1

Theories of Cultural Evolution and their Application to Language Change

Partha Niyogi
Bell Laboratories, 600 Mountain
Avenue, Murray Hill, NJ 07974, USA.

We discuss the problem of characterizing the evolutionary dynamics of linguistic populations over successive generations. Here we introduce the framework of Cavalli-Sforza and Feldman (1981) for the treatment of cultural evolution and show how to apply it to the particular case of language change. We relate the approach to that of Niyogi and Berwick (1995) and show how to map trajectories in one to those in the other. In both models, language acquisition serves as the mechanism of transmission of language from one generation to the next. For memory-less learning algorithms and and the case of two languages in contact, we derive particular dynamical systems under the assumptions of both kinds of models. As an application of such computational modeling to historical change, we consider the evolution of English from the 9th century to the 14th century A. D. and discuss the role of such modeling to judge the adequacy of competing linguistic accounts for historical phenomena.

1.1 The Problem of Language Change

A central concern for historical linguists is to characterize the dimensions along which human languages change over time and explain why they do so. Under the assumptions of contemporary linguistic theory, change in linguistic behavior of human populations must be a result of a change in the internal grammars that successive generations of humans employ. The question then becomes: why do the grammars of successive generations differ from each other? In order to answer this question, we need to know how these grammars are acquired in the first place and how the grammars of succeeding generations are related to each other. If such a relationship is uncovered, one might then be able to systematically predict the envelope of possible changes and relate them to actually observed historical trajectories.

Problems in historical or evolutionary linguistics have only recently begun to attract computational attention. This is in contrast to some other areas of linguistics, notably language acquisition, where for example, a significant body of work exists regarding computational models of grammatical inference under a variety of different assumptions (see, for example, Wexler and Culicover, 1980; Osherson, Stob and Weinstein, 1986). At the same time, computational and mathematical work in biological evolution has a long and rich tradition beginning with the pioneering work of Fisher, Wright, Haldane and continuing to the present day. In a treatise in 1981, Cavalli-Sforza and Feldman outlined a general model of cultural change that was inspired by models of biological evolution and has potential and hitherto unexploited applicability to the case of language.

Indeed many motivating examples in Cavalli-Sforza and Feldman (1981) were taken from the field of language change. However, the applicability of such models to language change was not formally pursued there. In this paper, we introduce their basic model and provide one possible way in which the principles and parameters approach to grammatical theory (construed in the broadest possible way) is amenable to their modeling framework.

More recently, a framework for the computational characterization of changing linguistic populations has also been developed in a series of papers by Niyogi and Berwick (1995,1997,1998). We explore here the formal connections between these two approaches for the case of two linguistic variants in competition. In particular, we show how evolutionary trajectories in one framework can be formally translated into the other

and discuss their similarities and differences. To ground the discussion in a particular linguistic context, we show the application of such models to generate insight into possible evolutionary trajectories for the case of diachronic evolution of English from the 9th century A.D. to the 15th century A.D. Finally, we provide some extensions of the basic Cavalli-Sforza and Feldman framework that might allow us to characterize the effect of spatial (geographical) location on the linguistic interactions between individuals in a population and the evolutionary consequences of such interactions.

The reader might ask – what is the role of formal modeling of the sort described in this paper in gaining insight in historical or evolutionary linguistics? From our perspective, these techniques provide research tools to increase our understanding about the range of explanations for historical phenomena. The formal model places constraints on the kinds of informal, largely descriptive accounts of attested historical changes which linguists develop. Tools of this sort therefore help us figure out the plausibility of various accounts and rule out logical inconsistencies that might be difficult to spot in a more informally developed treatment.

1.2 The Cavalli-Sforza and Feldman Theory of Cultural Transmission and Change

Cavalli-Sforza and Feldman (1981) outline a theoretical model for cultural change over generations. Such a model closely mimics the transmission of genetic parameters over generations: except now, we have "cultural" parameters that are transmitted from parents to children with certain probabilities. In the model (hereafter referred to as the CF model in this paper), the mechanism of transmission is unknown — only the probabilities of acquiring one of several possible variations of the trait are known.

We reproduce their basic formulation for *vertical* transmission (from one generation to the next) of a particular binary valued trait. Assume a particular cultural trait has one of two values. Some examples of traits they consider are political orientation (Democrat/Republican) or health habits (smoker/non-smoker) and so on. Let the two values be denoted by H and L. Each individual is assumed to have exactly one of these two values. However, such a value is presumably not innate but learned.

A child born to two individuals (mother and father) will acquire one of these two possible values over its lifetime. The probability with which it

Paternal Val.	Maternal Val.	$P(ChildVal. = L)$	$P(Types)$	Random Mating
L	L	b_3	p_3	u_t^2
L	H	b_2	p_2	$u_t(1-u_t)$
H	L	b_1	p_1	$u_t(1-u_t)$
H	H	b_0	p_0	$(1-u_t)^2$

Table 1: The cultural types of parents and children related to each other by their proportions in the population. The values depicted are for vertical transmission and random mating.

will acquire each of these traits depends upon its immediate environment – in the standard case of their model (though variations are considered[1]), its parents. Thus one can construct table 1.2.

The first three columns of table 1.2 are self-explanatory. As one can see easily enough, parental compositions can be one of 4 types depending upon the values of the cultural traits of each of the parents. We denote by b_i the probability with which a child of the ith parental type will attain the trait L (with $1 - b_i$, it attains H.) In addition, let p_i be the probability of the ith parental type in the population. Finally, we let the proportion of people having type L in the parental generation be u_t. Here t indexes the generation number and therefore proportion of L types in the parental generation is given by u_t and proportion of L types in the next generation (children who mature into adults) is given by u_{t+1}.

Under random mating, one sees that the proportion of parents of type (L, L) , i.e., male L types married to female L types is u_t^2. Similarly one can compute the probability of each of the other combinations.

Given this, they go on to show that the proportion of L types in the population will evolve according to the following quadratic update rule:

$$u_{t+1} = Bu_t^2 + Cu_t + D \tag{1.1}$$

where $B = b_3 + b_0 - b_1 - b_2$, $C = b_2 + b_1 - 2b_0$, and $D = b_0$. In this manner, the proportion of L types in generation $t + 1$ (given by u_{t+1})

[1] Pure vertical transmission involves transmission of cultural parameters from parents to children. They also consider (i) *oblique* transmission where members of the parental generation other than the parents affect the acquisition of the cultural parameters (ii) *horizontal* transmission where members of the same generation influence the individual child. We discuss in a later section the approach of Niyogi and Berwick (1995) that involves oblique transmission of a particular sort and different from the Cavalli-Sforza and Feldman (1981) treatment.

is related to the proportion of L types in generation t (given by u_t).

A number of properties and variations of this basic evolutionary behavior are then evaluated (Cavalli-Sforza and Feldman, 1981) under different assumptions.

Thus, we see that evolution (change) of the cultural traits within the population is essentially driven by the probabilities with which children acquire the traits given their parental types. The close similarity of this particular model[2] to biological evolution is clear: (1) trait values, like gene-types are discrete (2) their transmission from one generation to another depends (in a probabilistic sense) only on the trait-values (gene-types) of the parents.

The basic intuition they attempted to capture in their model is that cultural traits are acquired (learned) by children from their parents. Thus, by noting the population mix of different parental types and the probabilities with which they are transmitted one can compute the evolution of these traits within the population. They had hoped to apply this model to language. In the next section we show how to do this.

