A MARKOV LANGUAGE LEARNING MODEL FOR FINITE PARAMETER SPACES

Partha Niyogi and **Robert C. Berwick**
Center for Biological and Computational Learning
Massachusetts Institute of Technology
E25-201
Cambridge, MA 02139, USA
Internet: pn@ai.mit.edu, berwick@ai.mit.edu

Abstract

This paper shows how to *formally* characterize language learning in a finite parameter space as a *Markov structure*. Important new language learning results follow directly: explicitly calculated sample complexity learning times under different input distribution assumptions (including CHILDES database language input) and learning regimes. We also briefly describe a new way to formally model (rapid) *diachronic* syntax change.

BACKGROUND MOTIVATION: TRIGGERS AND LANGUAGE ACQUISITION

Recently, several researchers, including Gibson and Wexler (1994), henceforth GW, Dresher and Kaye (1990); and Clark and Roberts (1993) have modeled language learning in a (finite) space whose grammars are characterized by a finite number of parameters or n-length Boolean-valued vectors. Many current linguistic theories now employ such parametric models explicitly or in spirit, including Lexical-Functional Grammar and versions of HPSG, besides GB variants.

With all such models, key questions about sample complexity, convergence time, and alternative modeling assumptions are difficult to assess without a precise mathematical formalization. Previous research has usually addressed only the question of convergence in the limit without probing the equally important question of sample complexity: it is of not much use that a learner can acquire a language if sample complexity is extraordinarily high, hence psychologically implausible. This remains a relatively undeveloped area of language learning theory. The current paper aims to fill that gap. We choose as a starting point the GW Triggering Learning Algorithm (TLA). Our central result is that the performance of this algorithm and others like it is *completely* modeled by a Markov chain. We explore the basic computational consequences of this, including some surprising results about sample complexity and convergence time, the dominance of random walk over gradient ascent, and the applicability of these results to actual child language acquisition and possibly language change.

Background. Following Gold (1967) the basic framework is that of identification in the limit. We assume some familiarity with Gold's assumptions. The learner receives an (infinite) sequence of (positive) example sentences from some target language. After each, the learner either (i) stays in the same state; or (ii) moves to a new state (change its parameter settings). If after some finite number of examples the learner converges to the correct target language and never changes its guess, then it has correctly identified the target language in the limit; otherwise, it fails.

In the GW model (and others) the learner obeys two additional fundamental constraints: (1) the *single-value constraint*—the learner can change only 1 parameter value each step; and (2) the *greediness constraint*—if the learner is given a positive example it cannot recognize and changes one parameter value, finding that it can accept the example, then the learner retains that new value. The TLA essentially simulates this; see Gibson and Wexler (1994) for details.

THE MARKOV FORMULATION

Previous parameter models leave open key questions addressable by a more precise formalization as a Markov chain. The correspondence is direct. Each point i in the Markov space is a possible parameter setting. Transitions between states stand for probabilities b that the learner will move from hypothesis state i to state j. As we show below, given a distribution over $L(G)$, we can calculate the actual b's themselves. Thus, we can picture the TLA learning space as a directed, labeled graph V with 2^n vertices. See figure 1 for an example in a 3-parameter system.[1] We can now use Markov theory to describe TLA parameter spaces, as in Isaacson and

[1] GW construct an identical transition diagram in the description of their computer program for calculating local maxima. However, this diagram is not explicitly presented as a Markov structure and does not include transition probabilities.

Madsen (1976). By the single value hypothesis, the system can only move 1 Hamming bit at a time, either *toward* the target language or 1 bit away. Surface strings can force the learner from one hypothesis state to another. For instance, if state i corresponds to a grammar that generates a language that is a proper subset of another grammar hypothesis j, there can never be a transition from j to i, and there must be one from i to j. Once we reach the target grammar there is nothing that can move the learner from this state, since all remaining positive evidence will not cause the learner to change its hypothesis: an **Absorbing State** (AS) in the Markov literature. Clearly, one can conclude at once the following important learnability result:

Theorem 1 *Given a Markov chain C corresponding to a GW TLA learner, $\exists$ exactly 1 AS (corresponding to the target grammar/language) iff C is learnable.*

Proof. $\Leftarrow$. By assumption, C is learnable. Now assume for sake of contradiction that there is not exactly one AS. Then there must be either 0 AS or > 1 AS. In the first case, by the definition of an absorbing state, there is no hypothesis in which the learner will remain forever. Therefore C is not learnable, a contradiction. In the second case, without loss of generality, assume there are exactly two absorbing states, the first S corresponding to the target parameter setting, and the second S' corresponding to some other setting. By the definition of an absorbing state, in the limit C will with some nonzero probability enter S', and never exit S'. Then C is not learnable, a contradiction. Hence our assumption that there is not exactly 1 AS must be false.

