical) unidirectionality in change (e.g. Hopper and Traugott, 1990:94f) and there is evidence of skewing in the distribution of attested languages with respect to the range of possible grammars most theories of *UG* license (e.g. Hawkins, 1994), no one would argue that language change is this deterministic.

Lightfoot (1999:213f) uses similar arguments to criticise accounts of language change which rely on innate preferences or inductive bias, asking how the relevant variation might have arisen in the first place if bias is the explanation for the selection of a particular variant. He, in fact, goes on to endorse models, such as that of Kiparsky (1996), which explain change in terms of the interplay of acquisition preferences and changes in triggering data (as does the Bayesian version of *LP* outlined in section 3). Surprisingly, he also accepts the argument of Niyogi and Berwick (1997) that historical tendencies in language change may result from dynamic trajectories caused by learners misconverging. However, Kiparsky's position and that of Niyogi and Berwick are quite different. As we noted in section 1.2, Niyogi and Berwick explore a model in which change is inevitable even without initial 'external' perturbation of triggering data. Given a speech community fixated on a -V2 grammar some learners will misconverge initiating a process of change. This is essentially because the Trigger Learning Algorithm will not converge, or will converge with very low probability, given a random starting point and the theory of *UG* they consider (Niyogi and Berwick, 1996). Thus, the model is unstable and predicts inexorable change towards some fixed point. Robert Clark (1996) illustrates the instability of their model very clearly in a series of replicated simulations which explore the behaviour of the model when learners are given between 8 and 256 triggers. This account, then, is really an extreme version of the accounts of language change being driven by internal preferences (or failings) during acquisition. However, even accounts which predict stability under conditions of homogeneity, and require changes in triggering data as well as acquisition preferences in order for change to occur, still require some account of how apparently suboptimal variants ever arise in speech communities.

The idea that there are competing motivations or conflicting pressures deriving from the exigencies of production, comprehension and acquisition has been developed by linguists working from many different perspectives (e.g. Langacker, 1977; Fodor, 1981; Croft, 1990:192f). In linguistics little progress has been made in quantifying these pressures or exploring their interaction, except in the area of phonology where Lindblom (e.g. 1998) has adopted a similar evolutionary model to that advocated here. 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 a fitness landscape which itself inevitably changes. Conflicting selection pressures will cause the fitness 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 (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). A simple linguistic example is given in Briscoe (1998). There are (as in section 4) 17 parameters to be set during learning, and only setting a subset of these parameters yields a subset language. Given some degree of linguistic variation, the speech community will tend to fixate on a subset language, optimising learnability at the expense of expressivity. Regardless of whether there is natural selection for agents on the basis of communicative success, optimising learnability at the expense of expressivity is highly likely. To counteract such a tendency, we can either posit that *LP* requires all parameters to be set (somehow), or that there is a counteracting pressure for expressivity. This could be introduced into the model by positing some range of logical forms that must be realisable (possibly 'economically') and penalising agents' communicative success whenever their current grammar does not allow this.

Introducing such a conflicting or competing pressure (suitably weighted) prevents unconstrained optimisation for learnability. Now variants which are adaptive must improve learnability and maintain expressivity (or vice versa).

Another and perhaps better understood pressure on the evolution of grammatical systems derives from parsability (e.g. Gibson, 1998; Hawkins, 1994; Miller and Chomsky, 1963; Rambow and Joshi, 1994). A number of metrics of the relative parsability of different constructions have been proposed, both as accounts of the relative psychological complexity of sentence processing and of the relative prevalence of different construction types in attested languages. A metric of this type can be incorporated into an evolutionary linguistic model in a number of ways. Kirby (1999) argues, for example, that parsability equates to learnability, as triggers must be parsed before they can be used by a learner to acquire a grammar. By contrast, Hawkins (1994:83f) argues that parsability may influence generation so that more parseable variants will be used more frequently than less parseable ones (within the space of possibilities defined by a given grammar), and presents evidence concerning the relative frequency of constructions from several languages in support of this position. This would entail that less parseable constructions would be less frequent in potential triggering data, in any case. Briscoe (1998) demonstrates that either approach, alone or in combination, can account for linguistic selection in favour of more parseable variants.

