Chapter 7

An Application to Portuguese

In this chapter we consider a specific historical example of language change — the change in the grammatical system governing the placement of *clitics* in European Portuguese from the sixteenth to the eighteenth century A.D. The example illustrates the potential power of our computational approach in explicating the nature of the interaction of two important cognitive phenomena: language learning and language change. The first, language learning, occurs at the level of the individual — children acquire the language (grammar) of their caretakers, a cognitive ability that has been broadly investigated via a range of computational and experimental methodologies. The second, language change, occurs at the level of the population: it is individual language learners whose collective, ensemble properties constitute a *distribution* of linguistic knowledge and this distribution of linguistic knowledge might change over generational time scales.

Against the backdrop of Portuguese change, we will develop models to explore the interplay between language learning and language change. Two major insights emerge. First, that different language learning algorithms may have different evolutionary consequences. Second, that the learning derived dynamics is typically nonlinear giving rise to bifurcations as the parameters change continuously. Thus it may be possible for subtle changes in parameters to lead to discontinuous changes in the stability of linguistic systems. These insights were obtained in previous chapters already. Seeing them again in this new and concrete linguistic context reinforces our belief in their significance. We argue that these insights generalize across the specifics

of particular models and therefore provide some genuine understanding of the forces that shape language change and evolution.

7.1 Portuguese: A Case Study

In what follows, we will present certain aspects of the historical evolution of European (as distinct from Brazilian or Goan, for example) Portuguese and demonstrate how computational models allow us to sharpen our questions and clarify our reasoning in constructing explanations for the phenomena at hand. In particular, we will see how different learning algorithms have different evolutionary consequences some of which are incompatible with the historically observed trend. We will also see how the evolutionary consequences depend subtly upon parameter values and this dependence can only be worked out by mathematical analysis. It should be noted that a far more detailed investigation of Portuguese is being carried out as part of an interdisciplinary project coordinated at the University of Sao Paolo, Brazil. A corpus of historical Portuguese texts (the *Tycho Brahe* corpus) has been collected, is being linguistically annotated and statistical analyses of the major trends are being conducted. Central to the research effort associated with the Tycho Brahe corpus are mathematical models of learning and dynamical change that are similar in spirit to and motivated by the same philosophical concerns as those discussed in this book. The interested reader may consult http://www.ime.usp.br/~tycho for further details. Also see Galves and Galves (1995), Fernandez and Galves (1999), Cassandro et al (1999), Britto et al (1999) for research reports.

7.1.1 The Facts of Portuguese Language Change

We focus on a particular change involving an interaction of phonological and syntactic components of the grammar of Portuguese. The discussion that follows is adapted from Galves and Galves (1995) and elaborated in the publications associated with the Tycho Brahe project. Portuguese has always been an SVO language exemplified by the typical word order in the following sentence:

Paulo ama Virginia
Paulo loves Virginia.

Roughly, over a period of 200 years, starting from 1800, "classical" Portuguese (CP) underwent a change in clitic placement. Clitics are morpho-

logical items that attach to syntactic heads to form a lexical complex. In the cases discussed below, the pronominal clitic *a* attaches to the verb *ama* to form a lexical complex. However, the clitic may attach *before* ("a-ama") or *after* ("ama-a") the verb in question and is referred to as a proclitic or enclitic respectively. From the 16th century or before until the beginning of the 19th century, both proclitics and enclitics were possible in root declarative sentences (with nonquantified subjects), as given by examples (1) and (2) below (from Galves and Galves (1995); also referred to as G&G in later portions of this chapter).

(1) *Paulo a ama.*
Paulo her loves
'Paulo loves her'
(proclitic)

(2) *Paulo ama a*
Paulo loves-her
'Paulo loves her'
(enclitic)

In sentences with a quantified subject (containing a *wh* element, for example) proclisis was and continues to be the only option as in (3) below.

(3) *Quem a ama?*
Who her loves?
' Who loves her?'
(proclitic)

Galves and Galves (1995) summarize the relevant historical facts as follows: "During the 19^{th} century a change affecting the syntax of clitic-placement occurred in the language spoken in Portugal ... As a result, sentences like (1) became agrammatical and (2) remained as the only option for root affirmative sentences with non-quantified subjects. This change, however, did not concern sentences like (3) with quantified or *Wh*-subjects in which proclisis was, and continues to be, the only option."

Table 7.1 shows the percentage of enclitics in the writing of Portuguese authors in the classical period. In contexts where there is variation, proclisis is clearly dominant. We will refer to the language of this time as Classical Portuguese (CP) in our discussion in this chapter.

Author	Proclisis	Enclisis	%Enclisis
Gusmao (1695)	27	0	0%
Castro (1700)	15	1	7%
Oliviera (1702)	39	7	16%
Judeu (1705)	27	6	19%
Verney (1713)	14	11	44%
Marques (1728)	30	10	25%
Marquesa (1750)	34	23	40%

Table 7.1: Data extracted from the works of authors in the classical period from the end of the seventeenth century to the middle of the eighteenth. In contexts where there is variation, proclisis is clearly dominant.

Author	Proclisis	Enclisis	%Enclisis
Garrett (1799)	11	37	77%
Camilo (1825)	6	70	92%
Dinis (1839)	3	24	88%

Table 7.2: Data extracted from authors from the nineteenth century. Note the increasing percentage of enclisis in the same contexts. Modern European Portuguese uses exclusively enclitics.

In contrast, the writings of authors from 1799 onwards show a clear dominance of enclisis as summarized in Table 7.2. Modern European Portuguese has no proclisis in sentences with non quantified subjects at all. We refer to the language of modern European Portuguese as European Portuguese (EP) in our discussions.

Galves and Galves (1995) offer an explanation of this change by proposing a link between phonology and syntax. Roughly speaking, they argue that phonological changes in Portuguese altered the stress contours associated with the sentences, and consequently the probabilities with which these different sentence types occurred. This difference in stress is what learning hinges on, and so the historical change. Depending upon one's linguistic persuasion, one may argue with the details of such an explanation but for our purposes, we will accept it to illustrate how different learning algorithms

might have different evolutionary consequences for historical prediction. We will therefore ignore for the moment the linguistic implications of the various algorithms and concentrate only on their computational properties.

To each sentence we will assign (a) a morphological word sequence; (b) a stress contour; and (c) a syntactic structure. For example, again following G&G's analysis, sentence type (1) will remain only in Classical Portuguese (CP) while the two sentences (2) and (3) above will have *different* stress patterns for CP and EP.

We omit a detailed description of the stress[1] assignment and syntactic[2] properties, as they are not necessary for our analysis. All we need to know is that G&G assume that the stress contours corresponding to sentence types (1), (2), and (3), which we denote simply as c_1, c_2, c_3, follow a Markov chain description and, more importantly, govern the probability with which sentences are produced[3]. Thus, if two sentences have the same stress contour, then they will be produced with the same probability (given by the probability of the stress sequence according to Markov production rules).

In short, for the purposes of this chapter, it is sufficient to assume that there are simply two grammars (in accordance with the assumptions of

[1] In the treatment of G&G, given a bracketed clause, a stress mark S is assigned to each stressed word and a (non) stress mark U is assigned to each unstressed word according to the framework of metrical grid theory of Halle and Idsardi (1992). It follows that for Classical Portuguese, sentence type (1) has the stress assignment of SUS (2) has the stress assignment of SS where the morphological complex of 'ama-a' bears a single stress mark and (3) has the assignment of SUS. The probability of producing sentences is assumed to depend upon the stress assignment. Hence in Classical Portuguese, sentence types (1) and (3) are produced with equal probabilities. Similarly, the stress assignments according to the grammar of EP may be determined.

[2] Following Salvi (1990), Madeira (1992), and Manzini (1992), the following points are made in the syntactic treatment of Portuguese over the ages. (1) Only one functional category contains the clitic and the verb in both proclitic and enclitic positions. Proclisis corresponds to a structure in which the clitic has adjoined to the verb in *Infl*. (2) In Classical Portuguese enclitic constructions the subject lies outside the border of the clause, contrary to what happens in proclitic constructions. (3) The landing site for the subject in European Portuguese enclitic constructions is Spec/CP. (4) The specifier position which is the landing site of non interrogative subjects in Classical Portuguese is no more available in European Portuguese. (5) Enclisis appears in a position entering in complementary distribution with *wh* and *Focus*. Galves and Galves (1995) claim that the change from Classical to European Portuguese results from a reinterpretation of the position of the subject in enclitic constructions.

[3] We are of course aware that this assumption of G&G may also be questioned; one might substitute for it any other more plausible relation between stress and sentence types — if any; this assumption is simply designed as a bridge to get the child from a presumably observable surface fact to a sentence type.

G&G): G_{CP}, denoting the grammar of Classical Portuguese (earlier) and G_{EP}, denoting the grammar of European Portuguese. Furthermore, the only data that is relevant (ignoring other aspects of the grammar) is as follows:

Classical Portuguese

c_1 : produced with probability p; c_2 produced with probability $1-2p$; and c_3 produced with probability p.

European Portuguese

c_1: not produced; c_2 produced with probability $1-q$; and c_3 produced with probability q.

Any (historically changing) population will now by assumption contain a mix of speakers of Classical and European Portuguese. The Classical Portuguese speakers produce the sentence types shown above with probabilities that are parameterized by p. The European Portuguese speakers produce the sentence types shown above with probabilities that are parameterized by q.

Thus we have defined

1. the class of grammars (1) $\mathcal{G} = \{G_{EP}, G_{CP}\}$. Note that the language of $_{EP}$ is a proper subset of the language of G_{CP} in our setting.

2. the probabilities with which speakers of G_{EP} and G_{CP} produce sentences (parameterized by p and q).

We can therefore derive the evolutionary consequences at the population level for a variety of learning algorithms. We now proceed to do so.