$\Rightarrow$. Assume that there exists exactly 1 AS i in the Markov chain M. Then, by the definition of an absorbing state, after some number of steps n, no matter what the starting state, M will end up in state i, corresponding to the target grammar. ∎

Corollary 0.1 *Given a Markov chain corresponding to a (finite) family of grammars in a GW learning system, if there exist 2 or more AS, then that family is not learnable.*

DERIVATION OF TRANSITION PROBABILITIES FOR THE MARKOV TLA STRUCTURE

We now derive the transition probabilities for the Markov TLA structure, the key to establishing sample complexity results. Let the target language L_t be $L_t = \{s_1, s_2, s_3, ...\}$ and P a probability distribution on these strings. Suppose the learner is in a state corresponding to language L_s. With probability $P(s_j)$, it receives a string s_j. There are two cases given current parameter settings.

Case I. The learner can syntactically analyze the received string s_j. Then parameter values are unchanged. This is so only when $s_j \in L_s$. The probability of remaining in the state s is $P(s_j)$.

Case II. The learner cannot syntactically analyze the string. Then $s_j \notin L_s$; the learner is in state s, and has n neighboring states (Hamming distance of 1). The learner picks one of these uniformly at random. If n_j of these neighboring states correspond to languages which contain s_j and the learner picks any one of them (with probability n_j/n), it stays in that state. If the learner picks any of the other states (with probability $(n - n_j)/n$) then it remains in state s. Note that n_j could take values between 0 and n. Thus the probability that the learner remains in state s is $P(s_j)((n-n_j)/n)$. The probability of moving to each of the other n_j states is $P(s_j)(n_j/n)$.

The probability that the learner will remain in its original state s is the sum of the probabilities of these two cases: $\sum_{s_j \in L_s} P(s_j) + \sum_{s_j \notin L_s}(1 - n_j/n)P(s_j)$.

To compute the transition probability from s to k, note that this transition will occur with probability $1/n$ for all the strings $s_j \in L_k$ but not in L_s. These strings occur with probability $P(s_j)$ each and so the transition probability is: $P[s \rightarrow k] = \sum_{s_j \in L_t, s_j \notin L_s, s_j \in L_k}(1/n)P(s_j)$.

Summing over all strings $s_j \in (L_t \cap L_k) \setminus L_s$ (set difference) it is easy to see that $s_j \in (L_t \cap L_k) \setminus L_s \Leftrightarrow s_j \in (L_t \cap L_k) \setminus (L_t \cap L_s)$. Rewriting, we have $P[s \rightarrow k] = \sum_{s_j \in (L_t \cap L_k) \setminus (L_t \cap L_s)}(1/n)P(s_j)$. Now we can compute the transition probabilities between any two states. Thus the self-transition probability can be given as, $P[s \rightarrow s] = 1 - \sum_k$ is a neighboring state of s $P[s \rightarrow k]$.

Example.
Consider the 3-parameter natural language system described by Gibson and Wexler (1994), designed to cover basic word orders (X-bar structures) plus the verb-second phenomena of Germanic languages. Its binary parameters are: (1) Spec(ifier) initial (0) or final (1); (2) Compl(ement) initial (0) or final (1); and Verb Second (V2) does not exist (0) or does exist (1). Possible "words" in this language include S(ubject), V(erb), O(bject), D(irect) O(bject), Adv(erb) phrase, and so forth. Given these alternatives, Gibson and Wexler (1994) show that there are 12 possible surface strings for each (−V2) grammar and 18 possible surface strings for each (+V2) grammar, restricted to unembedded or "degree-0" examples for reasons of psychological plausibility (see Gibson and Wexler for discussion). For instance, the parameter setting [0 1 0]= Specifier initial, Complement final, and −V2, works out to the possible basic English surface phrase order of Subject–Verb–Object (SVO).