One type of common non-argument against the view that languages are adaptive is that languages exhibit dysfunctional or maladaptive properties. For example, many languages peripherally exhibit multiple centre- and self-embedding constructions (e.g. de Roeck *et al.*, 1982; Hudson, 1995). Yet such constructions are known to cause parsing problems (e.g. Miller and Chomsky, 1963; Gibson, 1998). If languages are (complex) adaptive systems why have such constructions survived? It is easy to devise models in which their survival is a mystery. For example, Kirby's (1999) equation of parsability with learnability predicts that relatively less parseable constructions will be less likely to function as triggers. As, his learning model does not involve predictive generalisation, it is inevitable that a less parseable construction will be replaced by a more parseable variant. However, a more complex and adequate model of grammar learning, such as that of Kirby (1999), may make very different predictions. For example, if learning involves setting parameters on the basis of degree-0 triggers (e.g. Lightfoot, 1992), then embedded constructions will be learnt indirectly as a predictive consequence of simpler triggers, and thus will continue to be 'inherited' by subsequent generations if the interaction of parameter settings generates centre- and/or self-embedded constructions, even though their acceptability in the arena of language use may well be marginal in view of the desire for communicative success. The models of *LP/UG* discussed in this paper are all potentially of this type since the interaction of functor categories acquired from degree-0 or degree-1 triggers will generate such constructions. Putative examples of dysfunctional or maladaptive features are not in themselves counter-evidence to the view of (E-)languages as complex adaptive systems, anymore than unanalysed strings are evidence for or against a specific syntactic theory.

Once we recognise that there are 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 fitness landscape. For instance, a canonical SOV grammar might evolve increasingly frequent extraposition because SOV clauses with long or 'heavy' object phrases are relatively unparsable (e.g. Hawkins, 1994:196f). However, SVO grammars will be less likely to do so since long object phrases will mostly occur postverbally anyway and will not create analogous parsing problems. Once such a change has spread, it may in turn create further parsability (or expressivity or learnability) issues, altering the fitness landscape; for example, by creating greater structural ambiguity, resulting perhaps in evolution of obligatory extraposition. (It is this locality or blindness in the search for good solutions that makes the evolutionary process more like tinkering than engineering.) In the framework advocated here, we can recognise that such historical pathways can be stereotypical responses to similar pressures arising in unrelated languages, in much the same way that eyes and wings have evolved independently in different lineages many times, without the need to posit a substantive theory of such changes or see them as deterministic (see e.g. Lightfoot, 1999:261f).³¹

³¹Lightfoot (1999:239f) argues that *UG* may be maladaptive because it incorporates a constraint against extraction of subjects from tensed clauses, while speakers / languages have developed idiosyncratic means to circumvent the constraint to fulfill expressive needs. He concludes: 'if maladaptive elements evolve, then we need something other than natural selection to drive evolutionary developments' (1999:248). However, as he points out, the constraint against extraction of subjects from tensed

6 Conclusion

Models of nonlinear dynamical systems can play several useful roles in the study of language change: to characterise population dynamics during a change, to characterise the dynamics of the linguistic change within the population, and to characterise the interaction of the population and linguistic dynamics. For instance, it is clear that creolisation is a type of rapid linguistic change which is also dependent on radical demographic changes. The model developed here allows characterisation of the linguistic change in the context of the demographic changes. The speed of linguistic change (i.e. the spread of the variant through the relevant part of the speech community) can then be seen to be faster in this case than in the case of a more stable population. This perspective supports an account of creolisation as rapid but otherwise normal selective linguistic change.