1.3 Instantiating the CF Model for Languages

In order to apply the model to the phenomena of language change, the crucial point to appreciate is that the mechanism of language transmission from generation to generation is "language learning", i.e., children learn the language of their parents as a result of exposure to the primary linguistic data they receive from their linguistic environment. Therefore, in this particular case, the transmission probabilities b_i's in the model above will depend upon the learning algorithm they employ. We outline this dependence for a simplified situation corresponding to two language types in competition.

1.3.1 One Parameter Models

Assume there are two languages in the world — L_1 and L_2. Such a situation might effectively arise if two languages differing by a linguistic parameter are in competition with each other and we will discuss later the historical example of syntactic change in English for which this is a

[2] To avoid misinterpretation, it is worthwhile to mention that extensions to continuous valued traits have been discussed. Those extensions have less relevance for the case of language since linguistic objects are essentially discrete.

reasonable approximation. We consider languages to be subsets of Σ^* in the usual sense where Σ is a finite alphabet. Furthermore, underlying each language L_i is a grammar g_i that represents the internal knowledge that speakers of the language possess of it.

Individuals are assumed to be native speakers of exactly one of these two languages. Furthermore, let speakers of L_1 produce sentences with a probability distribution P_1 and speakers of L_2 produce sentences with a distribution P_2. There are now four parental types and children born to each of these parental types are going to be exposed to different linguistic inputs and as a result will acquire a particular language with different probabilities.

In the abstract, let us assume that children follow some acquisition algorithm $\mathcal{A}$ (for a brief overview of the structure of learning theory, see appendix) that operates on the primary linguistic data they receive and comes up with a grammatical hypothesis – in our case, a choice of g_1 or g_2 (correspondingly L_1 or L_2). Formally, let $\mathcal{D}_k$ be the set of all subsets of Σ^* of cardinality k. Each subset of Σ^* of cardinality k is a candidate dataset consisting of k sentences that might constitute the primary linguistic data a child receives. Clearly $\mathcal{D}_k$ is the set of all candidate datasets of size k. Then $\mathcal{A}$ is a computable mapping from the set $\cup_{k=1}^{\infty}\mathcal{D}_k$ to $\{g_1, g_2\}$. We now make the following assumptions.

1. Children of parents who speak the same language receive examples only from the unique language their parents share, i.e., children of parents speaking L_1 receive sentences drawn according to P_1 and children of parents speaking L_2 receive examples drawn according to P_2.
2. Children of parents who speak different languages receive examples from an *equal* mixture of both languages, i.e., they receive examples drawn according to $\frac{1}{2}P_1 + \frac{1}{2}P_2$.
3. After k examples, children "mature" and whatever grammatical hypothesis they have, they retain for the rest of their lives.

Thus the learning algorithm $\mathcal{A}$ operates on the sentences it receives. These sentences in turn are drawn at random according to a probability distribution that depends on the parental type. We now define the following quantity:

$$g(\mathcal{A}, P, k) = \sum_{\{w \in \mathcal{D}_k : \mathcal{A}(w) = g_1\}} \prod_{i=1}^{k} P(w_i) \tag{1.2}$$

Paternal Language	Maternal Language	P	Prob. Child speaks L_1
L_1	L_1	P_1	$b_3 = g(\mathcal{A}, P_1, k)$
L_1	L_2	$\frac{1}{2}P_1 + \frac{1}{2}P_2$	$b_2 = g(\mathcal{A}, \frac{1}{2}P_1 + \frac{1}{2}P_2, k)$
L_2	L_1	$\frac{1}{2}P_1 + \frac{1}{2}P_2$	$b_1 = g(\mathcal{A}, \frac{1}{2}P_1 + \frac{1}{2}P_2, k)$
L_2	L_2	P_2	$b_0 = g\mathcal{A}, P_2, k)$

Table 2: The probability with which children attain each of the language types, L_1 and L_2 depends upon the parental linguistic types, the probability distributions P_1 and P_2 and the learning algorithm $\mathcal{A}$.

Recall that each element $w \in \mathcal{D}_k$ is a set of k sentences. In eq. 1.2 we denote by w_i the ith sentence of the set w. Therefore, $g(\mathcal{A}, P, k)$ is the probability with which the algorithm $\mathcal{A}$ hypothesizes grammar g_1 given a random i.i.d. draw of k examples according to probability distribution P. Clearly, g characterizes the behavior of the learning algorithm $\mathcal{A}$ if sentences were drawn according to P. It is worthwhile to note that learnability (in the limit, in a stochastic generalization of Gold, 1967) requires the following:

Statment 1. *If the support of P is L_1 then $\lim_{k \longrightarrow \infty} g(\mathcal{A}, P, k) = 1$ and if the support of P is L_2 then $\lim_{k \longrightarrow \infty} g(\mathcal{A}, P, k) = 0$.*

In practice, of course, we have made the assumption that children "mature" after k examples: so a reasonable requirement is that g be high if P has support on L_1 and low if P has support on L_2. Given this, we can now write down the probability with which children of each of the four parental types will attain the language L_1. These are shown in table 2.

Thus we can express the b_i's in the CF model of cultural transmission in terms of the learning algorithm. This is reasonable because after all, the $b_i's$ attempt to capture the fact that traits are "learned" —- in the case of languages, they are almost certainly learned from exposure to linguistic data.

Under random mating[3], we see that the population evolves according to equation 1.1. Substituting the appropriate $g's$ from table 2 above in place of the $b_i's$ we obtain an evolution that depends upon P_1,P_2,$\mathcal{A}$, and

[3] We have only considered the case of random mating here for illustrative convenience. The extension to more assortative forms of mating can be carried using the standard techniques in population biology.

k.

1.3.2 An Alternative Approach

In a recent attempt to explicitly characterize the problem of language change, Niyogi and Berwick (1995,1997,1998) develop a model (hereafter, we refer to this class of models as NB models) for the phenomenon making the following simplifying assumptions.

1. The population can be divided into children (learners) and adults (sources).
2. All children in the population are exposed to sentences drawn from the same distribution.
3. The distribution with which sentences are drawn depends upon the distribution of language speakers in the adult population.

The equations for the evolution of the population under these assumptions were derived. Let us consider the evolution of two-language populations. At any point, one can characterize the state of the population by a single variable ($s_t \in [0,1]$) denoting the proportion of speakers of L_1 in the population. Further assume, as before, that speakers of L_1 produce sentences with distribution P_1 on the sentences of L_1 and speakers of L_2 produce sentences with distribution P_2 on the sentences of L_2.

The evolution of s_t over time (the time index t denotes generation number) was derived in terms of the learning algorithm $\mathcal{A}$, the distributions P_1 and P_2, and the maturation time k. Essentially, this evolution turns out to be the following:

$$s_{t+1} = f(s_t) = g(\mathcal{A}, s_t P_1 + (1 - s_t)P_2, k)$$

The interpretation is clear. If the previous state was s_t, then children are exposed to sentences drawn according to $s_t P_1 + (1 - s_t)P_2$. The probability with which the average child will attain L_1 is correspondingly provided by g and therefore one can expect that this will be the proportion of L_1 speakers in the next generation, i.e., after the children mature to adulthood.

