# 1 Introduction

Human language acquisition, and in particular the acquisition of grammar, is a partially-canalized, strongly-biased but robust and efficient procedure.[1] For example, children prefer to induce lexically compositional rules (e.g. Wanner and Gleitman, 1982) despite the use, in every attested human language, of constructions, such as morphological negation or non-compositional idioms. And, most 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. Bickerton, 1984; Hyams, 1986; Lightfoot, 1992). A variety of explanations have been offered for the emergence of a partially-innate language acquisition device (LAD) with such properties based on saltation (Berwick, 1998; Bickerton, 1990, 1998) or genetic assimilation (Pinker and Bloom, 1990). But none provide a coherent detailed account of both the emergence and maintenance of a LAD in an evolving population.

The account proposed here is that a minimal LAD emerged via recruitment of general-purpose (Bayesian) learning mechanisms (e.g. Staddon, 1988; Cosmides and Tooby, 1996) to a specifically-linguistic mental representation capable of expressing mappings from the 'language of thought' to realizable, essentially linearized, encodings of propositions of the language of thought. However, the selective pressure favouring such a development, and its subsequent maintenance and refinement, is only coherent given a coevolutionary scenario in which a (proto)language supporting successful communication within a population had already itself evolved on a historical timescale (e.g. Hurford, 1987; Kirby, 1998; Steels, 1998) and continued to coevolve with the LAD (e.g. Briscoe, 1997, 1998, 2000a).

The model of the LAD presented here builds on and extends previous work in the parameter setting framework (e.g. Chomsky, 1981; Clark, 1992; Gibson and Wexler, 1994; Niyogi and Berwick, 1996; Briscoe, 1997, 1998, 2000a) by developing a Bayesian account of parameter setting, and embedding this within a more general theory of language acquisition in which it is not essential that the hypothesis space of grammars is finite. The Bayesian account of parameter setting can explain the robustness of acquisition in the face of noise and the indeterminacy of parameter expression in triggering input (e.g. Clark, 1992) as well as underlie a more insightful account of language change via differential acquisition of competing linguistic variants (e.g. Kroch, 1989). The extension of the theory of grammatical acquisition beyond parameter setting to one in which an infinite range of grammars can, in principle, be acquired would underpin a coevolutionary account of the development of human language and of the LAD (e.g. Kirby, 1998).

The paper begins by summarizing (§2) the model of the LAD described in Briscoe (1997, 1998, 2000a) and experiments with (evolving) populations of language agents (LAgts), defined in terms of this model of the LAD (§3). It then (§4) describes the Bayesian extensions to this model designed to address weaknesses of the earlier work. An implementation of this new model of the LAD is used to define a new LAgt and it is demonstrated in (§5) that such LAgts can acquire non-trivial grammars from finite positive samples of triggering input, even in the face of noise and well known examples of parameter indeterminacy. Experiments with an evolving population of LAgts (§6) show that, given the assumption that communicative success confers benefit to LAgts, LADs evolve via genetic assimilation to improve

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

the acquisition procedure, and languages evolve via linguistic selection for more learnable linguistic variants. In conclusion (§7), it is argued that only rather modest cognitive developments are required for the emergence of a LAD in agents already equipped with the capacity for social reasoning and reasoning with uncertainty.

## 2 The Language Acquisition Device

A model of the language acquisition device (LAD) must incorporate a theory of universal grammar (UG) with an associated finite set of finite-valued parameters (Chomsky, 1981) 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). The following subsections present such a model (see Briscoe, 1997, 1998, 2000a for further details and background).

### 2.1 The Grammar (set)

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

| | |
|---|---|
| | 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' | $\lambda$ y,x [love$'$(x y)] | sandy$'$ |
| | ————P | |
| | (S/NP)\NP | |
| | $\lambda$ x,y [love$'$(x y)] | |
| ————————BA | | |
| S/NP | | |
| $\lambda$ y [love$'$(kim$'$ y)] | | |
| ———————————————FA | | |
| S | | |
| love$'$(kim$'$ sandy$'$) | | |

Figure 2: GCG Derivation for *Kim loves Sandy*

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 ($PVS$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. A prespecified proper subset of $PVS$s constitute the parameters of the GCUG.[2] 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 rela-

[2]See Lascarides *et al.*, 1996; 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. The representation of P as a constraint is problematic. Instead it is 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 (1993) provides a detailed description of the encoding of categories for English verbs in a nonmonotonic CG setting.