Characterising (E-)languages as adaptive systems undergoing differential linguistic selection enriches and constrains their modelling as nonlinear dynamical systems. Adaptive systems are a subset of possible dynamical systems, so this step is only justified if we can demonstrate clear selection pressure as opposed to (random) drift. Logistic growth is predicted by selection rather than by random drift. Adaptive systems which change on the basis of interactions between conflicting selection pressures in unpredictable ways, involving positive or negative feedback, with no centralised control are increasingly termed *complex* adaptive systems. Viewing (E-)languages as complex adaptive systems promises new insights into old issues, such as prototypical unidirectionality, competing motivations, or internally-motivated change, ones which do not involve teleology or a substantive theory of linguistic change *per se*.

Serious exploration of this framework is only just beginning because the requisite understanding of dynamic systems and of the tools to study them are very recent developments. It would be unfortunate if old prejudices, or misunderstanding of modern evolutionary theory, precluded the full and proper exploration of languages as complex adaptive systems.

Acknowledgements

I would like to thank the DIGS-5 audience for insightful feedback which helped shape this paper as well as Jim Hurford, Simon Kirby, Susan Pintzuk, George Tsoulas and Anthony Warner for very helpful comments on earlier drafts of this paper. The remaining errors, infelicities and opinions are entirely my responsibility.

References

- Baldwin, J.M. (1896) ‘A new factor in evolution’, *American Naturalist*, vol.30, 441–451.
- Batali, J. (1998) ‘Computational simulations of the emergence of grammar’ in Hurford, J., Studdert-Kennedy, M., and Knight, C. (ed.), *Approaches to the Evolution of Language*, Cambridge University Press, Cambridge, pp. 405–426.
- Berwick, R. (1998) ‘Language evolution and the minimalist program: the origins of syntax’ in Hurford, J., Studdert-Kennedy, M., and Knight, C. (ed.), *Approaches to the Evolution of Language*, Cambridge University Press, Cambridge, pp. 320–340.
- Bickerton, D. (1981) *Roots of Language*, Karoma, Ann Arbor.
- Bickerton, D. (1984) ‘The language bioprogram hypothesis’, *The Behavioral and Brain Sciences*, vol.7.2, 173–222.
- Bickerton, D. (1988) ‘Creole languages and the bioprogram’ in F. Newmeyer, vol. 2 (ed.), *Linguistics: The Cambridge Survey*, Cambridge University Press, Cambridge, pp. 267–284.

clauses may be adaptive with respect to parsability, reducing ambiguity over the location of traces. If it is maladaptive with respect to expressivity, this is only an argument against a simplistic ‘one-dimensional optimisation’ view of evolution. Given the framework developed here, it is quite possible that such a constraint evolved in one or more languages (and either was or was not ultimately assimilated into *UG*) as a local step in evolutionary space, which in turn created expressivity problems requiring further local adaptations of a possibly idiosyncratic and language specific nature.