7.2 The Logical Basis of Language Change

In our model the logical basis of change is language learning: the possibility of *mis*learning the particular target grammar of one's caretakers. As we have argued, if children *always* converged on the language of their parents, then their language would be the same as that of their parents, and this would be true from each generation to the next. Consequently, for languages to change from one generation to the next it must be the case that children attain a language different from that of their parents.

In our setting, there are two different linguistic types – G_{CP} and G_{EP} that are represented in the population. On the basis of example sentences from the previous generation, children acquire either one of the two grammars. Some may acquire G_{CP} and others may acquire G_{EP}. The probability

with which they do so will depend upon (i) the distribution of the different grammatical types in the adult population (ii) the probability with which sentences are produced by each of the grammatical types in the population (iii) the learning algorithm children use to infer a grammar.

A complete analysis of the behavior of the individual learner will allow us to analyze the behavior of the population. In order to do this, we make the following idealizations for population modeling: (1) non-overlapping discrete generations: the populations consists of parents and children with parents being the source of linguistic data and children being the learners (2) no neighborhood effects: the mix of linguistic types in the entire adult population determines the source of sentences and this distributional source is identical for all children (3) adults do not change their grammar/language over their lifetime, i.e,. a monolingual maturation hypothesis (4) children have a finite time to acquire the grammar, i.e., a learning maturation hypothesis.

These assumptions may be systematically dropped and the consequences examined. We do this in a later chapter.

7.2.1 Galves Batch Learning Algorithm

Galves and Galves (1995) describe a maximum likelihood approach to grammatical inference that may be summarized as follows:

1. Draw N examples (sentences).
2. Compute likelihoods, i.e., $P(S_n|G_{CP})$ and $P(S_n|G_{EP})$.
3. Use the Maximum Likelihood principle to choose between the grammars.

In order to derive the evolutionary consequences of such a learning procedure, we first need to be able to analyze the behavior of the individual learner.

Analysis of Individual Learning Algorithm

For the analysis of the algorithm, we assume that sentences are drawn in i.i.d. fashion according to a probability distribution determined by the stress contours of the relevant sentences as indicated in the earlier section.

First, consider the form of the likelihoods. Let the example sentences be $S_n = \{s_1, s_2, \ldots, s_n\}$. Due to the i.i.d. assumption $P(S_n|G_{CP})$ is given by

$\prod_{i=1}^{n} P(s_i|G_{CP})$. Suppose that the set of n examples consists of a draws of c_1, b draws of c_3 and $n-a-b$ draws of c_2. Then the following is immediately clear:

$$P(S_n|G_{CP}) = p^a(1-2p)^{(n-a-b)}p^b$$

and

$$P(S_n|G_{EP}) = (0)^a(1-q)^{(n-a-b)}q^b$$

Consequently, the individual child, following the likelihood principle will choose the grammar EP (G_{EP}) only if (1) no instances of c_1 occur in its sample; and (2) the number of occurrences of c_2 and c_3 are such that $q^b(1-q)^{(n-b)} > (1-2p)^{(n-b)}p^b$. There are three cases to consider:

Case 1. $p < q < 2p$.

Decision Rule: For this case, it is possible to show that the child (following the maximum likelihood rule) always chooses G_{EP} if no instances of c_1 occur. This is simply because $1-q > 1-2p$ and $q > p$.

Population Update: Suppose that the proportion of speakers of G_{CP} in the ith generation is α_i. Then the probability of drawing c_1 is given by $\alpha_i p$. Consequently, the probability of drawing a set of n examples without a single draw of c_1 is $(1-\alpha_i p)^n$. This is of course the probability with which the individual child chooses the grammar of European Portuguese, G_{EP}. Thus the update rule has the following form:

$$\alpha_{i+1} = 1 - (1-\alpha_i p)^n$$

Case 2. $q < p < 2p$.

Decision Rule: In this case, the maximum likelihood decision rule reduces to the following. Choose G_{EP} *if and only if* (1) $a = 0$, i.e., no instances of c_1 occur; and (2) $b < n\gamma$ where $\gamma = \frac{\log(\frac{1-q}{1-2p})}{\log(\frac{1-q}{1-2p})+\log(\frac{p}{q})}$. For all other data sets, choose G_{CP}.

Population Update: As usual, let α_i be the proportion of the previous generation speaking G_{CP}. Using the fact that the numbers of each of the three sentence types has a multinomial distribution, it can be shown that events (1) and (2) above occur with a total probability equal to $\sum_{k=0}^{n\gamma} \binom{n}{k} P^k Q^{(n-k)}$ where $P = \alpha_i p + (1-\alpha_i)q$ and $Q = \alpha_i(1-2p) + (1-\alpha_i)(1-q)$. Thus the update rule has the following form:

$$\alpha_{i+1} = 1 - \sum_{k=0}^{n\gamma} \left(\begin{array}{c} n \\ k \end{array} \right) P^k Q^{(n-k)}$$

Note that since γ is a real number, $n\gamma$ is not integer valued. Therefore, in the binomial expression, the sum should be taken upto and including the largest integer less than $n\gamma$.

Case 3. $p < 2p < q$.

Decision Rule: The maximum likelihood decision rule reduces to: choose G_{EP} *if and only if* (1) $a = 0$; and $b > n\gamma$ where $\gamma = \frac{\log(\frac{1-2p}{1-q})}{\log(\frac{q}{p})+\log(\frac{1-2p}{1-q})}$. Otherwise, choose G_{CP}.

Population Update:

As usual, let α_i be the proportion of the previous generation speaking G_{CP}. It can be shown that the update rule has the following form:

$$\alpha_{i+1} = 1 - \sum_{k=n\gamma}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) P^k Q^{(n-k)}$$

where P and Q are as in case 2.

Again note that because $n\gamma$ in not integer valued, the sum in the binomial expression should be taken from the smallest integer larger than $n\gamma$.

System Evolution

We have shown above how the behavior of the population can be characterized as a dynamical system and have derived the update rules for such a system for a maximum likelihood learning algorithm. The dynamical system captures the evolutionary consequences of this particular learning algorithm. Let us elaborate on the evolutionary possibilities. In principle, these may then be matched against the observed historical trends.

Case 1.

1. $\alpha = 0$ is a fixed point, i.e., if the initial population consists entirely of European Portuguese (g_{EP}) speakers, it will always remain that way. Furthermore, if $np < 1$, then this is a *stable* fixed point. It is also the *only* fixed point between 0 and 1. Thus in this case a population speaking *entirely* Classical Portuguese would gradually be converted to one speaking *entirely* European Portuguese.

2. If $np > 1$, then $\alpha = 0$ remains a fixed point but now becomes *unstable*. For this case, an additional fixed point (*stable*) is now created between 0 and 1. All initial population compositions will tend to this particular *mix* of G_{CP} and G_{EP} speakers. Figure 7.1 shows the fixed (equilibrium) point as a function of n and p.

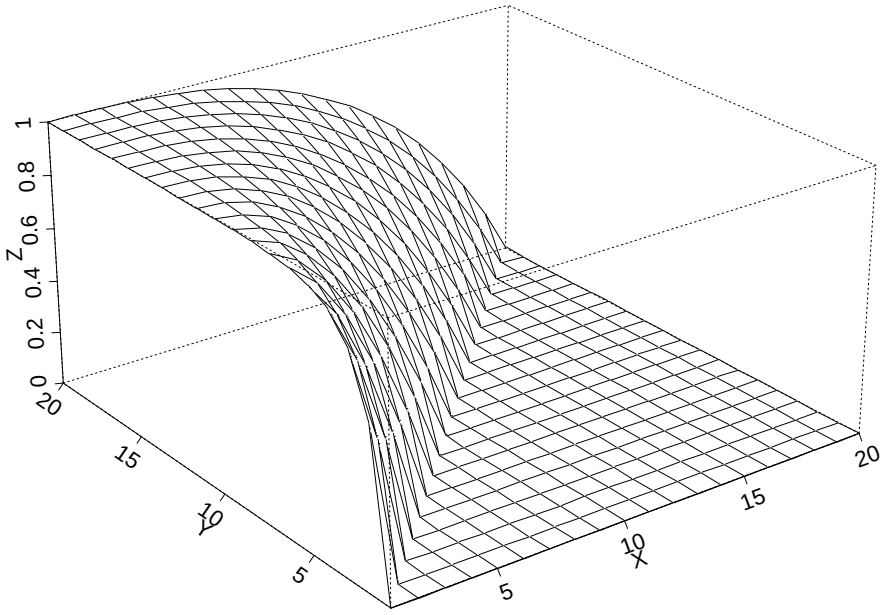

Figure 7.1: The fixed point of the dynamical system (on the Z axis) as a function of n (on the X axis) and $\frac{1}{p}$ (on the Y axis).

Case 2.

1. Unlike case 1, the dynamical evolution depends now upon both p and q in addition to n.

2. It is easily seen that $\alpha = 0$ is no longer a fixed (equilibrium) point (unless $p = q$). Consequently, irrespective of their initial composition, populations will *always* contain some speakers of Classical Portuguese.

3. It is possible to show that there is exactly one fixed (*stable*) point and all initial populations will tend to this value. Shown in Fig. 7.2 is a plot of the fixed point as a function of q and p for a fixed value of n. Notice the multiple ridges in the profile suggesting sensitivity to the value of q around some critical points. Shown in Fig. 7.3 is a plot of the fixed point as a function of p for various choices of n keeping q fixed at 0.1.

