The Computational Study of Diachronic Linguistics

Partha Niyogi

1 Introduction

Over the last five years, we have seen an explosion of interest and activity in computational models of language change and evolution (Niyogi and Berwick 1997; Clark and Roberts 1993; Yang 2000; Briscoe 2000; Hurford et al 1998; Galves and Galves 1996; Nowak and Krakauer 1999). The purpose of this paper is to take stock of this situation, review some of the basic models and consider extensions. Most significantly, it will be argued that such computational thinking, properly executed, will play a critical role in unraveling the complex processes that underlie the evolution of linguistic populations with time.

At the heart of these models is the subtle interplay between language

learning and language change. For the last forty years, synchronic linguistics has been driven by the so called "logical problem of language acquisition" – the problem of how children come to acquire the language of their parents. This has been investigated from several points of view. Linguists in the generative tradition have proposed theories of universal grammar that constrain the range of grammatical hypotheses that children might entertain during language learning. Psycholinguists, developmental psychologists and cognitive scientists have pursued a range of approaches from empirical studies of child language acquisition to formal analyses of the process (Pinker 1984; Wexler and Culicover 1980; Crain and Thornton 1998; Slobin 1985-97). Computational work on this question has a rich history. Beginning with the pioneering work of Gold (1967), computer scientists and mathematicians have considered the formal difficulty of inferring a grammar (language) from linguistic examples. Considerable analytical work exists on the subject and a summary can be found in Osherson et al (1986). Simulations in the style of artificial intelligence research have also been significant (see, for example, Berwick 1985; Siskind 1992; Feldman et al 1990; Brent 1996).

Language acquisition may be viewed as the mechanism by which language is transmitted from parent to child —- and in fact, from one generation of language users to the next. Perfect language acquisition would imply

perfect transmission. Children would acquire perfectly the language of their parents, language would be mirrored perfectly from one generation to the next and languages would not change with time. Therefore, for languages to change with time, children must do something differently from their parents. This insight has been around for at least a hundred years as evidenced by the following quote from the British phonetician Henry Sweet (1891).

> ....if languages were learnt perfectly by the children of each generation, then language would not change: English children would still speak a language as old at least as AngloSaxon and there would be no such languages as French or Italian.
>
> (Sweet, 1891, pg. 75)

Thus there is a tension between language learning on the one hand and language change on the other. Perfect language learning would imply no change. At the same time, language learning cannot be so imperfect that the learned language of the children does not resemble at all that of the parents (linguistic environment). If, due to slight imperfections of language learning, the linguistic composition of the population shifts just a bit, can this slight shift lead eventually to a significant change over long time scales?

While language learning has come under intense computational scrutiny

for a while, computational work in language change is fairly recent. Over the last five years, there has been a growing body of computational work that explores the relationship between language learning and language change. Let us begin by considering the simplest family of such models — those of two languages in "competition" with each other.

2 A Preliminary Model

Our discussion is motivated by a syntactic[1] view of the world where languages are viewed as sets of grammatically well formed expressions and therefore formally as subsets of Σ^* where Σ is a finite alphabet (denoting the lexical items). Imagine a world with only two possible languages L_1 and L_2 where each L_i is a subset of Σ^* in the usual way. In general, L_1 and L_2 are not disjoint. Sentences belonging to both L_1 and L_2 are ambiguous and may be parsed according to the underlying grammar of each language.

We consider a case where each individual is a user of precisely one language – this is the monolingual case. The language of the individual is acquired during a learning period (over childhood) on the basis of expo-

[1] The general methodology is applicable to phonology just as well. One may view a phonological grammar as defining a set of well formed phonological expressions. The set of such well formed expressions may be defined using a notational system that comprises of a phonological alphabet.

sure to linguistic examples provided by the ambient linguistic community. To make matters simple, we divide the population neatly into coincident generations and now consider two successive generations.

The state of any adult generation is simply described by a single variable α_t (the subscript t denoting generation number). Here α_t is the proportion of individuals speaking language L_1 in generation t — clearly, a proportion $1 - \alpha_t$ of the population consist of users of language L_2. Further, let users of L_1 produce sentences with a probability distribution P_1 on the sentences of L_1 and users of L_2 produce sentences with a probability distribution P_2 on the sentences of L_2. Thus a sentence $s \in \Sigma^*$ will be produced with probability $P_1(s)$ by a user of L_1 and with probability $P_2(s)$ by a user of L_2. If s is not an element of L_1, clearly $P_1(s) = 0$ and similarly, if s is not an element of L_2, then $P_2(s) = 0$.