- Bickerton, D. (1998) 'Catastrophic evolution: the case for a single step from protolanguage to full human language' in Hurford, J., Studdert-Kennedy, M., and Knight, C. (ed.), *Approaches to the Evolution of Language*, Cambridge University Press, Cambridge, pp. 341–358.
- Briscoe, E.J. (1997) 'Co-evolution of language and of the language acquisition device', *Proceedings of the 35th Meeting of Assoc. for Computational Linguistics*, Morgan Kaufmann, Palo Alto, Ca., pp. 418–427.
- Briscoe, E.J. (1998) 'Language as a complex adaptive system: co-evolution of language and of the language acquisition device' in Coppen, P., van Halteren, H. and Teunissen, L. (ed.), *8th Meeting of Computational Linguistics in the Netherlands*, Rodopi, Amsterdam, pp. 3–40.
- Briscoe, E.J. (1999) 'The acquisition of grammar in an evolving population of language agents' in Electronic Transactions in Artificial Intelligence. Special Issue: Machine Intelligence, 16 (ed.), *Muggleton, S.*, <http://www.etaij.org/>.
- Briscoe, E.J. (2000a, in press) 'Grammatical acquisition and linguistic selection' in Briscoe, E.J. (ed.), *Language Acquisition and Linguistic Evolution: Formal and Computational Approaches*, Cambridge University Press, Cambridge.
- Briscoe, E.J. (2000b, in press) 'Grammatical acquisition: inductive bias and coevolution of language and the language acquisition device', *Language*, vol. 76.
- Briscoe, E.J. and Copestake, A.A. (1999, in press) 'Lexical rules in constraint-based grammar', *Computational Linguistics*, vol. 25.4.
- Chomsky, N. (1965) *Aspects of the Theory of Syntax*, MIT Press, Cambridge, Ma..
- Chomsky, N. (1981) *Government and Binding*, Foris, Dordrecht.
- Clark, R. (1992) 'The selection of syntactic knowledge', *Language Acquisition*, vol. 2.2, 83–149.
- Clark, R.A.J. (1996) *Internal and External Factors Affecting language Change: A Computational Model*, MSc Dissertation, University of Edinburgh.
- Cosmides, L. and Tooby, J. (1996) 'Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgement under uncertainty', *Cognition*, vol. 58, 1–73.
- Croft, W. (1990) *Typology and Universals*, Cambridge University Press, Cambridge.
- Cziko, G. (1995) *Without Miracles: Universal Selection Theory and the Second Darwinian Revolution*, MIT Press, Cambridge, Ma..
- Dawkins, R. (1983) 'Universal Darwinism' in D.S. Bendall (ed.), *Evolution: From Molecules to Men*, Cambridge University Press, Cambridge, pp. 403–425.
- Dawkins, R. (1989) *The Selfish Gene*, Oxford University Press, Oxford, 2nd edition.
- Deacon, T. (1997) *The Symbolic Species: Coevolution of Language and Brain*, MIT Press, Cambridge, Ma..
- Dennett, D. (1995) *Darwin's Dangerous Idea: Evolution and the Meanings of Life*, Simon and Schuster, New York.
- De Roeck, A., Johnson, R., King, M., Rosner, M. and Sampson, G. (1982) 'A myth about center-embedding', *Lingua*, vol. 58, 327–340.
- Dowty, D., Wall, P. and Peters, S. (1981) *An Introduction to Montague Semantics*, Reidel, Dordrecht.
- Dresher, E.B. (1999) 'Charting the learning path: cues to parameter setting', *Linguistic Inquiry*, vol. 30.1, 27–67.
- Fodor, J.D. (1981) 'Does performance shape competence?', *Phil. Trans. Royal Society of London*, vol. B295, 285–295.
- Fodor, J.D. (1998) 'Unambiguous Triggers', *Linguistic Inquiry*, vol. 29.1, 1–36.
- Frank, R. and Kipur, S (1996) 'On the use of triggers in parameter setting', *Linguistic Inquiry*, vol. 27.4, 623–660.
- Gibson, E. (1998) 'Linguistic complexity: locality of syntactic dependencies', *Cognition*, vol. 68, 1–76.
- Gibson, E. and Wexler, K. (1994) 'Triggers', *Linguistic Inquiry*, vol. 25.3, 407–454.
- Gold, E.M. (1967) 'Language identification in the limit', *Information and Control*, vol. 10, 447–474.
- Hawkins, J.A. (1994) *A Performance Theory of Order and Constituency*, Cambridge University Press, Cambridge.
- Hopper, P. and Traugott, E. (1990) *Grammaticalization*, Cambridge University Press, Cambridge.
- Hudson, R. (1995) *Measuring syntactic complexity*, Ms., University College London.
- Hurford, J. (1987) *Language and Number*, Blackwell, Oxford.
- Hurford, J. (1999) 'The evolution of language and languages' in Dunbar, R., Knight, C., and Power, C. (ed.), *The Evolution of Culture*, Edinburgh University Press, pp. 173–193.