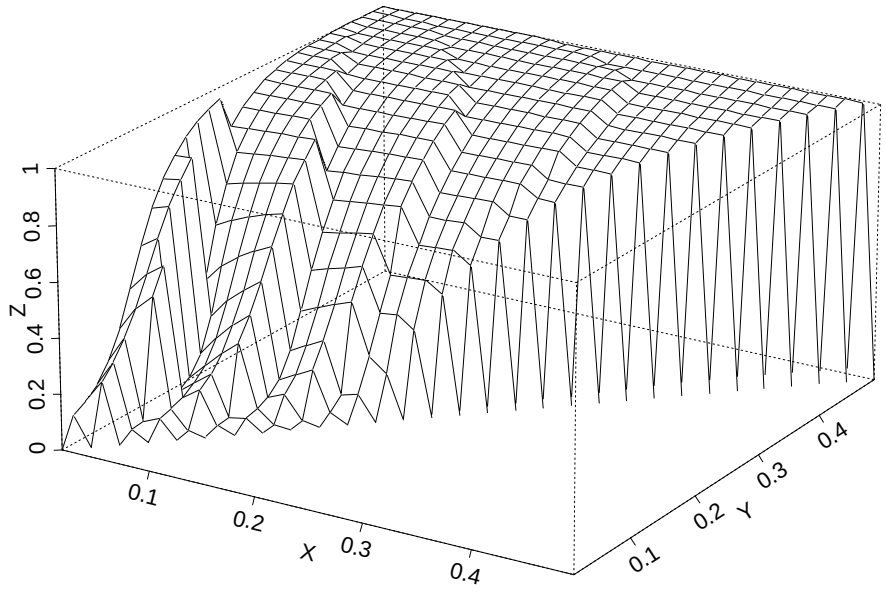

Figure 7.2: The fixed point of the dynamical system (on the Z axis) as a function of q (on the X axis) and p (on the Y axis). The value of n was held fixed at 5.

Case 3.

1. Like case 2, the dynamical evolution depends upon both p and q in addition to n.

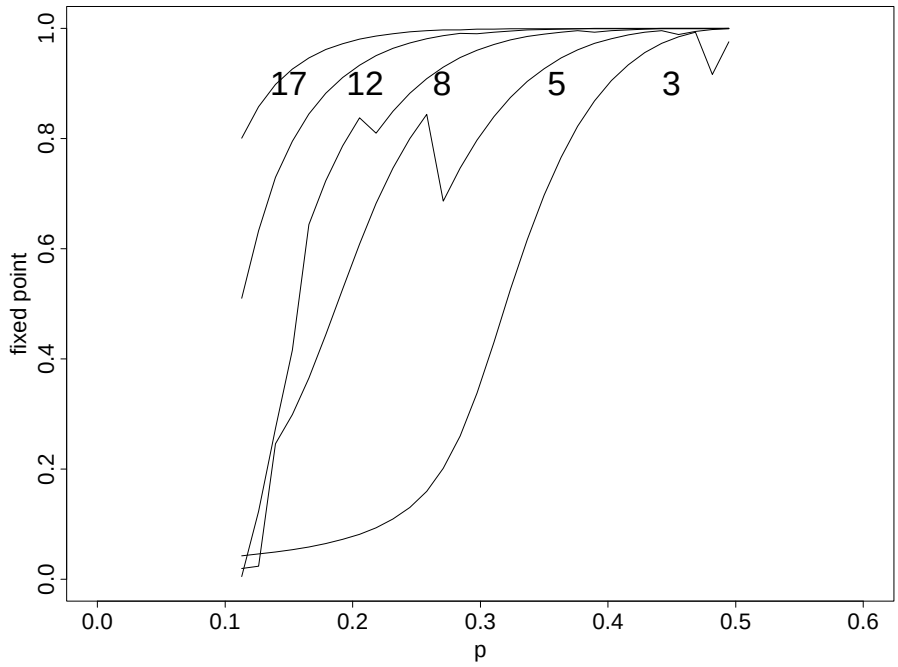

Figure 7.3: The fixed point of the dynamical system as a function of p (on the X axis) for various values of n. Here q was held fixed at 0.1 and p was allowed to vary from 0.1 to 0.5.

2. Again, it is easily seen that $\alpha = 0$ is no longer a fixed point. Therefore, the speakers of Classical Portuguese can never be eliminated altogether for p and q in this range.

We can again plot the fixed points of the resulting dynamical system as a function of q and p where n is held fixed at 5 or for various values of n, keeping p fixed. We omit the figures for reasons of space. The results are: again the ridges in the landscape suggest a great sensitivity of the final equilibrium point to slight changes in the values of p and q. Classical Portuguese speakers are never completely eliminated, although their frequency can get quite low in certain regions.

What are the important conclusions from this analysis? In short, children using the maximum likelihood rule will choose G_{EP} over G_{CP}. However, a dynamical systems analysis must be carried out to see if that will suffice to "wipe out" Classical Portuguese. Only in case 1 will Classical Portuguese be lost completely (provided $p < 1/n$). *In all other cases*, there will always remain some speakers of Classical Portuguese within the community. In fact, the evolutionary properties can be quite subtle. Consider the following three example cases.

Example 1. Let $p = 0.05, q = 0.02$ and $n = 4$. In this case, if the parental generation were *all* speaking Classical Portuguese ($\alpha = 1$) then a simple computation shows that the probability with which the child would pick G_{EP} (European Portuguese) is 0.66, i.e., it is greater than one-half. Thus, **in spite of the fact** that the **majority** of children choose the grammar of European Portuguese, the speakers of Classical Portuguese will never die out completely. In fact, the fixed point is 0.11. Roughly 11 percent of the population will continue to speak Classical Portuguese.

In the absence of this analysis, one might naively argue that since a majority of learners choose g_{EP}, this trend will continue generation after generation leading the population to lose all its CP speakers eventually. This argument, as we have just seen would be mistaken.

Example 2. Let $p = 0.05, q = 0.06$ and $n = 8$. If this were the case, and the parental generation *all* spoke Classical Portuguese, it turns out that the probability with which the individual child would pick G_{EP} would again be 0.66. However, now the speakers of classical Portuguese would all be lost and the population would move to its stable, fixed point containing only speakers of European Portuguese.

Example 3. If $p = 0.05, q = 0.06$ and $n = 21$ however, it is easily seen that Classical Portuguese speakers can never be completely lost.

7.2.2 Batch Subset Algorithm

Most importantly, we see that the above learning algorithm makes specific predictions about the change of the linguistic composition of the population as a whole. For purposes of exploration, let us turn our attention to a simple modification of the previous learning algorithm that we call the Batch Subset Algorithm because (i) it incorporates the Subset Principle (see discussion in Berwick, 1985) and (ii) all the data is processed "at once." Our aim is to demonstrate how readily one may carry out changes in the learning algorithm and investigate their model consequences.

1. Draw n examples.

2. If c_1 occurs even once, choose G_{CP}, otherwise choose G_{EP}.

Since European Portuguese is a subset of Classical Portuguese for the data at hand, such a learner would choose the grammar of European Portuguese G_{EP} as its default grammar unless it received contradictory data (in this case c_1: which informs it that the target is not G_{EP} but G_{CP}). Of course, such a learning algorithm is guaranteed to converge to the correct target as the data goes to infinity. A natural question to ask is whether it makes a different prediction about how the population would evolve.

Assume that a proportion α_i of the adult population speaks Classical Portuguese. Let us then calculate the probability with which a typical child internalizes g_{CP}. After n i.i.d. draws, the child would internalize g_{EP} if no examples of c_1 occur. Since examples of c_1 are produced for the child with probability $p\alpha_i$, the probability of not encountering a single example of c_1 in n draws is given by $(1 - p\alpha_i)^n$. Therefore the probability of internalizing g_{CP} is given by

$$\alpha_{i+1} = 1 - (1 - \alpha_i p)^n$$

One can already see that the evolutionary properties for this learning algorithm are different from the previous one. The dynamics is always given by the same update rule *irrespective* of the values of p and q. In fact, the evolution, which is totally independent of q, is identical to Case 1 of the dynamics of the previous learning algorithm. Naturally, it has the same equilibrium behavior as Figure 7.1.

It is worthwhile to reflect a bit on the bifurcations that take place as n changes. For very large n, the stable fixed point of the population is at a value close to $\alpha = 1$. In fact, as $n \to \infty$, this stable fixed point tends

to 1. In other words, for large n, the stable mode of the population is to speak mostly Classical Portuguese. For very small n on the other hand — in fact, for $n < \frac{1}{p}$, the only stable mode of the population is to speak entirely European Portuguese. Therefore a consistent explanation of the change from Classical to European Portuguese within the framework of such a subset like learning algorithm might invoke a sudden change of n from large to small values. Thus, if for some reason, the total usage of clitics in these contexts decreased dramatically so that children had far fewer number of examples on which to base their decision of which grammatical principle to internalize, one could move from a stable population of Classical Portuguese speakers to one of European Portuguese speakers. Of course, one can now examine more data to check whether this is indeed a plausible explanation.

This example illustrates how the computational analysis sharpens considerably possible explanatory scenarios leading to more precise questions that one may then try to resolve.

7.2.3 Online Learning Algorithm (TLA)

One might argue that a batch learning mechanism where all the data is processed at once is cognitively less plausible than an online procedure with memory limitations. Let us therefore briefly consider the evolutionary consequences of an example of a canonical online learning algorithm — the Triggering Learning Algorithm (TLA) described in previous chapters.

In this context, the TLA works as follows. Choose a hypothesis initially at random. Stay with the current hypothesis until a counter example comes at which point flip to the opposite hypothesis. Following the treatment in Chapter 5, we see that there are two languages L_{CP} and L_{EP}. The update rule (for proportion of L_{CP} speakers) may be derived as

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

where (i) a is the probability with which speakers of L_{CP} produce sentences in $L_{CP} \cap L_{EP}$ (ii) b is the probability with which speakers of L_{EP} produce sentences in $L_{CP} \cap L_{EP}$ (iii) B is the probability with which unique triggers of L_{CP} are presented to the learner (iv) A is the probability with which unique triggers of L_{EP} are presented to the learner. It is easy to check

$$b = 1; a = 1 - p; B = \alpha_i p; A = 0$$

Plugging this in, one one obtains the update rule as

$$\alpha_{i+1} = 1 - \frac{1}{2}(1 - \alpha_i p)^n$$

where α_i and α_{i+1} are the proportion of the population speaking Classical Portuguese in generation i and $i+1$ respectively. As usual, n is the number of examples drawn.