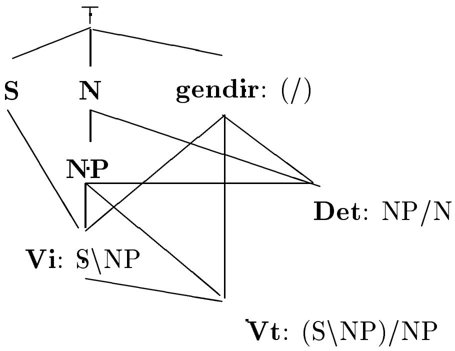


Figure 3: Fragment of an Inheritance Semi-Lattice

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

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

tions. **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 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 GCG and *P-setting* encodes information about the $PVS$s which can be varied. For the experiments below a GCG 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-valued sequential encoding (A=Absolute (principle), D=Default, ?=unset, 0=Rightward/False, 1=Leftward,True, ? = unset) where serial order encodes the specific $PVS$ and its (partial) specificity. Figure 4 shows a p-setting encoding of part of the network in Figure 3. **S** and **N** are treated as definitely invariant principles of $UG$. **NP** has an absolute specification in the p-setting and, therefore, is also a principle of UG. However, including **NP** in the p-setting makes it a potential parameter given an alternative p-setting specification. $CON(Type, \subseteq)$ 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 p-setting specification in Figure 4 reflects the fact that $PVS$s specifying directionality for the object of a transitive verb or argument of a determiner are redundant here, as directionality follows from **gendir**.

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 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. Some p-settings 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, 1987, 1998b 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 5. 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 sentence type or potential trigger in the dataset encodes a surface form and associated logical form as a sequence of determinate lexical syntactic categories when parsed with this algorithm.

## 2.3 Parameter Setting

The parameter setting procedure is an extension and modification of Gibson and Wexler's (1994) Trigger Learning Algorithm (TLA) to take account of the inheritance-based partial ordering, the role of memory in learning,

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 5: The Parsing Algorithm

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

unless
  $Parser_i(P\text{-}setting_i(UG))(S_j)$ = Success
then
  $P\text{-}setting_j(UG)$ = Update($P\text{-}setting_i(UG)$)
  if
  $\text{Parser}_j(P\text{-}setting_j(UG))(S_j)$ = Success
  then
    RETURN $P\text{-}setting_j(UG)$
  else
    RETURN $P\text{-}setting_i(UG)$

Update:
Reset the first $n$ default parameter(s) or set the first $n$ unset parameter(s) in a 'left-to-right' search of the p-settings (consistent with the partial order encoding their generality) according to the following table:

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

Figure 6: The Learning Algorithm

and so forth. The TLA is error-driven – parameter settings are altered in constrained ways when a learner cannot parse trigger input and when the alteration results in a successful parse. Trigger input is defined as primary linguistic data which, because of its structure or context of use, is determinately unparsable with the correct interpretation (i.e. logical form). A trigger is a sentence type (i.e. sequence of lexical syntactic categories) generated by a target grammar $g_t$ drawn from the finite set of grammars, $G$. A learner must converge to $g_t \in G$ with high probability on exposure to a feasible number, $n$, of triggers, $t \in L(g_t)$. In the modified algorithm each parameter can be updated once in the partial order defined by the inheritance network. However, because of the use of default specification in the grammatical representation language, this does not lead to strictly monotonic refinement of grammatical hypotheses. Up to $n$ parameters per trigger can be updated, if this results in a successful parse.

Each step for a learner can be defined in terms of two functions: *P-setting* and *Parser*, as in Figure 6. A *P-setting* defines a grammar which in turn defines a parser (where the subscripts indicate the output of each function given the previous trigger, $S_i$). A parameter is updated on parse failure and, if this results in a successful parse, the new setting is retained. The core of the algorithm is the update rule, which is applied to a ternary sequential p-setting encoding (see §1.1). A default parameter can be reset to its opposite value and the 'D' encoding changed to a 'R" to record that this default parameter has been reset, and unset parameters are randomly set to one possible value.

# 3 Experiments with Language Agents

The account of the LAD described in §2 is used to define a language agent (LAgt) capable of parsing or generating sentence types from the language defined by its current p-setting and capable of acquiring a grammar by altering p-settings on the basis of trigger input. An initial p-setting defines a starting point for grammar acquisition. The effectiveness of the parameter setting algorithm for a significant subset of the 70 full languages defined by the grammar set has been demonstrated experimentally via computational simulation for two initial p-settings, each defining a minimal UG consisting of **S**, **N** and Application: one unset learner for which the 17 remaining parameters were unset, and one default learner for which several further ordering parameters were default-valued to define a learner with an initial preference for SVO clause order and predominantly right-branching phrasal syntax. These learners converged to the target grammar on exposure to uniformly sampled triggers from it within a mean 32 triggers for all languages tested with high probability ($p \geq 0.99$).