Niyogi and Berwick (1995,1997,1998) derive the specific functional form of the update rule f (equivalently g) for a number of different learning algorithms. In the next section, we show how these two approaches to characterizing the evolutionary dynamics of linguistic pop-

ulations are related. Specifically, we show how the evolutionary update rule f in the NB framework is explicitly related to the update rule in the CF framework.

1.3.3 Transforming NB Models into the CF Framework

Let the NB update rule be given by $s_{t+1} = f(s_t)$. Then, we see immediately that:

1. $b_3 = f(1)$
2. $b_2 = b_1 = f(0.5)$
3. $b_0 = f(0)$.

The CF update rule is now given by eq. 1.1. The update as we have noted is quadratic and the coefficients can be expressed in terms of the NB update rule f. Specifically, the system evolves as

$$s_{t+1} = (f(1) + f(0) - 2f(0.5))\, s_t^2 + (2f(0.5) - 2f(0))\, s_t + f(0) \quad (1.3)$$

Thus we see that if we are able to derive the NB update rule, we can easily transform it to arrive at the CF update rule for evolution of the population. The difficulty of deriving both rules rests upon the difficulty of deriving the quantitiy g that appears in both update rules. Notice further that the CF update rule is *always* quadratic while the NB update rule is in general not quadratic.

The essential difference in the nature of the two update rules stems from the different assumptions made in the modeling process. Particularly, Niyogi and Berwick (1995,1997,1998) assume that all children receive input from the same distribution. Cavalli Sforza and Feldman (1981) assume that children can be grouped into four classes depending on their parental type. The crucial observation at this stage is that by dividing the population of children into classes that are different from each other, one is able to arrive at alternate evolutionary dynamics. In a later section we utilize this observation to divide children into classes that depend on their geographical neighbourhood. This will allow us to derive a generalization of the NB model for neighbourhoods. Before proceeding any further, let us now translate the update rules derived in Niyogi and Berwick (1995,1996) into the appropriate CF models. The update rules are dervied for memoryless learning algorithms operating on grammars. We consider an application to English with grammars

represented in the principles and parameters framework.

1.4 CF Models for Some Simple Learning Algorithms

In this section we consider some simple online learning algorithms (like the Triggering Learning Algorithm of Gibson and Wexler (1994); henceforth TLA) and show how their analysis within the NB model can be plugged into eq. 1.3 to yield the dynamics of linguistic populations under the CF model.

1.4.1 TLA and its Evolution

How will the population evolve if the learning algorithm $\mathcal{A}$ in question is the Triggering Learning Algorithm[4] (or related memoryless learning algorithms in general)? The answer is simple. We know how the TLA driven system evolves in the NB model (from an analysis of the TLA in Niyogi and Berwick, 1996). All we need to do is to plug such an evolution into eq. 1.3 and we are done.

Recall that the TLA is as follows:

1. **Initialize:** Start with randomly chosen input grammar.
2. Receive next input sentence, s.
3. If s can be parsed under current hypothesis grammar, go to 2.
4. If s cannot be parsed under current hypothesis grammar, choose another grammar uniformly at random.
5. **If** s can be parsed by new grammar, retain new grammar, **else** go back to old grammar.
6. Go to 2.

It is shown in Niyogi and Berwick, 1996, that such an algorithm can be analyzed as a Markov Chain whose state space is the space of possible grammars and whose transition probabilities depend upon the distribution P with which sentences are drawn. Using such an analysis, the function f can be computed. For the case of two grammars (languages)

[4] The Triggering Learning Algorithm has been chosen here for illustrative purposes to develop the the connections between individual acquisition and population change in a concrete manner in both NB and CF models. Replacing the TLA by another learning algorithms does not alter the spirit of the major points we wish to make in this paper but rather the details of some of the results we might obtain here. In general, acquisition algorithms can now be studied from the point of view of adequacy with respect to historical phenomena.

in competition under the assumptions of the NB model, this function f is seen to be:

$$f(s_t) = \frac{s_t(1-a)}{(1-b)+s_t(b-a)} + \frac{[b-s_t(b-a)]^k[(1-b)+s_t(a+b-2)]}{2[(1-b)+s_t(b-a)]} \quad (1.4)$$

In eq. 1.4, the evolving quantity s_t is the proportion of L_1 speakers in the community. The update rule depends on parameters a, b and k that need further explanation. The parameter a is the probability with which ambiguous sentences (sentences that are parsable by both g_1 and g_2) are produced by L_1 speakers, i.e., $a = \sum_{w \in L_1 \cap L_2} P_1(w)$; similarly, b is the probability with which ambiguous sentences are produced by L_2 speakers, i.e., $b = \sum_{w \in L_1 \cap L_2} P_2(w)$. Finally, k is the number of sentences that a child receives from its linguistic environment before maturation. It is interesting to note that the only way in which the update rule depends upon P_1 and P_2 is through the parameters a and b that are bounded between 0 and 1 by construction.

It is not obvious from eq. 1.4 but it is possible to show that f is a polynomial (in s_t) of degree k. Having obtained $f(s_t)$, one obtains the quadratic update rule of the CF model by computing the b_i's according to the formulae given in the earlier section. These are seen to be as follows:

$$b_3 = 1 - \frac{a^k}{2}; b_0 = \frac{b^k}{2}; b_1 = b_2 = \frac{(1-a)}{(1-a)+(1-b)} + (\frac{a+b}{2})^k \frac{(a-b)}{2[(1-a)+(1-b)]}$$

The following remarks are in order:

1. For $k = 2$, i.e., where children receive exactly two sentences before maturation, both the NB and CF models yield quadratic update rules for the evolution of the population. For the NB model, the following is true: (i) for $a = b$, there is *exponential* growth (or decay) to one fixed point of $p* = \frac{1}{2}$, i.e., populations evolve until both languages are in equal proportion and they coexist at this level; (ii) for $a \neq b$, there is *logistic* growth (or decay) and in particular, if $a < b$ then there is one stable fixed point $p^*(a, b)$ whose value depends upon a, b and is greater than $\frac{1}{2}$. If $a > b$ then there is again one stable fixed point $p^*(a, b)$ that is less than $\frac{1}{2}$. Populations tend to the stable fixed point from all initial conditions in logistic fashion. The value of p^* as a function of a and b is shown in fig. 1.
2. For $k = 2$, the evolution of the CF model is as follows: (i) for $a = b$, there is exponential growth (or decay) to one fixed point of $p^* = \frac{1}{2}$.

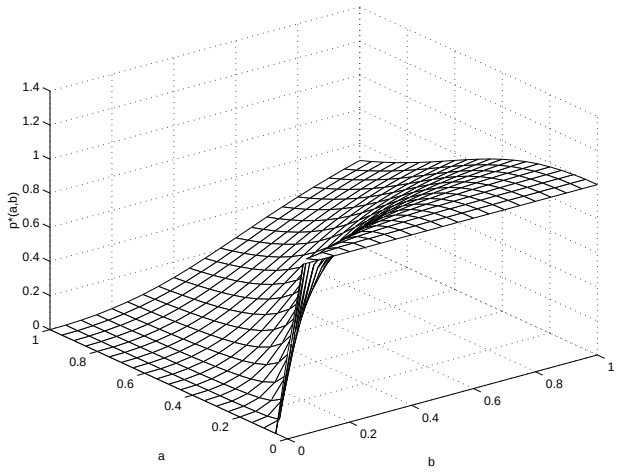

figure 1.1. The fixed point $p^*(a, b)$ for various choices of a and b for the NB model with $k = 2$.