2.1 *Learning By Individuals*

We begin by examining the acquisition of language by individuals in the population. Language acquisition is the process of developing grammatical hypotheses[2] on the basis of linguistic experience, i.e., exposure to linguistic

[2]We consider in general a grammatical hypothesis to be a set of rewrite or phrase structure rules. These hypotheses therefore define a set of valid expressions in the usual way. It is a matter of some debate whether children conjecture fully formed grammars

data during childhood. Within the purview of generative linguistics, this is conceptually regarded as choosing an appropriate grammar from a class of potential grammars $\mathcal{G}$ (Universal Grammar or UG) on the basis of primary linguistic data. In this example, we consider the case where there are only two potential grammars —- those underlying the languages L_1 and L_2 respectively.

Consider now a learning procedure (algorithm) to choose a grammar based on linguistic examples. This can, in general, be characterized as a mapping from linguistic data sets to the hypothesis set $\{L_1, L_2\}$. To fix notation, let D_1 be the set of all data sets with just one example sentence, and in general D_n be the set of all data sets containing n example sentences each. Clearly, $D_1 = \Sigma^*$ and $D_n = (\Sigma^*)^n$. We can then let $D = \cup_{i \geq 1} D_i$ to be the set of all finite-length data sets. The learning algorithm $\mathcal{A}$ is an effective procedure (partial recursive function) from D to $\{L_1, L_2\}$. This simply means that the learning algorithm is an effective procedure (computable by a Turing machine) that constructs linguistic hypotheses from example sentences. Consider a $d_l \in D_l$. Here D_l is the set of all possible data sets of size l and d_l is a particular data set of size l. Then if the learning algorithm

during the process of language acquistion or partial grammars. For our purposes, both partial and full grammars are notationally represented by sets of rewrite rules and therefore do not affect greatly the discussion that follows.

guesses L_1 when this particular data set is presented to it, we will say that $\mathcal{A}(d_l) = L_1$. For some other data set (say d'_l) in D_l, the algorithm might guess L_2 in which case we say $\mathcal{A}(d'_l) = L_2$.

Now fix a probability distribution P on Σ^* according to which sentences are drawn independently at random and presented to the learner. After k such examples are drawn, the learner's data set can be denoted by $d_k = \{s_1, s_2, \ldots, s_k\}$ where each s_i is drawn according to the distribution P. Clearly d_k is an element of D_k. In this setting, it is possible to define the following object

$$p_k = Prob[\mathcal{A}(d_k) = L_1]$$

In other words, p_k is the probability with which the learning algorithm will guess L_1 after k randomly drawn sentences are presented to it. Now p_k will in general depend upon the probability distribution P that generates the data as well as the learning algorithm $\mathcal{A}$. We will denote this dependence by $p_k(\mathcal{A}, P)$.

In this probabilistic setting, it is now possible to define the notion of *learnability*. Learnability is the property of the learning algorithm to converge to the target as the data goes to infinity. This simply means that if

the probability distribution P had support on L_1 so that only sentences of L_1 occured in the data sets of the learner (i.e., L_1 is the target language), then

$$p_k(\mathcal{A}, P = P_1) \rightarrow_{k\rightarrow\infty} 1$$

Similarly, if the probability distribution P had support only on L_2 so that only sentences of L_2 were presented to the learner, then $p_k(\mathcal{A}, P = P_2) \rightarrow_{k\rightarrow\infty} 0$.

Given this set up, a few remarks are worthwhile.

Remarks:

1. The mathematical framework has been created to make precise the notion of learnability and to study the difficulty of inductive inference of grammars from data. For example, Wexler and Culicover (1980), Osherson, Stob, and Weinstein (1986), Pinker (1984) all present an explication of this framework in differing degrees of mathematical detail and linguistic relevance.

2. Learners are treated as procedures that develop grammatical hypotheses on the basis of linguistic experience. Part of the goal of language learning theory is to explore various such procedures in order to dis-

cover the ones that are most like the ones that children presumably use. In the previous discussion, learners are viewed as deterministic procedures – this means that given the same data set (primary linguistic data), all children would develop the same hypothesis. The framework can easily be extended to consider randomized procedures without affecting the generality of the arguments presented. We do not consider such extensions here.

3. The basic framework for learning is applicable to situations where there are multiple languages and to cases where learners conjecture more than one language at a time.

2.2 *Population Dynamics*

The previous section focuses on how the individual learner develops grammatical hypotheses from example sentence to example sentence until it matures. The central question there is whether or not the learner's hypothesis gets closer and closer to the target and eventually converges to it as more and more data become available. Of course, convergence to the target only occurs as the number of data goes to infinity — and learners live only finite lives. As a matter of fact, learners do not endlessly update their hypotheses

but "mature" after a point and live with their mature hypothesis thereafter. Let us assume that maturation occurs after K examples have been presented to the learner.

