# Grammatical Acquisition and Linguistic Selection

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

March 15, 1999

**DRAFT**

**CUP Evolution Book**

## 1 Introduction

This paper is part of an ongoing research effort (Briscoe, 1997, 1998, 1999a,b,c,d) to develop a formal model of language acquisition, demonstrate that an innate language acquisition device could have coevolved with human (proto)language(s) given plausible assumptions, and explore the consequences of the resulting model of both language and the language faculty for theories of language change. The paper builds on the earlier work by examining the model's ability to account for the process of creolization (Bickerton, 1981; 1984; 1988; Roberts, 1998) within a selectionist theory of language change.

§1.1 and §1.2 describe the theoretical background to this research. §2 presents a detailed model of the LAD utilizing generalized categorial grammars embedded in a default inheritance network integrated with a Bayesian statistical account of parameter setting. §3 reports experiments with this model demonstrating feasible and effective acquisition of target grammars for a non-trivial fragment of UG. §4 describes the simulation of an evolving population of language learners and users. §5 reports experiments with the simulation model which demonstrate linguistic selection for grammatical variants on the basis of frequency and learnability. §6 reports further experiments demonstrating evolution of the LAD by genetic assimilation of aspects of the linguistic environment of adaptation. §7 describes the experiments modelling the demographic and linguistic context of creolization. §8 summarizes the main findings and outlines areas of further work.

### 1.1 Grammatical Acquisition

Human language acquisition, and in particular the acquisition of grammar, is a partially-canalized, strongly-biased but robust and efficient procedure. It is a near-universal feat, where (partial) failure appears to correlate more with genetic deficits (e.g. Gopnik, 1994) or with an almost complete lack of linguistic input during the critical period (e.g. Curtiss, 1988), than with measures of general intelligence (e.g. Smith and Tsimpli, 1991) or the quality or informativeness of the learning environment (e.g. Kegl and Iwata, 1989; Ochs and Sheiffelin, 1995). Yet children prefer, for example, to induce lexically compositional rules (e.g. Wanner and Gleitman, 1982:12f) in which atomic elements of meaning are mapped to individual words despite the use, in every attested human language, of constructions which violate this preference, such as morphological negation or non-compositional idioms. Despite or perhaps because of these often counterfactual biases, grammatical acquisition remains highly robust.

Within the parameter setting framework of Chomsky (1981), the Language Acquisition Device (LAD) is taken to consist of a partial genotypic specification of (universal) grammar (UG) complemented with a parameter setting procedure which, on exposure to a finite positive sample of triggers from a given language, fixes the values of a finite set of finite-valued parameters to select a single fully-specified grammar from within the space defined by UG. Many parameters of grammatical variation set during language acquisition appear to have default or so-called unmarked values retained in the absence of robust counter-evidence (e.g. Chomsky, 1981:7f; Hyams, 1986; Wexler and Manzini, 1987; Lightfoot, 1992).

Thus, the LAD incorporates both a set constraints defining a possible human grammar and a set of biases (partially) ranking possible grammars by markedness. A variety of explanations have been offered for the emergence of an innate LAD with such properties based on saltation (Berwick, 1998; Bickerton, 1990, 1998) or genetic assimilation (Pinker and Bloom, 1990; Kirby, 1998). Formal models of parameter setting (e.g. Clark, 1992; Gibson and Wexler, 1994; Niyogi and Berwick, 1996; Brent, 1996) have demonstrated that development of a psychologically-plausible and effective parameter setting algorithm, even for minimal fragments of UG, is not trivial. The account developed in Briscoe (1997, 1998, 1999a,b,c,d) and outlined here improves the account of parameter setting, and suggests that biases as well as constraints evolve through a process of genetic assimilation of properties of human (proto)language(s) in the environment of adaptation for the LAD, but these constraints and biases in turn influence subsequent development of language via linguistic selection.

## 1.2 Linguistic Selection

In recent generative linguistic work on diachronic syntax, language change is primarily located in parameter resetting (reanalysis) during language acquisition (e.g. Lightfoot, 1992, 1997; Clark and Roberts, 1993; Kroch and Taylor, 1997). Differential learnability of grammatical variants, on the basis of learners' exposure to triggering data from varying grammatical sources, causes change. Language can be viewed as a dynamic system which adapts to its niche – of human language learners and users (e.g. Cziko, 1995; Hurford, 1987; 1998; Keller, 1994). Thus, language *itself* is evolving, on a historical timescale, and the primary source of *linguistic* selection is the language acquisition 'bottleneck' through which successful grammatical forms must pass repeatedly with each generation of new language learners. Under this view, the core evolutionary concepts of (random) variation, (adaptive) selection and (differential) inheritance are being used in their technical 'universal Darwinist' sense (e.g. Dawkins, 1983; Cziko, 1995) and not restricted to evolution of biological organisms.

To study linguistic evolution, it is necessary to move from the study of individual (idealized) language learners and users, endowed with a LAD and acquiring an idiolect, to the study of *populations* of such generative language learners and users, parsing, learning and generating a set of idiolects constituting the language of a community. Once this step is taken, then the dynamic nature of language emerges more or less inevitably. Misconvergence on the part of language learners can introduce variation into a previously homogeneous linguistic environment. And fluctuations in the proportion of learners to adults, or migrations of different language users into the population can alter the distribution and nature of the primary linguistic data significantly enough to affect grammatical acquisition. Once variation is present, then properties of the LAD become critical in determining which grammatical forms will be differentially selected for and maintained in the language, with language acquisition across the generations of users as the primary form of linguistic inheritance.

# 2 The Language Acquisition Device

A model of the LAD must incorporate a theory of UG with an associated finite set of finite-valued parameters defining the space of possible grammars, a parser for these grammars, and an algorithm for updating initial parameter settings on parse failure during acquisition (e.g. Clark, 1992). It must also specify the starting point for acquisition; that is, the initial state of the learner in terms of the default or unset values of each parameter of variation (e.g. Gibson and Wexler, 1994).

## 2.1 The (Universal) Grammar

Classical (AB) categorial grammar uses one rule of application which combines a functor category (containing a slash) with an argument category to form a derived category (with one less slashed argument category). Grammatical constraints of order and agreement are captured by only allowing directed application to adjacent matching categories. Generalized categorial grammars (GCGs) extend the AB system with further rule schemata (e.g. Steedman, 1988, 1996). Each such rule 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. The rules of forward application (FA), backward application (BA), generalized weak permutation (P) and forward and backward composition (FC, BC) are given in Figure 1 (where X, Y and Z are category variables, | is a variable over slash and backslash, and ... denotes zero or more further functor arguments). Generalized weak permutation enables cyclical permutation of argument categories, but not modification of their directionality. Once permutation is included, several semantically equivalent derivations for simple clauses such as *Kim loves Sandy* become available, Figure 2 shows the non-conventional left-branching one. Composition also makes alternative non-conventional semantically-equivalent (left-branching) derivations available.

This set of GCG rule schemata represents a plausible kernel of UG; Hoffman (1995, 1996) explores the descriptive power of a very similar system, in which P is not required because functor arguments are interpreted as multisets. She demonstrates that this system can handle (long-distance) scrambling elegantly and generate some mildly context-sensitive languages (e.g. languages with cross-serial dependencies such as $a^n, b^n, c^n$, though not some MIX languages with arbitrarily intersecting dependencies, e.g. Joshi *et al*, 1991). The majority of language-particular grammatical differences are specified in terms of the category set, though it is also possible to parameterize the rule schemata by, for example, parameterizing the availability of P, FC or BC and whether P can apply post-lexically.

The relationship between GCG as a theory of UG (GCUG) and as a specification of a particular grammar is captured by defining the category set and rule schemata as a default inheritance network characterizing a set of (typed) feature structures. The network describes the set of possible categories, each represented as a feature structure, via type declarations on network nodes. It also defines the rule schemata in terms of constraints on the unification of feature structures representing the categories. Type declarations $CON(Type, \subseteq)$ consist of path value specifications

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

Figure 1: GCG Rule Schemata

```
Kim                  loves                  Sandy
NP                   (S\NP)/NP              NP
kim'                 λ y,x [love′(x y)]     sandy′
                     ———————P
                     (S/NP)\NP
                     λ x,y [love′(x y)]
————————————————————BA
S/NP
λ y [love′(kim′ y)]
————————————————————————————————————FA
S
love′(kim′ sandy′)
```

Figure 2: GCG Derivation for *Kim loves Sandy*

(*PVS*s). An inheritance chain of (super)type declarations (i.e. a set of *PVS*s) defines the feature structure associated with any given (sub)type (see Lascarides *et al.*, 1995; Lascarides and Copestake, 1999, for further details of the grammatical representation language, and Bouma and van Noord (1994) for the representation of a categorial grammar as a constraint logic grammar.[1] Figure 3 is a diagram of a fragment of one possible network for English categories in which *PVS*s on types are abbreviated informally, $\top$ denotes the most general type, and meets display the (sub)type / (default) inheritance relations. **Vi** inherits a specification of each atomic category from which the functor intransitive verb category is constituted and the directionality of the subject argument (hereafter **subjdir**) by default from a type **gendir**. For English, **gendir** is default 'rightward' (/) but the *PVS* in **Vi** specifying the directionality of subject arguments, overrides this to 'leftward', reflecting the

[1] In fact, the representation of P as a constraint may be problematic. Instead it may be better represented as a unary rule which generates further categories. See Briscoe and Copestake (1997) for a discussion of lexical and other unary rules in the nonmonotonic representation language assumed here. Sanfilippo (1994) provides a detailed description of the encoding of categories for English verbs.

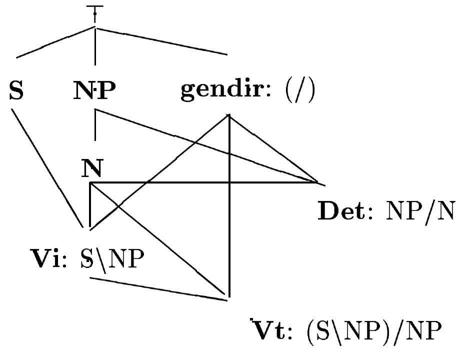

Figure 3: Fragment of an Inheritance Semi-Lattice

| NP | gendir | subjdir | objdir | ndir |
|---|---|---|---|---|
| A 1/T | D 0/R | D 1/L | ? ? | ? ? |

Figure 4: A p-setting encoding for the category fragment

fact that English is predominantly right-branching, though subjects appear to the left of the verb. Transitive verbs, **Vt**, inherit structure from **Vi** and an extra NP argument with default directionality specified by **gendir**. Nevertheless, an explicit $PVS$ in the type constraints for **Vt** could override this inherited specification. We will refer to this $PVS$ as **objdir** and the equivalent specification of the determiner category's argument as **ndir** below. A network allows a succinct definition of a set of categories to the extent that the set exhibits (sub)regularities.