One can now make the following two observations about this evolution.

1. Since $\alpha_i \geq \frac{1}{2} \forall i$, the proportion of CP speakers can *never* be eliminated to less than $\frac{1}{2}$ of the population. Consequently, one is able to see immediately that the TLA does *not* have the right evolutionary properties to explain the change from Classical to European Portuguese.

2. It is possible to show that there is exactly one stable fixed point between $\frac{1}{2}$ and 1 to which the system evolves. The exact value of this stable fixed point depends upon n and p. If $p = 0$, then the only stable point is $\alpha = \frac{1}{2}$.

7.3 Conclusions

In this chapter, we have grounded our discussion of language change in the context of the historically observed change of clitic placement in Portugal from the sixteenth to the eighteenth century. This is a case for which considerable data seems to exist and is being systematized as part of the Tycho Brahe collections.

Let us reflect on the two themes that are illustrated in this chapter and that constitute two of the central insights of this book. First, we note that different language learning strategies on the part of individual learners can lead to qualitatively different evolutionary trajectories over time. We saw this for the Galves Batch Algorithm, the Batch Subset Algorithm and the Online TLA. Second, we note that the dynamics is typically non-linear. As a result, the interaction between learning and change can be quite subtle. The dynamical systems mathematics therefore becomes essential because intuitions can lead one astray. For example, using the learning algorithm proposed by Galves and Galves (1995), we see that the evolutionary consequences depend subtly upon the parameter values p and q and the operating regime of the system. Bifurcations arise as p and q change continuously leading the system from one qualitative regime to another. Again and again, in a

number of different settings, we encounter such bifurcations pointing to the pervasiveness of such phase transition like phenomena in language evolution.

These considerations illustrate how the mathematical and computational approach discussed in this book allow us to probe more deeply, sharpen questions, correct fallacious reasoning and generally tease apart issues that arise in constructing explanations for these kinds of changes. While it is certainly too early to claim a successful resolution of the Portuguese puzzle, the effort associated with the Tycho Brahe project represents strong steps in this direction.

Chapter 8

Competition of Two Forms in a Population: The case of Chinese Phonology

In a number of cases of language change studied in historical linguistics, one observes two linguistic forms that have co-existed for a period of time in changing ratios leading ultimately to the complete loss of one form from the entire population.

In this chapter, we consider a population analysis of such a situation. To ground our discussion in a concrete linguistic example, we consider a case of phonological merger in Chinese (taken from Z. Shen (1993, 1997)). Over the course of this example, we develop the various issues that need to be resolved as linguists try to account for linguistic changes of the sort that occurred in Wenzhou over the last century. These issues are sometimes tricky to tease apart and the primary role of this chapter is to introduce a computational framework within which one can embed various linguistic accounts, examine their consequences, and generally reason about them.

Historical phenomena in language present a particularly interesting window on synchronic linguistics and language acquisition. While considerable computational work exists in the latter two areas, historical linguistics has received minimal computational attention in the past. Yet there is clearly an interplay between the range of possible synchronic variation, the modes of transmission of language from one generation to the next (effected by language acquisition, broadly construed) and diachronic variation over generational time scales.

This chapter explores and exploits such an interplay in the particular context of a documented instance of phonological change in China. It is part of our ongoing effort in this book to better understand the nature of the relationship between language acquisition by individuals and language change in populations. We will see over the course of this chapter how the various issues surrounding language change are brought into sharp focus as a result of our computational reasoning. It will also give us a welcome opportunity to depart from the syntax centric formulation of our models to consider an explicitly phonological example. This will allow us to see both the generality of our computational framework and the universality of the issues that arise in any treatment of language change. We will also see an interesting contrast between linguistic evolution and biological evolution. Our models indicate that in contrast to biological evolution, the particulate acquisition of language (equated with categorical monolingualism) *eliminates* linguistic diversity while the blending acquisition of language (equated with multilingualism) maintains linguistic diversity. Note that in biological evolution, particulate inheritance maintains diversity while blending inheritance eliminates it (see, e.g. Fisher (1930) for a discussion of this issue). Finally, we will reflect again on the role of bifurcations in language change and see how they may provide a novel solution to the actuation problem — the problem of why a change may come about in the first place.

8.1 Phonological Merger in Wenzhou province of China

Zhongwei Shen (1997) describes two detailed studies of phonological change in the Wu dialects of China. We consider here as an example the monophthongization of $/o^y/$ resulting in a phonological merger with the rounded front vowel $/o/$. This sound change is apparently not influenced by contact with Mandarin and is conjectured to be due to phonetic similarities between the two sounds. These two phonological categories were preserved as distinct by many speakers but over a period of time, the distinction was lost and their merger created many homophonous pairs.

Thus, the word for "cloth" – $/po^y/^{42}$ – now became homophonous with the word for "half" – $/po/^{42}$ and similarly, the word for "road" – $/lo^y/$ became homophonous with the word for "in disorder" – $/lo/^{11}$. A list of 35 words with the diphthong $/o^y/$ is presented in Z. Shen (1993) and some of these are reproduced in Table 8.1.

$/po^y/^{42}$	"cloth"
$/do^y/^{31}$	"graph"
$/mo^y/^{31}$	"to sharpen"
$/to^y/^{42}$	"jealous"
$/so^y/^{42}$	"to tell"

Table 8.1: A subset of the words that underwent change.

The phonetic difference between the two sounds lies in movements of the first and second formants. Both of the sounds in question are long vowels. The monophthong $/o/$ has a first formant at around 600 Hz. and a second formant at 2200 Hz. The diphthong $/o^y/$ has a first formant that starts around 600 Hz and gradually drops down to 350 Hz while the second formant increases slightly above 2200 Hz. The change from the diphthong to monophthong can in principle be gradual with no compelling phonetic reason to make this change abrupt.

During the process of this phonological merger, each word in this list had two alternative pronunciations (i) the *original* pronunciation with the diphthongized vowel (ii) the *changed* pronunciation with the diphthongized vowel replaced by the monophthongized equivalent.

In an effort to locate the rate at which the change spreads through the speakers in a population and through the words in the lexicon, Zhongwei Shen conducted a field study in the summer of 1990 and the results are reported in Shen (1993, 1997). A total of 363 subjects were questioned regarding the pronunciation of each of the 35 words from which it was possible to elicit the distribution of pronunciations in the population as a whole. The subjects were distributed in age from 15 to 77. A striking pattern was observed providing a veritable snapshot of the change in process. Among the older people the original pronunciation of the word was in vogue. The proportion of people using the original pronunciation decreased with age so that among the youngest people in the community, the original pronunciation was almost completely replaced by the new one. Fig. 8.1 shows the percentage of people using the original pronunciation as a function of age for two different words in the lexicon.

It is clear that different generations of speakers use preferentially different forms of the word. It is also worthwhile to observe that a true *diachronic* picture of the language changing over time would be observed by literally

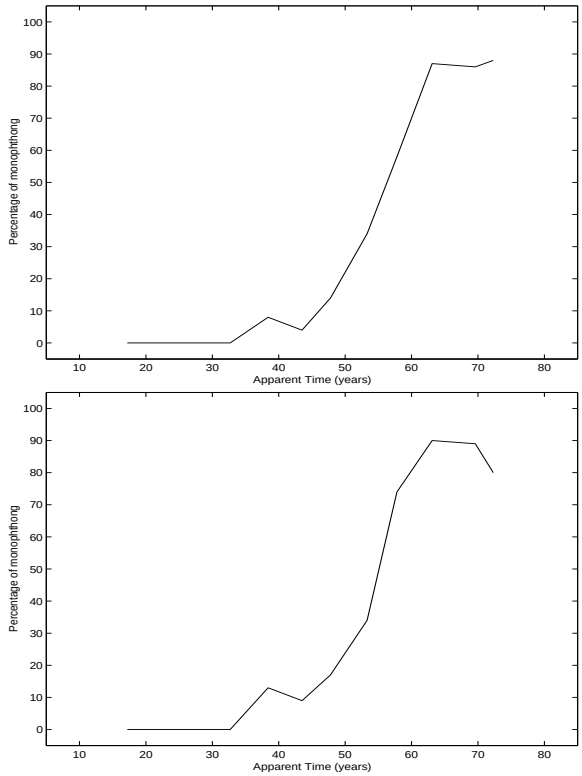

Figure 8.1: Gradual change in the percentage of speakers using the monophthong version of a word. The two figures correspond to two different words that underwent a change in pronunciation. The words are /po^y/ (top) and /lo^y/ (bottom) respectively. After Shen (1997).

sampling in time, i.e., sampling the average behavior of the population (represented by the percentage of speakers who use the original form) at different points in time. This would require cross sectional data over many years. This is usually infeasible and the technique of using the *synchronic* variation and factoring it by generations provides an estimate of "apparent time" (introduced by Labov, 1966) allowing us a window into the diachronic variation.

We thus see a phonological (sound) change in progress with two linguistic forms (i) a diphthongized form and (ii) a monophthongized form in competition. In almost characteristic fashion, one form is completely replaced by the other.

Why does such a change happen? What initiates it? Why does it go to completion? Why don't mixed populations remain stable over all time? These are the canonical questions that arise in many studies in historical linguistics. Let us consider some of the key issues that linguists have to resolve as they construct explanations of such phenomena:

1. *the role of learning:* Clearly language is acquired by children – most significantly from the input provided by the previous generation of speakers in the community. Now at one point in time, the two phonological forms were being distinguished and *only* the diphthongized form was being used by all speakers for each of the words in question. Children heard this form of the word and most of them should have acquired this form. Indeed, if all of them had acquired this form, the language would not have changed from one generation to the next in this regard. So some children must have acquired the monophthongized version and thus two successive generations differed from each other. A common explanation for the kind of phonological change discussed here is the phonetic similarity between the two phonological items in question. While phonological classes are discrete, their phonetic realizations are, of course, continuous. For phonetically similar sounds the overlap in the distributions of the two sounds might be considerable leading often to errors in production or perception. As a result, some child learners might conceivably end up learning an alternative pronunciation leading to phonological variation in the population and ultimately change.