As in figure 1 below, suppose the SVO ("English", setting #5=[0 1 0]) is the target grammar. The figure's shaded rings represent increasing Hamming distances from the target. Each labeled circle is a Markov state. Surrounding the bulls-eye target are the 3 other parameter arrays that differ from [0 1 0] by one binary digit: e.g., [0, 0, 0], or Spec-first, Comp-first, −V2, basic order SOV or "Japanese".

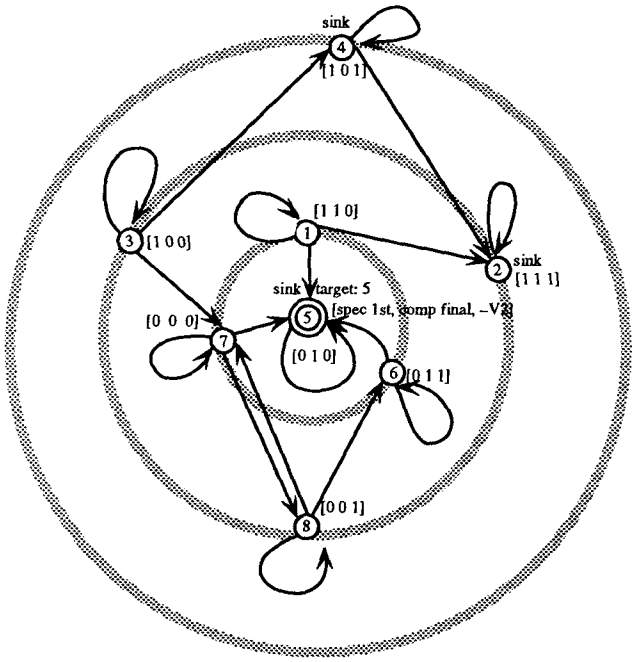

Figure 1: The 8 parameter settings in the GW example, shown as a Markov structure, with transition probabilities omitted. Directed arrows between circles (states) represent possible nonzero (possible learner) transitions. The target grammar (in this case, number 5, setting [0 1 0]), lies at dead center. Around it are the three settings that differ from the target by exactly one binary digit; surrounding those are the 3 hypotheses two binary digits away from the target; the third ring out contains the single hypothesis that differs from the target by 3 binary digits.

Plainly there are exactly 2 absorbing states in this Markov chain. One is the target grammar (by definition); the other is state 2. State 4 is also a *sink* that leads only to state 4 or state 2. GW call these two nontarget states *local maxima* because local gradient ascent will converge to these without reaching the desired target. Hence this system is *not* learnable. More importantly though, in addition to these local maxima, we show (see below) that there are *other* states (not detected in GW or described by Clark) from which the learner will never reach the target with (high) positive probability. Example: we show that if the learner starts at hypothesis VOS–V2, then with probability 0.33 in the limit, the learner will never converge to the SVO target. Crucially, we must use set differences to build the Markov figure straightforwardly, as indicated in the next section. In short, while it is possible to reach "English" from some source languages like "Japanese," this is not possible for other starting points (exactly 4 other initial states).

It is easy to imagine alternatives to the TLA that avoid the local maxima problem. As it stands the learner only changes a parameter setting *if* that change allows the learner to analyze the sentence it could not analyze before. If we relax this condition so that under unanalyzability the learner picks a random parameter to change, then the problem with local maxima disappears, because there can be only 1 Absorbing State, the target grammar. All other states have exit arcs. Thus, by our main theorem, such a system *is* learnable. We discuss other alternatives below.

CONVERGENCE TIMES FOR THE MARKOV CHAIN MODEL

Perhaps the most significant advantage of the Markov chain formulation is that one can calculate the number of examples needed to acquire a language. Recall it is not enough to demonstrate convergence in the limit; learning must also be *feasible*. This is particularly true in the case of finite parameter spaces where convergence might not be as much of a problem as feasibility. Fortunately, given the transition matrix of a Markov chain, the problem of how long it takes to converge has been well studied.