The parameter setting procedure utilizes a function *P-setting*($UG$) which encodes the range of potential variation defining $g \in G$ where $UG$ is an invariant underspecified description of a CCG and *P-setting* encodes information about the $PVS$s which can be varied. For the experiments below a CCG covering typological variation in constituent order (e.g. Greenberg, 1966; Hawkins, 1994) was developed, containing 20 binary-valued unset or default-valued potential parameters corresponding to specific $PVS$s on types which are represented as a ternary sequential encoding (A=Absolute (principle), D=Default, ?=unset, 0=Rightward/False, 1=Leftward,True, ? = unset) where position encodes the specific $PVS$ and its (partial) specificity. Figure 4 shows a p-setting encoding of part of the network in Figure 3, where **S** and **N** are the only definitely invariant principles of $UG$, though this p-setting also encodes **NP** as an absolute absolute specification and, therefore, effectively a principle of UG. This encoding reflects the fact that $PVS$s specifying directionality for the object of a transitive verb or argument of a determiner are redundant as directionality follows from **gendir**. $CON(Type, \subseteq)$ defines a partial ordering on $PVS$s in p-settings, which is exploited in the acquisition procedure. For example, **gendir** is a $PVS$ on a more general type than **subjdir** and thus has more global (default) consequences in the specification of the category set, but **subjdir** will inherit its specification from **gendir** in the absence of an explicit $PVS$ for **Vt**.

The eight basic language families in $G$ are defined in terms of the unmarked, canonical order of verb (V), subject (S) and objects (O). Languages within families further specify the order of modifiers and specifiers in phrases, the order of adpositions, and further phrasal-level ordering parameters. In this paper, familiar attested p-settings are abbreviated as "German" (SOVv2, predominantly right-branching phrasal syntax, prepositions, etc), and so forth. Not all of the resulting 300 or so languages are (stringset) distinct and some are proper subsets of other

|  | **gen** | **v1** | **n** | **subj** | **obj** | **v2** | **mod** | **spec** | **rcl** | **adpos** | **cpl** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| "English" | R | F | R | L | R | F | R | R | R | R | R |
| "German" | R | F | R | L | L | T | R | R | R | R | R |
| "Japanese" | L | F | L | L | L | F | L | L | L | L | ? |

Figure 5: The Constituent Ordering Parameters

languages. "English" without P results in a stringset-identical language, but the grammar assigns different derivations to some strings, though their associated logical forms are identical. Figure 5 shows the settings for the 11 constituent ordering parameters utilized for some familiar languages. In addition, there are 3 p-setting elements which determine the availability of application, composition and permutation, 5 which determine the availability of specific categories (S,N,NP,Prep,Compl), and one, **argorder**, whose marked value allows a subject argument to combine first with a verbal functor for 'true' VOS or OSV languages.[2] Some p-setting configurations do not result in attested grammatical systems, others yield identical systems because of the use of default inheritance. The grammars defined generate (usually infinite) stringsets of lexical syntactic categories. These strings are sentence types since each defines a finite set of grammatical sentences (tokens), formed by selecting a lexical item consistent with each lexical syntactic category.

## 2.2 The Parser

The parser uses a deterministic, bounded-context shift-reduce algorithm (see Briscoe, 1998, 1999b) for further details and justification). It represents a simple and natural approach to parsing with GCGs which involves no grammar transformation or precompilation operations, and which directly applies the rule schemata to the categories defined by a GCG. The parser operates with two data structures, an input buffer (queue), and an analysis stack (push down store). Lexical categories are shifted from the input buffer to the analysis stack where reductions are carried out on the categories in the top two cells of the stack, if possible. When no reductions are possible, a further lexical item is shifted onto the stack. When all possible shift and reduce operations have been tried, the parser terminates either with a single 'S' category in the top cell, or with one or more non-sentential categories indicating parse failure. The algorithm for the parser working with a GCG which includes all the rule schemata defined in §2.1 is given in Figure 6. This algorithm finds the most left-branching derivation for a sentence type because Reduce is ordered before Shift. The algorithm also finds the derivation involving the least number of parsing operations because only one round of permutation occurs each time application and composition fail. The category sequences representing the sentence types in the data for the entire grammar set are unambiguous relative to this 'greedy, least effort' algorithm, so it will always assign the correct logical form to each sentence type given an appropriate sequence of lexical syntactic categories. Thus each sen-

[2] The p-setting encoding of the non-ordering parameters does not correspond to a single *PVS* in the grammatical representation language. We make the assumption for the parameter setting model that the definitions of rule schemata and of complex categories are predefined in UG as sets of *PVS*s but that a single element must be switched 'on' for them to become accessible. This could correspond to the *PVS* that links these definitions to the rest of the inheritance network by, for example, specifying that **application** is a subtype of **binary-rule**. However, even this is a simplification in the case of complex categories since these will typically inherit subcomponents from several places in the part of the network defining the category set. These assumptions speed up learning but do not alter fundamental results concerning convergence, provided that a more direct encoding retained a finite number of such finite-valued 'parameters'.

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

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

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

Figure 6: The Parsing Algorithm

tence type or potential trigger in the dataset encodes a surface form and asociated logical form as a sequence of determinate lexical syntactic categories when parsed with this algorithm.

## 2.3 Parameter Setting

The parameter setting algorithm used here is a statistical extension of an $n$-local partially-ordered error-driven parameter setting algorithm utilizing limited memory. Briscoe (1997, 1998) discusses related proposals (e.g. Gibson and Wexler, 1994; Niyogi and Berwick, 1997) and Briscoe (1999b,c) motivates the statistical approach to parameter setting). The algorithm only adjusts p-settings on parse failure of trigger input given the learner's current grammar. If flipping the settings of $n$ parameters results in a successful parse, then these settings receive further support and after a few such consistent observations the learner's p-setting will be more permanently updated, though there is nothing to stop subsequent triggers reversing the settings. The setting of parameters is partially-ordered in the sense that the partial order on $PVS$s corresponding to particular parameters defined by the inheritance network determines the manner in which updating proceeds. More general supertype parameters inherit their evidence and settings from their more specific subtype parameters. A p-setting not only encodes the current settings of parameters but also the degree of evidence supporting that setting, as a probability, and this determines how easily a setting can be updated. The statistical approach adopted is Bayesian in the sense that initial p-settings encoded a prior probability for a parameter setting and posterior probabilities for settings are computed in accordance with Bayes' theorem.

Bayes' theorem, given in (1), adapted to the grammar learning problem states that the posterior probability of a grammar, $g \in G$, where $G$ defines the space of possible grammars, is determined by its likelihood given the triggering input, $t_n$, multiplied by its prior probability.

(1) $$p(g \in G \mid t_n) = \frac{p(g)p(t_n \mid g)}{p(t_n)}$$

The probability of an arbitrary sequence of $n$ triggers, $t_n$, is usually defined as in (2).

(2) $$p(t_n) = \sum_{g \in G} p(t_n \mid g)\ p(g)$$

Since we are interested in finding the most probable grammar in the hypothesis space, $G$, given the triggering data, this constant factor can be ignored and learning can be defined as (3).

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

A sentence type / trigger is a pairing of a surface form (SF), defined as an ordered sequence of words, and a logical form (LF) representing (at least) the correct predicate-argument structure for the surface form in some context: $t_i = \{< w_1, w_2, ...w_n >, LF_i\}$.[3] A valid category assignment to a trigger ($VCA(t)$) is defined as a pairing of a lexical syntactic category with each word in the SF of $t$, $< w_1 : c_1, w_2 : c_2, ...w_n : c_n >$ such that the parse derivation, $d$ for this sequence of categories yields the same LF as that of $t$.[4]

Each grammar, $g$, interpreted as a stochastic generator of $L(g)$, should yield a probability distribution over sentence types, which (indirectly) will define a distribution for triggers, $t$ for $g$, and $g$ must also have a prior probability defined in terms of the probabilities of its components. We augment the account of GCG from §2.1 with probabilities associated with path value specifications ($PVS$s) in type declarations on nodes in the default inheritance network, $CON(Type, \subseteq)$. The probability of a $PVS$ with an 'uninteresting' absolute value is simply taken to be 1 for the purposes of the experiments reported below. A $PVS$ which is specified in a p-setting and which, therefore, plays a role in differentiating the class of grammars $g \in G$ will be binary-valued so $p(PVS_i = 0)$ and $1 - p(PVS_i = 1)$. An unset $PVS_i =?$ is assigned a prior probability of 0.5. The probability of each such $PVS$ is assumed to be independent, so the prior probability of a category is defined as the product of the probabilities of the $PVS$s in the type declarations which define it $CON(c, \subseteq)$, in (4).

(4) $$p(c \in g) = \prod_{PVS \in CON(c, \subseteq)} p(PVS)$$

The prior probability of a grammar, $g$ is the product of the probabilities of all its $PVS$s, as in (5).

(5) $$p(g) = \prod_{PVS \in CON(Type, \subseteq)} p(PVS)$$

[3] The definition of a LF is not critical to what follows. However, we assume that a logical form is a (possibly underspecified) formula of a well-defined logic representing at least the predicate-argument structure of the sentence (see e.g. Alshawi, 1996). It is possible that the definition of a trigger could be further relaxed to allow underdetermined predicate-argument structure(s) to be associated with a SF.

[4] We assume that the parse recovered will be that yielded by the parser of §2.2; namely, the least effort, most left-branching derivation. Strict equivalence of LFs could be relaxed to a consistency / subsumption relation, but this would not affect the experiments described below.

$CON(Type, \subseteq)$ is the grammatical representation language which defines the default inheritance network, which in turn denotes a minimal set of feature structures representing the category set for a particular grammar. The probability of an unset $PVS$ is always 0.5, so grammars $g \in G$ are differentiated by the product of the probabilities of the default and absolute valued $PVS$s represented in a p-setting. Therefore, (5) defines a prior over $G$ which prefers succinctly describable maximally-regular and minimally-sized category sets.[5] These constraints are enough to ensure that prior probabilities will be assigned in such a way that $\sum_{(g \in G)} p(g) = 1$.

The likelihood, $p(t_n \mid g)$, is defined as the product of the probabilities of each trigger (6).

$$(6) \quad p(t_n \mid g) = \prod_{t \in t_n} p(t \mid g)$$

Where the probability of a trigger is itself the product of the probabilities of each lexical syntactic category in the valid category assignment for that trigger, $VCA(t)$, as in (7).

$$(7) \quad p(t \mid g) = \prod_{c \in VCA(t)} p(c \mid g)$$

And the probability of a lexical category, $c$, is the product of the probabilities of the $PVS$s in the type declarations which define it, (8).

$$(8) \quad p(c|g) = \prod_{PVS \in CON(c, \subseteq)} p(PVS)$$

This is sufficient to define a likelihood measure, however, it should be clear that it yields a deficient language model (Abney, 1997) in which the total probability mass assigned to sentences generated by $g$ will be less than one and some of the probability mass will be assigned to non-sentences (i.e. sequences of lexical syntactic categories which will not have a derivation or $VCA$ given $g$).

The assumption of independence of $PVS$s rests partly on the semantics of the grammatical representation language in which a feature structure is a conjunction of atomic path values each specified by a single $PVS$. It represents a claim about the cognitive representations and procedures involved in language acquisition, rather than a claim that the different aspects of grammatical information encoded in distinct $PVS$s show no dependencies. In other words, we are claiming that language learners utilize an approximate statistical model of language. Similarly, the use of such a deficient model amounts to the cognitive claim that learners are sensitive to lexical probabilities but not the derived probabilities of phrases or clauses (see e.g. Merlo 1994).[6]

[5] Because the parameters of variation are a set of binary-valued $PVS$s with uniform probability assigned to unset $PVS$s, the product of these $PVS$s effectively defines an informative prior on $G$ consistent with the Minimum Description Length Principle (Rissanen, 1989). A more sophisticated encoding of the grammar would be required to achieve this if the parameters of variation differed structurally or 'unset' / unused $PVS$s were not assigned a uniform probability.