In the experiments outlined above a single learning LAgt parses output generated by a non-learning LAgt initialized to speak one of the full languages defined by the grammar set. Thus, the learner is exposed to data from a single source of randomly generated sentence types. To explore the behaviour of an evolving population of LAgts, the model of a LAgt is extended to include, an age, a method of reproduction, and a fitness defined in terms of communicative success; that is, the proportion of successful interactions with other LAgts. An interaction takes place between a speaking LAgt and a listening LAgt. The speaking LAgt randomly generates a sentence type compatible with its current p-settings (i.e. its grammar). An interaction is successful if the sentence type generated by the speaking LAgt can be parsed by the listening LAgt to yield the same logical form that the speaking LAgt associates with this sentence type. So, their p-settings and associated grammars need to be consistent with respect to this sentence type, though not necessarily identical.

A population of LAgts participates in a sequence of interaction cycles consisting of a specified number of random interactions between its members. 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. They are removed from the population after 10 interaction cycles. 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-settings. 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. LAgts either reproduce randomly or in proportion to their fitness. The fitness of a LAgt is defined by 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. Populations are typically initialized with LAgts speaking a specific full language. However, linguistic heterogeneity can be introduced and maintained by regular migrations of further adults speakers with identical initial p-settings but speaking a distinct full language. The simulation model and typical values for its variables are outlined in Figure 7. Further details and motivation are given in Briscoe (1998a,b).

Evolutionary simulations of an evolving population of LAgts with dif-

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

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

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

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

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

LAgt Fitness:

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

| **Variables** | **Typical Values** | |
|---|---|---|
| $\text{POP}_n$ | Initially | 32 |
| Interaction Cycle | Mean Ints./LAgt | 15-65 |
| Simulation Run | Int. Cycles | 300-2k |
| Crossover Probability | | 0.9 |
| Mutation Probability | | 0/0.05 |
| Migrations | per cycle | 2 |
| | dominant lg | 90% |

Figure 7: The Evolutionary Simulation

ferent initial p-settings demonstrate that, if LAgts' fitness is determined by communicative success, LAgts will evolve initial p-settings which make grammar acquisition faster. That is, variant p-settings which are 'closer' to the dominant language of the population will be selected, because they incorporate default initial settings compatible with that language. If there is linguistic heterogeneity, then LAgts preferentially learn linguistic variants which are more compatible with their current initial p-settings, so there is linguistic selection for more easily learnable languages; where learnability is itself relative to the current state of initial p-settings in the population of LAgts. Furthermore, even when the rate of linguistic change is as high as is compatible with a mean 90% or greater percentage of communicative success within the population, genetic assimilation for default initial parameter settings is still observed (see Briscoe 1997, 1998a,b for further details).

These results suggest that a progressively more canalized and robust LAD will evolve once it has emerged, and that language change will occur, in part, as a result of linguistic selection for more learnable variants relative to a specific state of the LAD. The model of the LAD developed extends previous work in several ways: by relating parameter setting to a more articulated theory of UG, by developing a more effective parameter setting algorithm in terms of the default inheritance network in which the grammar is specified, and by demonstrating convergence for a larger number of parameters on a more complex grammar set. Nevertheless, there are several weaknesses to the model. Firstly, parameters are (re)set on exposure to a single relevant trigger and cannot be updated again. This makes the learning model inadequate in the face of noisy input and also makes the selection between linguistic variants with competing parameter settings depend very heavily on which variant the learner happens to be exposed to first. Secondly, the issue of the indeterminacy or ambiguity of parameter expression in triggers is largely finessed by representing triggers as determinate sequences of lexical syntactic categories. And thirdly, the restriction of language learning to parameter setting makes it difficult to explore the growth of grammatical complexity or expressiveness, and any corresponding growth of complexity in the LAD. The Bayesian view of language learning presented in the next section is designed to rectify these weaknesses.

# 4 A Bayesian LAD

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

The definition of a sentence type / trigger is relaxed from a determinate sequence of lexical syntactic categories to 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]

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 ternary-valued so we must ensure $p(PVS_i = 0) + p(PVS_i = 1) + p(PVS_i =?) = 1$. An unset $PVS_i =?$ is always assigned a prior probability of 0.5 and the rest of the mass is distributed evenly between the two values. For $PVS$s which are set to a value the unset case is assigned a probability of zero, so $p(PVS_i = 0)$ and $1 - p(PVS_i = 1)$. The probability of each such $PVS$ is taken to be independent.[5] The prior probability of a grammar, $g$ is thus the product of the probabilities of all its $PVS$s, as in (4).

