

A Selectionist Theory of Language Acquisition

Charles D. Yang*
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
charles@ai.mit.edu

Abstract

This paper argues that developmental patterns in child language be taken seriously in computational models of language acquisition, and proposes a formal theory that meets this criterion. We first present developmental facts that are problematic for statistical learning approaches which assume no prior knowledge of grammar, and for traditional learnability models which assume the learner moves from one UG-defined grammar to another. In contrast, we view language acquisition as a population of grammars associated with “weights”, that compete in a Darwinian selectionist process. Selection is made possible by the *variational* properties of individual grammars; specifically, their differential compatibility with the primary linguistic data in the environment. In addition to a convergence proof, we present empirical evidence in child language development, that a learner is best modeled as multiple grammars in co-existence and competition.

1 Learnability and Development

A central issue in linguistics and cognitive science is the problem of language acquisition: How does a human child come to acquire her language with such ease, yet without high computational power or favorable learning conditions? It is evident that any adequate model of language acquisition must meet the following empirical conditions:

- *Learnability*: such a model must converge to the target grammar used in the learner’s environment, under plausible assumptions about the learner’s computational machinery, the nature of the input data, sample size, and so on.
- *Developmental compatibility*: the learner modeled in such a theory must exhibit behaviors that are analogous to the actual course of language development (Pinker, 1979).

* I would like to thank Julie Legate, Sam Gutmann, Bob Berwick, Noam Chomsky, John Frampton, and John Goldsmith for comments and discussion. This work is supported by an NSF graduate fellowship.

It is worth noting that the developmental compatibility condition has been largely ignored in the formal studies of language acquisition. In the rest of this section, I show that if this condition is taken seriously, previous models of language acquisition have difficulties explaining certain developmental facts in child language.

1.1 Against Statistical Learning

An empiricist approach to language acquisition has (re)gained popularity in computational linguistics and cognitive science; see Stolcke (1994), Charniak (1995), Klavans and Resnik (1996), de Marcken (1996), Bates and Elman (1996), Seidenberg (1997), among numerous others. The child is viewed as an inductive and “generalized” data processor such as a neural network, designed to derive structural regularities from the statistical distribution of patterns in the input data without *prior* (innate) specific knowledge of natural language. Most concrete proposals of statistical learning employ expensive and specific computational procedures such as compression, Bayesian inferences, propagation of learning errors, and usually require a large corpus of (sometimes pre-processed) data. These properties immediately challenge the psychological plausibility of the statistical learning approach. In the present discussion, however, we are not concerned with this but simply grant that someday, someone might devise a statistical learning scheme that is psychologically plausible and also succeeds in converging to the target language. We show that even if such a scheme were possible, it would still face serious challenges from the important but often ignored requirement of developmental compatibility.

One of the most significant findings in child language research of the past decade is that different aspects of syntactic knowledge are learned at different rates. For example, consider the placement of finite verb in French, where inflected verbs precede negation and adverbs:

Jean *voit* souvent/pas Marie.
Jean *sees* often/not Marie.

This property of French is mastered as early as

the 20th month, as evidenced by the extreme rarity of incorrect verb placement in child speech (Pierce, 1992). In contrast, some aspects of language are acquired relatively late. For example, the requirement of using a sentential subject is not mastered by English children until as late as the 36th month (Valian, 1991), when English children stop producing a significant number of subjectless sentences.

When we examine the adult speech to children (transcribed in the CHILDES corpus; MacWhinney and Snow, 1985), we find that more than 90% of English input sentences contain an overt subject, whereas only 7-8% of all French input sentences contain an inflected verb followed by negation/adverb. A statistical learner, one which builds knowledge purely on the basis of the distribution of the input data, predicts that English obligatory subject use should be learned (much) earlier than French verb placement – exactly the opposite of the actual findings in child language.