[6] The definition of probabilistic GCGs given here could be straightforwardly extended to define 'lexically-probabilistic' GCGs in which the probability of a trigger is conditioned on the lexical items, $w$ which occur in the trigger $p(w \mid c)$. However, we do not do so here since in the experiments which follow we assume that valid category assignments, $VCA(t)$, are given, and thus abstract away from the lexicon and lexical probabilities. Extending the model in this fashion would be critical if we wanted to deal with (probabilistic) selection between valid category assignments in order to resolve ambiguity.

| P-setting Type | Prior | Posterior | Setting |
|---|---|---|---|
| Principle | $\frac{1}{50}$ | $\frac{1}{50}$ | 0 |
| | $\frac{49}{50}$ | $\frac{49}{50}$ | 1 |
| Default Parameter | $\frac{1}{5}$ | $\frac{1}{5}$ | 0 |
| | $\frac{4}{5}$ | $\frac{4}{5}$ | 1 |
| Unset Parameter | $\frac{1}{2}$ | $\frac{1}{2}$ | ? |

Table 1: Probabilities of Parameter Types

## 2.4 Implementation

The Bayesian account of parameter setting has been partially implemented as an on-line, incremental grammar acquisition procedure which updates probabilities associated with the subset of $PVS$s which define (potential) parameters as each trigger is parsed. Though acquisition is restricted to the space defined by P-setting($UG$), the preference for the most succinct descriptions within this space requires that settings on more general types are updated to reflect the bulk of the probability mass of subtypes which potentially inherit settings from them. The resulting learner finds the locally maximally probable grammar given the specific sequence of triggers, $t_n$, seen so far, (9).

(9) $g = locmax_{g \in G}\ p(g)\ p(t_n \mid g)$

Each element of a p-setting is associated with a prior probability, a posterior probability and a current setting, as shown in Table 1 for the different types of possible initial p-setting (before exposure to data). The current setting is 1 iff the posterior probability associated with the parameter is >0.5, 0 iff it is <0.5 and unset (?) iff $p = 0.5$. Probabilities are stored as fractions so that incremental updates based on new observations can be expressed as additions to denominators and/or numerators, and larger denominators can be used to represent stronger priors. In the experiments reported below the values shown in Table 1 are used to initialize simulations, but values of numerators and denominators in priors can be modified by mutation and crossover operators during the reproduction of new language agents (see §4 below).

The Bayesian approach to incrementally updating the posterior probability of each parameter is approximated by incrementally computing the maximum likelihood estimate for each parameter but smoothing this estimate with the prior probability.[7] Firstly, the posterior probability is initialized to the (inherited) prior probability and these values are used to compute the parameter settings which define the starting point for learning. Then, as the learner successfully parses sentence types, the posterior probability of each parameter expressed in the sentence type is updated, reinforcing the probabilities of the parameter settings required to assign

[7] Strictly smoothing the maximum likelihood estimate with the prior does not conform to Bayes theorem in the limit because, given (9), if the likelihood is 1 or 0 then the prior has no effect. Very similar and strictly Bayesian results could be had in the implementation by using a Laplace-corrected estimate of the likelihood (that never goes to 1 or 0), corresponding to the assumption that in incremental updating of likelihood probabilities the data observed so far may not constitute a representative sample. The simpler and perhaps more psychologically-plausible (Cosmides and Tooby, 1996) approximation used here only differs in assigning more weight to the prior than would be achieved by multiplying the prior by a Laplace-corrected likelihood.

them the correct LF. However, when a sentence type cannot be successfully parsed, the acquisition procedure flips the settings of $n$ parameters in a p-setting, and, if this results in a successful parse, updates posterior probabilities according to these revised settings. The effect of this acquisition procedure is that a trigger does not usually cause an immediate switch to a different grammar. Rather the learner is more conservative and waits for enough evidence to shift a posterior probability through the $p = 0.5$ threshold before changing a setting more permanently.[8] For unset parameters at the beginning of the learning period, a single trigger, $t$, will suffice to set the parameter appropriately for $VCA(t)$, but incorrect default parameters will require a few more consistent observations, as will initially unset parameters which become inappropriately set as a result of noise or misanalysis.

Categorized triggers, $VCA(t)$, are encoded in terms of the most specific p-settings required to parse them successfully. However, each time posterior probabilities of most specific parameters are updated, it is necessary to examine the probabilities of their supertypes, and the pattern of default inheritance from them to subtype categories, in order to determine the most probable grammar $p(g \in$ P-setting$(UG))$ for these settings. The probability of a supertype $PVS$ is defined as the sum of the probabilities of those subtypes which inherit that $PVS$. Since inheritance is default, not all subtypes will necessarily inherit a given $PVS$ from a supertype, they may instead override it with an explicit specification on the subtype. Both the value of the supertype $PVS$ and its probability are determined by the amount of evidence supporting specific values for that $PVS$ on subtypes. For example, in the grammar fragment introduced above the $PVS$ for **gendir** is a supertype of **subjdir**, **objdir** (subject and object argument direction for verbal functors, respectively) and of **ndir** (general direction of arguments in nominal functors). The value of the $PVS$ for **gendir** (right / left) is determined by the values required on its subtypes and the probabilities associated with the subtype values. For example, if both **objdir** and **ndir** are 'right' (0) (i.e. their posterior probabilities are both $< 0.5$) but **subjdir** is 'left' (1), then the $PVS$ for **gendir** will be set to 'right' with probability derived from the sum of the probabilities of these two inheriting subtypes. However, **subjdir** will override the supertype with an explicit $PVS$ whose probability will not affect that of the supertype since the inheritance chain has been broken. This will ensure that the resulting grammar has the minimal number of explicit $PVS$s on types required to specify a grammar consistent with the data observed (so far) and thus that this is the most probable grammar *a priori*.[9] If subsequent evidence favours a 'left' setting for **ndir** or **objdir** then the $PVS$ for **gendir** will be revised to 'left' and the remaining rightward subtype will become the one requiring an explicit $PVS$ to override the default. Similarly, if **subjdir** in the above example had an unset (?, $p = 0.5$) value, then the setting of **gendir** rightward on the basis of the evidence from **ndir** and **objdir** would cause the learner to adopt a default rightward setting for **subjdir** too.

Figure 7 summarizes the algorithm used to find the most probable grammar compatible with the evidence for $PVS$s on the most specific types, where $PVS_j$ denotes a path value specification in a potential inheritance chain of type declarations

[8] For example, suppose parameter $i$ has a prior and initial posterior probability of 1/5, and thus a default value of 0. A single successful parse of sentence type expressing $i$ as 0 will cause the denominator of the posterior probability to be incremented by 1, yielding a new posterior of 1/6. A single observation of a sentence type expressing $i$ as 1 which gets a successful parse when $n$ parameter settings are flipped, including that for $i$, will cause the numerator and denominator to be incremented by 1, yielding a new posterior probability of 2/6. Thus, it will take at least 4 such observations to take the posterior past $p = 0.5$ and cause the learner to change the parameter setting.

[9] In the implementation the redundancy of a $PVS$ is modelled by assigning it a probability of 0.5 (i.e. by treating it as unset). Equivalent results would be obtained if the probability of $g \in G$ was computed by removing such $PVS$s altogether.

```
$\forall supertype_i \in g$
$\forall PVS_j \in subtypes_k \ of \ supertype_i$
if
   $\mid PVS_j = 1 \in subtypes_k \mid > \mid PVS_j = 0 \in subtypes_k \mid$
then
      $p(PVS_j) \in supertype_i = \sum p(PVS_j = 1) \in subtypes_k$
(and vice-versa)
else
   if
      $\sum p(PVS_j = 1) \in subtypes_k > \sum p(PVS_j = 0) \in subtypes_j$
   then
      $p(PVS_j) \in supertype_i = \sum p(PVS_j = 1) \in subtypes_k$
   (and vice-versa)
   else
      $p(PVS_j) \in supertype_i = 0.5$
```

Figure 7: Algorithm for computing posterior probabilities of supertypes

which may or may not need to be explicitly specified to override inheritance. The complete learning algorithm is summarized in Figure 8. Potential triggers, $t$ of $g^t$ are encoded in terms of p-schemata inducing $VCA(t)$, following Clark (1992). This obviates the need for on-line parsing of triggers during computational simulations. It also means that flip can be encoded deterministically by examining the parameter settings expressed by a trigger in the p-schemata and computing whether any resetting of $n$ parameters will yield a successful parse. If so, then these parameters are deemed to have been flipped and posterior probabilities are updated. The use of a deterministic flip speeds up convergence considerably and amounts to the strong assumption that learners are always able to determine an appropriate $VCA(t)$ for a trigger outside their current grammar if it is reachable with $n$ parameter changes. However, as their are finite finite-valued parameters, relaxing this assumption and, say, making random guesses without examining the trigger encoding would still guarantee eventual convergence.

# 3 Feasible and Effective Grammatical Acquisition

Two learners were defined on the basis of the grammar acquisition procedure described in §2. Both learners can flip up to 4 parameters per trigger and differ only in terms of their initial p-settings. Unset learners were initialized with p-settings consistent with a minimal inherited CGUG consisting of Application with the **NP** and **S** categories already present. All the remaining p-settings were genuine parameters for both learners. The unset learner was initialized with all these unset, while the default learner had default settings for the parameters **argorder**, **gendir**, **subjdir**, **v1** and **v2** which specify a minimal SVO right-branching grammar.[10]. The initialization of p-settings is in terms of their prior probabilities, as in Table 1, in accordance with the probabilistic model defined in §2.3, so that the prior probability of supertype PVSs is calculated from the priors associated with their subtypes.

Each variant learner was tested against a source grammar generating one of seven full languages in the grammar set (§2.1) which are close to an attested language; namely, "English" (SVO, predominantly right-branching), "Welsh" (SVOv1, mixed

[10] For a more detailed description of the effect of these five parameters in the model see Briscoe (1998, 1999b).

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

if
  $VCA(S_j) \in P\text{-}setting_i(UG)$
then
  $P\text{-}setting_j(UG) = \mathrm{Update}(P\text{-}setting_i(UG))$
else
  $P\text{-}setting_j(UG) = \mathrm{Flip}(P\text{-}setting_i(UG))$
  unless
  $VCA(S_j) \in P\text{-}setting_j(UG)$
  then
    RETURN $P\text{-}setting_i(UG)$
  else
    RETURN $\mathrm{Update}(P\text{-}setting_j(UG))$

Flip:
Flip or set the values of the first $n$ default or unset most specific parameter(s)
in a left-to-right search of the p-schemata representation of $VCA(t)$.