In this section, we consider the evolutionary implications at the population level of learning procedures at the individual level. Let us begin by considering a completely homogeneous population where all adult speakers speak the language L_1. Consider now the generation of children in this community who attempt to learn the language of the adults. A typical child will receive examples drawn according to a probability distribution P_1. Over its learning period, it will receive K examples and with probability $p_K(\mathcal{A}, P_1)$ the typical child will acquire the language L_1. With probability $1-p_K(\mathcal{A}, P_1)$, however, the child might acquire the language L_2. Therefore, when the generation of children mature into adulthood, the population of new adults will no longer be homogeneous. In fact, a proportion $p_K(\mathcal{A}, P_1)$ will be L_1 users and a proportion $1 - p_K$ will be L_2 users. In this fashion, the linguistic composition of two successive generations may be related to each other.

We need not have started with a homogeneous adult population. Imagine now that the state of the adult population is denoted by α_a where α_a is the proportion of L_1 users in the adult population. Now consider the

generation of children. They will receive example sentences from the entire adult population —- which in this case is a mixed population. In particular, they will receive examples drawn according to the distribution

$$P = \alpha_a P_1 + (1 - \alpha_a) P_2$$

On receiving example sentences from this distribution, children proceed as before. A proportion $p_K(\mathcal{A}, P)$ will acquire L_1. Letting α_c be the proportion of children who grow up to be L_1 speakers, we see that

$$\alpha_c = p_K(\mathcal{A}, \alpha_a P_1 + (1 - \alpha_a) P_2)$$

In this manner, we see that α_c can be expressed in terms of α_a.

In this example, the linguistic composition of the population can be characterized by a single variable α_t. This denotes the proportion of the population that consists of L_1 users in generation t. By considering the behavior of the typical child and then averaging over the entire population of children, we have related the linguistic composition of two successive

generations as follows:

$$\alpha_{t+1} = p_K(\mathcal{A}, \alpha_t P_1 + (1 - \alpha_t) P_2) \tag{1}$$

In order to do this, we assumed

1. The population could be isolated into coincident generations.
2. Children receive data drawn from the *entire* adult population in a manner that reflects the distribution of languages in the adult population.
3. The probability of drawing sentences P_1 and P_2 do not change with time.
4. The learning algorithm $\mathcal{A}$ constructs a single hypothesis language after each example and after maturation ends up with a single language.
5. Population sizes are infinite.

We will return to a discussion of these assumptions later. Let us now consider some examples where we make specific choices regarding the learning algorithm and derive the evolutionary consequences. In particular, the functional relationship between α_t and α_{t+1}, which is not transparent from eqn. 1 will be derived for a number of different settings.

2.3 *Some Examples*

A variety of dynamical maps can be obtained by different particular choices for (i) the maturation time K and (ii) the learning algorithm $\mathcal{A}$. We consider three different examples here.

2.3.1 $\mathcal{A}$: *Memoryless Learners*

A memoryless learner is one whose hypothesis at every stage depends only upon the current input sentence and the previous hypothesis it had. There are a wide class of such algorithms and one in particular has received considerable attention in the linguistic parameter setting literature. This is the triggering learning algorithm (TLA) of Gibson and Wexler 1994. While the algorithm works for any finite parameter space in general, the particular instantiation for the two language case is as follows:

TLA (Triggering Learning Algorithm)

- [Initialize] Step 1. Start with an initial hypothesis (either L_1 or L_2) chosen uniformly at random.
- [Process input sentence] Step 2. Receive a positive example sentence s_i at the ith time step.
- [Learnability on error detection] Step 3. If the current grammatical

hypothesis parses (generates) s_i, then go to Step 2 to receive next example sentence; otherwise, continue.

- [Single-step hill climbing] Step 4. Flip the current hypothesis and go to Step 2 to receive next example sentence.

The population at any point in time can be characterized by a single quantity (α_t for the tth generation) that describes the proportion of L_1 users in the population. *If children were TLA learners, then how would the population evolve?*

The precise nature of the evolution will depend not only upon the the algorithm $\mathcal{A}$ (in this case, the TLA) but also the probability distribution with which sentences are produced by L_1 and L_2 users respectively. For this case, it turns out that it is sufficient to characterize P_1 and P_2 by two parameters a and b given as follows:

$$a = P_1[L_1 \cap L_2]; 1 - a = P_2[L_1 \setminus L_2]$$

and similarly

$$b = P_2[L_1 \cap L_2]; 1 - b = P_2[L_2 \setminus L_1]$$

Here $L_1 \cap L_2$ refers to the set of ambiguous sentences —- those that can

be parsed (generated) by the underlying grammars of both languages. Thus a is the probability with which such ambiguous sentences are produced by L_1 users and b is the same for L_2 users. If we now assume that $K = 2$, i.e., the maturation time is short, it is fairly easy to show that the evolution occurs according to the following update rule:

Theorem 1 *The linguistic composition in the* $(t+1)$*th generation* (α_{t+1}) *is related to the linguistic composition of the* t*th generation* (α_t) *in the following way:*

$$\alpha_{t+1} = A\alpha_t^2 + B\alpha_t + C$$

where $A = \frac{1}{2}((1-b)^2 - (1-a)^2)$*;* $B = b(1-b) + (1-a)$ *and* $C = \frac{b^2}{2}$.