Further evidence against statistical learning comes from the Root Infinitive (RI) stage (Wexler, 1994; *inter alia*) in children acquiring certain languages. Children in the RI stage produce a large number of sentences where matrix verbs are not finite – ungrammatical in adult language and thus appearing infrequently in the primary linguistic data if at all. It is not clear how a statistical learner will induce *non-existent* patterns from the training corpus. In addition, in the acquisition of verb-second (V2) in Germanic grammars, it is known (e.g. Haegeman, 1994) that at an early stage, children use a large proportion (50%) of verb-initial (V1) sentences, a marked pattern that appears only sparsely in adult speech. Again, an inductive learner purely driven by corpus data has no explanation for these disparities between child and adult languages.

Empirical evidence as such poses a serious problem for the statistical learning approach. It seems a mistake to view language acquisition as an inductive procedure that constructs linguistic knowledge, directly and exclusively, from the distributions of input data.

1.2 The Transformational Approach

Another leading approach to language acquisition, largely in the tradition of generative linguistics, is motivated by the fact that although child language is different from adult language, it is different in highly *restrictive* ways. Given the input to the child, there are logically possible and computationally simple inductive rules to describe the data that are never attested in child language. Consider the following well-known example. Forming a question in English involves inversion of the auxiliary verb and the subject:

Is the man *t* tall?

where “is” has been fronted from the position *t*, the position it assumes in a declarative sentence. A possible inductive rule to describe the above sentence is this: front the *first* auxiliary verb in the sentence. This rule, though logically possible and computationally simple, is never attested in child language (Chomsky, 1975; Crain and Nakayama, 1987; Crain, 1991): that is, children are never seen to produce sentences like:

* Is the cat that the dog *t* chasing is scared?

where the first auxiliary is fronted (the first “is”), instead of the auxiliary following the subject of the sentence (here, the second “is” in the sentence).

Acquisition findings like these lead linguists to postulate that the human language capacity is constrained in a finite *prior* space, the Universal Grammar (UG). Previous models of language acquisition in the UG framework (Wexler and Culicover, 1980; Berwick, 1985; Gibson and Wexler, 1994) are *transformational*, borrowing a term from evolution (Lewontin, 1983), in the sense that the learner moves from one hypothesis/grammar to another as input sentences are processed.¹ Learnability results can be obtained for some psychologically plausible algorithms (Niyogi and Berwick, 1996). However, the developmental compatibility condition still poses serious problems.

Since at any time the state of the learner is identified with a particular grammar defined by UG, it is hard to explain (a) the inconsistent patterns in child language, which cannot be described by any single adult grammar (e.g. Brown, 1973); and (b) the smoothness of language development (e.g. Pinker, 1984; Valiant, 1991; *inter alia*), whereby the child gradually converges to the target grammar, rather than the abrupt jumps that would be expected from binary changes in hypotheses/grammars.

Having noted the inadequacies of the previous approaches to language acquisition, we will propose a theory that aims to meet language learnability and language development conditions simultaneously. Our theory draws inspirations from Darwinian evolutionary biology.

2 A Selectionist Model of Language Acquisition

2.1 The Dynamics of Darwinian Evolution

Essential to Darwinian evolution is the concept of variational thinking (Lewontin, 1983). First, differ-

¹Note that the transformational approach is not restricted to UG-based models; for example, Brill’s influential work (1993) is a corpus-based model which successively revises a set of syntactic rules upon presentation of partially bracketed sentences. Note that however, the state of the learning system at any time is still a single set of rules, that is, a single “grammar”.

ences among individuals are viewed as “real”, as opposed to deviant from some idealized archetypes, as in pre-Darwinian thinking. Second, such differences result in variance in operative functions among individuals in a population, thus allowing forces of evolution such as natural selection to operate. Evolutionary changes are therefore changes in the *distribution* of variant individuals in the population. This contrasts with Lamarckian *transformational* thinking, in which individuals *themselves* undergo direct changes (transformations) (Lewontin, 1983).