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

$CON(Type, \subseteq)$ 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. 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, (4) defines a prior over $G$ which prefers succinctly describable maximally-regular and minimally-sized category sets.[6] These constraints are enough to ensure that prior probabilities will be assigned in such a way that $\sum_{(g \in G)} p(g) = 1$.

[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.

[5]This assumption of independence of $PVS$s rests partly on the semantics of the representation language in which a feature structure is a conjunction of atomic path values each specified by a single $PVS$. As with any such model assumption though, one can question whether the phenomenon modelled justifies it. In this case, the model is cognitive so a demonstration that in language there are dependencies between phenomena treated by distinct $PVS$s is at best only indirect evidence against the psychological claim that this is the appropriate cognitive model. For further discussion of probabilistic interpretation of similar representation languages see Abney (1997) and Goodman (1997).

[6]Because the parameters of variation are a set of ternary-valued $PVS$s with uniform probability assigned to unset $PVS$s, the product of these $PVS$s effectively

If we assume, as above, that $G$ is defined by P-setting($G$), then G is finite and defined by alternative ternary-valued $PVS$s (i.e. P-setting($CON(Type, \subseteq)$) for a given description of UG). However, if we also utilize a function, *Extension*(*P-setting*($CON(Type, \subseteq)$)), which defines all possible extensions to a particular grammar obtainable by adding further $PVS$s, this will expand the hypothesis space, $G$, and without further stipulation, render this space infinite. However, without practical loss of generality we could stipulate limits to *Extension* by requiring that the atomic category set be finite and the complex category set contain functors requiring no more than, say, 5 arguments.[7] This would still allow straightforward normalization of probabilities in $G$ but would require a definition of the prior probability of grammars which took account of the number, type and probability of the set of $PVS$s which defined them. We do not pursue this issue here as the experiments reported below do not require *Extension*.

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 normalized with respect to the entire category set in UG, as in (5).

$$(5) \qquad p(c) = \frac{\prod_{PVS \in CON(c,\subseteq)} p(PVS)}{\sum_{c \in CON(Type,\subseteq)} \prod_{PVS \in CON(c,\subseteq)} p(PVS)}$$

The prior probability of a category from a particular grammar can be defined similarly by restricting the normalization to specific grammars, as in (6).

$$(6) \qquad p(c \mid g) = \frac{\prod_{PVS \in CON(c,\subseteq)} p(PVS)}{\sum_{c \in g} \prod_{PVS \in CON(c,\subseteq)} p(PVS)}$$

The likelihood, $p(t_n \mid g)$, is defined as the product of the probabilities of each trigger (7).

$$(7) \qquad p(t_n \mid g) = \prod_{t \in t_n} p(t \mid g)$$

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 (8).

$$(8) \qquad p(t \mid g) = \prod_{c \in VCA(t)} p(c \mid g)$$

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 $g$).[8]

---

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.

[7] The 'highest-arity' functor in English is *bet* which (arguably) takes 5 arguments in *I bet you £5 for Red Rum to win.* The atomic category set can uncontroversially be kept finite under the assumption that the $PVS$s which define it, and morphological variation within it, are themselves finite-valued.

[8] The use of such a deficient model amounts to the (psychological) claim that learners are sensitive to the probabilities of lexical categories (see e.g. Merlo 1994)

## 4.1 Implementation

The Bayesian model has been 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. In the current implementation learning is still 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, in (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.[9] 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 LAgts.

A 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. 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 LAgts successfully parse sentence types, the posterior probability of each parameter expressed in the sentence type is updated, reinforcing the probabilities of the parameter settings required to successfully parse them (i.e. assign 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

but not the derived probabilities of phrases or clauses. Given the equivalence of probabilistic and compression perspectives exploited in minimum description length approaches (e.g. Li and Vitanyi, 1993; Osborne and Briscoe, 1997) 'likelihood' is being defined in terms of the degree of compression of the data achieved by grammar, $g$. These definitions can be straightforwardly extended to define a 'lexically-stochastic' GCG 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.

[9] Cosmides and Tooby (1996) present experimental evidence which suggests that humans utilize this type of representation in reasoning about uncertainty. This encoding of probabilities requires that we use the reciprocal of $p(PVS = 0)$ when computing priors for $g \in G$.

| 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

for enough evidence to shift a posterior probability through the $p = 0.5$ threshold before changing a setting more permanently.[10] 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 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. Principles are not updated during acquisition. However, during LAgt reproduction, principles can become default or unset parameters via crossover or mutation of prior probabilities.