Update:
Adjust the posterior probabilities of the $n$ successfully flipped parameters and
of all their supertypes so that they represent the most probable grammar given
the data so far (see Figure 7 etc.).
```

Figure 8: The Parameter Setting Algorithm

| **Learner** | **Language** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | SVO | SVOv1 | VOS | VSO | SOV | SOVv2 | OVS | OSV |
| Unset (n4) | 33 | 32 | 34 | 32 | 34 | 32 | 32 | 32 |
| Default (n4) | 19 | 32 | 21 | 39 | 20 | 21 | 22 | 23 |

Table 2: Convergence Times for Two Learners

order), “Malagasy” (VOS, right-branching), “Tagalog” (VSO, right-branching), “Japanese” (SOV, left-branching), “German” (SOVv2, mixed branching), “Hixkaryana” (OVS, mixed branching), and a hypothetical OSV language with left-branching phrasal syntax. In these tests, a single learner parsed and, if necessary, updated parameters from a randomly drawm sequence of unembedded or singly embedded (potential) triggers, $t$ from $L(g^t)$ with $VCA(t)$ preassigned. The predefined proper subset of triggers used constituted a uniformly-distributed fair sample capable of distinguishing each $g \in G$ (e.g. Niyogi and Berwick, 1996). The first figure in Table 2 shows the mean number of potential triggers required by the learners to converge on each of the eight languages. These figures are each calculated from 1000 trials and rounded to the nearest integer. Presentation of 150 sentence types for each trial ensured convergence with $p \geq 0.99$ on all languages tested for both learners. As can be seen, the unset learner converges equally effectively on all eight languages, however, the preferences incorporated into the default learner's initial p-setting make languages compatible (e.g. SVO) or partially compatible (e.g. VOS, SOV, etc) with these settings relatively faster to learn, and ones largely incompatible with them (e.g. VSO) a little slower than the unset learner. Thus, the initial configuration of a learner's p-setting (i.e. the prior probabilities) can alter the relative learnability of different languages. Many experiments of this kind with these and other predefined variant learners demonstrate experimentally that convergence is possible, under these assumptions, for the 70 full and over 200 subset languages defined by *P-setting(UG)* (see Briscoe, 1997, 1998, 1999a,b).

The mean number of potential triggers required for convergence may seem unrealistically low, however, this figure is quite arbitrary as it is effectively dictated by the number of $n$ flippable parameters, the distribution and size of the trigger set, $t$, preassignment of $VCA(t)$ and the deterministic flipping of parameters (as well as the encoding of p-settings). The more general requirement for convergence is that their be a trigger path from the learners' initial settings which allows the (re)setting of all parameters for $g^t$ in $n$-local steps. For this the trigger set must constitute a fair sample capable of uniquely identifying $g^t \in G$ and the sequence of triggers in a trigger path supporting a $n$-local algorithm must be observed frequently enough during the learning period to support the $n$ parameter updating steps at each stage. The number of triggers required will depend, primarily, on the proportion of triggers for which $VCA(t)$ is hypothesized by the learner. A demonstration of the feasibility of the algorithm depends on replacing these optimal assumptions with more empirically motivated ones. Such modifications would be unlikely to alter the relative learnability results of Table 2, though they could increase the mean number of potential triggers required for convergence by several orders of magnitude (see Niyogi and Berwick, 1996). Here we focus on exploring the consequences of allowing some miscategorizations of trigger input and of allowing spurious triggers not from $g^t$ in the learner's input.

The results of Table 2 are computed on the basis that the learner is always able to assign the appropriate lexical syntactic categories to a sentence type / trigger (i.e. that $VCA(t)$ is always given). However, this is an unrealistic assumption. Even if we allow that a learner will only alter parameter settings given a trigger, that is, a determinate SF:LF pairing, there will still be indeterminacy of parameter expression. For example, Clark (1992) discusses the example of a learner acquiring "German" (SOVv2) in which triggers such as S-V, S-V-O, S-V-$O_1$-$O_2$, S-Aux-V will occur (where S denotes subject, O for object, and so forth, indicating informally a SF:LF pairing). These triggers are all compatible with a SVO grammar, though if "German" is the target language, then SVO triggers such as Aux-S-V-O will not occur, while other non-SVO ones such as O-V-S, S-Aux-O-V, O-Aux-S-V, and so forth will (eventually) occur. That is, neither SVO or SVOv2 is a subset of the other, but they share a proper subset of triggers. Thus, for a trigger like S-V-0 there is indeterminacy over the setting of the **objdir** parameter: it might be 'right' in which case VO grammars will be hypothesised, or 'left' with **v2** 'on' in which case OVv2 grammars will be hypothesised, and under either hypothesis the correct LF will be recovered. Thus, depending on the precise order in which specific triggers are seen by a learner, a deterministic learner might converge to an incorrect target grammar.

In the Bayesian framework parameters can, in principle, be repeatedly reset during the critical period for learning and their setting is conservative, based on observing a consistent *series* of triggers supporting a specific setting. The robustness of the acquisition procedure in the face of examples of such indeterminacies of parameter expression can be explored by exposing a learner to sentence types from the proper subset of SVO triggers (with $VCA(t)$ predefined) which overlap with SOVv2, as well as to SOVv2 triggers. This simulates the effect of a learner miscategorizing a proportion of the triggers compatible with SVO (i.e. assigning a $VCA(t)$ valid given the current state of the learner, but incorrect with respect to $g^t$). In these circumstances, the Bayesian parameter setting procedure should converge reliably to SOVv2 provided that the proportion of miscategorized triggers (to their

| **Lner**/$L(g^t)$ | **Trigger Proportions** | | | | |
|---|---|---|---|---|---|
| SVO-N/$L(g^t)$ | 15/85 | 30/70 | 40/60 | 50/50 | 60/40 |
| **SOVv2** | | | | | |
| Unset (n4) | 100 | 97.7 | 86.8 | 50.8 | 22.6 |
| Default (n4) | 100 | 97.6 | 87.9 | 62.2 | 28.8 |
| **SVOv1** | | | | | |
| Unset (n4) | 100 | 97.7 | 89.9 | 57.2 | 23.8 |
| Default (n4) | 100 | 96.6 | 90.1 | 59 | 25.1 |

Table 3: Percentage Convergence to SOVv2 / SVOv1 with SVO Miscategorizations

correctly categorized counterparts) does not cause any particular parameter to be expressed incorrectly in around 50% of all relevant triggers. The precise proportion will depend, of course, on whether the initial value is unset or default-valued and, if the latter, the relative strength of the prior. For a default-valued parameter with a strong prior probability whose correct value is marked, it is possible that quite a low proportion of miscategorized triggers could prevent its resetting within the learning period.

The two learners were tested on a mixture of 150 triggers randomly drawn from SVO-N-PERM-COMP and SOVv2 or SVOv1 in various proportions. SVO-N-PERM-COMP is the language corresponding to the proper subset of ambiguous triggers between "English" (SVO) and "German" (SOVv2), and also to a proper subset of ambiguous triggers between SVO and "Welsh" (SVOv1).[11] In each case, SVO-N-PERM-COMP triggers conflict with SOVv2 and SVOv1 in two parameters: **objdir** and bf v2, and *argorder* and **v1**, respectively. The percentage convergence to the 'target' SOVv2 or SVOv1 grammars over 1000 trials is given in Table 3. The first column gives percentage convergence when a miscategorized trigger was randomly drawn 15% of the time, the second 30% of the time, and so on until the proportion of miscategorized triggers exceeds that of the target grammar 60/40. By this stage most trials for both learners are converging to a SVO subset language, usually with some features determined by the full source grammar.

The percentages given in Table 3 include cases where the learner initially converged to the target and then switched to SVO. These accounted for from 4% up to 50% of the overall convergence rate, increasing as the proportion of SVO miscategorized triggers increased. One could posit, that the proportion of miscategorized triggers would decrease or cease over the learning period. Or that the $n$ updatable parameters per trigger over the learning period decrements; that is, the learner becomes more conservative towards the end of the learning period. Or that the learner knows when every parameter has been (re)set or 'reinforced' and then terminates learning. In each case, similar exploratory experiments indicate that the incidence of such 'postconvergence' to a different language, not actually exemplified in the source can be drastically reduced or eliminated.

The experiments suggest that the Bayesian approach to parameter setting, in principle, provides a robust and general solution to the indeterminacy of parameter expression. However, contingent details such as the frequency and order of specific (mis)categorized triggers, the weighting of priors, and so forth will determine the detailed behaviour and effectiveness of such learner in practice. The differences between the unset and default learners are minor in the results in Table 3 and only

[11] Subset languages are denoted by mnemonic names, where –F indicates that property F is missing, so –N indicates no multiword NPs, and –PERM and –COMP that permutation and composition are not available in derivations.

emerge, as expected, when the data has least influence on the initial settings; that is, when the proportion of miscategorized subset triggers to correctly categorized full language triggers is higher, so though two of the parameters receive more incorrect support, the majority are simply observed less frequently. Therefore, on balance, prior probabilities have a greater effect on posterior probabilities because the likelihood probabilities are less informative overall. Then the default learner converges slightly more successfully to either target grammar because, provided that the learner sees enough correctly categorized triggers to reset the two conflicting parameters with respect to each target, the rest of the directional parameter settings for each target grammar are correctly set by **gendir**'s default 'right' value, so these need less exemplification. Nevertheless, the probability that the two conflicting parameters will be correctly set for either target declines more for the default learner than the unset learner, as their initial default values also conflict with that required by each target. Therefore, increasing the prior weight of all the default valued parameters, or just of the two conflicting parameters, and rerunning the experiment would almost certainly yield a worse convergence rate for the default learner.

The non-statistical acquisition procedure of Briscoe (1997, 1998), as well as those of Gibson and Wexler (1994) and Niyogi and Berwick (1996), are excessively sensitive to miscategorizations of triggers or to other forms of noise in triggering input. If the learner is exposed to an extragrammatical or miscategorized trigger given the target grammar at a critical point, this can be enough to prevent convergence to the correct grammar. For example, given the deterministic parameter setting procedure of Briscoe (1997, 1998), a learner who has converged to a SVO grammar with right-branching phrasal syntax will, by default, assume the target grammar utilizes postnominal relative clauses. However, at this point exposure to a single trigger (mis)categorizable as containing a prenominal relative clause will be enough to override the default assumption of rightward looking nominal functors and, for the specific case of nominal functors taking relative clauses, permanently define these to be prenominal. On the other hand, a memoryless parameter setting procedure like the Trigger Learning Algorithm (Gibson and Wexler, 1994) will continue to switch between grammars, as mutually inconsistent sequences of triggers are observed, until the learning period ends. So effectively the final trigger and its (mis)categorization will determine the grammar selected (e.g. Niyogi, this volume). Clearly, the problem here is a special case of that of the indeterminacy of parameter expression. In the Bayesian framework, small proportions of noisy triggers encountered at any point in the learning period will not suffice to permanently set a parameter incorrectly. More systematic miscategorizations based on the indeterminacy of parameter expression will only result in misconvergence if the distribution of the triggering data allows the learner to miscategorize a high proportion of all triggers expressing a given parameter.

# 4 Populations of Language Agents

A language agent (LAgt) is minimally defined as a language learner, generator and parser endowed with the model of the LAD described above and a simple generation algorithm. The latter outputs a sentence type generated by the LAgt's current grammar (if any) drawn randomly according to a uniform distribution. In addition, LAgts have an age, which is used to determine the length of the learning period, and a fitness which can be used to determine their reproductive success and time of death.