2.2 A population of grammars

Learning, including language acquisition, can be characterized as a sequence of states in which the learner moves from one state to another. Transformational models of language acquisition identify the state of the learner as a *single* grammar/hypothesis. As noted in section 1, this makes difficult to explain the inconsistency in child language and the smoothness of language development.

We propose that the learner be modeled as a *population* of “grammars”, the set of all principled language variations made available by the biological endowment of the human language faculty. Each grammar G_i is associated with a *weight* p_i , $0 \leq p_i \leq 1$, and $\sum p_i = 1$. In a linguistic environment E , the weight $p_i(E, t)$ is a function of E and the time variable t , the time since the onset of language acquisition. We say that

Definition: Learning *converges* if

$$\forall \epsilon, 0 < \epsilon < 1, \forall G_i, |p_i(E, t+1) - p_i(E, t)| < \epsilon$$

That is, learning converges when the composition and distribution of the grammar population are stabilized. Particularly, in a monolingual environment E_T in which a target grammar T is used, we say that learning *converges to* T if $\lim_{t \rightarrow \infty} p_T(E_T, t) = 1$.

2.3 A Learning Algorithm

Write $E \rightarrow s$ to indicate that a sentence s is an utterance in the linguistic environment E . Write $s \in G$ if a grammar G can analyze s , which, in a narrow sense, is parsability (Wexler and Culicover, 1980; Berwick, 1985). Suppose that there are altogether N grammars in the population. For simplicity, write p_i for $p_i(E, t)$ at time t , and p'_i for $p_i(E, t+1)$ at time $t+1$. Learning takes place as follows:

The Algorithm:

Given an input sentence s , the child with the probability p_i , selects a grammar G_i

- if $s \in G_i$ $\begin{cases} p'_i = p_i + \gamma(1 - p_i) \\ p'_j = (1 - \gamma)p_j \end{cases}$ if $j \neq i$
- if $s \notin G_i$ $\begin{cases} p'_i = (1 - \gamma)p_i \\ p'_j = \frac{\gamma}{N-1} + (1 - \gamma)p_j \end{cases}$ if $j \neq i$

Comment: The algorithm is the Linear reward-penalty (L_{R-P}) scheme (Bush and Mostellar, 1958), one of the earliest and most extensively studied stochastic algorithms in the psychology of learning. It is real-time and on-line, and thus reflects the rather limited computational capacity of the child language learner, by avoiding sophisticated data processing and the need for a large memory to store previously seen examples. Many variants and generalizations of this scheme are studied in Atkinson et al. (1965), and their thorough mathematical treatments can be found in Narendra and Thathachar (1989).

The algorithm operates in a selectionist manner: grammars that succeed in analyzing input sentences are rewarded, and those that fail are punished. In addition to the psychological evidence for such a scheme in animal and human learning, there is neurological evidence (Hubel and Wiesel, 1962; Changeux, 1983; Edelman, 1987; *inter alia*) that the development of neural substrate is guided by the exposure to specific stimulus in the environment in a Darwinian selectionist fashion.

2.4 A Convergence Proof

For simplicity but without loss of generality, assume that there are two grammars ($N = 2$), the target grammar T_1 and a *pretender* T_2 . The results presented here generalize to the N -grammar case; see Narendra and Thathachar (1989).

Definition: The *penalty probability* of grammar T_i in a linguistic environment E is

$$c_i = \Pr(s \notin T_i \mid E \rightarrow s)$$

In other words, c_i represents the probability that the grammar T_i fails to analyze an incoming sentence s and gets punished as a result. Notice that the penalty probability, essentially a fitness measure of individual grammars, is an *intrinsic* property of a UG-defined grammar relative to a particular linguistic environment E , determined by the distributional patterns of linguistic expressions in E . It is not explicitly computed, as in (Clark, 1992) which uses the Genetic Algorithm (GA).²

The main result is as follows:

Theorem:

$$\lim_{t \rightarrow \infty} p_1(t) = \frac{c_2}{c_1 + c_2} \quad \text{if } |1 - \gamma(c_1 + c_2)| < 1 \quad (1)$$

Proof sketch: Computing $E[p_1(t+1) \mid p_1(t)]$ as a function of $p_1(t)$ and taking expectations on both

²Clark's model and the present one share an important feature: the outcome of acquisition is determined by the differential compatibilities of individual grammars. The choice of the GA introduces various psychological and linguistic assumptions that can not be justified; see Drescher (1999) and Yang (1999). Furthermore, no formal proof of convergence is given.

sides give

$$E[p_1(t+1)] = [1 - \gamma(c_1 + c_2)]E[p_1(t)] + \gamma c_2 \quad (2)$$

Solving [2] yields [1].

Comment 1: It is easy to see that $p_1 \rightarrow 1$ (and $p_2 \rightarrow 0$) when $c_1 = 0$ and $c_2 > 0$; that is, the learner converges to the target grammar T_1 , which has a penalty probability of 0, by definition, in a monolingual environment. Learning is *robust*. Suppose that there is a small amount of noise in the input, i.e. sentences such as speaker errors which are not compatible with the target grammar. Then $c_1 > 0$. If $c_1 \ll c_2$, convergence to T_1 is still ensured by [1].

Consider a non-uniform linguistic environment in which the linguistic evidence does not unambiguously identify any single grammar; an example of this is a population in contact with two languages (grammars), say, T_1 and T_2 . Since $c_1 > 0$ and $c_2 > 0$, [1] entails that p_1 and p_2 reach a stable equilibrium at the end of language acquisition; that is, language learners are essentially bi-lingual speakers as a result of language contact. Kroch (1989) and his colleagues have argued convincingly that this is what happened in many cases of diachronic change. In Yang (1999), we have been able to extend the acquisition model to a population of learners, and formalize Kroch's idea of grammar competition over time.

Comment 2: In the present model, one can directly measure the rate of *change* in the weight of the target grammar, and compare with developmental findings. Suppose T_1 is the target grammar, hence $c_1 = 0$. The expected increase of p_1 , Δp_1 is computed as follows:

$$E[\Delta p_1] = c_2 p_1 p_2 \quad (3)$$

Since $p_2 = 1 - p_1$, Δp_1 [3] is obviously a quadratic function of $p_1(t)$. Hence, the growth of p_1 will produce the familiar S-shape curve familiar in the psychology of learning. There is evidence for an S-shape pattern in child language development (Clahsen, 1986; Wijnen, 1999; *inter alia*), which, if true, suggests that a selectionist learning algorithm adopted here might indeed be what the child learner employs.

2.5 Unambiguous Evidence is Unnecessary

One way to ensure convergence is to assume the existence of *unambiguous evidence* (cf. Fodor, 1998): sentences that are *only* compatible with the target grammar but not with any other grammar. Unambiguous evidence is, however, not necessary for the proposed model to converge. It follows from the theorem [1] that even if no evidence can unambiguously identify the target grammar from its competitors, it is still possible to ensure convergence as long as all competing grammars fail on *some* proportion of input sentences; i.e. they all have positive penalty

probabilities. Consider the acquisition of the target, a German V2 grammar, in a population of grammars below:

1. German: SVO, OVS, XVSO
2. English: SVO, XSVO
3. Irish: VSO, XVSO
4. Hixkaryana: OVS, XOVS

We have used X to denote non-argument categories such as adverbs, adjuncts, etc., which can quite freely appear in sentence-initial positions. Note that none of the patterns in (1) could conclusively distinguish German from the other three grammars. Thus, no unambiguous evidence appears to exist. However, if SVO, OVS, and XVSO patterns appear in the input data at positive frequencies, the German grammar has a higher overall "fitness value" than other grammars by the virtue of being compatible with *all* input sentences. As a result, German will eventually eliminate competing grammars.