Sentence types are represented 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 parameters, in order to determine the most probable grammar $p(g \in \text{P-setting}(UG))$ for these settings. The probability of a supertype $PVS$ is defined as the mean 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 is 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. the posterior probabilities associated with them are $< 0.5$), but **subjdir** is 'left' (1), then the $PVS$ for **gendir** will be set to 'right' with probability derived from the probabilities of these two inheriting subtypes. However,

[10]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. On the other hand, a first 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.

$\forall supertype_i \in \ CON(Type, \subseteq)$
$\forall PVS_j \in \ subtypes_k \ of \ supertype_i$
if
$\mid PVS_j = 1 \in subtypes_k \mid > \mid PVS_j = 0 \in \ subtypes_k \mid$
then
$p(PVS_j \ = \ 1) \ \in \ \ supertype_i \ = \ \frac{\sum p(PVS_j=1) \in \ subtypes_k}{|PVS_j=1 \in \ subtypes_k|}$
(and vice-versa)
else
if
$\frac{\sum p(PVS_j=1) \in \ subtypes_k}{|PVS_j=1 \in \ subtypes_k|} > 1 - \frac{\sum p(PVS_j=0) \in \ subtypes_k}{|PVS_j=0 \in \ subtypes_k|}$
then
$p(PVS_j \ = \ 1) \ \in \ \ supertype_i \ = \ \frac{\sum p(PVS_j=1) \in \ subtypes_k}{|PVS_j=1 \in \ subtypes_k|}$
(and vice-versa)
else
$p(PVS_j) \in \ supertype_i \ is \ 0.5$

Figure 8: Algorithm for computing posterior probabilities of supertypes

**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 ensures 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*. 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 8 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 which may or may not need to be explicitly specified to override inheritance. The learner keeps track of the relative frequency with which specific lexical categories are used to parse triggers and updates the probabilities of supertype PVSs which lie on a potential inheritance path to those categories, and thus potentially play a role in their definition. The setting of a supertype parameter is revised when a more compact, more probable grammatical description compatible with the data seen so far is possible.

The complete learning algorithm is summarized in Figure 9. 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

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

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

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 8 etc.).

Figure 9: The New Parameter Setting Algorithm

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 there are finite finite-valued parameters, relaxing this assumption and, say, making random guesses without examining the trigger encoding would still guarantee eventual convergence.

# 5 Noise, Indeterminacy and Linguistic Selection in Acquisition

Two versions of LAgt learners were predefined on the basis of the revised grammar acquisition procedure described in §4. Both learners have a minimal inherited GCUG consisting of Application with the **N** and **S** categories already present. Both can flip up to 4 parameters per trigger and differ only in terms of their initial p-settings. The unset learner was initialized with all these unset, whilst the default learner had default settings for the parameters **argorder**, **gendir**, **subjdir**, **v1** and **v2** which specify a minimal SVO right-branching grammar.[11] The initialization of p-settings is in terms of their prior probabilities, as in Table 1, in accordance with the probabilis-

[11] For a more detailed description of the effect of these five parameters in the model see Briscoe (1998, 2000a).

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

Table 2: Effectiveness of Acquisition Procedures

tic model defined in §4, so that the prior probability of supertype PVSs is calculated from the priors associated with their subtypes.

Each variant learner was tested against a non-learning adult LAgt initialized to generate one of seven full languages in the set which are close to an attested language; namely, "English" (SVO, predominantly right-branching), "Welsh" (SVOv1, mixed order), "Malagasy" (VOS, right-branching), "Tagalog" (VSO, right-branching), "Japanese" (SOV, left-branching), "German" (SOVv2, 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 is equally effective 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.

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 there 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, Muggleton, 1996).[12]

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 stands for 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, whilst 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-O 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 found. The parameter setting procedure of §2.3 only provided very limited means for a learner to recover from 'premature' incorrect setting of a parameter based on exposure to a trigger with indeterminacies of parameter expression at a relevant stage in acquisition. Recovery depended entirely on there being a non-monotonic path from hypothesized to target grammar in terms of the overriding of default inheritance. In the current framework though, parameters can, in principle, be repeatedly reset during the critical period for learning and their setting is more conservative, based on observing a consistent *series* of triggers supporting a specific setting.