A population of LAgts participates in a sequence of interaction cycles consisting of a specified number of random linguistic interactions between its members.

A linguistic interaction consists of a randomly chosen generating LAgt emitting a sentence type to a randomly chosen distinct parsing agent. The interaction is successful if their p-settings are compatible. Compatibility is defined in terms of the ability to map from a given SF to the same LF, rather than in terms of the sharing of an identical grammar. Populations are sometimes initialized with LAgts speaking a specific full language. Linguistic heterogeneity can then be introduced and maintained by regular migrations of further adults speakers with identical initial p-settings, but speaking a distinct full language. Alternatively, populations can be initialized to speak a variety of languages so that the range of variation can be controlled directly.

A LAgt's age is defined in terms of interaction cycles. LAgts can learn from age one to four; that is, during the first four interaction cycles. If a LAgt is a learner and cannot parse a sentence type during a linguistic interaction, then it is treated as a potential trigger and the LAgt applies the parameter setting procedure given its current p-setting. LAgts are removed from the population, usually at age 10. Two LAgts can reproduce a third at the end of an interaction cycle, if they are both aged four or over, by single point crossover and single point mutation of their p-setting encodings. The crossover and mutation operators are designed to allow variant initial p-settings to be explored by the population. For example, they can with equal probability flip the initial value of a default parameter, make a parameter into a principle or vice versa, and so forth, by altering the prior probabilities inherited by a new LAgt. LAgts either reproduce randomly or in proportion to their fitness. The fitness of a LAgt is defined by its communicative success; that is, the ratio of its successful interactions over all its interactions for the previous interaction cycle. The rate of reproduction is controlled so that a population always consists of >60% adult LAgts.

The simulation model and typical values for its variables are outlined in Figure 9. Further details and motivation are given in Briscoe (1998, 1999a). The mean number of interactions per LAgt per cycle are fixed so that acquisition of the target grammar in a linguistically homogeneous population is reliable ($p > 0.99$) for either of the predefined learners. The simulation can be used to study the process of learning and consequent linguistic selection for grammatical variants, or to study the interaction of linguistic selection with natural selection for more effective learners defined in terms of variant initial p-settings.

# 5 Linguistic Selection Experiments

Linguistic selection can be seen as a population level, and therefore dynamic, counterpart to the learner's problem of the indeterminacy of parameter expression. For example, if we initialize a population of LAgts so that some speak the SVO-N-PERM-COMP subset language, corresponding to the proper subset of triggers which overlap with "German" (SOVv2), and the remainder speak "German", then learners should reliably converge to "German", even when exposed to triggers from all the population, provided that SVO-N-PERM-COMP triggers do not much exceed 15% of all triggers (see the results of §3). On the other hand, if the initial proportion of SVO-N-PERM-COMP speakers is higher, but still below 50%, then we would expect a minority of learners to converge to SVO subset languages or mixtures of the two sources. However, as the simulation run continues, SVO (subset) speakers will disappear because the relative frequency of SOVv2 speakers will increase with each new batch of learners and as the original SVO-N-PERM-COMP adults die out.

A series of simulations were run to test these predictions, in which an initial population of either 32 default or unset learner LAgts reproduced randomly and

LAgt: <P-setting($UG$),Parser,Generator,Age,Fitness>

POP$_n$: {LAgt$_1$, LAgt$_2$, $\ldots$ LAgt$_n$}

INT(LAgt$_i$,LAgt$_j$), $i \neq j$, Gen(LAgt$_i$, t$_k$),Parse(LAgt$_j$, t$_k$)

SUCC-INT: Gen(LAgt$_i$, t$_k$) $\mapsto$ LF$_k$ $\wedge$ Parse(LAgt$_j$, t$_k$) $\mapsto$ LF$_k$

REPRO: (LAgt$_i$,LAgt$_j$), $i \neq j$,
Create-LAgt(Mutate(Crossover(P-setting(LAgt$_i$,P-setting(LAgt$_j$)))))

LAgt Fitness:

1. Generate cost: 1 (GC)
2. Parse cost: 1 (PC)
3. Success benefit: 1 (SI)
4. Fitness function: $\frac{SI}{GC+PC}$

LAgt Death: Age 10

| **Variables** | **Typical Values** | |
|---|---|---|
| POP$_n$ | Initially | 32 |
| Interaction Cycle | Mn. Ints/LAgt | 65 |
| Simulation Run | Int. Cycles | 1k |
| Crossover Probability | | 0.9 |
| Mutation Probability | | 0 |
| Migrations | per cycle | 2 |
| | dominant lg | 90% |

Figure 9: The Evolutionary Simulation

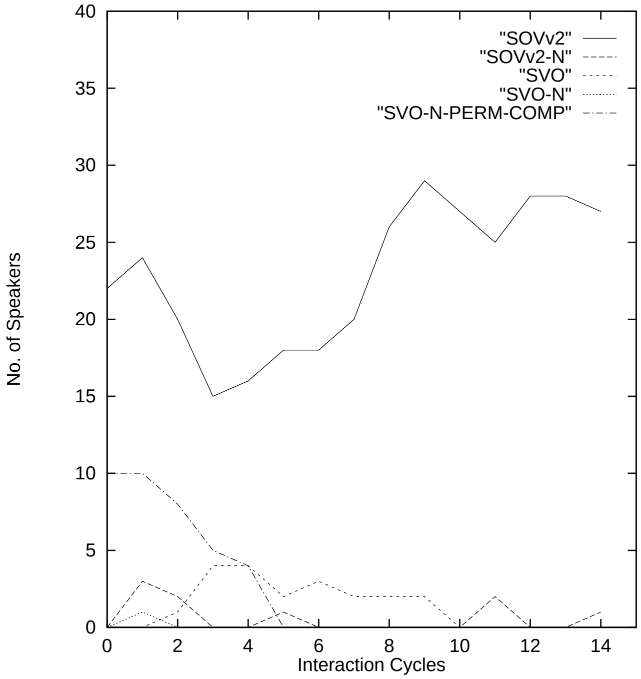

Figure 10: Linguistic Selection between Languages

the number of SVO-N-PERM-COMP, SOVv2 and SOVv2 subset language speakers was tracked through interaction cycles. In these simulations there is no variation amongst LAgts, and so no evolution at the 'genetic' (initial p-setting) level – all learners are either default or unset n4 learners as defined in §3. However, there is linguistic selection between the languages, where the ultimate units of selection / inheritance are competing parameter values. The selection pressure comes from two conflicting sources: learnability and relative frequency. SVO-N-PERM-COMP is easier to learn than SOVv2 because it requires the setting of fewer parameters, but it may be less frequently exemplified in the primary linguistic data than SOVv2, depending on the proportions of speakers in the initial population.

Figure 10 plots the languages spoken across interaction cycles for a population of default learners initialized with 10 SVO-N-PERM-COMP and 22 SOVv2 adult LAgts with ages varying randomly from 5-9; so the first generation of learners will be exposed, on average, to 30% SVO-N-PERM-COMP triggers. This plot is typical: the SVO-N-PERM-COMP speakers dwindle rapidly, though a few SVO superset language learners emerge briefly, until cycle 10 when only SOVv2 speakers and SOVv2-N learners remain. In 19 out of 20 such runs, the population converged fully on SOVv2 in a mean 9.8 interaction cycles (with the exception of learners speaking a SOVv2 subset language); that is, within two and a half full generations of LAgts. (It is not possible for full convergence to occur in less than 6 interaction cycles unless no initial SVO-N-PERM-COMP adult is aged 5.) After the first interaction cycle in which learners were present, no subsequent learner converged to a SVO (subset) language in any of these runs. However, in the one other run, the population converged on a full SVO language after 11 interaction cycles, and in around half the other runs, a few learners briefly spoke the full SVO language. In runs with a lower proportion of initial SVO-N-PERM-COMP speakers, linguistic selection for SOVv2 was 100%. In the runs initialized with unset learners, the results were similar except that there was much less tendency for learners to visit the full SVO language on the path to SOVv2. However, in an otherwise identical series of runs initialized with 16 SVO-N-PERM-COMP and 16 SOVv2 speakers, an equally clear opposing result was obtained: populations nearly always converged on SVO-N-PERM-COMP within 15 interaction cycles. Here ease of learnability swayed the balance in favour of the subset language when each was exemplified equally in the learners' data, regardless of whether the population comprised default or unset learners.

These experiments examine the interplay of the relative frequency with which linguistic variants are exemplified and the relative learnability of languages in determining what learners acquire. They are sufficient to demonstrate that linguistic selection is a viable approach to accounting for some types of language change. In Briscoe (1998, 1999b) more experiments are reported which look at the role of parsability and expressibility in linguistic selection and also at the potential impact of natural selection for LAgts on this process. Whenever there is linguistic heterogeneity in speech community, a learner is likely to be exposed to sentence types deriving from more than one source grammar. In reality this is the norm rather than the exception during language acquisition. Learners are typically exposed to many speakers, none of whose idiolects will be entirely identical, some of them may themselves be learners with an imperfect command of the target grammar of their speech community, and some may come from outside this speech community and speak a different dialect / language. The Bayesian approach to parameter setting predicts that learners will track the frequency of competing variants in terms of the posterior probabilities of the parameters associated with the variation. This accords with the empirical behaviour of learners in such situations (e.g. Kroch, 1989; Kroch and Taylor, 1997; Lightfoot, 1997). They appear to acquire both variants and choose which to produce on broadly sociolinguistic grounds in some cases, and to converge preferentially to one variant in others. This behaviour could be modelled, to a first approximation in the current framework, by assigning varying weights to prior default-values and postulating that parameters are set permanently if their posterior probabilities reach a threshold value (say, $> 0.95$ for 1, and $< 0.05$ for 0). In this case, parameters which never reached threshold might be accessible for sociolinguistically-motivated register variation, while those which did reach threshold within the learning period would not.[12]

# 6 Coevolution of the LAD and of Language

The acquisition experiments of §3 demonstrated the effectiveness of the Bayesian parameter setting procedure with several initial p-settings on some full languages, even in the presence of noise and indeterminacy of parameter expression. The simulations of populations of default learner LAgts of §5 demonstrated linguistic selection on the basis of learnability and the relative frequency of conflicting triggers without any variation at the genetic, initial p-setting level. Introducing variation in the initial p-settings of LAgts, allows for the possibility of selection for better initial settings, at the same time as languages, or their associated grammars, are themselves being selected.

Variation amongst LAgts can be introduced in two ways. Firstly, by initializing the population with LAgts with variant p-settings, and using a crossover operator during LAgt reproduction to explore the space defined by this initial variation. And

[12] A modification of this type might also form the basis of a less stipulative version of the critical period for learning in which LAgts simply ceased to track posterior probabilities of parameters once they reached threshold; see, e.g. Hurford and Kirby, 1997 for discussion and putative explanations of the critical period for language acquisition.

secondly, by also using a mutation operator during reproduction which can introduce variation during a simulation run, with reproduction via crossover propagating successful mutations through the population. Single point crossover with a prespecified probability of 0.9 is utilized on a flat list of the numerators and denominators representing the prior probabilities of each p-setting. The mutation operator can modify a single p-setting during reproduction with a prespecified probability (usually $p = 0.05$). Mutation alters an element of a p-setting, with equal probability, from its existing type (absolute principle, default or unset parameter) and initial setting (1, 0, ?) to a new type and/or initial setting. Thus, no evolutionary bias is introduced at this level, but mutation can alter the definition of UG by making a principle a parameter or vice-versa, and alter the starting point for learning by altering the prior probabilities of parameters.

Briscoe (1998, 1999a,b) argues in detail that, under the assumption that communicative success confers an increase in fitness, we should expect the learning period to be attenuated by selection for more effective acquisition procedures in the space which can be explored by the population; that is, we should expect genetic assimilation (e.g. Waddington, 1942, Pinker and Bloom, 1990). In the context of the Bayesian acquisition procedure, genetic assimilation corresponds to the evolution of the prior probabilities which define the starting point for learning to more accurately reflect properties of the environment (during the period of adaptation). Staddon (1988) and Cosmides and Tooby (1996) independently argue that many aspects of animal and human learning behaviour can be accounted for under the assumption that learning is Bayesian, priors evolve, and this general learning mechanism can be exapted to specific problems with domain-specific representational and inferential components. Nevertheless, the selection for better language acquisition procedures will be relative to the dominant language(s) in the environment of adaptation (i.e. the period before the genetic specification of the LAD has gone to (virtual) fixation in the population). And these languages will themselves be subject to changing selective pressures as their relative learnability is affected by the evolving LAD, creating reciprocal evolutionary pressures, or *coevolution*. However, the selective pressure favouring genetic assimilation, and its subsequent maintenance and refinement, is only coherent given a coevolutionary scenario in which (proto)language(s) supporting successful communication within a population had already itself evolved on a historical timescale (e.g. Hurford, 1987; Kirby, 1998), probably with many of the constraints and biases subsequently assimilated already present in the (proto)language(s) as a consequence of *linguistic* selection, perhaps initially driven by quite general cognitive constraints such as working memory limitations.