(ii) for $a \neq b$, there is still one stable fixed point whose value can be seen as a function of a and b in fig. 2. For $b > a$, the value of this fixed point is greater than $\frac{1}{2}$, for $a > b$, the value is less than $\frac{1}{2}$. While the overall qualitative behavior of the two models for this value of k, are quite similar, the value of $p^*(a, b)$ is not identical. This can be seen from fig. 3 where we plot the difference (between p^*_{NB} and p^*_{CF}) in values of the fixed point obtained for each choice of a and b.

3. If one considers the limiting case where $k \longrightarrow \infty$, i.e., where children are given an infinite number of examples to mature, then the evolution of both the NB and the CF models have the same qualitative character. There are three cases to consider: (i) for $a = b$, we find that $s_{t+1} = s_t$, i.e., there is no change in the linguistic composition; (ii) for $a > b$, the population composition s_t tends to 0 (iii) for $a < b$, the population composition s_t tends to 1. Thus one of the languages *drives the other out* and the evolutionary change proceeds to completion. However the rates at which this happens differs under the differing assumptions of the NB and the CF models. This difference is explored in a later section as we consider the application of the models to the historical evolution of English syntax. It is worthwhile to add that in real life, $a = b$ is unlikely to be exactly true — therefore language contact between populations is likely to drive one out of existence.

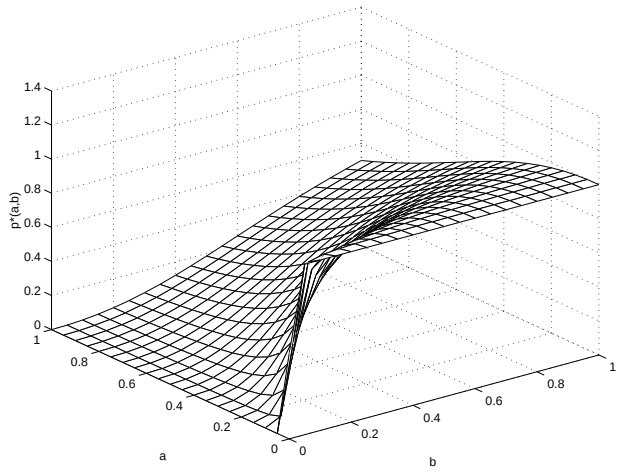

figure 1.2. The fixed point $p^*(a, b)$ for various choices of a and b for the CF model with $k = 2$.

Additionally, the limiting case of large k is also more realistic since children typically get adequate primary linguistic data over their learning years in order to acquire a unique target grammar with high probability in homogeneous linguistic communities where a unique target grammar exists. In the treatment of this paper, we have always assumed that learners attain a single target grammar. Often, when two languages come in contact, learners typically attain both grammars in addition to a reasonable understanding of the social and statistical distribution of the two grammars in question. This can be handled within the framework we discuss here by requiring the learner to actually learn (estimate) a mixture factor ($\lambda \in [0, 1]$, say) that decides in what proportion the two grammars are to be used. A value of $\lambda = 0$ or $\lambda = 1$ would then imply that the learner had actually attained a unique grammar. One can then analyze a population of such learners to characterize their evolutionary consequences. We do not discuss such an analysis here.

4. We have not yet been able to characterize the evolutionary behavior of populations for arbitrary values of k.

From the preceding discussion we see that the evolutionary characteristics of a population of linguistic agents can be precisely derived under certain simplifying assumptions. We show how the differing assump-

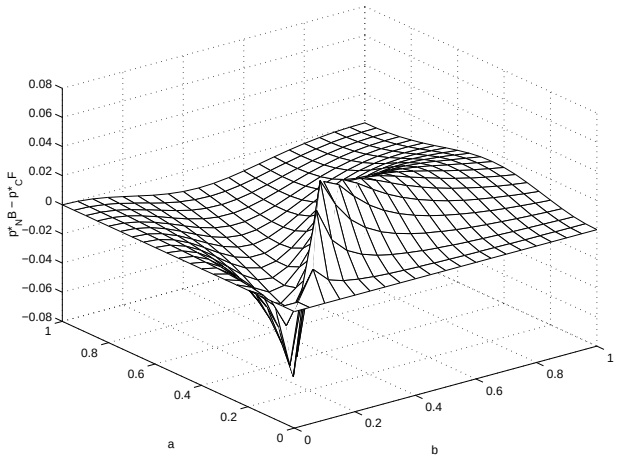

figure 1.3. The difference in the values of $p*(a,b)$ for the NB model and the CF model $p^*_{NB} - p^*_{CF}$ for various choices of a and with $k = 2$. A flat surface taking a value of zero at all points would indicate that the two were identical. This is not the case.

tions of the NB model and the CF model yield dynamical systems with different behaviors and how these models relate to each other.

1.4.2 A Historical Example

So far the development has been fairly abstract. To ground our discussion in a particular context, let us consider the phenomena surrounding the evolution of Old English to Modern English and its treatment within both kinds of models.

One of the significant changes in the syntax of English as it evolved from the 9th century to the 14th century is the change in word order. Consider, for example, the following passage taken from the Anglo Saxon Chronicles (878 A.D.) and reproduced in Trask (1996):

> *Her ... AElfred cyning ... gefeaht wid ealne here, and hine*
> Here ... Alfred king ... fought against whole army and it
> *geflymde, and him aefter rad od pet geweorc, and paer saet*
> put to flight and it after rode to the fortress and there camped
> *XIIII niht, and pa sealde se here him gislas and myccle*
> fourteen nights and then gave the army him hostages and great
> *adas, pet hi of his rice woldon, and him eac geheton*

oaths that they from his kingdom would [go] and him also promised
pet heora cyng fulwihte onfon wolde, and hi paet gelaston ...
and their king baptism receive would and they that did

The original text is in italics and a word for word translation (gloss) provided immediately below each line of the passage. Some phrases have been underlined to indicate the unusual word order prevalent in the writing of the times. Sampling the historical texts over the period from Old to Middle English, one finds that the early period shows three major alternations (i) verb phrases (VP) may show Object-Verb (OV) or Verb-Object (VO) order (ii) the inflectional head (I) may precede (I-medial) or follow (I-final) the verb phrase (iii) there may or may not be movement of the inflected verb to head of CP (complementizer position in clauses) (following the notation of Government and Binding theory; see Haegeman, 1991).

For the purposes of the discussion in this paper, we will collapse the OV/VO and I-final/I-medial distinctions into a single head-complement parameter in accordance with commonly made assumptions of the principles and parameters approach to grammatical theory. The movement of the finite verb to second position is related to the V2 parameter — modern German and Dutch are +V2 while modern English is -V2. Therefore, the two grammatical parameters at issue are:

1. The **head-complement** parameter: this denotes the order of constituents in the underlying phrase-structure grammar. Recall from X-bar theory that phrases XP have a **head** (X) and **complement**, e.g. the verb phrase *ate with a spoon* and the prepositional phrase *with a spoon* have as a head the verb *ate* and the preposition *with* respectively. Grammars of natural languages could be **head**-first or **head**-final. Thus X-bar phrase structure rules have the form (X and Y are arbitrary syntactic categories in the notation below):
   head-first: (i) $XP \longrightarrow X'\ YP$ (ii) $X' \longrightarrow X$
   head-final: (i) $XP \longrightarrow Y\ P\ X'$ (ii) $X' \longrightarrow X$
2. The **V2** parameter: this denotes the tendency in some languages of the finite verb to move from its base position to the head of the complementizer (C of CP) by V to I to C raising. The specifier of CP has to be filled resulting in the verb appearing to be in the second position in linear order of constituents. Grammars of natural languages could be **+V2** or **-V2**. Thus
   +V2: Obligatory movement of V to I to C and specifier of CP filled.