A few remarks concerning this dynamical system are in order:

Remark 1. When $a = b$, the system has exponential growth. When $a \neq b$, the dynamical system is a quadratic map (which can be reduced by a transformation of variables to the logistic, and shares the same dynamical properties). We note that Cavalli-Sforza and Feldman (1981), using a different formulation, also obtain a quadratic map in such cases for the example of general 'vertical' cultural change.

Remark 2. The scenario $a \neq b$ is much more likely to occur in practice — consequently, we are more likely to see logistic change rather than

exponential change.

Remark 3. Logistic maps are known to be chaotic. However, in our system it is possible to show that:

Theorem 2 *Due to the fact that* $a, b \leq 1$, *the dynamical system never enters the chaotic regime.*

Remark 4. We obtain a class of dynamical systems. The quadratic nature of our map comes from the fact that $K = 2$. If we choose other values for K we would get cubic and higher order maps. In general, it is possible to show that for a fixed, finite K, the evolutionary dynamics is given by

Theorem 3 *If individual learners in a population of TLA learners have a maturation time* K, *the population evolves according to*

$$\alpha_{t+1} = \frac{B + \frac{1}{2}(A - B)(1 - A - B)^K}{A + B}$$

where α_t *is the proportion of* L_1 *users in the* t*th generation and* $A = (1 - \alpha_t)(1 - b)$ *and* $B = \alpha_t(1 - a)$.

Remark 5. The parameters a and b determine the evolution of the population and its stable modes. As we have mentioned before, they represent

respectively the proportion of L_1 and L_2 sentences respectively that are ambiguous. It is conceivable that one might be able to estimate these parameters from synchronic or diachronic corpora.

2.3.2 $\mathcal{A}$: *Batch Error Based Learner*

In contrast to the memoryless learner, a batch learner waits until the entire data set of K examples has been received. Then, it simply chooses the language that is more consistent with the data received.

For each language L_i, one can define an error measure (denoted by $e(L_i)$) as

$$e(L_i) = \frac{k_i}{K}$$

where k_i is the number of example sentences in the data set that is not analyzable according to the grammar of L_i. Then a simple decision rule is

$$\hat{L} = \arg \min_{L_1, L_2} e(L_i)$$

This amounts to the following rule. Group the K example sentences of the data set (PLD) into three classes: (A) those sentences that belong to L_1 alone and are not analyzable by the underlying grammar of L_2; (B) those sentences that belong to L_2 alone and are not analyzable by the underly-

ing grammar of L_1; (C) those sentences that are ambiguous and and are analyzable (with different interpretations, perhaps) under the grammars of both L_1 and L_2. Let n_1, n_2, n_3 be the number of examples of type A,B,C respectively. Clearly, $n_1 + n_2 + n_3 = K$. Choose L_1 if $n_1 \geq n_2$, otherwise choose L_2.

For this learning algorithm, it is possible to show that the proportion of L_1 users in two successive generations (α_t and α_{t+1}, respectively) is related by the following update rule.

$$\alpha_{t+1} = \sum_{(n_1,n_2,n_3)|n_1>=n_3;\sum_i n_i=K} C^K_{n_1,n_2,n_3} p_1(\alpha_t)^{n_1} p_2(\alpha_t)^{n_2} p_3(\alpha_t)^{n_3}$$

Here $C^K_{n_1,n_2,n_3} = \frac{K!}{n_1!n_2!n_3!}$ is the multinomial coefficient and $p_1(\alpha_t) = \alpha_t(1-a)$; $p_2(\alpha_t) = \alpha_t a + (1-\alpha_t)b$; and $p_3(\alpha_t) = (1-\alpha_t)(1-b)$.

In general, the nature of the population dynamics is different in this case from the previous one.

2.3.3 $\mathcal{A}$: *Cue-Based Learner*

A cue based learner examines the data set for cues to a linguistic parameter setting. Let a set $C \subset (L_1 \setminus L_2)$ be a set of examples that are cues to the learner that L_1 is the target language. If such cues occur often enough in the

learner's data set, the learner will choose L_1, otherwise the learner chooses L_2. This follows the cue-driven approach advocated in Lightfoot (1997).

This approach is instantiated in the following procedure. Let the learner receive K examples. Out of the K examples, say k are from the cue set. Then, if

$$\frac{k}{K} > \tau$$

the learner chooses L_1, otherwise the learner chooses L_2.