SOME TRANSITION MATRICES AND THEIR CONVERGENCE CURVES

Consider the example in the previous section. The target grammar is SVO–V2 (grammar #5 in GW). For simplicity, assume a uniform distribution on L_5. Then the probability of a particular string s_j in L_5 is 1/12 because there are 12 (degree-0) strings in L_5. We directly compute the transition matrix (0 entries elsewhere):

	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8
L_1	$\frac{1}{2}$	$\frac{1}{6}$			$\frac{1}{3}$			
L_2		1						
L_3			$\frac{3}{4}$	$\frac{1}{12}$			$\frac{1}{6}$	
L_4		$\frac{1}{12}$		$\frac{11}{12}$				
L_5					1			
L_6					$\frac{1}{6}$	$\frac{5}{6}$		
L_7					$\frac{5}{18}$		$\frac{2}{3}$	$\frac{1}{18}$
L_8						$\frac{1}{12}$	$\frac{1}{36}$	$\frac{8}{9}$

States 2 and 5 are absorbing; thus this chain contains local maxima. Also, state 4 exits only to either itself or to state 2, hence is also a local maximum. If T is the transition probability matrix of a chain, then the corresponding i, j element of T^m is the probability that the learner moves from state i to state j in m steps. For learnability to hold irrespective starting state, the probability of reaching state 5 should approach 1 as m goes to infinity, i.e., column 5 of T^m should contain all 1's, and 0's elsewhere. Direct computation shows this to be false:

	L_1	L_2	L_3	L_4	L_5	L_6	L_7	L_8
L_1		$\frac{1}{3}$			$\frac{2}{3}$			
L_2		1						
L_3		$\frac{1}{3}$			$\frac{2}{3}$			
L_4		1						
L_5					1			
L_6					1			
L_7					1			
L_8					1			

We see that if the learner starts out in states 2 or 4, it will *certainly* end up in state 2 in the limit. These two states correspond to local maxima grammars in the GW framework. We also see that if the learner starts in states 5 through 8, it will *certainly* converge in the limit to the target grammar.

States 1 and 3 are much more interesting, and constitute new results about this parameterization. If the learner starts in either of these states, it reaches the target grammar with probability 2/3 and state 2 with probability 1/3. Thus, local maxima are *not* the only problem for parameter space learnability. To our knowledge, GW and other researchers have focused exclusively on local maxima. However, while it is true that states 2 and 4 will, with probability 1, not converge to the target grammar, it is *also* true that states 1 and 3 will not converge to the target, with probability 1/3. Thus, the number of "bad" initial hypotheses is significantly larger than realized generally (in fact, 12 out of 56 of the possible source-target grammar pairs in the 3-parameter system). This difference is again due to the new probabilistic framework introduced in the current paper.

Figure 2 shows a plot of the quantity $p(m) = \min\{p_i(m)\}$ as a function of m, the number of examples. Here p_i denotes the probability of being in state 1 at the end of m examples in the case where the learner started in state i. Naturally we want

$$\lim_{m\to\infty} p_i(m) = 1$$

and for this example this is indeed the case. The next figure shows a plot of the following quantity as a function of m, the number of examples.

$$p(m) = \min\{p_i(m)\}$$

The quantity $p(m)$ is easy to interpret. Thus $p(m) = 0.95$ means that for every initial state of the learner the probability that it is in the target state after m examples is at least 0.95. Further there is one initial state (the worst initial state with respect to the target, which in our example is L_8) for which this probability is exactly 0.95. We find on looking at the curve that the learner converges with high probability within 100 to 200 (degree-0) example sentences, a psychologically plausible number.

We can now compare the convergence time of TLA to other algorithms. Perhaps the simplest is random walk: start the learner at a random point in the 3-parameter space, and then, if an input sentence cannot be analyzed, move 1-bit randomly from state to state. Note that this regime cannot suffer from the local maxima problem, since there is always some finite probability of exiting a non-target state.

Computing the convergence curves for a random walk algorithm (RWA) on the 8 state space, we find that the convergence times are actually faster than for the TLA; see figure 2. Since the RWA is also superior in that it does not suffer from the same local maxima problem as TLA, the conceptual support for the TLA is by no means clear. Of course, it may be that the TLA has empirical support, in the sense of independent evidence that children do use this procedure (given by the pattern of their errors, etc.), but this evidence is lacking, as far as we know.

DISTRIBUTIONAL ASSUMPTIONS: PART I

In the earlier section we assumed that the data was uniformly distributed. We computed the transition matrix for a particular target language and showed that convergence times were of the order of 100-200 samples. In this section we show that the convergence times depend crucially upon the distribution. In particular we can choose a distribution which will make the convergence time as large as we want. Thus the distribution-free convergence time for the 3-parameter system is infinite.