- Hurford, J. (2000, in press) 'Expression/induction models of language evolution: dimensions and issues' in Briscoe, E.J. (ed.), *Language Acquisition and Linguistic Evolution: Formal and Computational Approaches*, Cambridge University Press, Cambridge.
- Hyams, N. (1986) *Language Acquisition and the Theory of Parameters*, Reidel, Dordrecht.
- Kauffman, S. (1993) *The Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press, New York.
- Keller, R. (1994) *On Language Change: The Invisible Hand in Language*, Routledge, London.
- Kiparsky, P. (1996) 'The shift to head-initial VP in Germanic' in Thrainsson, H., Epstein, S. and Peters, S. (ed.), *Studies in Comparative Germanic Syntax, Vol II*, Kluwer, Dordrecht.
- Kirby, S. (1997) 'Competing motivations and emergence: explaining implicational hierarchies', *Language Typology*, vol. 1, 5–32.
- Kirby, S. (1998) 'Fitness and the selective adaptation of language' in Hurford, J., Studdert-Kennedy, M., and Knight, C. (ed.), *Approaches to the Evolution of Language*, Cambridge University Press, Cambridge, pp. 359–383.
- Kirby, S. (1999) *Function, Selection and Innateness: The Emergence of Language Universals*, Oxford University Press, Oxford.
- Kirby, S. (2000, in press) 'Learning, bottlenecks and the evolution of recursive syntax' in Briscoe, E.J. (ed.), *Language Acquisition and Linguistic Evolution: Formal and Computational Approaches*, Cambridge University Press, Cambridge.
- Kirby, S. and Hurford, J. (1997) 'Learning, culture and evolution in the origin of linguistic constraints' in Husbands, P. and Harvey, I. (ed.), *4th European Conference on Artificial Life*, MIT Press, Cambridge, MA., pp. 493–502.
- Kroch, A. (1989) 'Reflexes of grammar in patterns of language change', *Language Variation and Change*, vol. 1, 199–244.
- Langacker, R. (1977) 'Syntactic reanalysis' in Mechanisms of Syntactic Change (ed.), *Li, C.*, University of Texas Press, Austin, pp. 57–139.
- Lieberman, P. (1991) 'Preadaptation, natural selection and function', *Language and Communication*, vol. 11.1, 67–69.
- Lightfoot, D. (1979) *Principles of Diachronic Syntax*, Cambridge University Press, Cambridge, UK.
- Lightfoot, D. (1992) *How to Set Parameters: Arguments from language Change*, MIT Press, Cambridge, Ma..
- Lightfoot, D. (1999) *The Development of Language: Acquisition, Change, and Evolution*, Blackwell, Oxford.
- Lindblom, B. (1998) 'Systemic constraints and adaptive change in the formation of sound structure' in Hurford, J., Studdert-Kennedy, M., and Knight, C. (ed.), *Approaches to the Evolution of Language*, Cambridge University Press, Cambridge, pp. 242–264.
- Mayley, G. (1996) 'Landscapes, learning costs and genetic assimilation' in Turney, P., Whitley, D., and Anderson, R. (ed.), *Evolution, Learning and Instinct: 100 Years of the Baldwin Effect*, MIT Press, Cambridge, Ma..
- Maynard-Smith, J. (1998) *Evolutionary Genetics*, Oxford University Press, Oxford, 2nd ed..
- McColl, J.H. (1995) *Probability*, Edward Arnold, London.
- McMahon, A. (1994) *Understanding Language Change*, Cambridge University Press, Cambridge.
- Miller, G.A. and Chomsky, N. (1963) 'Finitary models of language users' in Handbook of Mathematical Psychology, vol II (ed.), Luce, R.D., Bush, R.R. and Galanter, E., Wiley, New York, pp. 419–491.
- Muggleton, S. (1996) 'Learning from positive data', *Proceedings of the 6th Inductive Logic Programming Workshop*, Stockholm.
- Newport, E. (1977) 'Motherese: the speech of mothers to young children' in Castellan, N., Pisoni, D. and Potts, G. (ed.), *Cognitive Theory, Vol II*, Lawrence Erlbaum, Hillsdale, NJ, pp. 177–217.
- Niyogi, P. (1996) *The Informational Complexity of Learning from Examples*, MIT AI Lab TR-1587.
- Niyogi, P. (2000, in press) 'Theories of Cultural Change and their Application to Language Evolution' in Briscoe, E.J. (ed.), *Language Acquisition and Linguistic Evolution: Formal and Computational Approaches*, Cambridge University Press, Cambridge.
- Niyogi, P. and Berwick, R.C. (1996) 'A language learning model for finite parameter spaces', *Cognition*, vol. 61, 161–193.