   The idea that language change is contingent on language learning has been a long standing one and indeed Shen (1997) spends some time on this connection. We will examine this in a mathematically precise manner over the course of this chapter.

2. *populations versus idiolects:* Isolated instances of mislearning or idiosyncratic linguistic behavior are clearly of little consequence unless they spread through the community over time to result in large scale language change. Central to our point of view in this book is the distinction between the population and the individuals in it — the distinction between the behavior of the individual language users and the group linguistic characteristics of the population as a whole. Shen (1997) provides the source of the Wenzhou data discussed here and explicitly samples multiple people in the population for each generation and the curves of Fig. 8.1 are plots of average speaker behavior with respect to pronunciation. Another goal of this chapter is to explore the relationship between change at the individual level and change at the population level in this context.

3. *lexical diffusion versus neogrammarian hypothesis:* In phonological changes of the sort described in Wenzhou, there are multiple words in the lexicon where each word has two (or more) forms at any given point in time. The neogrammarian position has been that the change (phonetically gradual and phonologically abrupt) occurs in all words at the same rate. In contrast, the lexical diffusion (Wang, 1969) theory suggests that the change is initiated in some words and gradually "diffuses" through the lexicon to completion.

4. *monolingualism versus bilingualism:* At any given point in time when the two linguistic forms are competing, do speakers choose one form versus the other or do they essentially become bilingual users of both forms? These two different cases might have very different consequences for languages changing with time and whether such a change goes to completion. We will elaborate on this later in the chapter.

5. *the role of frequency (statistical) versus rule (categorical) effects:* While language is typically conceptualized as categorical (algebraic), there are clearly statistical effects at the margins of categorical behavior. Linguistic expressions be they words, phrases, or sentences of different forms do not all occur with the same frequency in language use. The probability with which various forms are heard might affect the acquisition of those forms and therefore the transmission of forms from one generation to the next. This accumulated effect might ultimately lead to categorical change.

As we construct explanations for various historically observed trends, we need to tease these issues apart. We need tools to help us reason about such issues to separate the plausible from the implausible. In the next section, we explicitly consider a series of models of word learning and examine the long term evolutionary consequences at the population level of various word learning models at the individual level.

8.2 Two forms in a Population

We develop some models of word learning in the context of the phonological merger in Wenzhou. Clearly, there are many words in the lexicon and let us focus on one particular word with its two alternative pronunciations.

In generation t, let a proportion α_t of the population use the changed form of a particular word. Correspondingly, a proportion $1 - \alpha_t$ of the population use the original form of the word.

The next generation of children will now obviously hear both forms of the word being used by adults at large. We will now consider the following four situations in turn.

1. Children hear both forms of the word. Each child receives a different random draw of words from the *entire* adult population. Each child acquires the form that occurs more often in a certain sense.

2. Children hear both forms of the word. Each child receives a different random draw of words. However, these words are heard *only* from the parents and therefore reflect the linguistic form of each parent. Each child acquires the form that occurs more often in a certain sense.

3. Children hear both forms of the word. Each child acquires *both* forms. However, it uses the two forms in a ratio that reflects the ratio of the two forms in the data set it received (from the entire adult population) during the learning phase.

4. Children hear both forms of the word. These words are heard *only* from the parents. Each child acquires *both* forms. Moreover, it uses the two forms in a ratio that reflects the ratio of the two forms in the data set it received during the learning phase.

As we shall see in the analysis that follows, these different situations lead to different evolutionary behavior at the population level. Linguistic

explanations of the ultimate loss of one form will therefore need to take these differences into account.

8.2.1 Case I

Let us assume that each child hears N words after which it acquires one or other of the two linguistic forms of the word. Since the words are randomly drawn from the entire adult population of speakers, the probability with which an arbitrary child will hear exactly k words of form 1 is given by

$$\binom{N}{k}(\alpha_t)^k(1-\alpha_t)^{(N-k)}$$

Let us further assume that a child acquires form 1 if it occurs at least K times during its learning phase. Therefore the probability with which an arbitrary child will acquire form 1 is clearly

$$\sum_{k=K}^{N}\binom{N}{k}(\alpha_t)^k(1-\alpha_t)^{(N-k)}$$

Naturally there will be variation in the population of children. This variation will arise due to particular differences in the primary linguistic data set that each individual child is exposed to. Thus some children will have acquired form 1, others will have acquired form 2. Crucially, however, in the population at large, the proportion of children who have acquired form 1 is also given by

$$\alpha_{t+1} = \sum_{k=K}^{N}\binom{N}{k}(\alpha_t)^k(1-\alpha_t)^{(N-k)} \tag{8.1}$$

Thus we have immediately related the proportion of form 1 users in two successive generations. This gives us an iterated map that we examine in some detail below.

8.2.2 Analysis

Note that Eq. 8.1 represents a non linear (polynomial) iterated map from the interval $I = [0, 1]$ to itself whose properties now need to be examined. It is possible to show the following

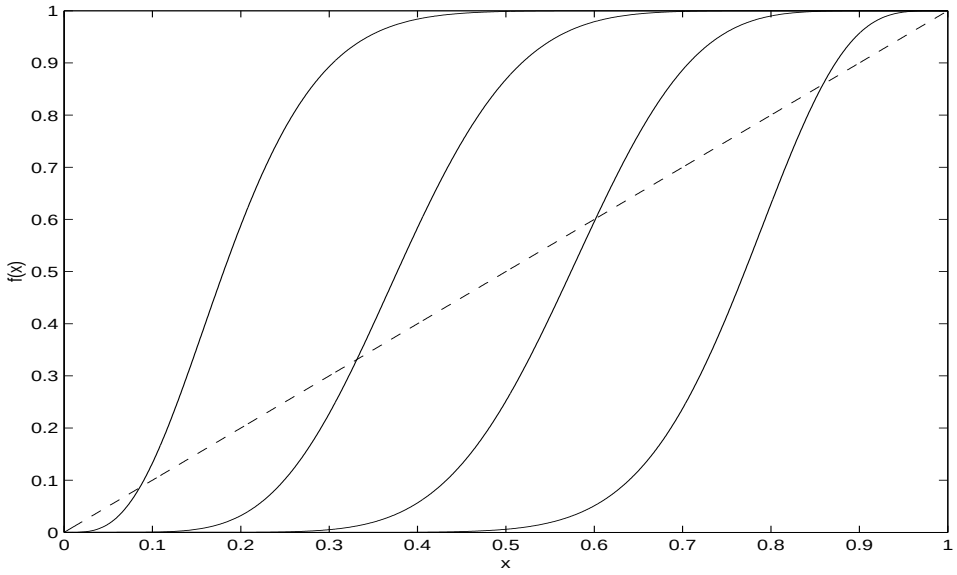

Figure 8.2: The binomial map $f(x) = \sum_{i=K}^{N} \binom{N}{i} x^i (1-x)^{N-i}$ plotted against x. The dotted line is the identity map $f(x) = x$. N is chosen to be 20 and the four curves correspond to choices of $K = 4, 8, 12, 16$ respectively.

1. There are two *stable* equilibrium points that are given by $\alpha = 0$ and $\alpha = 1$ respectively.

2. There is *exactly* one *unstable* equilibrium point obtained in the interior $\alpha^* \in (0, 1)$.

3. There are no further equilibria.

To see this, consider the map $f(\alpha) = \sum_{k=K}^{N} \binom{N}{k} \alpha^k (1-\alpha)^{N-k}$. Shown in Fig. 8.2 is a graph of this function for an arbitrary choice of $N = 20$ and various choices of K.

By inspection of $f(\alpha)$, it is easy to see that $f(0) = 0$ and $f(1) = 1$. To show stability, it is sufficient to show that $|f'(\alpha)| < 1$ for those values of $\alpha = 0, 1$ respectively. Differentiating f term by term, we get

$$f'(\alpha) = \sum_{k=K}^{N-1} \binom{N}{k} \left[k\alpha^{k-1}(1-\alpha)^{N-k} - (N-k)\alpha^k (1-\alpha)^{N-k-1} \right] + N\alpha^{N-1}$$

Expanding this out, we see

$$\begin{aligned} f'(\alpha) = & \sum_{k=K}^{N-1} \frac{N!}{(N-k)!k!} k\alpha^{k-1}(1-\alpha)^{N-k} - \sum_{k=K}^{N-1} \frac{N!}{(N-k)!k!}(N-k)\alpha^{k}(1-\alpha)^{N-k-1} \\ & + N\alpha^{N-1} \end{aligned}$$

Factoring N out of the expression, we have

$$\begin{aligned} f'(\alpha) = & N\left[\sum_{k=K}^{N-1} \frac{(N-1)!}{(N-k)!(k-1)!}\alpha^{k-1}(1-\alpha)^{N-k} - \sum_{K}^{N-1} \frac{(N-1)!}{k!(N-k-1)!}\alpha^{k}(1-\alpha)^{N-k-1} \right. \\ & \left. + \alpha^{N-1}\right] \end{aligned}$$

After cancelling terms, we see

$$f'(\alpha) = N\left(\binom{N-1}{K-1}\alpha^{K-1}(1-\alpha)^{N-K}\right) \tag{8.2}$$

Clearly, $f'(0) = f'(1) = 0$. Stability is shown.