One can again determine the evolutionary dynamics of the population based on such a learner. Let $P_1(C) = p$, i.e., p is the probability with which an L_1 user produces a cue. If a proportion α_t of adults use L_1, then we see that the probability with which a cue is presented to a typical child is given by $p\alpha_t$ and so the probability with which $k > K\tau$ is given by

$$\sum_{K\tau \leq i \leq K} (p\alpha_t)^i (1 - p\alpha_t)^{(K-i)}$$

and therefore, we get

$$\alpha_{t+1} = \sum_{K\tau \leq i \leq K} (p\alpha_t)^i (1 - p\alpha_t)^{(K-i)}$$

Here, α_t is the proportion of L_1 users in the tth generation. Interestingly,

for such learners, it is possible to show that the only stable configuration of the population is $\alpha = 0$, i.e., a homogeneous population of L_2 users. Thus a homogeneous population of L_2 users will always remain stable and can *never* change to a population of L_1 users. A homogeneous population of L_1 users will never be stable and will *always* drift over time to a population of L_2 users. A change from L_2 to L_1 is possible —- a change the other way is never possible. As we shall see later, this raises immediate complications for Lightfoot's explanation for the change from Old English +V2 grammars to Modern -V2 grammars. A more nuanced form of the explanation will become necessary.

3 Further Directions

The simple two-language models of the preceding sections are not without significant linguistic applications. In many cases of language change, one finds that there are two variants (dialects, grammars) differing by a significant linguistic parameter that coexist in a population in varying proportions at different points in time. Often, linguistic change leads to the gradual loss of one variant from the population entirely (often following an S-shaped pattern over time). For example, the loss of verb-second from grammars of

Old English to that of Modern English is a much studied instance of precisely such a change (Lightfoot 1997; Kroch and Taylor 1997; Pintzuk 1991). Other examples include the loss of verb-second (V2) from Old to Modern French (Clark and Roberts 1993), the change in subordinate clause word order in Yiddish considered by Santorini (1992) and so on.

At the same time, one needs to recognize the drastic simplifications that have been made in order to formulate this first coherent model. It is worthwhile to reflect on some of these simplifying assumptions and the possibility of relaxing them in more complex models of this process.

1. *Multiple Languages:* Clearly, there are more than two languages in the world. The space $\mathcal{G}$ represents the hypothesis space that learning algorithms operate on and draw grammatical hypotheses from over their learning period. While this space $\mathcal{G}$ is in principle all of universal grammar, in linguistic applications of a more specific nature, one might consider a subset of $\mathcal{G}$ to be the more appropriate object to model and study. For example, in studies of syntax, its acquisition and change, it is meaningful to ignore the phonological components of UG that might have no bearing on the phenomena at hand. Depending upon the submodule of syntax under study, other "irrelevant" modules might also

be usefully ignored. Thus, the space $\mathcal{G}$ that is formulated in models of language change is really a very low-dimensional projection of the high dimensional space of UG. It is the linguist's intuition and understanding of the phenomena that provides the appropriate low-dimensional projection. Indeed, linguists might differ on this matter, and the consequences then need to be worked out. Having thus argued that in most useful applications, the space $\mathcal{G}$ will be low dimensional, it might still consist of more than two grammars (languages) and it is important to extend such models to multilingual settings. The extension to n-language families has already been considered (Niyogi and Berwick 1997; Yang 2000). Setting up the models for such cases is easy enough – analytical solutions are harder to come by and one might need to resort to simulations.

2. *Finite Populations:* One reason we have been able to derive deterministic dynamical maps relating successive generations to each other is the assumption of infinite population size that allows us to take ensemble averages of individual behavior over the entire population. In practice, of course, populations are always finite. If the population sizes are large, then the assumption of infinite sizes may not be too

bad. If, on the other hand, population sizes are very small, then one might need to consider the implications of such small sizes more carefully. Let us consider briefly the effect of finite population sizes on the two language models discussed in this paper. Recall that each individual child attains L_1 with probability $p_K(\mathcal{A}, \alpha_a P_1 + (1-\alpha_a)P_2)$. From this we concluded that a proportion p_K of the children would end up as L_1 users. This statement is exactly true if there were an infinite number of children. Imagine, instead, there were only N children in the population. Each child could end up either as an L_1 speaker or as an L_2 speaker. In fact, with probability $(p_K)^N$, all children would acquire L_1; with probability $(1-p_K)^N$ all would acquire L_2; and different intermediate mixes are possible with probabilities given by the binomial distribution. Thus all evolutionary trajectories are *possible*, the question is – which ones are *likely* or *probable*? The evolution is characterized now as a stochastic process rather than a deterministic dynamical system. The consequences of this can be worked out. The details are beyond the scope of this paper.