-V2: V2 movement absent.

Modern English is exclusively head-first and -V2. Old English seems to be largely head-final and +V2. How did such remarkable changes in grammars occur? There are several competing accounts for these changes (see chapters by Kroch and Taylor; Lightfoot; and Warner in Van Kemenade (1997) for discussions) but there seems to be some agreement that there were two competing grammars — a northern Scandinavian based +V2 grammar and a southern indigenous -V2 grammar. The first of these grammars was lost as the populations came into contact. Invoking learnability arguments as an explanation for such a change, Lightfoot (1997) writes: "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 % of matrix clauses with initial non-subject in Spec of CP, 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-initial to trigger a V2 grammar." [implying that the +V2 grammar was therefore lost over time] These are the kinds of arguments that can be modeled precisely and tested for plausibility within the framework we have discussed here.

We will not attempt in this section to do justice to the various accounts of the historical change of English in a serious manner as the subject of such a discussion is well beyond the scope of the current paper. However, for illustrative purposes, we discuss below the evolutionary trajectories of populations with two competing grammar types that come into contact. The grammar types have been chosen to capture the parametric oppositions that played themselves out over the course of the historical evolution of English.

1.4.2.1 Case I: +V2/-V2 for head-first grammars

Imagine that two linguistic populations came together and the two languages in competition differed only by one parameter — the $V2$ parameter. Further assume that all other grammatical parameters of these two languages were identical to modern English. Children growing up in the mixed communities would hear sentences from both grammatical types. Suppose they set (learnt) all other grammatical parameters correctly and it was only in the V2 parameter that children differed from each other in how they set it — i.e., some acquired the $+V2$ grammar and some

acquired the $-V2$ grammar. How would the population evolve? Would the $+V2$ grammar die out over time? What conditions must exist for this to happen?

These questions can be addressed within the framework that we have developed in this paper. To begin with, we need to identify the sets L_1 and L_2. Following Gibson and Wexler (1994), we derive the set of degree-0 sentences[5] (with no recursion) that are associated with the +V2 and -V2 grammars. These are listed below where S = subject, V = verb, O1 = direct object; O2 = indirect object; Aux = auxiliary; Adv = adverb.

g_1: *-V2; Head-first; Spec-first*

L_1 = { S V, S V O, S V O1 O2, S Aux V, S Aux V O, S Aux V O1 O2, Adv S V, Adv S V O, Adv S V O1 O2, Adv S Auz V, Adv S Aux V O, Adv S Aux V O1 O2 }

The grammar underlying these sentences corresponds to that of modern English. For example, the sentence type (S Aux V O1 O2) maps to actual sentences like *John will eat beef in London.*

g_2: *+V2; Head-first; Spec-first*

L_2 = { S V, S V O, O V S, S V O1 O2, O1 V S O2, O2 V S O1, S Aux V, S Aux V O, O Aux S V, S Aux V O1 O2, O1 Aux S V O2, O2 Aux S V O1, Adv S V, Adv V S O, Adv V S O1 O2, Adv Aux S V, Adv Aux S V O, Adv Aux S V O1 O2}

This grammar requires obligatory movement of the inflected verb to second position (actually to C and the specifier of CP must be filled). Thus, an example of an actual sentence (not following English word order of course) corresponding to the sentence type Adv V S O1 O2 is *often saw we many students in London.*

Given these lists of sentences, one can obtain by taking an intersection of the two languages, the set of ambiguous types, i.e., sentence types that may have different but valid parses under the two assumptions. We see that

[5] Of course, both L_1 and L_2 have infinite sentences each. Recall that the evolutionary properties of the population will depend upon the probability distributions P_1 and P_2 with which sentences are produced. In practice, due to cognitive limitations, speakers produce sentences with bounded recursion. Therefore P_1 and P_2 will have effective support on a finite set only. Furthermore, the learning algorithm of the child $\mathcal{A}$ operates on sentences and a common psycholinguistic premise is that children learn only on the basis of degree-0 sentences (Gibson and Wexler, 1994) and all sentences with recursion are ignored in the learning process. We have adopted this premise for the purposes of this paper. Therefore only degree-0 sentences are considered in this analysis.

$L_1 \cap L_2$ = { S V, S V O, S V O1 O2, S Aux V, S Aux V O, S Aux V O1 O2}

We have considered several variants of both the CF and NB models for two languages in competition in the previous sections. Recall that for large k, the qualitative behavior of the two models is similar and L_1 would drive L_2 out from all initial conditions if and only if $a < b$. Here a is the probability measure on the set of ambiguous sentences produced by speakers of L_1 and b is the probability measure on the set of ambiguous sentences produced by speakers of L_2. This situation would lead to the loss of +V2 grammar types over time.

Under the unlikely but convenient assumption that P_1 and P_2 are uniform distributions on degree-0 sentences of their respective languages (L_1 and L_2), we see that

$$a = \frac{1}{2} > b = \frac{1}{3}$$

Therefore, the +V2 grammar, rather than being lost over time would tend to be gained over time. Shown in fig. 1.4.2.1 are the evolutionary trajectories in the CF and NB models for various choices of a and b. Some further remarks are in order:

1. The directionality of change is predicted by the relationship of a with b. While, uniform distributions of degree-0 sentences predict that the V2 parameter would be gained rather than lost over time, the empirical validity of this assumption needs to be checked. From corpora of child-directed sentences in synchronic linguistics and aided perhaps by some historical texts, one might try to empirically assess the distributions P_1 and P_2 by measuring how often each of the sentence types occur in spoken language and written texts. These empirical measures are being conducted at the present time and will be the subject of a later paper.
2. The dynamcal systems that we have derived and applied to this particular case hold only for the case of memoryless learning algorithms like the TLA. For other kinds of algorithms and their evolutionary consequences, see Niyogi and Berwick, 1997.

1.4.2.2 Case II: OV/VO for +V2 grammars

Here we consider a head-first (comp-final) grammar in competition with a head-final (comp-first) grammar where both are +V2 grammars that have the same settings for all other parameters — settings that are the

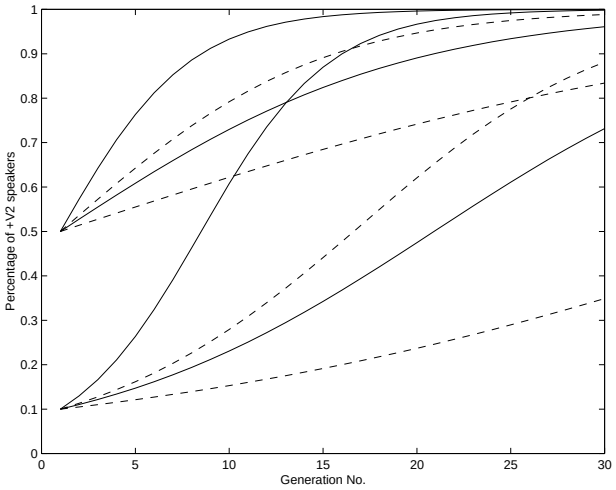

figure 1.4. Trajectories of V2 growth. Shown in the figure are the evolving trajectories of s_t = proportion of +V2 grammars in the population over successive generations. The solid curves denote the evolutionary trajectories under the NB model; the dotted curves denote the trajectories under the CF model. Two different initial population mixes are considered (a) 0.1 initial +V2 speakers (b) 0.5 initial +V2 speakers. For each initial mix and each model (CF and NB) the upper curve (faster change) corresponds to a choice of $a = 0.5$ and $b = 0.33$ and $a = 0.4$ and $b = 0.33$ respectively. Notice that the NB model has a faster rate of change than the CF model.

same as that of modern English. Therefore, one of the two grammars (head-first setting) is identical to modern English except for the V2 parameter. It is also the same as g_2 of the previous section. The other grammar differs from modern English by two parameters.