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 known examples of such indeterminacies of parameter expression can be explored by exposing a learner to sentence types from the proper subset of SVO triggers which overlap with SOVv2, as well as to unambiguous SOVv2 triggers. This simulates the effect of a learner miscategorizing some 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 the target grammar). In these circumstances, the new parameter setting procedure will converge reliably to SOVv2 provided that the proportion of miscategorized triggers (to their correctly categorized counterparts) does not approach 50% with respect to any given parameter value. Conflicting triggering data will tend to delay any resetting of a parameter from its prior value, depending on its prior

[12] Since learners in this framework can have or acquire supertype settings which, by default, control the settings of subtypes not necessarily expressed in the data seen by the learner, there is a general question concerning the learnability of proper subset languages. For example, a learner might hypothesize an unseen syntactic category as a result of setting a supertype parameter on the basis of other data, and thus converge to a superset language. In fact, when exposed exclusively to a proper subset language, learners converge to that language. However, the pidgin-creole transformation (e.g. Bickerton, 1984) suggests that in special circumstances, children can converge to superset languages (see Briscoe, 2000b for detailed discussion and an account of creolisation within the framework presented here).

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

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

weighting. However, at some point the dominant trigger (or categorization of it) will determine the value of the posterior probability associated with the relevant parameter. The only circumstances in which this will not occur is if a default-valued parameter has such a strong prior weighting that the proportion of correctly categorized triggers is not great enough to override the prior before the end of 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).[13] In each case, SVO-N-PERM-COMP triggers conflict with SOVv2 and SVOv1 in two parameters: **objdir** and **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 decreases; 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 old acquisition procedure of §2.3 is also excessively sensitive to noise in the triggering input. For example, 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 target gram-

[13] 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.

mar. For example, 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)analyzable 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. 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.

A final case of potentially inconsistent triggering input occurs when the learner is exposed to sentence types deriving from more than one target 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. In the experiments, summarized in §3, involving selection between linguistic variants, random factors in the simulation, concerning the proportion of learners who happened to be exposed first to one competing variant form, tended to dominate this selection process. However, the Bayesian approach to parameter setting entails that learners will track the frequency of competing variants in terms of the posterior probabilities of the parameters associated with the variation. This accords better 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 on 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, whilst those which did reach threshold within the learning period would not.[14]

Linguistic selection can be seen as a population level 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[15] subset language corresponding to a proper subset of triggers which overlap with "German" (SOVv2) and the remainder speak "German", then learners should reliably converge to "German" when exposed to triggers from all the population, provided that SVO triggers do not approach 50% for the relevant parameters (see above). A series of sim-

[14] 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. Kirby and Hurford, 1997 for discussion and putative explanations of the critical period for language acquisition.

[15] Languages which do not have attested counterparts are given mnemonic names in which −F means the absence of parameter F and +F indicates a marked parameter value with respect to the main attested full language(s) with that canonical constituent order; for example SVO+Pleft names a language like "English" except that it uses postpositions.