It is easy to see that there is at least one fixed point in the interior. For that, we need simply to consider

$$h(x) = f(x) - x$$

and notice that (i) $h(0) = h(1) = 0$ and (ii) $h'(0) = h'(1) = -1$. From (i) and (ii) we have that there must exist $1/2 > \epsilon_1 > 0$ and $1/2 > \epsilon_2 > 0$ such that $h(\epsilon_1) < 0$ and $h(1-\epsilon_2) > 0$. Now since $h(x)$ is continuous, there must exist some α^* such that $h(\alpha^*) = 0$, i.e., α^* is an equilibrium point. Further, it is easy to see that since $h(\epsilon_1) < 0$ and $h(1-\epsilon_2) > 0$, and h is continuous, there are an odd number of interior roots in $(0,1)$. These roots correspond to equilibrium points that are alternately unstable and stable. Let the roots be given by $\alpha_1, \ldots, \alpha_{2m+1}$. Then α_1 is unstable, α_2 is stable, α_3 is unstable and so on. However, the structure of $f'(x)$ (having only one local maxima in $(0,1)$) as derived in Eq. 8.2 shows that there is only one root and it has to be unstable.

8.2.3 Case II

As a starting point, we consider the case of random mating between individuals in the population. This gives rise to four types of parental groups (AA refers to the type where the father uses form A and the mother uses

Pat. Form	Mat. Form	P(Ch. Form =A)	$P(Types)$	Rand. Mating
A	A	b_3	p_3	α_t^2
A	B	b_2	p_2	$\alpha_t(1-\alpha_t)$
B	A	b_1	p_1	$\alpha_t(1-\alpha_t)$
B	B	b_0	p_0	$(1-\alpha_t)^2$

Table 8.2: The linguistic forms of each parent (A and B indicate the two kinds of forms). The third column indicates the probability with which a child of each parental type will acquire the form A. Column 4 indicates the probability of each parental type and its value under random mating (Column 5).

form A; the other three types arise from the other possible combinations). If a proportion α_t of the population use form A, and there is no difference between males and females in this regard, one can compute the proportions of each of the four parental types in the population. This is indicated in Table 8.2.

Assume that children are born into each parental type equally often. Recall that each child hears N words drawn at random from the parents. One can now list the source distributions for each of the parental types. These are

1. Children of AA hear only the form A and consequently acquire the form A.

2. Children of BB hear only the form B and consequently acquire the form B.

3. Children of AB and BA hear both forms. Each form has probability $\frac{1}{2}$ of occuring. The child acquires either form A or form B using the same criterion as in Case 1, i.e., it chooses A if this form occurs at least K times in the set of N words.

Thus, according to the notation of Table 8.2, we see that $b_3 = 1$ and $b_0 = 0$. Clearly, $b_2 = b_1$ and this is given by

$$b_2 = b_1 = \sum_{k=K}^{N} \binom{N}{k} (\frac{1}{2})^k (\frac{1}{2})^{(N-k)}$$

We can now compute the proportion of children who have acquired form A. This is given by

$$\sum_{i=1}^{4} b_i p_i = 1.(\alpha_t^2) + b_2.(2\alpha_t(1-\alpha_t)) + 0.(1-\alpha_t)^2$$

$$= \alpha_t(2b_2 - (2b_2 - 1)\alpha_t)$$

Thus we have again related the proportion of children acquiring form A in two successive generations. This yields the logistic map and the analysis of this classic iterated map is well known. We can make the following observations.

1. The evolution depends upon the value of b_2. If $b_2 = \frac{1}{2}$ then $\alpha_{t+1} = \alpha_t$. Therefore the population never changes in linguistic composition.

2. If $b_2 \neq \frac{1}{2}$, there are two fixed points $\alpha = 0$ and $\alpha = 1$ corresponding to the two homogeneous populations. Exactly one of these is stable. If $b_2 > \frac{1}{2}$ then $\alpha = 1$ is stable, if $b_2 < \frac{1}{2}$ then $\alpha = 0$ is stable.

Variations of the Parental Model

The model discussed above closely follows that developed in Niyogi (2002) and is intimately related to cultural models developed in Cavalli-Sforza and Feldman (1981). This connection is discussed at length in the next chapter. Several variations have been considered in Cavalli-Sforza and Feldman (1981) and we briefly consider a few in the current context.

First, note that we have assumed that both parents have an equal role in producing primary linguistic data for the child. In general, the update rule is given by

$$\alpha_{t+1} = \alpha_t(b - (b-1)\alpha_t)$$

where $b = b_1 + b_2$. This, does not affect the qualitative nature of the results.

Second, consider relaxing the assumption of random mating. After Cavalli-Sforza and Feldman (1981) one might model this as follows. Let a proportion p of the population mate in an assortative way (i.e., the type AB does not occur) and a proportion $1-p$ mate at random. In that case, the update rule is given by

$$\alpha_{t+1} = \alpha_t[p + (1-p)b + \alpha_t(1-p)(1-b)]$$

where $b = b_2 + b_1$ as before. Again, this does not affect the qualitative nature of the results.

A greater variety of models may be developed by considering the neighborhood structure of the population or oblique transmission (across generations). These are considered in the next chapter and their applicability to the Wenzhou case is not considered any further.

8.2.4 Case III

First consider generation number t of adult speakers. Let an arbitrary adult in this population use the two forms in the ratio λ_t. The distribution $P(\lambda_t)$ of possible λ_t-values captures the variation in the adult population with respect to the usage of these two forms. The mean of this distribution is given by

$$E[\lambda_t] = \int \lambda_t P(\lambda_t) d\lambda_t.$$

This mean characterizes the average use of each of these two forms in the population as a whole.

Now consider the arbitrary child. This child hears words uttered by random speakers with random usage of each of the two forms. With probability $P(\lambda_t)$ it comes across a speaker who uses form A with probability λ_t. Therefore, averaging over all speakers, the total probability of hearing a word of form A is given by

$$\int P(\lambda_t)\lambda_t d\lambda_t = E[\lambda_t]$$

This arbitrary child hears N words in all. Some of these (say n) are of form A and the rest $(N - n)$ are of form B. Let us assume that the child matures into adulthood using both forms in a manner that reflects its childhood learning experience. Thus the child uses form A with probability $\frac{n}{N}$. Clearly n is a random variable and different children will have different experiences and there will be a variation in the population of speakers of the next generation – $t + 1$. However, the average over this population is given by

$$E[\lambda_{t+1}] = E[\frac{n}{N}] = \frac{E[n]}{N} = \frac{NE[\lambda_t]}{N} = E[\lambda_t]$$

Interestingly, the average use of the two forms in the population does not change at all! There is no evolution and the usage of both forms in the population is preserved.

8.2.5 Case IV

We consider this variant of case III for completeness. Here the child is exposed only to data from his or her parents. Assume in generation t the father has a lambda-value of λ_F and the mother has a lambda-value of λ_M. Both values of λ are samples from a distribution of λ-values in the population that can be characterized by $P(\lambda)$ – a probability distribution over λ-values in the usual way.

Assuming random mating, as before, the proportion of children whose fathers have $\lambda = \lambda_F$ and mothers have $\lambda = \lambda_M$ is given by $P(\lambda_M)P(\lambda_F)$. Such children will receive N instances of the word each. Assuming that parents are equally likely to speak to the children (as before) we have that with probability $\frac{1}{2}$ the father will provide a word and with probability $\frac{1}{2}$ the mother will. Therefore, the probability with which the child will hear form 1 is given by

$$p = \frac{1}{2}\lambda_F + \frac{1}{2}\lambda_M$$

Let us assume that of N instances of the word, n are of form 1 and the remaining $N - n$ are of form 2. The child would estimate λ as $\frac{n}{N}$. Denoting by $\hat{\lambda}(\lambda_M, \lambda_F)$ the estimate of λ that such a child (whose parental λ-values are denoted by λ_M and λ_F respectively) constructs, we have,

$$\hat{\lambda}(\lambda_M, \lambda_F) = \frac{n}{N}$$

Clearly different children of the same parental type will actually get different random draws and therefore there might be some variation in the population of such children. However, their mean value is given by

$$E[\hat{\lambda}(\lambda_M, \lambda_F)] = \frac{NE[n]}{N} = p = \frac{1}{2}\lambda_F + \frac{1}{2}\lambda_M$$

Thus the average λ-value that such children will internalize is given by $\frac{1}{2}\lambda_F + \frac{1}{2}\lambda_M$. Averaging over children of all parental types, we have

$$E[\hat{\lambda}] = \int_{\lambda_M}\int_{\lambda_F}[\frac{1}{2}\lambda_F + \frac{1}{2}\lambda_M]P(\lambda_M)P(\lambda_F)d\lambda_M d\lambda_F = E[\lambda]$$

Clearly, the average value of λ in the generation of children (after learning) is the same as the average value of λ in the generation of adults. Like in case III, there is no evolution or change in the population over time.

8.2.6 Remarks and Discussion

The four cases we have considered fall into two broad categories (i) Cases I and II correspond to a situation where children make a categorical choice in the linguistic variable of interest, i.e., they choose *exactly* one of the two linguistic forms that exist in the population as a whole (ii) Cases III and IV, in contrast, correspond to a situation where children adopt both forms but in a ratio that reflects the frequency of occurrence of the two linguistic forms in the population. Two different modes of transmission were considered for each of these two subcases.

There are two interesting results of the analysis.

First, we observe that in cases I and II, mixed populations are inherently unstable. In case I, we see that an unstable equilibrium exists for a mixed population but if the population mix shifts even slightly from this unstable balance, one of the two forms is driven to extinction. In case II, the parameter b_2 has to be exactly equal to $\frac{1}{2}$ for a population mix to remain that way, else one of the two variants is driven to extinction. In cases III and IV, populations don't have any inherent tendency to change. Both forms are preserved in the population over generational time.