- Niyogi, P. and Berwick, R. (1997) 'A dynamical systems model of language change', *Linguistics and Philosophy*, vol.17.
- Peak, D. and Frame, M. (1994) *Chaos Under Control: The Art and Science of Complexity*, Freeman, New York.
- Pinker, S (1989) *Learnability and Cognition: The Acquisition of Argument Structure*, MIT Press, Cambridge, MA.
- Pinker, S. and Bloom, P. (1990) 'Natural language and natural selection', *Behavioral and Brain Sciences*, vol.13, 707–784.
- Rambow, O. and Joshi, A. (1994) 'A processing model of free word order languages' in C. Clifton, L. Frazier and K. Rayner (ed.), *Perspectives on Sentence Processing*, Lawrence Erlbaum, Hillsdale, NJ., pp. 267–301.
- Ridley, M. (1990) 'Reply to Pinker and Bloom', *Behavioral and Brain Sciences*, vol.13, 756.
- Roberts, S. (1998) 'The role of diffusion in the genesis of Hawaiian creole', *Language*, vol.74.1, 1–39.
- Sapir, E. (1921) *Language: An Introduction to the Study of Speech*, Harcourt, Brace, Jovanovich.
- Schütze, H. (1997) *Ambiguity in language learning: computational and cognitive models*, CLSI Series, Cambridge University Press, Stanford.
- Sigmund, K. (1993) *Games of Life: Explorations in Ecology, Evolution and Behaviour*, Penguin, London.
- Sivia, D.S. (1996) *Data Analysis: A Bayesian Tutorial*, Clarendon Press, Oxford.
- Steels, L. (1998) 'Synthesizing the origins of language and meaning using coevolution, self-organization and level formation' in Hurford, J., Studdert-Kennedy, M., and Knight, C. (ed.), *Approaches to the Evolution of Language*, Cambridge University Press, Cambridge, pp. 384–404.
- Staddon, J.E.R. (1988) 'Learning as inference' in Evolution and Learning (ed.), *Bolles, R. and Beecher, M.*, Lawrence Erlbaum, Hillside NJ..
- Waddington, C. (1942) 'Canalization of development and the inheritance of acquired characters', *Nature*, vol.150, 563–565.
- Waddington, C. (1975) *The Evolution of an Evolutionist*, Edinburgh University Press, Edinburgh.
- Wanner, E. and Gleitman, L. (1982) 'Introduction' in Wanner, E. and Gleitman, L. (ed.), *Language Acquisition: The State of the Art*, MIT Press, Cambridge, Ma., pp. 3–48.
- Wexler, K. and Culicover, P. (1980) *Formal Principles of Language Acquisition*, MIT Press, Cambridge, Ma..
- Wexler, K. and Manzini, R. (1987) 'Parameters and learnability in binding theory' in T. Roeper and E. Williams (ed.), *Parameter Setting*, Reidel, Dordrecht, pp. 41–76.
- Wood, M.M. (1993) *Categorial Grammars*, Routledge, London.
- Wright, S. (1931) 'Evolution in Mendelian populations', *Genetics*, vol.16, 97–126.