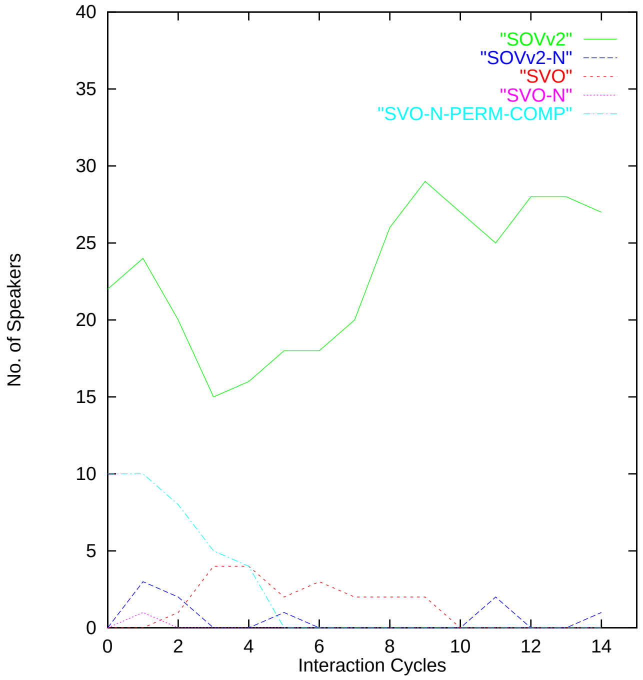


Figure 10: Linguistic Selection between Languages

ulations were run to test this prediction in which a population of 32 default (SVO) learner LAgts reproduced according to communicative success (but with no variation in initial p-settings) and 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, however, there is linguistic selection between the competing languages, where the ultimate units of selection / inheritance are competing parameter values. The linguistic selection pressure comes from two conflicting sources: learnability and expressiveness. SVO-N-PERM-COMP is easier to learn than SOVv2 because it requires the setting of fewer parameters, but it is less expressive than SOVv2 because it generates less sentence types. Therefore, in a linguistically heterogeneous environment LAgts may converge faster to SVO-N-PERM-COMP but may also communicate less successfully than SOVv2 speakers over their lifetime. However, these pressures are frequency-dependent: if enough subset language speakers emerge, the SOVv2 LAgts will be disadvantaged because the majority of their interactions will be with subset language speakers. Figure 10 plots the languages spoken across interaction cycles for a population initialized with 10 SVO-N-PERM-COMP and 22 SOVv2 LAgts. This plot is typical: the SVO-N-PERM-COMP speakers dwindle rapidly, a few SVO superset language learners emerge, but also disappear as they are outnumbered by SOVv2 speakers, 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). However, in one case, the population converged on a full SVO language after 11 interaction cycles, and in many others a few full SVO language speakers emerged briefly during the run. However, in an identical series of runs initialized with 16 SVO-N-PERM-ASSOC and 16 SOVv2 speakers, an equally clear opposing result was obtained: populations always converged on the subset language within 15 interaction cycles.[16]

These experiments illustrate a number of phenomena. Firstly, linguistic selection, even when LAgt fitness is simply based on communicative success, is a complex process with strong potential interactions between the proportions of speakers of competing variants, their (relative) learnability, the (relative) expressiveness of languages learnt, and their degree of overlap with other languages extant in the population. However, linguistic selection is now a more predictable ('damped') process because learners rely, on average, on the relative frequency with which competing parameter settings are exemplified in the arena of language use. Finally, in a heterogeneous linguistic environment it is not the case that learners will only converge to the languages directly exemplified in the arena of use – in this case, SVO-N-PERM-COMP or SOVv2 and their subset languages. Learning LAgts can also converge to grammars incorporating mixtures of the exemplifying languages and also to grammars of superset languages. For example, the initial speakers of SVO full languages in the above simulations have acquired languages containing sentence types (and associated parameter settings) not exemplified in the arena of use at all. This possibility becomes increasingly likely in very heterogeneous environments because some parameters will, in effect, be set by default on the basis of the posterior or even prior probabilities of more general (supertype) parameters. Briscoe (2000b) presents an account of the pidgin-creole transformation (e.g. Bickerton, 1984) which exploits this property of the Bayesian account of language acquisition.

# 6 Coevolution of the LAD and of Language

The acquisition experiments of §5 demonstrated the effectiveness of the new parameter setting procedure with some prior settings on some full languages, even in the presence of noise and indeterminacy of parameter expression, whilst the evolutionary simulations of populations of default (SVO) learning LAgts demonstrated linguistic selection on the basis of learnability without any variation at the genetic, initial p-setting level. Introducing variation in the initial p-settings of LAgts, allows for the possibility of coevolution of LAgt's acquisition procedures, at the same time as languages or their associated grammars are themselves being selected (see Briscoe 1997, 1998a,b for extensive experiments of this kind with the old acquisition procedure).

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 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 moves a p-setting from its existing type (absolute

[16] These experiments still utilize LAgt fitness to determine reproductive success. Briscoe (1998a,b) reports experiments which demonstrate linguistic selection for learnability with random reproduction of LAgts.

principle, default or unset parameter) and initial setting (1, 0, ?) to a new type and/or initial setting with equal probability. Thus, no 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 (1997, 1998a,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). However, this selection for better 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). 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*. 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% 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 §5. Simulation runs lasted for 2000 interaction cycles (about 500 generations of LAgts). 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 p-settings consistent with the dominant settings already extant in the population; that is, migrations are designed to introduce linguistic, not genetic, variation.

The mean increase in the proportion of default parameters in all such runs was 46.7%. The mean increase in principles was 3.8%. These accounted for an overall decrease of 50.6% in the proportion of unset parameters in the initial p-settings of LAgts. Figure 11 shows the relative proportions of default parameters, unset parameters and principles for one such 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 is truncated, though there are fluctuations caused by migrations or by an increased proportion of learners. These results, which are replicated for different languages, different learners, and so forth (see Briscoe, 1997, 1998a,b) 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.