Here we report the results of a series of simulation experiments designed to demonstrate that the LAD evolves towards a more specific UG (more principles) with more informative initial parameter settings (more default-values) consistent with the dominant language(s) in the environment of adaptation, even in the face of the maximum rate of language change consistent with maintenance of a language community (defined as mean 90% adult LAgt communicative success throughout a simulation run). Populations of LAgts were initialized to be unset learners all speaking one of the seven attested languages introduced in §3. Simulation runs lasted for 2000 interaction cycles (about 500 generations of LAgts) and each condition was run ten times. Reproduction was proportional to communicative success and was by crossover and mutation of the initial p-settings of the 'parent' LAgts. Constant linguistic heterogeneity was ensured by migrations of adult LAgts speaking a distinct full language with 1-3 different parameter settings at any point where the dominant (full) language utilized by the population accounted for over 90% of interactions in the preceding interaction cycle. Migrating adults accounted for approximately one-third of the adult population and were initialized to have initial

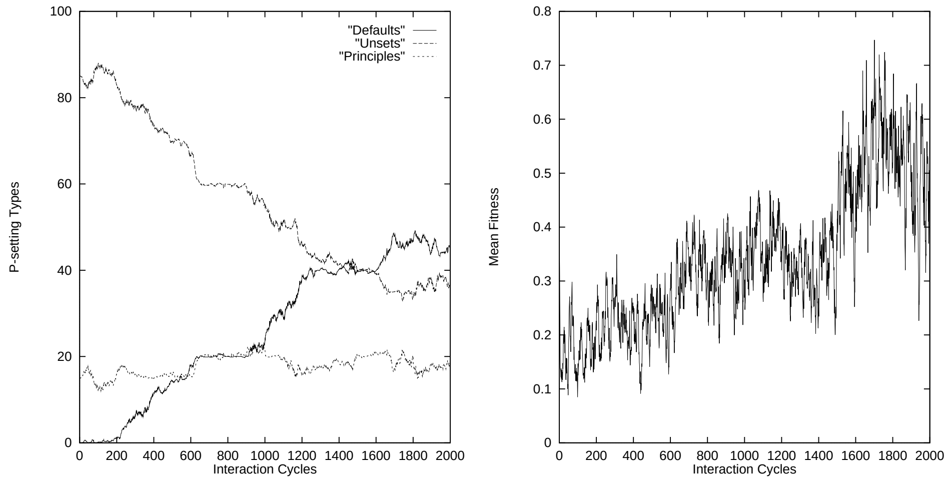

Figure 11: Proportions of p-setting types and Mean fitness

p-settings consistent with the dominant settings already extant in the population; that is, migrations are designed to introduce linguistic, not genetic, variation. Populations typically sampled about 100 languages with the dominant language changing about 40 times during a run.

The mean increase in the proportion of default parameters in all such runs was 46.7%. The mean increase in principles was 3.8%. These accounted for an overall mean decrease of 50.6% in the proportion of unset parameters in the initial p-settings of LAgts. Figure 11 shows the relative proportions of default parameters, unset parameters and principles for one such typical run with the population initialized to unset n4 learners. It also shows the mean fitness of LAgts over the same run; overall this increases as the learning period gets shorter, though there are fluctuations caused by migrations or by an increased proportion of learners. These results, which have been replicated for different languages, different learners, and so forth (see Briscoe, 1997, 1998, 1999a) are clear evidence that a minimal LAD, incorporating a Bayesian learning procedure, could evolve the prior probabilities and UG configuration which define the starting point for learning in order to attenuate the acquisition process by making it more canalized and robust. On average, a shorter learning period will result in increased communicative success because learners will be able to parse the full range of sentence types from the dominant language by an earlier age.

In these experiments, linguistic change (defined as the number of interaction cycles taken for a new parameter setting to go to fixation in the population) is about an order of magnitude faster than the speed with which a genetic change (new initial p-setting) can go to fixation. Typically, 2-3 grammatical changes occur during the time taken for a principle or default parameter setting to go to fixation. Genetic assimilation remains likely, however, because the space of grammatical variation (even in the simulation) is great enough that typically the population is only sampling about 5% of possible variation in the time taken for a single p-setting variant to go to fixation (or in other words, 95% of the selection pressure is constant during this period).

Though many contingent details of the simulation are arbitrary and unverifiable, such as the size of the evolving population, size of the set of grammars, $G$, and relative speed at which both can change, it seems likely that the simulation model massively *underestimates* the size of the potential space of grammatical possibilities. Few linguists would baulk at 30 independent parameters of variation, defining a space of billions of grammars, for an adequate characterization of a parameter setting model of the LAD, while even fewer would argue that the space of possibilities could be finitely characterized at all prior to the emergence of a LAD (e.g. Pullum, 1983). Thus, although there is a limit to the rate at which genetic evolution can track environmental change (e.g. Worden, 1995a), while the speed limit to major grammatical change before effective communication is compromised will be many orders of magnitude higher, it is very likely that 95% of this space would *not* be sampled in the time taken for fixation of any one parameter of variation in the LAD, given plausible ancestor population sizes (e.g. Dunbar, 1993). Nevertheless, there is a limit to genetic assimilation in the face of ongoing linguistic change, in simulation runs with LAgts initialized with all default parameters, populations evolve away from such 'fully-assimilated' LADs (e.g. Briscoe, 1998) when linguistic variation is maintained.

# 7 Creolization

The abrupt transition from pidgin to creole, which Bickerton (1981, 1984, 1988) argues occurs in one generation, constitutes one of the most dramatic and radical attested examples of language change. In more recent work, Roberts (1998), using a large database of Hawaiian pidgin and creole utterances, has revised Bickerton's original claim slightly, by arguing that some aspects of the transition in Hawaii took two generations. Nevertheless, this careful empirical work by-and-large confirms Bickerton's original claims that the creole emerges very abruptly and embodies a much richer grammatical system than the pidgin, whose properties are not directly exemplified in the super- or sub-stratum languages to which learners might be exposed and are very similar to the properties of other creoles which emerged at geographically and historically unrelated points. Bickerton (1984:173) has described the process by which learners acquire a creole grammar as one of invention in terms of an innate bioprogram.

Creolization represents a potential challenge for the account of language learning and language change presented here. Though we claim that language learning is partially innate and that many of the constraints and biases incorporated into the LAD have evolved via genetic assimilation, the parameter setting algorithm is purely selectionist and largely data-driven, and the associated account of change is thus also selectionist. Confronted with variation, a language learner will preferentially acquire variants which are more learnable or more robustly exemplified in the primary linguistic data.[13] If there is an element of 'invention' in creolization how could this arise? The account that we will pursue here is that in some respects the primary linguistic data that creole learners are exposed to is so uninformative that they retain their prior default-valued parameter settings as a direct consequence of the Bayesian parameter setting procedure. However, this is not enough to ensure that a rich and full grammatical system will emerge if the data never exemplifies, however indirectly, a particular grammatical phenomenon. When exposed exclusively to a subset language, the Bayesian parameter setting procedure reliably acquires that subset language and does not go 'beyond the evidence' to predict a full language 'extension' of the subset language learners have been exposed to. Critically, although the p-setting encoding adopted assumes that a supertype parameter, **gendir**, determines ordering of arguments to all functors by default, the

[13] In addition, parsability and expressibility may also play a role either by affecting learnability or by influencing speakers' choice of sentence types in order to optimize communicative success; see Briscoe (1998, 1999b) for further discussion.

use of the 5 category parameters (see §2.1, means that complex catgories, such as those associated with a complementizer or nominal modifier, are only accessible to a learner if expressed in a trigger.

Plantation creoles arise as a result of unusual and radical demographic conditions (Baker and Corne, 1982; Bickerton, 1988). In the initial preparatory phase of plantation colonization, the speech community consists of European managers and technicians and some native labourers. At this stage, the labourers may learn the European language with reasonable proficiency on the basis of frequent contact with the Europeans. However, in the second exploitative stage, when the plantation is up and running, the (indentured or slave) labour population increases five to tenfold within a single generation as successive waves of immigrants are brought in from diverse parts of the world to increase the labour force and also to compensate for the typically high mortality rates. These new immigrants do not share a native language and have much reduced exposure to the European superstratum language as the proportion of colonialists to labourers decreases radically and the original native population or earlier arrivals take on much of the day-to-day management of the plantation. In these circumstances, the *lingua franca* of the labouring community rapidly develops into an extremely impoverished pidgin language, consisting of a limited vocabulary, learnt indirectly via the original native population from the European superstratum language, and virtually no grammatical system (see Bickerton, 1990:122f) for a summary of the properties of pidgins). Children born to labourers during the third stage of the community are predominantly exposed to the pidgin language – contact with the Europeans is limited, many parents are of mixed descent and do not share a native language, and the children are mostly brought up by a few older women in large groups, while all able bodied men and women labour for long hours in the fields (Bickerton, 1984:215). The birthrate in most plantation communities was not particularly high, with under twelves typically accounting for no more than 25% of the population, except in Hawaii where birthrates were higher (Bickerton, personal communication). The mixture and composition of substratum native languages spoken by the labourers varied widely between communities. Nevertheless, in the third stage when native creole speakers emerge, remarkably similar grammatical extensions of the impoverished pidgins, which provide the bulk of the learners' primary linguistic data, have been documented (see Roberts, 1998 for recent discussion and argument that such similarities cannot be the result of substratum language influences).

Bickerton (1984:179) describes the prototypical creole grammar, based on Saramaccan, as a minimal SVO right-branching grammar with distinct syntactic categories for determiners, adjectives, numerals, nouns, verbs, auxiliaries and complementizers.[14] The predefined default learner of §3 incorporates prior probabilities favouring a SVO right-branching grammar consistent with a Saramaccan-like grammar but underdetermining all the properties of the creole language. The questions that we will attempt to answer experimentally in the remainder of this section are: What distributions of primary linguistic data would cause learners hypothetically endowed with this initial p-setting via coevolution (see §6) to converge to a Saramaccan-like grammar? And how well do these distributions accord with the known demographic conditions governing the emergence of creoles? We will make no attempt to model the emergence of an impoverished pidgin language but concentrate entirely on the stage three pidgin-creole transition in which learners exposed to predominantly pidgin data rapidly converge to a creole language.

In all the experiments which follow, the initial population contains 64 LAgts,

[14] Many similarities, such as the tense-modality-aspect morphosyntactic systems of creoles or choices for lexicalization of specific grammatical properties, are not modelled in the set of grammars used in the current simulation so cannot be investigated directly here. Nevertheless, the main points about the acquisition process made below should carry over to these phenomena too.

who live for 20 interaction cycles and reproduce randomly without mutation. All LAgts are either unset of default learners in the initial population, and thus all new LAgts reproduced during these runs are also either default or unset learners. The birth rate is set to six new LAgts per interaction cycle which grows quickly to a stable population of around 115 LAgts always containing 18 learners. The highest proportion of learners to adults occurs around the sixth interaction cycle, before the adult population has peaked, when learners constitute about 28% of the total population. We do not model high mortality rates or influxes of new immigrants directly but rather keep the original adult population constant over the first twelve interaction cycles. This represents the assumption that the pidgin and sub-/super-stratum trigger distribution heard by learners remains constant over the first three generations. However, the fact that new learners are added at the end of each interaction cycle means that the first learners reach the end of the learning period in the fourth interaction cycle, and that the proportion of learners to adults grows over the first few cycles. Thus, the overall linguistic distribution of triggers does change, as learners and new adults begin to form a significant proportion of the population and participate in interactions; and therefore, the degree to which early learners converge consistently to a specific language significantly affects the probability that later learners will follow suit by skewing the distribution of triggers in favour of this language..