As in the previous section, following Gibson and Wexler (1994) we can derive the degree-0 sentences associated with each of the two languages. We do this below:

g_1: *+V2; Head-first; Spec-first*

L_1 = { S V, S V O, O V S, S V O1 O2, O1 V S O2, O2 V S O1, S Aux V, S Aux V O, O Aux S V, S Aux V O1 O2, O1 Aux S V O2, O2 Aux S V O1, Adv S V, Adv V S O, Adv V S O1 O2, Adv Aux S V, Adv Aux S V O, Adv Aux S V O1 O2}

This grammar is the same as g_2 of the previous section.

g_2: *+V2; Head-final; Spec-first*

L_2 = { S V, S V O, O V S, S V O2 O1, O1 V S O2, O2 V S O1, S

Aux V, S Aux O V, O Aux S V, S Aux O2 O1 V, O1 Aux S O2 V, O2 Aux S O1 V, Adv V S, Adv V S O, Adv V S O2 O1, Adv Aux S V, Adv Aux S O V, Adv Aux S O2 O1 V}

An example of a sentence type corresponding Adv V S O2 O1 is *often saw we in London many students.*

We can therefore straightforwardly obtain the set $L_1 \cap L_2$ as

$L_1 \cap L_2$ = { S V, S V O, O V S, O1 V S O2, O2 V S O1, S Aux V, O Aux S V, Adv V S O, Adv Aux S V}

Assuming P_1 and P_2 are uniform distributions on the degree-0 sentences of their respective languages, we see that

$$a = \frac{1}{2} = b$$

Therefore, under the assumptions of both the NB and the CF models there is no particular tendency for one grammar type to overwhelm the other. Language mixes would remain the same. If for some reason, a became slightly less than b, we see that the head-final (comp-first) language would be driven out and only the head-first language would remain. This would replicate the historically observed trajectory for the case of English. The rate is faster for the NB model than it is for the CF model.

1.4.2.3 A Final Note

Taking stock of our modeling results, we see that when a +V2 and a -V2 grammar come together (other parameters being the same) there is an inherent asymmetry with the -V2 grammar being more likely to lose out in the long run. On the other hand when a head-first and head-final grammar come together, there is no particular proclivity to change — the directionality could go either way. The reason for this asymmetry is seen to be in the asymmetry in the number of surface degree-0 sentences that are compatible with each of the grammars in question with +V2 grammars giving rise to a larger variety of surface sentences and therefore ambiguous sentences (those parsable with both +V2 and -V2 constraints) constitute a smaller proportion of the total sentence types of such grammars.

In conclusion, however, it is worthwhile to reiterate again our motivation in working through this particular example of the syntactic change in English. There are many competing accounts of how English changed over the years. Among other things, these accounts differ in (i) the precise grammatical characterization of the two grammars in compe-

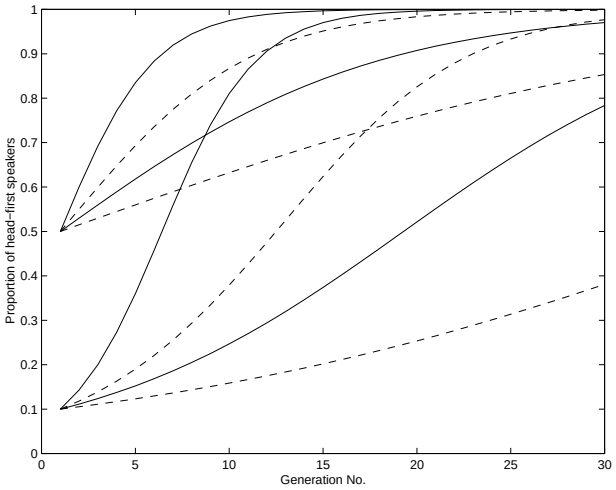

figure 1.5. Trajectories of head-first growth. Shown in the figure are the evolving trajectories of s_t = proportion of head-first grammars in the population over successive generations. The solid curves denote the evolutionary trajectories under the NB model; the dotted curves denote the trajectories under the CF model. Two different initial population mixes are considered (a) 0.1 initial head-first speakers (b) 0.5 initial head-first speakers. For each initial mix and each model (CF and NB) the upper curve (faster change) corresponds to a choice of $a = 0.4$ and $b = 0.6$ and $a = 0.47$ and $b = 0.53$ respectively. Notice that the NB model has a faster rate of change than the CF model.

tition (ii) the number of parametric changes that happened and their description in the context of a grammatical theory (iii) the nature of the learning mechanism that children employ in learning grammmars (e.g. monolingual versus bilingual acquisition) and so on. However, these factors can be modeled and the plausibility of any particular account can then be verified. To give the reader a sense of how this might happen in a linguistically grounded manner, we worked through these examples — not to make a linguistic point but to demonstrate the applicability of this kind of computational thinking to historical problems.

1.5 A Generalized NB Model for Neighbourhood effects

The Cavalli-Sforza and Feldman model described in this paper assumes that children can be divided into four classes depending upon their parental types. The children of each class then receive input sentences from a different distribution depending upon their parental type. The Niyogi and Berwick approach on the other hand assumes that all children in the population receive inputs from the same distribution that depends on the linguistic composition of the entire parental generation. In this section, we consider a generalization of both approaches with a particular view to modeling "neighbourhood" effects in linguistic communities.

The key idea here is that in multiple language communities speakers tend to cluster in linguistically homogeneous neighbourhoods. Consequently children growing up in the community might receive data drawn from different distributions depending upon their spatial location within the community. Imagine as usual a two-language population consisting of speakers of L_1 or L_2. We now let the parental generation of speakers reside in adjacent neighbourhoods. Children receive sentences drawn from different distributions depending upon their location in this neighbourhood. At one end of the scale, children receive examples drawn only from L_1. At the other end of the scale, children receive examples drawn only from L_2. In the middle — at the boundary between the two neighborhoods as it were — are children who receive examples drawn from both sources.

Let us develop the notion further. Let children of type α be those who receive examples drawn according to a distribution $P = \alpha P_1 + (1-\alpha)P_2$. Here P_1 is the probability with which speakers of L_1 produce sentences and P_2 is the probability with which speakers of L_2 produce sentences. The quantity $\alpha \in [0, 1]$ is the proportion of L_1 speakers that an α-type child is effectively exposed to. Children will be of different α types depending upon their spatial location.