As before, we consider the situation where the target language is L_1. There are no local maxima problems for this choice. We begin by letting the distribution be parametrized by the variables a, b, c, d where

$$\begin{aligned} a &= P(A = \{\text{Adv(erb)Phrase V S}\}) \\ b &= P(B = \{\text{Adv V O S, Adv Aux V S}\}) \\ c &= P(C = \{\text{Adv V O1 O2 S, Adv Aux V O S,} \\ &\quad \text{Adv Aux V O1 O2 S}\}) \\ d &= P(D = \{\text{V S}\}) \end{aligned}$$

Thus each of the sets A, B, C and D contain different degree-0 sentences of L_1. Clearly the probability of the set $L_1 \setminus \{A \cup B \cup C \cup D\}$ is $1-(a+b+c+d)$. The elements of each defined subset of L_1 are equally likely with respect to each other. Setting positive values for a, b, c, d such that $a + b + c + d < 1$ now defines a unique probability for each degree(0) sentence in L_1. For example, the probability of $AdvVOS$ is $b/2$, the probability of $AdvAuxVOS$ is $c/3$, that of VOS is $(1-(a+b+c+d))/6$ and so on; see figure 3. We can now obtain the transition matrix corresponding to this distribution. If we compare this matrix with that obtained with a uniform distribution on the sentences of L_1 in the earlier section. This matrix has non-zero elements (transition probabilities) exactly where the earlier matrix had non-zero elements. However, the value of each transition probability now depends upon a, b, c, and d. In particular if we choose $a = 1/12, b = 2/12, c = 3/12, d = 1/12$ (this is equivalent to assuming a uniform distribution) we obtain the appropriate transition matrix as before. Looking more closely at the general transition matrix, we see that the transition probability from state 2 to state 1 is $(1 - (a + b + c))/3$. Clearly if we make a arbitrarily close to 1, then this transition probability is arbitrarily close to 0 so that the number of samples needed to converge can be made arbitrarily large. Thus choosing large values for a and small values for b will result in large convergence times.

This means that the sample complexity cannot be bounded in a distribution-free sense, because by choosing a highly unfavorable distribution the sample complexity can be made as high as possible. For example, we now give the convergence curves calculated for different choices of a, b, c, d. We see that for a uniform distribution the convergence occurs within 200 samples. By choosing a distribution with $a = 0.9999$ and $b = c = d = 0.000001$, the convergence time can be pushed up to as much as 50 million samples. (Of course, this distribution is presumably not psychologically realistic.) For $a = 0.99, b = c = d = 0.0001$, the sample complexity is on the order of 100,000 positive examples.

Remark. The preceding calculation provides a worst-case convergence time. We can also calculate *average* convergence times using standard results from Markov chain theory (see Isaacson and Madsen, 1976), as in table 2. These support our previous results.

There are also well-known convergence theorems derived from a consideration of the eigenvalues of the transition matrix. We state without proof a convergence result for transition matrices stated in terms of its eigenvalues.

Table 1: Complete list of problem states, i.e., all combinations of starting grammar and target grammar which result in non-learnability of the target. The items marked with an asterisk are those listed in the original paper by Gibson and Wexler (1994).

Initial Grammar	Target Grammar	State of Initial Grammar (Markov Structure)	Probability of Not Converging to Target
(SVO-V2)	(OVS-V2)	Not Sink	0.5
(SVO+V2)*	(OVS-V2)	Sink	1.0
(SOV-V2)	(OVS-V2)	Not Sink	0.15
(SOV+V2)*	(OVS-V2)	Sink	1.0
(VOS-V2)	(SVO-V2)	Not Sink	0.33
(VOS+V2)*	(SVO-V2)	Sink	1.0
(OVS-V2)	(SVO-V2)	Not Sink	0.33
(OVS+V2)*	(SVO-V2)	Not Sink	1.0
(VOS-V2)	(SOV-V2)	Not Sink	0.33
(VOS+V2)*	(SOV-V2)	Sink	1.0
(OVS-V2)	(SOV-V2)	Not Sink	0.08
(OVS+V2)*	(SOV-V2)	Sink	1.0