2.6 Learning in a Parametric Space

Suppose that natural language grammars vary in a parametric space, as cross-linguistic studies suggest.³ We can then study the dynamical behaviors of grammar *classes* that are defined in these parametric dimensions. Following (Clark, 1992), we say that a sentence s *expresses* a parameter α if a grammar must have set α to some definite value in order to assign a well-formed representation to s . Convergence to the target value of α can be ensured by the existence of evidence (s) defined in the sense of parameter expression. The convergence to a single grammar can then be viewed as the intersection of parametric grammar classes, converging in parallel to the target values of their respective parameters.

3 Some Developmental Predictions

The present model makes two predictions that cannot be made in the standard transformational theories of acquisition:

1. As the target gradually rises to dominance, the child entertains a number of co-existing grammars. This will be reflected in distributional patterns of child language, under the null hypothesis that the grammatical knowledge (in our model, the population of grammars and their respective weights) used in production is that used in analyzing linguistic evidence. For grammatical phenomena that are acquired relatively late, child language consists of the output of more than one grammar.

³Although different theories of grammar, e.g. GB, HPSG, LFG, TAG, have different ways of instantiating this idea.

2. Other things being equal, the rate of development is determined by the penalty probabilities of competing grammars relative to the input data in the linguistic environment [3].

In this paper, we present longitudinal evidence concerning the prediction in (2).⁴ To evaluate developmental predictions, we must estimate the the penalty probabilities of the competing grammars in a particular linguistic environment. Here we examine the developmental rate of French verb placement, an early acquisition (Pierce, 1992), that of English subject use, a late acquisition (Valian, 1991), that of Dutch V2 parameter, also a late acquisition (Haegeman, 1994).

Using the idea of parameter expression (section 2.6), we estimate the frequency of sentences that unambiguously identify the target value of a parameter. For example, sentences that contain finite verbs preceding adverb or negation (“Jean *voit* souvent/pas Marie”) are unambiguous indication for the [+] value of the verb raising parameter. A grammar with the [-] value for this parameter is incompatible with such sentences and if probabilistically selected for the learner for grammatical analysis, will be punished as a result. Based on the CHILDES corpus, we estimate that such sentences constitute 8% of all French adult utterances to children. This suggests that unambiguous evidence as 8% of all input data is sufficient for a very early acquisition: in this case, the target value of the verb-raising parameter is correctly set. We therefore have a direct explanation of Brown’s (1973) observation that in the acquisition of fixed word order languages such as English, word order errors are “triflingly few”. For example, English children are never to seen to produce word order variations other than SVO, the target grammar, nor do they fail to front Wh-words in question formation. Virtually all English sentences display rigid word order, e.g. verb almost always (immediately) precedes object, which give a very high (perhaps close to 100%, far greater than 8%, which is sufficient for a very early acquisition as in the case of French verb raising) rate of unambiguous evidence, sufficient to drive out other word order grammars very early on.

Consider then the acquisition of the subject parameter in English, which requires a sentential subject. Languages like Italian, Spanish, and Chinese, on the other hand, have the option of dropping the subject. Therefore, sentences with an overt subject are not necessarily useful in distinguishing English

⁴In Yang (1999), we show that a child learner, en route to her target grammar, entertains multiple grammars. For example, a significant portion of English child language shows characteristics of a topic-drop optional subject grammar like Chinese, before they learn that subject use in English is obligatory at around the 3rd birthday.

from optional subject languages.⁵ However, there exists a certain type of English sentence that *is* indicative (Hyams, 1986):

There is a man in the room.
Are there toys on the floor?

The subject of these sentences is “there”, a non-referential lexical item that is present for purely structural reasons – to satisfy the requirement in English that the pre-verbal subject position must be filled. Optional subject languages do not have this requirement, and do not have expletive-subject sentences. Expletive sentences therefore express the [+] value of the subject parameter. Based on the CHILDES corpus, we estimate that expletive sentences constitute 1% of all English adult utterances to children.