A second observation is worth making. In cases I and II, there is variation in the population that is forced by the categorical nature of language use, i.e., which of the two forms speakers use. In cases III and IV, the speakers are all potentially bilingual (i.e., use both forms) and in a proportion λ which varies from individual to individual. One can compute the variance of λ in the population and this characterizes the variation present in language use in the population. Interestingly, we see

$$Var(\lambda_{t+1}) = Var(\frac{n}{N}) = \frac{1}{N^2} Var(n)$$

where each individual child simply estimates the λ-value as $\frac{n}{N}$. We saw earlier that $n = \sum_{i=1}^{N} X_i$ where X_i is a $0-1$ valued random variable that takes on the value 1 if form 1 occurs and 0 otherwise. Now form 1 occurs with probability $E[\lambda_t]$. Therefore,

$$Var(n) = NVar(X_i) = NE[\lambda_t](1 - E[\lambda_t])$$

Therefore

$$Var(\lambda_{t+1}) = \frac{E[\lambda_t](1 - E[\lambda_t])}{N}$$

But we have seen that $E[\lambda_t]$ remains the same from generation to generation. Thus $E[\lambda_{t+1}] = E[\lambda_t] = \alpha$ (say). Both the mean and the variance in the population is fixed for all time.

Thus we see that in cases I and II, the variation in the population is gradually lost over time as one of the two forms is driven out of existence. In contrast, in cases III and IV the variation is maintained. This leads us to conjecture that it is the categorical nature of language that forces change. We will explore this in greater detail in the later sections.

8.3 Examining the Wenzhou Data Further

Let us reexamine the Wenzhou data keeping in mind the issues of section 2 and the analyses of the previous section. The data has been presented and studied in great detail by Shen (1993) and we present certain aspects of that study that are particularly interesting in the light of our current analysis.

A central conclusion of our analysis is that categorical behavior by language users is more likely to result in change than bilingual or blending behavior which tends to maintain the variation in the population. The first interesting question for us was to figure out which of these two kinds of linguistic behavior was closer to reality in the case of Wenzhou.

An examination of the data in Shen (1993) suggests that for a particular word with its two forms, people tend to always use one form consistently. A single individual never uses both forms for the same word. Of course, there is variation in the population with a higher percentage of older people using form 1 (diphthongized) and a higher percentage of younger people using form 2 (monophthongized). This is consistent with our analysis since a categorical use of forms tends to be stable only in homogeneous populations with only one form present.

Shown below are the percentage of speakers who use the monopthongized form of each word in 11 age brackets (from Shen, 1993) for the five words presented in the earlier section.

The thirty five different words do not all change at the same time nor do they change at the same rate. If one studies the distribution of the two forms of each word in the population, then the oldest people all use form 1 (diphthongized) and the youngest mostly use form 2 (monophthongized) version of the word. The speakers in the middle (intermediate age) show a variation along two directions (i) the number of words they have changed and (ii) which words they have changed.

Word	Meaning	Age									
(diph.)		17.7	21.4	26.9	32.2	36.7	42.2	46.5	51.6	57.3	61.0	72.8
		24	28	30	43	47	65	54	39	19	8	6
$/po^y/^{42}$	"cloth"	88	86	87	58	34	14	4	8	0	0	0
$/do^y/^{31}$	"graph"	70	86	93	65	40	20	9	8	0	13	0
$/mo^y/^{31}$	"to sharpen"	77	93	93	70	43	23	11	10	0	0	0
$/to^y/^{42}$	"jealous"	88	86	87	58	34	14	4	8	0	0	0
$/so^y/^{42}$	"to tell"	88	86	87	58	34	14	4	8	0	0	0

Table 8.3: Percentage of speakers using the monophthongized form for each of five words. The first column provides the word, the second its meaning, and the columns after that indicate the percentage of speakers using the changed form for each of 11 age groups. The 11 age groups are indicated by their average age in the first row (with the number of people sampled in each age group in the second row). Adapted from Shen (1993).

The change in some of the words occurs before the change in other words. Further, Shen (1993) finds that the rate of change is higher for words that have started changing later. This lends support to the lexical diffusion theory put forward by Wang (1969).

The preceding discussion suggests that cases I and II are closer to the underlying phenomena than cases III and IV. Consider now cases I and II in some greater detail. Is it possible to disambiguate these two cases from the data? In other words, in the acquisition of the appropriate pronunciation for each word (form) do speakers learn only (mostly) from parents or is it more likely that they learn from the entire population. Furthermore, is it possible to evaluate the threshold from the data? Are these thresholds different for different words?

Consider the analysis under case I. We see that both $\alpha = 0$ and $\alpha = 1$ are stable equilibrium conditions where α is the proportion of form 1 (monophthong form) users. Let the unstable interior equilibrium condition be α'. The initial condition was $\alpha = 0$ and under the modeling assumptions this would remain stable over all time. Therefore, internally driven change is not possible. Language contact with another population consisting of mostly form 1 users is necessary for change. In particular, the number of new speakers must be enough to move the effective population mix from one basin of attraction to another, i.e, $\alpha > \alpha'$. In other words, a slight introduction of form 1 users in the population is not sufficient – it must actually be greater than α'.

Now consider the analysis under case II. In this case, exactly one of the conditions $\alpha = 0$ and $\alpha = 1$ is stable. The other is an unstable equilibrium.

Clearly, $\alpha = 0$ can not be a stable situation – for the system would never be driven to change. Therefore, $\alpha = 1$ must be the stable situation where all speakers tend to produce monophthongized versions of the word and this is what the system ultimately tends to anyway. However, for this to be the case, $b_2 > \frac{1}{2}$ which means that $K < \frac{N}{2}$. In other words it is easier to learn form 1 (monophthongized) than it is to learn form 2. In this case, even the slightest contact with form 2 users will drive the entire population to change completely. Of course, this raises the question – how come the population found itself in an unstable equilibrium in the first place?

It is only an examination now of the social setting in which the change comes about that will yield further insight into this issue. If a substantial language contact came about, both case I and case II are likely scenarios. However, in the absence of a substantial language contact, case II is the more likely candidate.

It should also be said that on the face of it the dynamics of case I seems more satisfying since it is desirable that both linguistic systems be stable. It is reasonable to think that Wenzhou speakers maintained the diphthongized form of the words for many generations in a relatively stable mode until change came about over the last century. It is hard to imagine how a basically unstable linguistic system would be maintained for so long.

A puzzling question for us is – what initiates the change (the so called *actuation problem*)? It is reasonable to assume that at one point in time only the diphthongized form of the word was used by the population. Such a population is actually in (stable) equilibrium – so what caused it to change? And why didn't the change get initiated earlier? The analysis so far has not shed any light on this question. In the next section, we consider error driven models. In particular, we see that if there are asymmetries in the errors during the learning period, the resulting population dynamics may have bifurcations. These bifurcations have the potential to successfully resolve the actuation problem — a problem that has resisted a satisfactory explanatory account in the scientific literature so far.

8.4 Error Driven Models

In the analysis above, it has been assumed that transmission between speaker and hearer is essentially error free. Each learner internalizes a particular form of the word and produces it upon adulthood. Listeners perceive this particular form upon receiving it. In reality, the transmission is likely to

have some errors. This is especially so in phonetic and phonological communication where the speaker's intent might be misperceived by the listener. For example, in Wenzhou, it is conceivable that due to phonetic similarity, the diphthong was perceived as a monophthong (and vice versa) on occasion by listeners.

Let us examine the consequences of such errors in the kinds of models discussed above. Let the probability with which form 1 is misperceived (i.e., perceived as form 2) be ϵ_1 and the probability with which form 2 is misperceived as form 1 be ϵ_2. Then if form 1 was uttered with probability α, one can compute the probability with which a random utterance was perceived as form 1. This is given by

$$\alpha' = \alpha(1 - \epsilon_1) + (1 - \alpha)\epsilon_2$$

Clearly, if $\epsilon_1 = \epsilon_2 = 0$, then we have the error free case and $\alpha' = \alpha$ – that is the probability of hearing form 1 is the same as the probability with which form 1 is produced by speakers. Thus, if a learning child hears N instances of the word directed at it, the probability of perceiving more than K of them as form 1 is simply given by

$$\sum_{i=K}^{N} \binom{N}{i} (\alpha')^i (1 - \alpha')^{(N-i)} = f(\alpha')$$

In the models of case I and case II, the population contains a mix of type I (users of form 1) and type II (users of form 2) speakers in proportion α to $1 - \alpha$. Consider case I. The source probability of form 1 is simply α. As a result of communication errors along the channel, the probability with which form 1 is perceived is given by α'. The probability with which the typical learning child acquires (internalizes) form 1 as the preferred pronunciation is given by $f(\alpha')$. The update rule for case I is simply

$$\alpha_{t+1} = f(\alpha'_t) = f(\alpha_t(1 - \epsilon_1) + (1 - \alpha_t)\epsilon_2)$$

One can now conduct an equilibrium analysis of this update rule in the usual way. When $\epsilon_1 = \epsilon_2$, i.e, the errors are symmetric, an analysis yields the following:

1. The stable fixed points at $\alpha = 0$ and $\alpha = 1$ are now lost. Instead, all fixed points are interior.

2. There are two stable fixed points at $\alpha = \alpha_1$ and $\alpha = \alpha_2$.

3. There is one unstable interior point at $\alpha = \alpha^* \in (\alpha_1, \alpha_2)$.

Thus we see that the effect of noise does not change the spirit of the results in case I. The two stable modes of the population are still *largely* homogeneous. Due to errors in the channel, the stable modes of the population contain a *small* fraction of the "other" linguistic type. The proportion of this type in the stable mode changes with the error rate of the channel. The mixed interior mode is still unstable.

8.4.1 Asymmetric Errors

The case when $\epsilon_1 \neq \epsilon_2$ is the most interesting for our purposes. For simplicity, let us assume that $\epsilon_2 = 0$ while $\epsilon_1 > 0$. In this case, form 1 may be misperceived by the listener on occasion while form 2 is never misperceived. In the context of our case study, this amounts to assuming, for example, that diphthongs are occasionally misperceived as monophthongs while the reverse never occurs.