3. *Generational Structure:* In attempting to derive the relationship between successive generations, we have assumed that generations move

in clean time steps. In practice, of course, the generational structure is a little more complex than this. One might therefore need to divide time into finer intervals and consider the cohort of learning children at each such time interval. The primary linguistic data that this cohort receives is now drawn from a more diverse group of older people in the population. This group would consist of parents, grandparents, older cohorts and so on. For example, in the two language models described earlier, one might proceed as follows:

Let the state of cohort t be described by a variable α_t (as before, where α_t denotes the proportion of the cohort using the language L_1). Consider now the $(t+1)$th cohort of learning children. Assume that they receive data drawn from the previous three cohorts (who may, for example, be characterized as the cohort of young adults, parents, and grandparents respectively) in equal proportions. Then the probability distribution with which data is presented to the $(t+1)$th cohort of learners is given by

$$P = \frac{1}{3}(\alpha_t P_1 + (1-\alpha_t)P_2) + \frac{1}{3}(\alpha_{t-1} P_1 + (1-\alpha_{t-1})P_2) + \frac{1}{3}(\alpha_{t-2} P_1 + (1-\alpha_{t-2})P_2)$$

where we have assumed that all cohorts are equal in size and influence.

Given this set up, it is easy to see that α_{t+1} is now going to be given by

$$\alpha_{t+1} = p_K(\mathcal{A}, P)$$

and in this manner, α_{t+1} will depend upon α_t,α_{t-1}, and α_{t-2} respectively. This too is a dynamical system and can easily be analysed using the traditional tools.

4. *Spatial Population Structure:* We have assumed in the models that speakers of both language types are evenly distributed throughout the population. Further, the child learners all receive data from the *entire* adult population. In other words, all children receive data drawn from the same probability distribution and this distribution reflects the mix of L_1 and L_2 speakers in the adult population as a whole.

   Reality, as always, might be more complicated. Speakers of different linguistic types might reside in different "neighborhoods". Children born in different neighborhoods might receive data drawn from different probability distributions that reflect the linguistic composition of their neighborhood. For example, one might imagine an L_1 speaking neighborhood and an L_2 speaking neighborhood whose population sizes are in the ratio $\alpha_t : (1 - \alpha_t)$. Children born in the L_1 speaking

neighborhood might receive data drawn mostly according to P_1 while those born in the L_2 speaking neighborhood might receive data mostly drawn according to P_2. The evolutionary consequences of such a spatial structure in the population need to be worked out and represents an important direction of future research.

5. *Multilingual Acquisition:* The learning algorithm $\mathcal{A}$ realizes a mapping from linguistic data sets to grammatical hypotheses. In particular, we have restricted the learner to having precisely one grammatical conjecture at each point in time. Furthermore, at the end of the learning period (i.e., after receiving K examples) it is assumed that the learner will end up with precisely one language.

   If the target distribution corresponds to a unique grammar, it is certainly reasonable to expect the learner to end up with exactly one language. The case when the target distribution is mixed, i.e., not consistent with a single unique target grammar, natural models of the learning process should allow the possibility of multilingual rather than monolingual acquisition. Thus, for example, in the two-language case of this paper, one might allow the possibility that the learner acquires both languages (in some ratio, perhaps). For the case of English, for

example, Kroch and Taylor (1997) argue that learners were effectively bilingual having acquired both dialectal variations in different proportions. Yang (2000) considers such a learning algorithm and explores the evolutionary consequences.

6. *Non-vertical and Other modes of Transmission:* We have considered vertical modes (from parent to child or from one generation to the next) as the primary mode of transmission of language over time. It is often remarked that the interaction of a cohort of language users with each other in a social setting shapes the way language develops in children and therefore the way it evolves over time. The effect of children of the same generation on each other might be viewed as a non-vertical (horizontal) mode of transmission of language. It might therefore become necessary to consider such alternative modes of transmission for a more complete understanding of the complexities involved in such processes.

The effect of each of these assumptions can be systematically explored. Together they constitute important directions of future work in this nascent field of computational studies of language change. Some of these directions have already begun to be explored. Others await further explication.

4 Conclusions

What do we hope to learn from modeling of the sort presented in this paper? Why do we expect that a phenomenon as complex as language change over historical time is amenable to mathematical or computational analysis?

First, let us begin by accepting that the phenomena of linguistic change are real, pervasive, and in many cases, present a linguistic and cognitive puzzle of sorts. A compelling example for me is the case of syntactic change in English. Examining, for instance, the following examples from English of the ninth and tenth century leaves one in no doubt that there has been a deep change in the syntactic grammars of English users over time. It is not a word here, a nuance there that can be simply explained away by notions of changing fads or sociological circumstance. Clearly, linguistic populations restructured grammars over generational time.

pa Darius geseah paet he oferwunnen beon wolde

then Darius saw that [he conquered be would]

(Orosios 128.5)

& him aefterfylgende waes

and [him following was]

(Orosius 236.29)

Nu ic wille eac paes maran Alexandres gemunende beon

now I will also [the great Alexander considering be]

What precisely is the nature of the syntactic change? What initiated the change? Under what circumstances would the change have proceeded to completion? These are the kinds of questions that need to be untangled. Many complex cognitive phenomena tend to present puzzles and it is difficult to reason our way through such puzzles by verbal argument alone. For example, consider Lightfoot's discussion of parameter resetting leading to language change over time from Old to Modern English.