How do we characterize location? Let location be indicated by a one-dimensional real-valued variable n in the interval $[0, 1]$. Let speakers be uniformly distributed on this interval so that speakers of L_1 are close to $n = 0$ and speakers of L_2 are close to $n = 1$. Let the proportion of L_1 speakers in the population be s_t. Therefore, all children located in $[0, s_t]$ are in the L_1 speaking neighbourhood and all children located in $[s_t, 1]$ are in the L_2 speaking neighbourhood. Let us now define the mapping

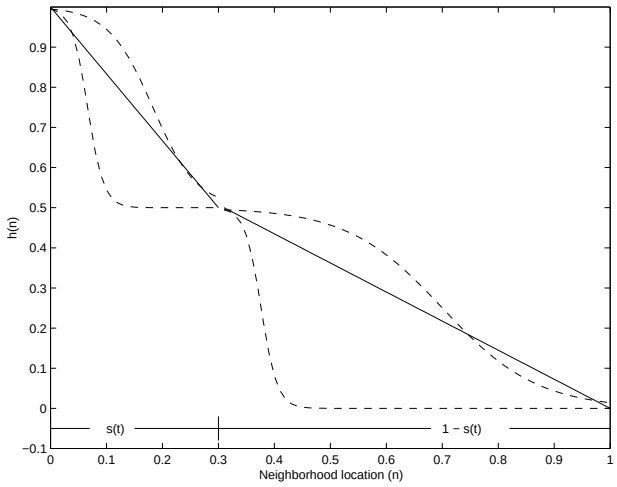

figure 1.6. Examples of h mappings between the location n and the α type of the children occupying that location. Here the value of s_t (proportion of L_1 speakers) is taken to be 0.3 for illustrative purposes. Therefore the interval $[0, 0.3]$ is the L_1 speaking neighbourhood; the interval $[0.3, 1]$ is the L_2 speaking neighbourhood. For any location n the value of $h(n)$ represents the proportion of L_1 speakers the child occupying that location is exposed to.

from neighbourhood to α-type by $\alpha = h(n)$ where $h : [0, 1] \longrightarrow [0, 1]$. We leave undefined the exact form of h except noting that it should possess certain reasonable properties, e.g., $h(0)$ should be close to 1, $h(1)$ should be close to 0, $h(s_t)$ should be close to $\frac{1}{2}$ and h be monotonically decreasing.

Shown in fig. 1.5 are some plausible mappings h that mediate the relation between location of the child in the neighbourhood and its α-type. The x-axis denotes location. The y- axis denotes the α-type of a learner. We now have learners distributed uniformly in location and a mapping from location to α-type provided by h. One can therefore easily compute the probability distribution of children by α-type. This is just the probability distribution function for the random variable $\alpha = h(n)$ where n is uniform. Let this distribution be $P_h(\alpha)$ over $[0, 1]$. Now a child (learner) of type α receives sentences drawn according to $P = \alpha P_1 + (1 - \alpha)P_2$. According to our notation developed earlier, we see that it therefore has a probability $f(\alpha)$ of attaining the grammar of L_1. (This is provided by an analysis of the learning algorithm in the usual way, i.e., $f(\alpha) = g(\mathcal{A}, \alpha P_1 + (1-\alpha)P_2, k)$). Therefore, if children of type

α are distributed in the community according to distribution $P_h(\alpha)$ and each child of type α attains L_1 with probability $f(\alpha)$, we see that in the next generation, the percentage of speakers of L_1 is provided by eq. 1.5:

$$s_{t+1} = \int_0^1 P_h(\alpha) f(\alpha) d\alpha \tag{1.5}$$

1.5.1 A Specific Choice of Neighbourhood Mapping

For purposes of illustration, let us choose a specific form for h. In particular, let us assume that it is piecewise linear in the following way (eq. 1.6; the solid line of fig. 1.5):

$$h(0) = 1; h(1) = 0; h(s_t) = \frac{1}{2}; h(n) = 1 - \frac{1}{2s_t} n \text{ for } n < s_t; h(n) = \frac{1-n}{2(1-s_t)} \text{ for } n > s_t \tag{1.6}$$

Thus, clearly, h is parameterized by s_t. For such an h, it is possible to show that P_h is piecewise uniform — given by the following:

$$P_h(\alpha) = 2s_t \text{ if } \alpha > \frac{1}{2}; P_h(\alpha) = 2(1-s_t) \text{ if } \alpha < \frac{1}{2}; P_h(\alpha) = 0 \text{ if } \alpha \notin [0,1]. \tag{1.7}$$

In previous sections, we discussed the form of the NB update rule $f = g(\mathcal{A}, s_t P_1 + (1-s_t)P_2, k)$ for memoryless learning algorithms like the TLA. From eq. 1.4, we see that it is a polynomial of degree k. Putting this into eq. 1.5, we get the update rule with neighbourhood effects to be

$$s_{t+1} = 2(1-s_t)(\int_0^{1/2} f(\alpha) d\alpha) + 2s_t(\int_{1/2}^1 f(\alpha) d\alpha) \tag{1.8}$$

Since α is a dummy variable in the above integral, the effect of the neighbourhood is to reduce the update rule to a linear one. This is in striking contrast to the original NB update rule (kth order polynomial) and the Feldman update rule (quadratic). It is worthwhile to reflect on a few aspects of such behavior.

1. The linear map implies an exponential growth (or decay) to a stable fixed point whose value is given by

$$s^* = \frac{2\int_0^{1/2} f(\alpha) d\alpha}{1 + 2(\int_0^{1/2} f(\alpha) d\alpha - \int_{1/2}^1 f(\alpha) d\alpha)}$$

2. Notice that $s^* = 0$ requires $\int_0^{1/2} f(\alpha)d\alpha = 0$. Correspondingly, $s^* = 1$ requires $\int_{1/2}^{1} f(\alpha)d\alpha = \frac{1}{2}$. Neither is very likely — therefore, no language is likely to be driven out of existence completey. If one chooses the update rule f for large k $(= \infty)$ one can compute these quantities exactly. It is then possible to show that the fixed point s^* is never 0 or 1. In contrast, both NB and CF models result in one language becoming extinct if $a \neq b$.

Needless to say, the particular form of the update rule obtained with such neighbourhood effects actually depends upon the functional mapping h. In general, however, this approach allows us to compute the evolutionary trajectories of populations where children have arbitrary α-types. It is worthwhile to recall the original CF and NB models of the previous sections in this light. The CF models are derivable from this perspective with a particular choice of $P_h(\alpha)$ which happens to be a probability mass function with $P_h(\alpha = 0) = s_t^2; P_h(\alpha = \frac{1}{2}) = 2s_t(1 - s_t); P_h(\alpha = 1) = (1 - s_t)^2$. The NB model of previous sections is equivalent to choosing $P_h(\alpha)$ to be a delta function, i.e., $P_h(\alpha) = \delta(\alpha - s_t)$.

Remark It is important to recognize two aspects of the neighbourhood model introduced here. First, the function h is not a fixed function but depends upon the proportion s_t of the L_1 speakers at any time. Therefore, h changes from generation to generation (as s_t evolves). Second, the population of mature adults is *always* organized into two linguistically homogeneous neighbourhoods in *every* generation. Ofcourse, children in a particular neighbourhood might acquire different languages. It is implicitly assumed that on maturation, the children (now adults) re-organize themselves into homogeneous neighbourhoods. It is this reorganization into homogeneous neighbourhoods that prevents the elimination of any one language from the system.