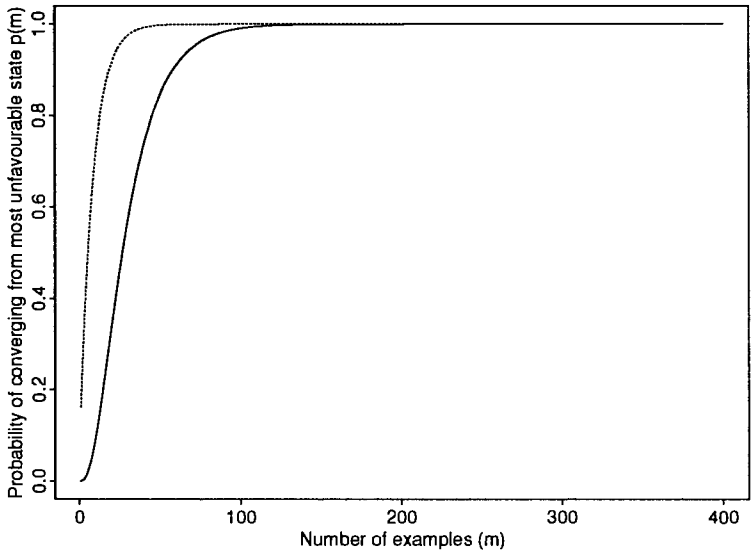

Figure 2: Convergence as a function of number of examples. The probability of converging to the target state after m examples is plotted against m. The data from the target is assumed to be distributed uniformly over degree-0 sentences. The solid line represents TLA convergence times and the dotted line is a random walk learning algorithm (RWA) which actually converges *faster* than the TLA in this case.

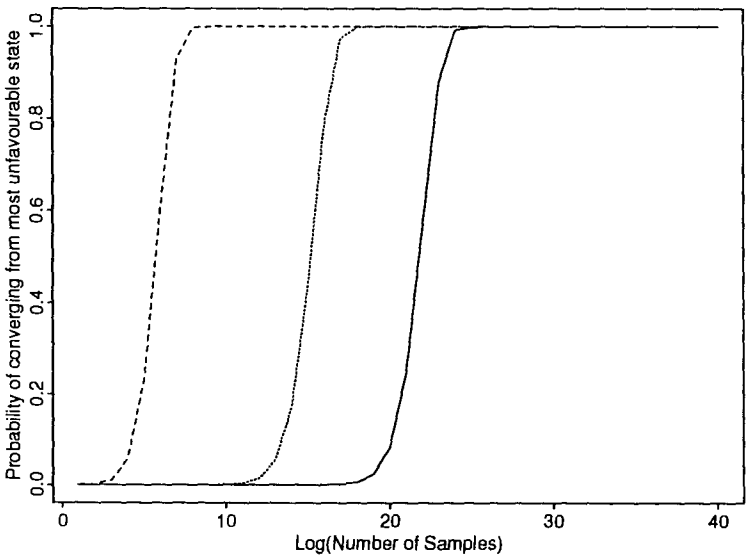

Figure 3: Rates of convergence for TLA with L_1 as the target language for different distributions. The probability of converging to the target after m samples is plotted against $\log(m)$. The three curves show how unfavorable distributions can increase convergence times. The dashed line assumes uniform distribution and is the same curve as plotted in figure 2.

Table 2: Mean and standard deviation convergence times to target 5 (English) given different distributions over the target language, and a uniform distribution over initial states. The first distribution is uniform over the target language; the other distributions alter the value of a as discussed in the main text.

Learning scenario	Mean abs. time	Std. Dev. of abs. time
TLA (uniform)	34.8	22.3
TLA ($a = 0.99$)	45000	33000
TLA ($a = 0.9999$)	4.5×10^6	3.3×10^6
RW	9.6	10.1

Theorem 2 *Let T be an $n \times n$ transition matrix with n linearly independent left eigenvectors $\mathbf{x}_1, \ldots \mathbf{x}_n$ corresponding to eigenvalues $\lambda_1, \ldots, \lambda_n$. Let $\mathbf{x}_0$ (an n-dimensional vector) represent the starting probability of being in each state of the chain and π be the limiting probability of being in each state. Then after k transitions, the probability of being in each state $\mathbf{x}_0 T^k$ can be described by*

$$\| \mathbf{x}_0 T^k - \pi \| = \| \sum_{i=1}^{n} \lambda_i^k \mathbf{x}_0 \mathbf{y}_i \mathbf{x}_i \| \leq \max_{2 \leq i \leq n} |\lambda_i|^k \sum_{i=2}^{n} \| \mathbf{x}_0 \mathbf{y}_i \mathbf{x}_i \|$$

where the $\mathbf{y}_i$'s are the right eigenvectors of T.