> As somebody adopts a new parameter setting, say a new verb-object order, the output of that person's grammar often differs from that of other peoples's. This in turn affects the linguistic environment, which may then be more likely to trigger the new parameter setting in younger people. Thus a chain reaction may be created, which may gradually permeate the speech community.
>
> (Lightfoot, 1991)

While the overall framework of analysis presented in Lightfoot (1991) is powerful, there are many "sub-theories" within the same framework that

need to be precisely investigated. The parameter resetting argument fits in quite naturally in our two-language model where the the two languages in question differ by a linguistic parameter. Therefore, one will have to work out the precise details of how often ambiguous sentences are uttered, whether the drift resulting from parameter resetting can take the population all the way so that one of the language types dies out altogether. Further, Kroch and Taylor (1997) have argued that grammatical variation resides not just in the population but in the individual as well. This suggests a move towards bilingual populations with perhaps native language and second language effects providing the necessary asymmetry for populations to gradually change over time.

One might assume (i) multilingual acquisition (ii) trigger-based learning (iii) cue-based learning. Each of these different acquisition algorithms possibly leads to different evolutionary consequences some of which might not be compatible with the historical trends observed. Additionally, linguists might also differ with respect to the grammatical characterization of the difference between the variants of English in the tenth and eleventh century. The different grammatical characterization would have different extensional consequences, a different set of ambiguous sentences, leading to different evolutionary consequences for the same learning algorithm.

For example, Lightfoot (1997) argues for a cue-based learner in explaining the loss of V2 from Old English to Modern English. The cue for +V2 grammars is taken to be sentences in non-subject-Vf (Vf being the finite verb) form. Thus, he argues: "Children in Lincolnshire and Yorkshire, as they mingled with southerners, would have heard sentences whose initial elements were non-subjects followed by a finite verb less frequently than the required threshold; if we take seriously the statistics from the modern V2 languages and take the threshold to be about 30 percent non-subject-Vf, then southern XP-Vf forms, where the Vf is not I-final and where the initial element is not a *wh* item or negative, are too consistently subject-V to trigger a V2 grammar" (Lightfoot 1997). As we have seen in the previous section, populations of cue-based learners are inherently asymmetric. If cues existed only for a +V2 grammar but not for a -V2 grammar, then +V2 populations would never remain stable. Contact with a foreign population is not necessary to instigate change. One must then ask, why would +V2 populations arise in the first place? Why have the modern +V2 languages not been driven out under the drift of misconvergence of their own learners? These inconsistencies are possible to spot only after employing some more precise reasoning. One is therefore led to suggest that a cue-based learning theory has to allow for cues for both parameter values (+V2 and -V2) so

that case 2 (Batch Error Based Learner) of the learning models discussed earlier is applicable. A more serious discussion of this issue is beyond the scope of this paper.

The goal of computational modeling is to serve as a research tool with which one might reason through the possibilities. Therefore, while empirical studies of the sort conducted by historical linguists in the field remain an essential component of the research program to clarify the data and the phenomena at hand, such computational studies will become increasingly important as we explore various plausible explanations for the phenomena.

As an example of similar reasoning tools, it is worthwhile to consider again the mathematical landscape in evolutionary biology or in language acquisition. At the outset, it is by no means clear that mathematical or computational modeling is necessary or worthwhile in evolutionary biology or language acquisition. Nevertheless, both fields have become mathematized.

The acquisition of language by children is a complex task involving the interplay between the genetic endowment the child brings to the table and the linguistic and extra-linguistic input it receives from the environment. One of the central facts that bears explanation is the ability of children to generalize from finite experience to novel linguistic examples. Explana-

tory positions vary widely from highly innatist (exemplified, perhaps, by generative linguistics) to more learning oriented (exemplified, perhaps, by connectionist accounts). To understand the difficulty of inductive inference, to tease apart the influences of genetic endowment and environmental inputs, it has been necessary to employ computational and mathematical tools in our reasoning. It is almost impossible to make progress otherwise.

Evolutionary biology is perhaps even more complex. The principles of heredity and natural selection, the adaptive forces of competition and cooperation in ecological systems, the emergence of global order from local interactions in biological populations present an array of forces and principles that are difficult a to understand by verbal reasoning. The drive to remove seeming tautologies and internal inconsistencies led to the mathematization of the discipline. The path from the highly theoretical but verbal account presented in Darwin's origin of species through the early pioneering mathematical work of Fisher, Haldane, and Wright to the more computational simulations of recent times was highly successful in clarifying the principles of the discipline. The similarities between evolutionary biology and historical linguistics are many. The analogy of language with species, the presence of variation in the population, the transmission of information from one generation to the next are all notions that have been invoked in

the historical linguistics literature. If we are to go beyond metaphor and make precise the arguments of linguists as they study historical phenomena, we have no choice but to pursue a computational path.