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

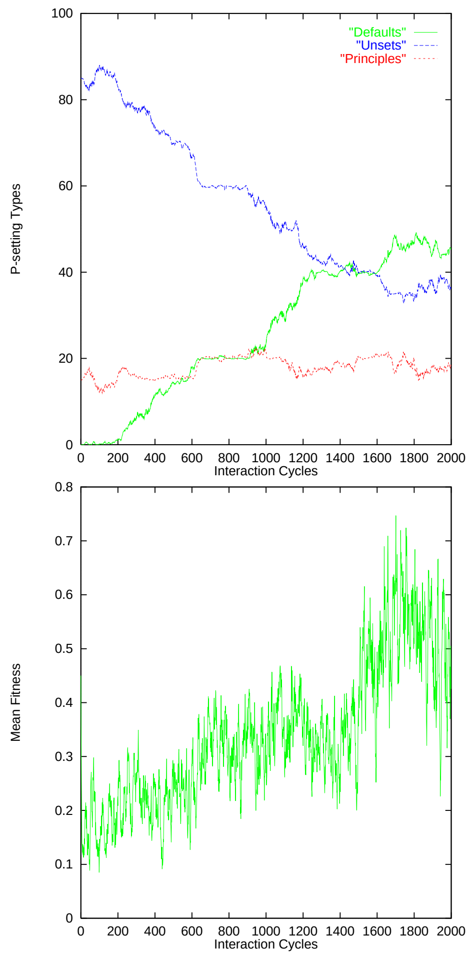


Figure 11: Proportions of p-setting types and Mean fitness

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 grammar set, and relative speed at which both can change, it seems likely that the simulation model massively *underestimates* the size of the potential space of grammatical possibilities. 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, whilst 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), whilst 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 (Briscoe, 1998a,b) when linguistic variation is maintained.

# 7 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 a GCG, 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 potentially noisy or indeterminate input. Human language learners in special circumstances converge to grammars different from that of the preceding generation (e.g. Bickerton, 1984; Clark and Roberts, 1993). The model proposed has the same behaviour, though further work is needed to characterize the exact circumstances under which such behaviour will occur and whether this appears realistic with respect to attested cases of major and rapid grammatical change. Nevertheless, the need for such behaviour in a (psycho)linguistically realistic model of language learning casts doubt on the relevance of learnability results, in general. Gold's 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. Gibson and Wexler, 1994; Muggleton, 1996; Niyogi and Berwick, 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. 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 scenario developed here 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, as in the model presented, learners may extend the grammatical specification beyond that given in UG by adding path value specifications to the default inheritance network to create new grammatical categories. An implementation of this aspect of the model is a priority since it would allow such innovations to be incorporated into the specification of UG via genetic assimilation, and this in turn would underpin a better evolutionary account of the development and refinement of the LAD.

The model of a LAgt assumes the existence of a minimal LAD since agents come equipped with a (potentially-minimal) UG, associated learning procedure, and parser. Thus the simulations demonstrate that an effective, robust but biased acquisition procedure specialized for/on specific grammars could emerge by genetic assimilation. However, they do not directly address the question of how such an embryonic LAD 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 LAD 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 a GCG's 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 reasoning. Worden (1995b) argues that Bayesian reasoning 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 reasoning, then evolution could construct a minimal LAD 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. Crucially though, the mutations producing a minimal LAD would only be selected for in an environment where more efficient and robust acquisition and extension of a pre-existing (proto)language conferred benefit (e.g. Kirby, 1998).

More generally, the model behind the simulations suggests a dynamic systems framework for the study of communication systems, and human language in particular, which both incorporates the insights gained from formalizing a language as a static well-formed stringset (Chomsky, 1957) and extends them by embedding this model in an evolving population of distributed language agents, yielding a characterization of language itself as an adaptive system. The broad practical implication of this framework for automated natural language processing is that development of static hand-coded systems should be replaced by development of autonomous software agents capable of adapting to their linguistic environment. Nevertheless, considerable further work will be required to translate the theoretical model of language acquisition developed here into such a practical engineering tool.

## Acknowledgements

I would like to thank the audience at Machine Intelligence 16 for helpful feedback, Stephen Muggleton for the invitation to participate, and Miles Osborne, Aline Villavicencio and Ben Waldron for comments on earlier drafts, which have considerably improved this one. All remaining errors and obscurities are my responsibility.

## References

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

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

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.

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,* Las Cruces, NM, pp. 147–154.

Briscoe, E.J. (1987) *Modelling Human Speech Comprehension: A Computational Approach,* Ellis Horwood, Chichester / John Wiley, New York.

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

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

Briscoe, E.J. (2000a, in press) 'Grammatical Acquisition: Coevolution of Language and the Language Acquisition Device', *Language.*

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

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.

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

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

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

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. 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.

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 (1996) 'Order-independent and persistent default unification', *Linguistics and Philosophy, vol.19.1,* 1–89.

Lascarides, A. and Copestake A.A. (1999) 'Order-independent typed default unification', *Computational Linguistics, vol.25.1,* 55–106.

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.

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

Pinker, S. (1994) *The Language Instinct,* Morrow, New York.

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.

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

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..

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

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.

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.