This theorem bounds the convergence rate to the limiting distribution π (in cases where there is only one absorption state, π will have a 1 corresponding to that state and 0 everywhere else). Using this result we bound the rates of convergence (in terms of number k of samples). It should be plain that these results could be used to establish standard errors and confidence bounds on convergence times in the usual way, another advantage of our new approach; see table 3.

DISTRIBUTIONAL ASSUMPTIONS, PART II

The Markov model also allows us to easily determine the effect of distributional changes in the input. This is important for either computer or child acquisition studies, since we can use corpus distributions to compute convergence times in advance. For instance, it can be easily shown that convergence times depend crucially upon the distribution chosen (so in particular the TLA learning model does not follow any distribution-free PAC results). Specifically, we can choose a distribution that will make the convergence time as large as we want. For example, in the situation where the target language is L_1, we can increase the convergence time arbitrarily by increasing the probability of the string {Adv(verb) V S}. By choosing a more unfavorable distribution the convergence time can be pushed up to as much as 50 million samples. While not surprising in itself, the specificity of the model allows us to be precise about the required sample size.

CHILDES DISTRIBUTIONS

It is of interest to examine the fidelity of the model using real language distributions, namely, the CHILDES database. We have carried out preliminary direct experiments using the CHILDES caretaker English input to "Nina" and German input to "Katrin"; these consist of 43,612 and 632 sentences each, respectively. We note, following well-known results by psycholinguists, that both corpuses contain a much higher percentage of aux-inversion and wh-questions than "ordinary" text (e.g., the LOB): 25,890 questions, and 11, 775 wh-questions; 201 and 99 in the German corpus; but only 2,506 questions or 3.7% out of 53,495 LOB sentences.

To test convergence, an implemented system using a newer version of deMarcken's partial parser (see deMarcken, 1990) analyzed each degree-0 or degree-1 sentence as falling into one of the input patterns SVO, S Aux V, etc., as appropriate for the target language. Sentences not parsable into these patterns were discarded (presumably "too complex" in some sense following a tradition established by many other researchers; see Wexler and Culicover (1980) for details). Some examples of caretaker inputs follow:

this is a book ? what do you see in the book ?

how many rabbits ?

what is the rabbit doing ? (...)

is he hopping ? oh . and what is he playing with ?

red mir doch nicht alles nach !

ja , die schwätzen auch immer alles nach (...)

When run through the TLA, we discover that convergence falls roughly along the TLA convergence time displayed in figure 1–roughly 100 examples to asymptote. Thus, the feasibility of the basic model is confirmed by actual caretaker input, at least in this simple case, for both English and German. We are continuing to explore this model with other languages and distributional assumptions. However, there is one very important new complication that must be taken into account: we have found that one must (obviously) add patterns to cover the predominance of auxiliary inversions and wh-questions. However, that largely begs the question of whether the language is verb-second or not. Thus, as far as we can tell, we have not yet arrived at a satisfactory parameter-setting account for V2 acquisition.

VARIANTS OF THE LEARNING MODEL AND EXTENSIONS

The Markov formulation allows one to more easily explore algorithm variants. Besides the TLA, we consider the possible three simple learning algorithm regimes by dropping either or both of the Single Value and Greediness constraints. The key result is that *almost any other* regime works faster than local gradient ascent and avoids problems with local maxima. See figure 4 for a representative result. Thus, most interestingly, parameterized language learning appears particularly robust under algorithmic changes.

EXTENSIONS, DIACHRONIC CHANGE AND CONCLUSIONS

We remark here that the "batch" phonological parameter learning system of Dresher and Kaye (1990) is susceptible to a more direct PAC-type analysis, since their system sets parameters in an "off-line" mode. We state without proof some results that can be given in such cases.

Table 3: Convergence rates derived from eigenvalue calculations.