Note that before the learner eliminates optional subject grammars on the cumulative basis of expletive sentences, she has probabilistic access to multiple grammars. This is fundamentally different from stochastic grammar models, in which the learner has probabilistic access to *generative rules*. A stochastic grammar is not a developmentally adequate model of language acquisition. As discussed in section 1.1, more than 90% of English sentences contain a subject: a stochastic grammar model will overwhelmingly bias toward the rule that generates a subject. English children, however, go through long period of subject drop. In the present model, child subject drop is interpreted as the presence of the true optional subject *grammar*, in co-existence with the obligatory subject grammar.

Lastly, we consider the setting of the Dutch V2 parameter. As noted in section 2.5, there appears to no unambiguous evidence for the [+] value of the V2 parameter: SVO, VSO, and OVS grammars, members of the [-V2] class, are each compatible with certain proportions of expressions produced by the target V2 grammar. However, observe that despite of its compatibility with with *some* input patterns, an OVS grammar can not survive long in the population of competing grammars. This is because an OVS grammar has an extremely high penalty probability. Examination of CHILDES shows that OVS patterns consist of only 1.3% of all input sentences to children, whereas SVO patterns constitute about 65% of all utterances, and XVSO, about 34%. Therefore, only SVO and VSO grammar, members of the [-V2] class, are “contenders” alongside the (target) V2 grammar, by the virtue of being compatible with significant portions of input data. But notice that OVS patterns do penalize *both* SVO and VSO grammars, and are only compatible with the [+V2] gram-

⁵Notice that this presupposes the child’s prior knowledge of and access to both obligatory and optional subject grammars.

mars. Therefore, OVS patterns are *effectively* unambiguous evidence (among the contenders) for the V2 parameter, which eventually drive SVO and VSO grammars out of the population.

In the selectionist model, the rarity of OVS sentences predicts that the acquisition of the V2 parameter in Dutch is a relatively late phenomenon. Furthermore, because the frequency (1.3%) of Dutch OVS sentences is comparable to the frequency (1%) of English expletive sentences, we expect that Dutch V2 grammar is successfully acquired roughly at the same time when English children have adult-level subject use (around age 3; Valian, 1991). Although I am not aware of any report on the timing of the correct setting of the Dutch V2 parameter, there is evidence in the acquisition of German, a similar language, that children are considered to have successfully acquired V2 by the 36-39th month (Clahsen, 1986). Under the model developed here, this is not an coincidence.

4 Conclusion

To capitulate, this paper first argues that considerations of language development must be taken seriously to evaluate computational models of language acquisition. Once we do so, both statistical learning approaches and traditional UG-based learnability studies are empirically inadequate. We proposed an alternative model which views language acquisition as a selectionist process in which grammars form a population and compete to match linguistic expressions present in the environment. The course and outcome of acquisition are determined by the relative compatibilities of the grammars with input data; such compatibilities, expressed in penalty probabilities and unambiguous evidence, are quantifiable and empirically testable, allowing us to make direct predictions about language development.

The biologically endowed linguistic knowledge enables the learner to go beyond unanalyzed distributional properties of the input data. We argued in section 1.1 that it is a mistake to model language acquisition as directly learning the probabilistic distribution of the linguistic data. Rather, language acquisition is guided by particular input evidence that serves to disambiguate the target grammar from the competing grammars. The *ability* to use such evidence for grammar selection is based on the learner's linguistic knowledge. Once such knowledge is assumed, the actual *process* of language acquisition is no more remarkable than generic psychological models of learning. The selectionist theory, if correct, show an example of the interaction between domain-specific knowledge and domain-neutral mechanisms, which combine to explain properties of language and cognition.