Only time will tell what lessons will emerge from such an enterprise. We are only at the beginning of this subject of evolutionary linguistics and many aspects of the problem will need to come under analysis before general principles as well as particular details get sorted out. As we proceed, both mathematical analysis and computational modeling will have a role to play. However, it is critical to not be seduced by computational or mathematical niceties, to retain a close touch with linguistic reality, and to focus on those aspects of the problem for which we believe a linguistic rather than extralinguistic explanation exists. Linguists are surely the best judge of this matter.

References

[1] M. R. Brent, 1996. *Computational Approaches to Language Acquisition.* Elsevier Press.

[2] E. J. Briscoe, 2000. 'Grammatical Acquisition: Inductive Bias and Coevolution of Language and the Language Acquisition Device.' *Language*,

76.22.

[3] L. Cavalli-Sforza and M. W. Feldman, 1981. *Cultural Transmission and Change: A Quantitative Approach.* Princeton University Press.

[4] R. Clark and I. Roberts, 1993. 'A Computational Model of Language Learnability and Language Change'. *Linguistic Inquiry,* 24.2.

[5] S. Crain and R. Thornton, 1998. *Investigations in Universal Grammar: A Guide to Experiments on the Acquisition of Syntax and Semantics.* MIT Press.

[6] J. Feldman, G. Lakoff, A. Stolcke, S. Weber, 'Miniature Language Acquisition: A Touchstone for Cognitive Science', 1990. *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, pp. 683-693. MIT Press.

[7] A. Galves and C. Galves, 1995. 'Statistical Physics, Pattern Recognition, and Language Change: A Model for European Portuguese'. manuscript. University of Sao Paolo, Brazil.

[8] E. Gibson and K. Wexler, 1994. 'Triggers.' *Linguistic Inquiry,* 25.

[9] J. Hurford, C. Knight, and M. Studdert-Kennedy (ed.), 1998. *Approaches to the Evolution of Language: Social and Cognitive Bases. Cambridge University Press.*

[10] A. van Kemenade and N. Vincent, 1997. *Parameters of Morphosyntactic Change.* Cambridge University Press.

[11] A. W. Kroch and A. Taylor, 1997. 'Verb Movement in Old and Middle English: Dialect Variation and Language Contact,' in van Kemenade and Vincent (ed).

[12] D. Lightfoot, 1997. 'Shifting Triggers and Diachronic Reanalyses.' in van Kemenade and Vincent (ed).

[13] D. Lightfoot, 1991. *How to Set Parameters.* MIT Press.

[14] P. Niyogi, 1997. *The Informational Complexity of Learning: Perspectives on Neural Networks and Generative Grammar.* Kluwer Academic Press, Boston.

[15] P. Niyogi and R. C. Berwick, 1997. 'A Dynamical Systems Model for Language Change'. *Complex Systems,* 11, pp. 161-204.

[16] M. A. Nowak and D. C. Krakauer, 1999. 'The Evolution of Language'. *PNAS,* 96:14464-14469.

[17] D. Osherson, M. Stob, S. Weinstein, 1986. *Systems that Learn.* MIT Press, Cambridge, MA.

[18] S. Pinker, 1984. *Language Learnability and Language Development.* Harvard University Press.

[19] S. Pintzuk, 1991. *Phrase Structures in Competition: Variation and Change in Old English Word Order.* Doctoral Dissertation, University of Pennsylvania.

[20] B. Santorini, 1992. 'Variation and change in Yiddish subordinate clause word order'. *Natural Language and Linguistic Theory,* 10:595-640, 1992.

[21] J. M. Siskind, 1992. *Naive Physics, Event Perception, Lexical Semantics and Language Acquisition.* Doctoral Dissertation, MIT.

[22] Slobin, D. I. (Ed.), *The crosslinguistic study of language acquisition*, (1985) Vol. 1. The Data; (1985) Vol. 2. Theoretical issues; (1992) Vol. 3; (1997) Vol. 4; (1997) Vol. 5. Expanding the contexts. Mahwah, N.J: Erlbaum Associates.

[23] H. Sweet, 1891. *A New English Grammar. Part I: Introduction, Phonology, Accidence.* Oxford: Clarendon Press.

[24] K. Wexler and P. Culicover, 1980. *Formal Principles of Language Acquisition.* MIT Press, Cambridge, MA.

[25] C. D. Yang, 2000. *A Variational Theory of Language Acquistion and Change.* Doctoral Dissertation, MIT.