We model a pidgin as a subset language without embedded clauses or multi-word NPs in which a wide variety of constituent orders is possible, perhaps partly influenced by the substratum native languages of the individual speakers and/or by pragmatic factors. In a first series of experiments, populations of adult LAgts were initialized to speak such subset languages with between three and five of the six basic word orders available (SVO, SOV, VSO, VOS, OSV, OVS) in equal proportions (as before, all conditions were run 10 times). Thus learners were exclusively exposed to subset language input, either exemplifying SVO order or not, with a variety of other orders also present. This corresponds to the hypothesis that learners are exclusively exposed to pidgin triggers and either do not hear sub-/super-stratum utterances or do not treat them as triggering data. When SVO subset triggers are present, even if this only constitutes one fifth of the triggering experience of learners on average, default learners reliably converge to the SVO subset language. By the fourth interaction cycle – the end of the first generation – a mean 95% of learners are speaking a SVO subset language. From that point, new learners all converge to SVO subset grammars. The results for similar runs with unset learners show a similar overall preference for SVO subset grammars, but a lower proportion of learners speak SVO in the early cycles and in a minority of runs the learner population still contains non-SVO subset language speakers beyond the third generation. When SVO triggers are not present learners converge to non-SVO subset languages. The picture that emerges, then, is largely expected given the Bayesian learning paradigm. The great majority of default learners, faced with conflicting triggering input, converge to SVO order because the prior probabilities of their inherited p-settings tend to dominate over the likelihood probabilities acquired during learning, as these are inconsistent and broadly 'uninformative'. Nevertheless, if SVO is never exemplified in the data, learners never converge to it though they frequently converge to 'close' right-branching grammars. Default learners do not overgeneralize and converge to a full language as the triggering data does not express parameters for complex nominal categories, and so forth. On the other hand, the tendency for unset learners to converge to SVO subset grammars, though weaker, was not expected and must be a consequence of the 'topology' of the hypothesis space created by the encoding of *P-setting(UG)*. However, the most important point is that no population of LAgt learners converged to a superset language in any of these runs. Thus, no process of creolization occurred.

In further experiments, populations were initialized with adult LAgts speaking a variety of languages exemplifying five of the six basic constituent orders, but also some sub-/super-stratum language utterances. In a first series of runs one fifth of the LAgts were full SVO right-branching speakers and the remaining four fifths spoke four non-SVO subset languages. Thus the average trigger distribution for initial learners consisted of 80% non-SVO pidgin utterances, 7% SVO pidgin or superstratum language utterances, and 13% SVO superstratum utterances. Corresponding to the hypothesis that creole learners are exposed to a small minority of "English" superstratum language triggers. In these runs, populations of default learners converged rapidly to SVO. By the second interaction cycle when twelve learners were present, ten or more were speaking a SVO subset language with the remainder, if any, currently speaking a subset language with VOS or SOV order. By the end of fourth interaction cycle, most of the earliest learners had converged to "English" with the minority speaking SVO subset languages compatible with it. After this point, all subsequent learners converged to "English". Similar runs with unset learners also mostly converged to "English" by the third generation but in about 40% of cases a non-SVO subset language was also being spoken by a significant proportion of the new population. Furthermore, in the crucial early interaction cycles a higher proportion of learners converged to non-SVO subset languages. On average, by the end of the first interaction cycle only one learner had converged to "English". These runs demonstrate that the default learner, but not the unset learner, predicts that creolization (that is, convergence to a SVO superset language) will occur essentially within a generation with minimal exposure to a superstratum language which is compatible with the acquired creole grammar.

The previous experiments ignored the role, if any, of the substratum languages. In a further otherwise identical series of runs, populations were initialized with SVO, two non-SVO subset languages, and two non-SVO languages with randomly-defined full language extensions in equal proportions. This meant that initial learners were exposed, on average, to 54% pidgin-like triggers with four equally-frequent non-SVO orders, 26% non-SVO richer triggers further exemplifying two of these non-SVO orders but otherwise randomly exemplifying more complex syntax, 7% SVO subset pidgin or superstratum triggers, and 13% richer "English" superstratum triggers. The non-SVO extensions are intended to model the rich and often conflicting variety of fragments of substratum languages that creole learners might hear uttered amongst the adult labouring population (e.g. Bickerton, 1984:182f). The broad effect of adding this degree of substratum data (or interference) is to slow down convergence to SVO for both types of learner. However, in the case of the default learner the difference is negligible. By the fourth interaction cycle half the earliest learners have converged to "English" and a mean 91.3% of all learners present are speaking a SVO (subset) language. By the twelfth interaction cycle a full SVO language is spoken by virtually all the new population, with "English" predominant (though in some runs a second full SVO language with some substratum influences is also present at this stage). With the unset learners, only a mean 60% of new adults and learners are speaking a SVO subset language by the sixth interaction cycle. By the twelfth interaction cycle the new population typically speaks a mixture of SVO languages, with "English" dominant but other full SVO language and some adult SVO subset language speakers present. These experiments suggest that positing substratum interference does not affect the basic conclusion that creolization will occur rapidly with default learners. Figure 12 plots the growth of SVO (subset) language speakers in two typical runs with unset and default learners respectively. The runs with default learners show more rapid and comprehensive selection of SVO languages, than those with unset learners.

Though creoles display SVO right-branching syntax, it is not the case that the superstratum language is always English or even SVO. It seems reasonable to as-

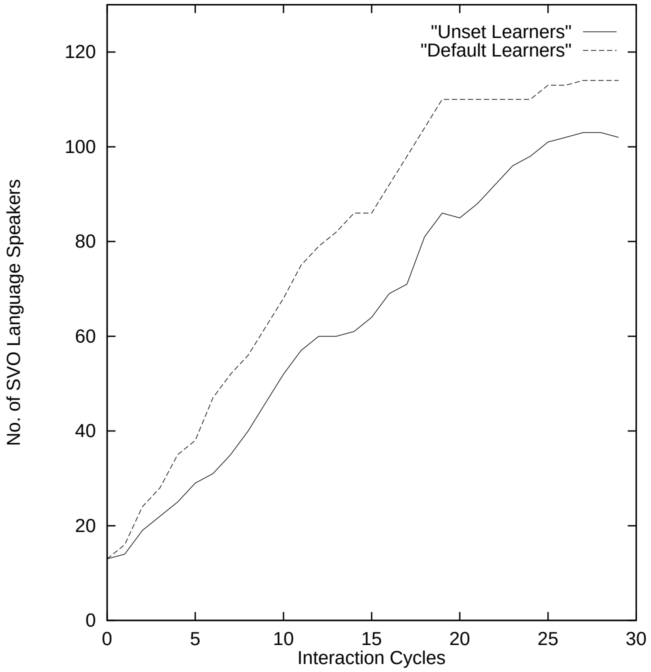

Figure 12: SVO (Subset) Language Speakers

sume that SVO order will always be exemplified to some extent in the pidgin data to which learners are exposed, but the account we have developed so far relies on initial learners' exposure to 13% of richer "English" superstratum triggers. While this might be plausible for Hawaii it is not for Berbice where Dutch, with a SOVv2 grammar, was the superstratum language. In a final series of experiments otherwise identical to those above, SOVv2 was substituted for the superstratum language, though SVO order remained one pidgin language, variant order. The initial trigger distribution was 89% pidgin-like subset languages with 4% random non-SVO substratum extensions and 7% SOVv2 superstratum extensions. 28% of the pidgin-like triggers had SVO order. The early dynamics of these runs are almost identical to those described above: learners predominantly speak SVO (subset) languages from the beginning and do so exclusively after the first two or three interaction cycles. By the end of the twelfth interaction cycle in runs with unset learners, the new population was speaking the SVO pidgin-like subset language in a mean 91.2% of cases; the remainder spoke "English" or its subset without permutation. At the same point in runs with default learners, the new population was speaking the SVO pidgin-like subset language in a mean 68% of cases, 28.3% were speaking "English" without permutation, and the rest "English". Given that 18 of those currently speaking the SVO pidgin-like subset were still learners and that "English" without permutation was the best represented language by the twelfth interaction cycle with populations of default learners, we would expect most if not all of these learners to converge to this language. In 80% of these runs, the new population had converged to "English" without permutation (and in one case "English") by the twentieth interaction cycle. Runs with a slightly lower proportion of initial SVO pidgin-like triggers resulted in a slower convergence to SVO languages. These results suggest a slower convergence rate to a SVO creole superset language when the superstratum language is SOVv2. Nevertheless, "English" without permutation is the closest grammar to Saramaccan available in the set $G$ used in the experiments.

The experiments suggest, then, that creolization could result as a consequence of a Bayesian parameter setting learner adopting default settings for some parameters, acquired via the coevolutionary scenario outlined in §6. Prior probabilities, and thus initial parameter settings, will play a bigger role in the acquired grammar whenever the data the learner is exposed to are inconclusive. It seems plausible that pidgin data are inconclusive about constituent order because pidgin speakers order constituents in inconsistent ways. Nevertheless, order is necessarily expressed in pidgin data, so the learner defaults to SVO order, and also predicts, by default, that right-branching, head-first order will extend to more complex categories. The encoding of *P-setting(UG)* and predefined definition of the default learner we utilize does not allow the learner to hallucinate or invent more complex categories for nominal modification, complementizers, and so forth. However, if such categories are reliably expressed somewhere in the triggering data for each learner, even with inconsistent ordering, then the default learner will 'switch on' a generic unordered form of these categories, and predict their ordering behaviour by default. This account does not require that the superstratum language be SVO, or that substratum languages consistently exemplify properties of the creole; merely, that richer triggers expressing parameters for more complex categories be present in the primary linguistic data. Thus, the learning procedure is different from that outlined by Bickerton (1984): there is no invention or other special mechanism at work, rather the grammar acquired is a consequence of the distribution of triggers and the prior probabilities of the Bayesian learner.

The timing of creolization for the default learner runs with SVO superstratum input is remarkably consistent with the timecourse documented by Roberts (1998), especially given that the prior probabilities of the default learner were not modified at all for these experiments. The Bayesian parameter setting framework and the population model are quite capable of simulating variant accounts in which, for example, prior probabilities are stronger and the data exemplifies some parameters less, or the proportion and growth rate of learners during the third stage of plantation communities is different. To refine the account developed here will require both a better understanding of the language learning procedure and a more precise and detailed account of demographic change and speed of creolization in different plantation communities. For instance, Bickerton (1984:178) suggests that sub-/super-stratum influence cannot be important because some communities of pidgin speakers were 'marooned' and learners did not have access to any speakers of either. If this is accurate, then 'invention', or at least a propensity to acquire superset grammars with default parameter settings on the basis of no triggering evidence, will need to be reconsidered. However, such a model would conflict with prevailing assumptions derived from learnability criteria, like the subset principle (e.g. Berwick, 1985), that predict that learners are conservative and do not overgeneralize superset grammars because no parse failure could force subsequent convergence to the target grammar. Moreover, the data concerning such marooned communities is very sparse, so it is difficult to know whether learners did completely lack sub-/super-stratum input.