Another (more complete) way to characterize neighbourhood effects is to treat the proportion of L_1 speakers in the tth generation as a function that varies continuously with distance (n) in the neighbourhood. It is this function that evolves from generation to generation. Without additional simplifying assumptions, this treatment requires techniques well beyond the scope of this paper and will be subject of future work.

1.6 Conclusions

In this paper, we have discussed the basic model of Cavalli-Sforza and Feldman (1981) for cultural transmission and change. We have shown how this provides us with a framework in which to think about problems of language evolution and change. Language acquisition serves as the mechanism of transmission of language from parents to children. By suitably averaging over a population we are then able to derive the population dynamics, i.e, the evolutionary trajectories of the linguistic composition of the population as a whole from generation to generation.

We have shown how the approach of Cavalli-Sforza and Feldman (1981) relates to that of Niyogi and Berwick (1995, 1997) and how to go back and forth between the two models. For the particular case of two languages in competition, we have derived several particular dynamical systems under varying assumptions. We have also considered the generalization of such models to explicitly take into account the effect of spatial clustering of speakers into linguistic neighbourhoods and have investigated the consequences of such neighbourhood effects.

The case of two languages in competition is of some significance since historical cases of language change and evolution are often traceable to a point in time when speakers of two language types came into contact with each other. As a particular case of this, we considered the evolution of English syntax from Old to Middle to Modern English. While the various linguistic explanations for such a change were not considered in a serious fashion, we demonstrated in this paper, how one might apply the computational framework developed here to test the plausibility of various accounts. In general, the possibility of pursuing such a strategy in a serious manner for the study of language evolution and change remains our main motivation for the future.

.1 Appendix: Language Learning

The problem of language learning ("logical problem of language acquisition") is typically formulated as a search by a learning algorithm for a grammar that is close to the one that generates the sentences the learner is exposed to. To make matters concrete, let us define the following objects that play an important role in the theory of language learning:

1. $\mathcal{G}$: *Target Class* The target class consists of a class of grammars $g \in \mathcal{G}$. Each grammar gives rise to a corresponding language $L \in \mathcal{L}$ where

all languages are subsets of Σ^* in the usual way. A unique target grammar $g_t \in \mathcal{G}$ is the grammar to which the learner is exposed via examples and which the leaner must "learn".

2. $\mathcal{S}$: *Examples* Examples are sentences $s \in L_t$ where L_t is the target language. The learner is provided with a stream of examples drawn in some manner. For our purposes here we will assume that examples are drawn in i.i.d. fashion according to a distribution P on the sentences of the target language L_t. In other words, P is a distribution on Σ^* that has support on L_t — it puts zero measure on all $s \notin L_t$ and non-zero measure on all $s \in L_t$.
3. $\mathcal{H}$: *Hypotheses Class* The hypothesis class consists of a class of grammars that the learning algorithm uses in order to approximate elements of the target class. For our purposes, we will assume that $\mathcal{H} = \mathcal{G}$.
4. $\mathcal{A}$: *Learning Algorithm* The learning algorithm is an effective procedure that maps sets of examples into elements of $\mathcal{H}$, i.e., it develops hypotheses on the basis of examples. Formally, let $\mathcal{D}_k$ be the set of all subsets of Σ^* of cardinality k. Each subset of Σ^* of cardinality k is a candidate dataset consisting of k example sentences that a learner might receive. Clearly $\mathcal{D}_k$ is the set of all candidate datasets of size k. Then $\mathcal{A}$ is a computable mapping from the set $\cup_{k=1}^{\infty} \mathcal{D}_k$ to $\mathcal{H}$.

Given this setup, the central question of learnability theory is whether or not the hypothesis of the learning algorithm converges to the target grammar as the number of examples k goes to infinity. Specifically, let $h_k \in \mathcal{H}$ be the grammar that the learner hypothesizes after exposure to k examples. Since the examples are randomly drawn and the learning algorithm itself might be randomized, it is clear that h_k is a random variable. One can then define the probability that the learner's hypothesis h_k is the same as the target grammar. Let us call this p_k as below:

$$p_k = P(h_k = g_t)$$

The target grammar g_t is said to be **learnable** if $\lim_{k \longrightarrow \infty} p_k = 1$ for any distribution P with which examples are drawn. This simply implies that "in the limit" as the number of examples tends to infinity, the learner's hypothesis will be the same as the target grammar, i.e., the learner will converge to the target. This notion of convergence in the limit was first introduced by Gold (1967) in a non-probabilisitic framework which required that the learner converge to the target on all sequences of examples that included all sentences of the target. The treat-

ment here is probabilistic in nature. For more information on this see Osherson, Stob, Weinstein (1986); Wexler and Culicover (1980); Niyogi (1997). If every grammar $g \in \mathcal{G}$ is learnable, then the **class** of grammars ($\mathcal{G}$) is said to be learnable.

In a certain sense, linguistic theory attempts to describe and formulate classes of grammars $\mathcal{G}$ that contain the grammars of the natural languages of the world. Empirically, it is observed and believed that all such naturally occuring languages are learnable. Therefore, any class $\mathcal{G}$ that is proposed must be learnable. Learning theory investigates the questions that are associated with the learnability of classes of grammars. It is important to recognize that the framework developed for learning theory is actually very broad and therefore a wide variety of grammatical theories and learning algorithms can be accommodated within the same framework of analysis.

In the analysis of the TLA developed in Niyogi and Berwick (1996), the class $\mathcal{H} = \mathcal{G}$ consists of a finite number of grammars. The learning algorithm can be modeled as a Markov chain with as many states as there are grammars in $\mathcal{H}$. Transition probabilities from state to state depend upon the probability distribution P with which sentences are drawn and set differences between the different languages in the family $\mathcal{L}$ (equivalently $\mathcal{G}$). Probabilities like p_k can then be computed as a function of the transition probability matrix and this is done in Niyogi and Berwick (1996).

Bibliography

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

E. Gibson and K. Wexler, "Triggers," *Linguistic Inquiry,* 25, 1994.

L. Haegeman, *Introduction to Government and Binding Theory,* Blackwell, 1991.

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

A. W. Kroch and A. Taylor, "Verb Movement in Old and Middle English: Dialect Variation and Language Contact," in *Parameters of Morphosyntactic Change,* CUP, 1997.

D. Lightfoot, "Shifting Triggers and Diachronic Reanalyses," in *Parameters of Morphosyntactic Change,* CUP, 1997.

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

P. Niyogi and R. C. Berwick, "The Logical Problem of Language Change," MIT AI Memo No. 1516, 1995.

P. Niyogi and R. C. Berwick, "Evolutionary Consequences of Language Learning," *Linguistics and Philosophy,* Vol. 17, 1997.

P. Niyogi and R. C. Berwick, "The Logical Problem of Language Change: A Case study of European Portuguese," *Syntax: A Journal of Theoretical, Experimental, and Interdisciplinary Research,* Vol. 1, 1998.

P. Niyogi and R. C. Berwick, "A Language Learning Model for Finite Parameter Spaces," *Cognition,* 61(1):161-193, 1996.

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

R. L. Trask, *Historical Linguistics,* Arnold, U.K., 1996

A. Warner, "The Structure of Parametric Change and V movement in the history of English," in *Parameters of Morphosyntactic Change,* CUP, 1997.

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