References

- Atkinson, R., G. Bower, and E. Crothers. (1965). *An Introduction to Mathematical Learning Theory*. New York: Wiley.
- Bates, E. and J. Elman. (1996). Learning rediscovered: A perspective on Saffran, Aslin, and Newport. *Science* 274: 5294.
- Berwick, R. (1985). *The acquisition of syntactic knowledge*. Cambridge, MA: MIT Press.
- Brill, E. (1993). Automatic grammar induction and parsing free text: a transformation-based approach. ACL Annual Meeting.
- Brown, R. (1973). *A first language*. Cambridge, MA: Harvard University Press.
- Bush, R. and F. Mostellar. *Stochastic models for learning*. New York: Wiley.
- Charniak, E. (1995). *Statistical language learning*. Cambridge, MA: MIT Press.
- Chomsky, N. (1975). *Reflections on language*. New York: Pantheon.
- Changeux, J.-P. (1983). *L'Homme Neuronal*. Paris: Fayard.
- Clahsen, H. (1986). Verbal inflections in German child language: Acquisition of agreement markings and the functions they encode. *Linguistics* 24: 79-121.
- Clark, R. (1992). The selection of syntactic knowledge. *Language Acquisition* 2: 83-149.
- Crain, S. and M. Nakayama (1987). Structure dependency in grammar formation. *Language* 63: 522-543.
- Dresher, E. (1999). Charting the learning path: cues to parameter setting. *Linguistic Inquiry* 30: 27-67.
- Edelman, G. (1987). *Neural Darwinism: The theory of neuronal group selection*. New York: Basic Books.
- Fodor, J. D. (1998). Unambiguous triggers. *Linguistic Inquiry* 29: 1-36.
- Gibson, E. and K. Wexler (1994). Triggers. *Linguistic Inquiry* 25: 355-407.
- Haegeman, L. (1994). Root infinitives, clitics, and truncated structures. *Language Acquisition*.
- Hubel, D. and T. Wiesel (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology* 160: 106-54.
- Hyams, N. (1986) *Language acquisition and the theory of parameters*. Reidel: Dordrecht.
- Klavins, J. and P. Resnik (eds.) (1996). *The balancing act*. Cambridge, MA: MIT Press.
- Kroch, A. (1989). Reflexes of grammar in patterns of language change. *Language variation and change* 1: 199-244.
- Lewontin, R. (1983). The organism as the subject and object of evolution. *Scientia* 118: 65-82.
- de Marcken, C. (1996). Unsupervised language acquisition. Ph.D. dissertation, MIT.

- MacWhinney, B. and C. Snow (1985). The Child Language Data Exchange System. *Journal of Child Language* 12, 271-296.
- Narendra, K. and M. Thathachar (1989). *Learning automata*. Englewood Cliffs, NJ: Prentice Hall.
- Niyogi, P. and R. Berwick (1996). A language learning model for finite parameter space. *Cognition* 61: 162-193.
- Pierce, A. (1992). *Language acquisition and syntactic theory: a comparative analysis of French and English child grammar*. Boston: Kluwer.
- Pinker, S. (1979). Formal models of language learning. *Cognition* 7: 217-283.
- Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.
- Seidenberg, M. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science* 275: 1599-1604.
- Stolcke, A. (1994) Bayesian Learning of Probabilistic Language Models. Ph.D. thesis, University of California at Berkeley, Berkeley, CA.
- Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. *Cognition* 40: 21-82.
- Wexler, K. (1994). Optional infinitives, head movement, and the economy of derivation in child language. In Lightfoot, D. and N. Hornstein (eds.) *Verb movement*. Cambridge: Cambridge University Press.
- Wexler, K. and P. Culicover (1980). *Formal principles of language acquisition*. Cambridge, MA: MIT Press.
- Wijnen, F. (1999). Verb placement in Dutch child language: A longitudinal analysis. Ms. University of Utrecht.
- Yang, C. (1999). The variational dynamics of natural language: Acquisition and use. Technical report, MIT AI Lab.