## 8 Conclusions and further work

The experimental results reported above suggest that a robust and effective account of parameter setting, broadly consistent with Chomsky's (1981) original proposals, can be developed by integrating generalized categorial grammars, embedded in a default inheritance network, with a Bayesian learning framework. In particular, such an account seems, experimentally, to be compatible with local exploration of the search space and robust convergence to a target grammar given feasible amounts of partly noisy or indeterminate input. It extends recent work in parameter setting by integrating the learning procedure more closely with a fully-specified grammatical representation language, by using a Bayesian statistical approach to resolve indeterminacies of parameter expression, and by demonstrating convergence for a more substantial language fragment containing around 300 grammars / languages.

Human language learners, in certain circumstances, converge to grammars different from that of the preceding generation. Linguistic selection for more learnable variant constructions during language acquisition offers a promising formal framework to account for this type of language change. Creolization represents a particularly radical version of such change which is potentially challenging for a selectionist and essentially data-driven account. However, given assumptions about the starting point for learning, the initial distribution of triggers, and the changing constitution of the plantation community, the model of the language acquisition device developed here predicts that creolization will occur within the timeframe identified by Roberts (1998) for SVO superstratum languages. The highly-biased nature of language learning is a consequence of the coevolutionary scenario outlined in §6 in which there is reciprocal interaction between natural selection for more efficient language learners and linguistic selection for more learnable grammars. The range of distributions of triggers to creole learners is compatible with the known linguistic and demographic data for the better studied cases, though it does require that creole learners are influenced, albeit somewhat indirectly, by sub-/super-stratum language triggers. The growth of the native learner and adult population during the third stage of plantation communities partly determines the speed of creolization and thus ideally requires more detailed examination.

Gold's (1967) negative 'learnability in the limit' results have been very influential in linguistic theory, accounting for much of the attraction of the parameter setting framework and for much of its perceived inadequacy (e.g. Niyogi and Berwick, 1996; Gibson and Wexler, 1994; Muggleton, 1996). Within the framework explored here, even a much weaker result, such as that of Horning (1969), that stochastic context-free grammars are learnable from positive finite evidence is only of heuristic relevance, since all such results rest crucially on the assumption that the input comes from a single stationary source (i.e. a static and given probability distribution over a target stochastic language). However, from the current evolutionary perspective, contingent robustness or local optimization in an irreducibly historical manner is the most that can be expected. The coevolutionary account suggests that the apparent success of language learning stems more from the power of our limited and biased learning abilities to select against possible but less easily learnable grammatical systems, than from the omnipotence of the learning procedure itself. Given this perspective, there is little reason to retain the parameter setting framework. Instead, learners might extend the model of universal grammar by adding path value specifications to the default inheritance network to create new grammatical categories when triggering data warranted it. An implementation of this aspect of the model is a priority since it would also allow such innovations to be incorporated into universal grammar via genetic assimilation, and this in turn would underpin a better evolutionary account of the development and refinement of the language acquisition device.

The model of a language agent assumes the existence of a minimal language acquisition device, since agents come equipped with a universal grammar, associated learning procedure, and parser. Simulation runs demonstrate that an effective, robust but biased variant learning procedures specialized for/on specific grammars could emerge by genetic assimilation / coevolution. However, they do not directly address the question of how such an embryonic language acquisition device might emerge. Evolutionary theory often provides more definitive answers to questions concerning the subsequent maintenance and refinement of a trait than to ones concerning its emergence (e.g. Ridley, 1990). However, other work suggests that the emergence of a minimal language acquisition device might have required only minor reconfiguration of cognitive capacities available in the hominid line. Worden (1998) and Bickerton (1998) argue that social reasoning skills in primates provide the basis for a conceptual representation and reasoning capacity. In terms of the model presented here, this amounts to claiming that the categorial logic underlying generalized categorial grammars' semantics was already in place. Encoding aspects of this representation (i.e. logical form) in a transmittable language would only involve the comparatively minor step of linearizing this representation by introducing directionality into functor types. Parsing here is, similarly, a linearized variant of logical deduction with a preference for more economical proofs / derivations. Staddon (1988), Cosmides and Tooby (1996) and others have argued that many animals, including primates and homo sapiens, exhibit reasoning and learning skills in conditions of uncertainty which can be modelled as forms of Bayesian learning. Worden (1995b) argues that Bayesian learning is the optimal approach to many tasks animals face, and therefore the approach most likely to have been adopted by evolution. If we assume that hominids had inherited such a capacity for Bayesian learning, then evolution could construct a minimal language acquisition device by applying this capacity to learning grammar, conceived itself as linearization of a pre-existing language of thought. Given this scenario, much of the domain-specific nature of language acquisition, particularly grammatical acquisition, would follow not from the special nature of the learning procedure *per se*, as from the specialized nature of the morphosyntactic rules of realization for the language of thought.

## Acknowledgements

??

## References

Abney, S. (1997) 'Stochastic attribute-value grammars', *Computational Linguistics, vol.23.4,* 597–618.

Alshawi, H. (1996) 'Underspecified first-order logics' in van Deemter, K. and Peters, S. (ed.), *Semantic Ambiguity and Underspecification,* CSLI Publications, Stanford, Ca., pp. 145–158.

Baker, P. and Corne, C. (1982) *Isle-de-France Creole,* Karoma, Ann Arbor.

Berwick, R. (1985) *The Acquisition of Syntactic Knowledge,* MIT Press.

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. (1990) *Language and Species,* University of Chicago Press, Chicago.

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.

Bouma, G. and van Noord, G (1994) 'Constraint-based categorial grammar', *Proceedings of the 32nd Assoc. for Computational Linguistics,* Morgan Kaufmann, Palo Alto, Ca., pp. 147–154.

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

Briscoe, E.J. (1997) 'Co-evolution of language and of the language acquisition device', *Proceedings of the 35th Assoc. for Comp. Ling.,* 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 ', *Proceedings of the 8th Meeting of Comp. Linguistics in the Netherlands,* Rodopi, Amsterdam, pp. 3–40.

Briscoe, E.J. (1999a, in press) 'The Acquisition of Grammar in an Evolving Population of Language Agents', *Proceedings of the Machine Intelligence, 16,* Oxford University Press, Oxford.

Briscoe, E.J. (1999b, submitted) 'Grammatical Acquisition: Co-evolution of Language and the Language Acquisition Device', *Language,*

Briscoe, E.J. (1999c, in press) 'Evolutionary perspectives on diachronic syntax', *Proceedings of the Diachronic Generative Syntax, 5,* Oxford University Press, Oxford.

Briscoe, E.J. (1999d) 'Resolving indeterminacies of parameter expression', *Proceedings of the 9th Eur. Assoc. for Comp. Ling.,* Morgan Kaufmann, Palo Alto, Ca..

Briscoe, E.J. and A. Copestake (1999, submitted) 'Lexical rules in constraint-based grammar', *Computational Linguistics,*

Chomsky, N. (1957) *Syntactic Structures,* Mouton, The Hague.

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. and Roberts, I. (1993) 'A computational model of language learnability and language change', *Linguistic Inquiry, vol.24.2,* 299–345.

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.

Curtiss, S.R. (1988) 'Abnormal language acquisition and the modularity of language' in F. Newmeyer, vol. 2 (ed.), *Linguistics: The Cambridge Survey,* 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.

Dunbar, R. (1993) 'Coevolution of neocortical size, group size and language in humans', *Behavioral and Brain Sciences, vol.16,* 681–735.

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.

Goodman, J. (1997) 'Probabilistic feature grammars', *Proceedings of the 5th Int. Workshop on Parsing Technologies,* Morgan Kaufmann, Palo Alto, Ca., pp. 89–100.

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

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

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

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

Hoffman, B. (1996) 'The formal properties of synchronous CCGs', *Proceedings of the ESSLLI Formal Grammar Conference,* Prague.

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

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

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

Hurford, J. and Kirby, S. (1997) *The evolution of incremental learning: language, development and critical periods,* Edinburgh Occasional Papers in Linguistics, 97-2.

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

Joshi, A., Vijay-Shanker, K. and Weir, D. (1991) 'The convergence of mildly context-sensitive grammar formalisms' in Sells, P., Shieber, S. and Wasow, T. (ed.), *Foundational Issues in Natural Language Processing,* MIT Press, pp. 31–82.

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

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

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.

Kroch, A. (1991) 'Reflexes of grammar in patterns of language change', *Language Variation and Change, vol.1,* 199–244.

Kroch, A. and Taylor, A. (1997) 'Verb movement in Old and Middle English: dialect variation and language contact' in van Kemenade, A. and N. Vincent (ed.), *Parameters of Morphosyntactic Change,* Cambridge University Press, pp. 297–325.

Lascarides, A., E.J. Briscoe, A.A. Copestake and N. Asher (1995) 'Order-independent and persistent default unification', *Linguistics and Philosophy, vol.19.1,* 1–89.

Lascarides, A. and Copestake A.A. (1999, in press ) 'Order-independent typed default unification', *Computational Linguistics,*

Lightfoot, D. (1992) *How to Set Parameters: Arguments from language Change,* MIT Press, Cambridge, Ma..

Lightfoot, D. (1997) 'Shifting triggers and diachronic reanalyses' in van Kemenade, A. and N. Vincent (ed.), *Parameters of Morphosyntactic Change,* Cambridge University Press, pp. 253–272.

Merlo, P. (1994) 'A corpus-based analysis of verb continuation frequencies', *Journal of Psycholinguistic Research, vol.23.6,* 435–457.

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

Niyogi, P. and Berwick, R.C. (1996) 'A language learning model for finite parameter spaces', *Cognition, vol.61,* 161–193.

Ochs, E. and Shieffelin, B. (1995) 'The impact of language socialization on grammatical development' in Fletcher, P. and MacWhinney, P. (ed.), *The Handbook of Child Language,* Blackwell, Oxford, pp. 73–94.

Osborne, M. and E.J. Briscoe (1997) 'Learning stochastic categorial grammars', *Proceedings of the Assoc. for Comp. Linguistics, Comp. Nat. Lg. Learning (CoNLL97) Workshop,* Morgan Kaufmann, Palo Alto, Ca., pp. 80–87.

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

Pullum, G.K. (1983) 'How many possible human languages are there?', *Linguistic Inquiry, vol.14.3,* 447-467.

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

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

Roberts, S. (1998) 'The role of diffusion in the genesis of Hawaiian creole', *Language, vol.74.1,* 1–39.

Sanfilippo, A. (1994) 'LKB encoding of lexical knowledge' in Defaults, Inheritance and the Lexicon (ed.), *Briscoe, E.J., A.A. Copestake and V. de Paiva (eds.),* Cambridge University Press, pp. 190–222.

Smith, N.V. and Tsimpli, I.A. (1991) 'Linguistic modularity? A case study of a 'savant' linguist', *Lingua, vol.84,* 315–351.

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

Steedman, M. (1988) 'Combinators and grammars' in R. Oehrle, E. Bach and D. Wheeler (ed.), *Categorial Grammars and Natural Language Structures,* Reidel, Dordrecht, pp. 417–442.

Steedman, M. (1996) *Surface Structure and Interpretation,* MIT Press, Cambridge, Ma..

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

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 Manzini, R. (1987) 'Parameters and learnability in binding theory' in T. Roeper and E. Williams (ed.), *Parameter Setting,* Reidel, Dordrecht, pp. 41–76.

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

Worden, R.P. (1995b) *An optimal yardstick for cognition,* Psycoloquy (electronic journal).

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