Learning scenario	Rate of Convergence
TLA (uniform)	$O(0.94^k)$
TLA($a = 0.99$)	$O((1-10^{-4})^k)$
TLA($a = 0.9999$)	$O((1-10^{-6})^k)$
RW	$O(0.89^k)$

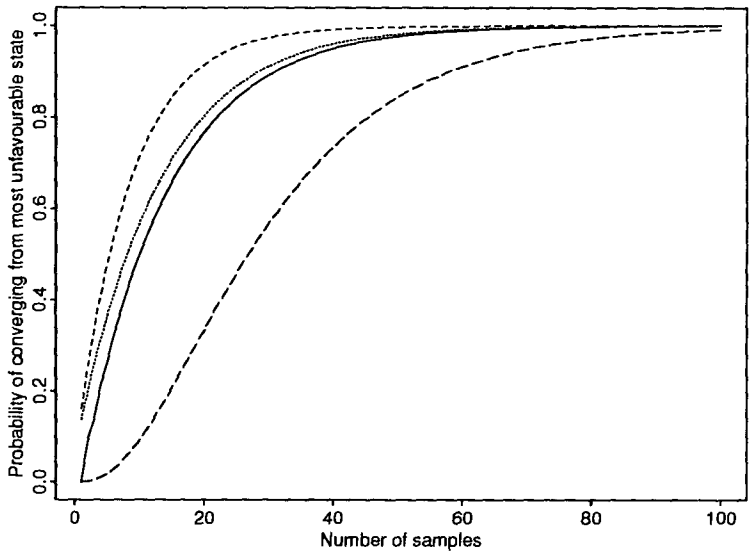

Figure 4: Convergence rates for different learning algorithms when L_1 is the target language. The curve with the slowest rate (large dashes) represents the TLA, the one with the fastest rate (small dashes) is the Random Walk (RWA) with no greediness or single value constraints. Random walks with exactly one of the greediness and single value constraints have performances in between.

Theorem 3 *If the learner draws more than* $M = \frac{1}{\ln(1/(1-b_t))} \ln(1/\delta)$ *samples, then it will identify the target with confidence greater than* $1 - \delta$. *(Here* $b_t = P(L_t \setminus \cup_{j \neq t} L_j)$*).*

Finally, the Markov model also points to an intriguing new model for syntactic change. One simply has to introduce two or more target languages that emit positive example strings with (probably different) frequencies: each corresponding to difference language sources. If the model is run as before, then there can be a large probability for a learner to converge to a state different from the highest frequency emitting target state: that is, the learner can acquire a different parameter setting, for example, a −V2 setting, even in a predominantly +V2 environment. This is of course one of the historical changes that occurred in the development of English. Space does not permit us to explore all the consequences of this new Markov model; we remark here that once again we can compute convergence times and stability under different distributions of target frequencies, combining it with the usual dynamical models of genotype fixation. In this case, the interesting result is that the TLA actually boosts diachronic change by orders of magnitude, since as observed earlier, it can permit the learner to arrive at a different convergent state even when there is just *one* target language emitter. In contrast, the local maxima targets are stable, and never undergo change. Whether this powerful "boost" effect plays a role in diachronic change remains a topic for future investigation. As far as we know, the possibility for formally modeling the kind of saltation indicated by the Markov model has not been noted previously and has only been vaguely stated by authors such as Lightfoot (1990).

In conclusion, by introducing a formal mathematical model for language acquisition, we can provide rigorous results on parameter learning, algorithmic variation, sample complexity, and diachronic syntax change. These results are of interest for corpus-based acquisition and investigations of child acquisition, as well as pointing the way to a more rigorous bridge between modern computational learning theory and computational linguistics.

ACKNOWLEDGMENTS

We would like to thank Ken Wexler, Ted Gibson, and an anonymous ACL reviewer for valuable discussions and comments on this work. Dr. Leonardo Topa provided invaluable programming assistance. All residual errors are ours. This research is supported by NSF grant 9217041-ASC and ARPA under the HPCC program.

REFERENCES

Clark, Robin and Roberts, Ian (1993). "A Computational Model of Language Learnability and Language Change." *Linguistic Inquiry*, 24(2):299–345.

deMarcken, Carl (1990). "Parsing the LOB Corpus." *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*. Pittsburgh, PA: Association for Computational Linguistics, 243–251.

Dresher, Elan and Kaye, Jonathan (1990). "A Computational Learning Model For Metrical Phonology." *Cognition*, 34(1):137–195.

Gibson, Edward and Wexler, Kenneth (1994). "Triggers." *Linguistic Inquiry*, to appear.

Gold, E.M. (1967). "Language Identification in the Limit." *Information and Control*, 10(4): 447–474.

Isaacson, David and Masden, John (1976). *Markov Chains*. New York: John Wiley.

Lightfoot, David (1990). *How to Set Parameters*. Cambridge, MA: MIT Press.

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