In this case, we see that the update rule is given by

$$\alpha_{t+1} = f(\alpha'_t) = f(\alpha_t(1 - \epsilon_1))$$

In other words,

$$\alpha_{t+1} = \sum_{i=K}^{N} \binom{N}{i} (p\alpha_t)^i (1 - p\alpha')^{(N-i)}$$

where $p = 1 - \epsilon_1$. We encountered this map in Chapter 5. Recall that the dynamics corresponding to this depends upon the value of p. For large values of p close to 1, the system is *bistable* with two stable fixed points $\alpha = 0$ and $\alpha = \alpha_2 \leq 1$. There is one unstable fixed point $\alpha_* \in (0, \alpha_2)$. Below a critical threshold, i.e., when $p < p_c$, the system has only one stable fixed point given by $\alpha = 0$.

Thus there are two *regimes* the system can be in. One corresponding to a bistable regime when there are two possible stable modes and the other corresponding to a regime when there is only one possible stable mode. The bifurcation between these two regimes is determined by the critical value of p_c.

8.4.2 Bifurcations and the Actuation Problem

The bifurcation noted here serves as a useful explanatory construct for the actuation problem. Recall our discussion of Sec. 8.3. The various symmetric

models of Cases I through IV had no bifurcations and left us puzzling as to *why* a population would move from one stable state to another.

Let us now offer an explanation for this puzzle. Assume that errors are asymmetric, and that $\frac{K}{N}$ and $p = 1 - \epsilon_1$ are fixed from generation to generation. In other words, (i) the learning algorithm used by the learner, and (ii) the perceptual discriminability does not change from one generation to the next. It is possible to compute the value of the critical point p_c and it is seen that p_c depends upon N, i.e., the total number of times the word in question was heard (in either form) by the learner. In particular it is seen that p_c increases as N decreases.

One can now imagine the following situation. Consider a particular word with two possible forms. At one point in time, the word was used often in discourse so that learners had many examples of that word on which to base their learning decision. Thus N was large and $p > p_c$. In this situation, there are two stable modes and one of these $\alpha = \alpha_2 \approx 1$ corresponds to a situation where most (all) of the speakers use form 1 (the diphthong).

Now consider what happens if the word starts being used less frequently in subsequent generations. The value of N decreases, p_c correspondingly increases, and at some point, if the word is used infrequently enough, a bifurcation occurs. At this point, the evolutionary dynamics enters the regime when $p < p_c$. Now there is only one stable point ($\alpha = 0$) corresponding to a population of form 2 users. A population of form 1 users is unstable and gradually the population drifts from users of diphthongs to users of monophthongs for the word in question.

In this manner, a change may come about simply because of a drift in the frequency with which the word is used during the learning period. This posits an entirely different mechanism for initiating language change. Under the assumption of our error-free models of Cases I through IV, the only way in which we could explain language change was by invoking language contact. Language contact would bring about a dramatic change in the mix of linguistic types in the population thereby driving the population from one equilibrium state to the basin of attraction of another. In the absence of such a dramatic reconstitution of the population, it was difficult to see how change might come about.

Our current explanatory framework differs from the theory of language contact in two ways. First, it is possible for change to be internally driven with no dramatic reconstitution of the population due to language contact. Second, it is possible for only a slight change in the value of N to bring about the bifurcation. Let us denote the dependence of p_c on N explicitly

by writing it as $p_c(N)$. Since p_c depends continuously on N, we see that adjacent values of N may let $p_c(N)$ be greater than or less than p. Hence, a subtle and gradual change in frequency based behavior may lead to a dramatic change in the stable language of subsequent generations.

The discovery of bifurcations in this context must count as one of the central insights obtained by the mathematical approach to the study of language evolution embodied in this book.

8.5 Discussion

As we have discussed earlier, the case where two linguistic forms are in competition is almost ubiquitous in historical linguistics.

8.5.1 Sound Change

Sound change is one of the early and well studied examples of historical linguistics for which there is often extensive fieldwork available. Indeed Wang (1977), Labov (1994), Kiparsky (1982) have all concerned themselves with changes of this nature.

Explanatory frameworks for such changes often reside at the phonological level in notions of phonological rule restructuring (see Kiparsky, 1982) and at the phonetic level in notions of speech production and perception. For example, Ohala (1989) discusses how "diachronic variation is drawn from a pool of synchronic variation". It is argued that there is synchronic variation among the speakers of a language and a language learner will need to deal with this variation. Further, miscommunication may occur between speaker and hearer due to articulatory sloppiness on the part of the speaker or perceptual confusion on the part of the hearer (learner). Several phonetic sources of such miscommunication were considered in an admirable discussion on potential sources of sound change.

To take an example from Ohala (1989), consider the classic sound change of Indo-European labio-velar stops to labial stops in Classical Greek.

Proto IE	*Classical Greek*	
ekwos	hippos	"horse"
gwiwos	bios	"life"

It is believed that at one point, the word for "horse" had a labio-velar stop and later it had only a labial stop. Again we have two forms in competition just like the much better documented case of Wenzhou we examined in this chapter. One form gradually gave way to the other. While the two sounds are

confusible and learners might incorrectly learn the wrong word, the analysis in the current chapter shows that this by itself is not enough to drive an entire population to change. Much more subtle population level analysis must be conducted to see how how sounds may be transmitted from speaker to speaker across generations before one can explain the clear directionality of the change.

8.5.2 Connections to Population Biology

Population biology has long considered mathematical models of gene transmission to characterize the evolution of gene frequencies in populations. In many ways, the models considered here are very similar in spirit to population biology models — what is evolving is the frequency of linguistic forms rather than biological forms. The laws of transmission of linguistic forms from one generation to the next are assumed to be governed by the language acquisition process. The macroscopic (population) consequences of such transmission over generational time scales need to be explored and this chapter represents a step in such a direction. In contrast to the case of language, the laws of transmission of genes (or other biological properties) from one generation to the next are usually governed by reproduction.

There are some further interesting connections to population biology. We see a distinction in evolutionary consequences for "categorical" as opposed to "blending" behavior in language learners. This difference is analogous to the distinction between evolutionary consequences of particulate versus blending inheritance. As is well known (see, for example, Fisher, 1930), particulate inheritance maintains variation in the population while blending inheritance eliminates it. In contrast, we see that categorical behavior in language eliminates variation and drives linguistic change to completion while blending behavior maintains such variation and tends to suppress change. It is also noteworthy that simple and abstract mathematical models played an important role in reasoning about the distinction between particulate and blending inheritance by Fisher and other writers of that period.

8.6 Conclusions

Using the case of lexical diffusion in the Wenzhou province of China as a motivating example for which concrete data exists, this chapter has explored models of language change when two linguistic forms co-exist in a population

and are transmitted from one generation to the next through the process of language acquisition.

A number of different models were developed leading to four different cases that were systematically studied. Interestingly enough, none of these four cases provides a satisfactory account of the change. The careful study of these four cases and their failure to account for the change in Wenzhou province highlights the power of our computational approach. It allows us to work out the consequences of various assumptions about language and learning and demonstrates the falsifiability of the models constructed. It brings into sharper focus the mystifying question that has motivated much of this entire book. Why does a linguistic community change from one stable mode to another? One answer to this question may be that language contact between two different groups as a result of migration may dramatically alter the linguistic composition of the population and move it into a basin of attraction of the other stable point. A second answer may be that a drift in certain parameters of learning may cause a bifurcation in the population dynamics. This bifurcation could explain why a previously stable linguistic system becomes unstable. The asymmetric error driven model of this chapter illustrates this point by showing how the change in the frequency with which a word is used during the learning period crucially affects the evolutionary dynamics. A third answer may be that change comes about because of random drift caused by finite population sizes. This is a possibility we did not explore in this chapter but we will take it up at a later point in this book.

Finally it is also worth reiterating an important difference between two modes of individual linguistic behavior as far as long term evolutionary consequences are concerned. It was found that categorical behavior on the part of learners results in an inherent tendency of linguistic populations to change with time to a homogeneous stable mode with only one linguistic form surviving. Blending behavior on the part of the learner leads to both forms being preserved in the population at large. This is in contrast to models of inheritance in evolutionary biology where blending inheritance eliminates variation while particulate inheritance preserves it. It will become important to incorporate this insight in an account of language change over time.

Chapter 9

A Model of Cultural Evolution and its Application to Language

9.1 Background

The evolutionary paradigm has applicability beyond the particular case of genetic reproduction to a wide variety of situations — from ecology (May, 1973) to cooperation and conflict in populations (Axelrod, 1984; Nowak and Sigmund, 1993) to the evolution of culture and cognition (Boyd and Richerson, 1985; Cavalli-Sforza and Feldman, 1981). In this chapter, we take a closer look at models of cultural evolution and outline the relationship between such models and the models of language change that we have discussed so far.

As we have seen, 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.

It is worthwhile to note that in considering the evolution of the linguistic

system, we have not invoked the genetic changes that may be going on in human populations at the same time. The explanatory paradigm does not rely on notions of "linguistic" fitness, genetic transmission, or differential reproduction. In this sense, the evolution of language over historical time scales may be viewed as a certain kind of neutral cultural evolution. At the same time, this does not preclude the study of the correlation between languages and genes (see, e.g., Cavalli-Sforza, 2001) or the intriguing possibility of coevolution of genes and languages.

In a remarkable treatise in 1981, the evolutionary biologists, L. Cavalli-Sforza and M. 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 chapter, 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[1].

The framework for the computational characterization of changing linguistic populations discussed in this book so far has evolved from an original series of papers by Niyogi and Berwick (1995,1997). 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 utilize the insights of the Cavalli-Sforza and Feldman framework to develop an extended model to characterize the effect of spatial (geographical) location. This allows us to study spatial effects on the linguistic interactions

[1] It is best to state upfront that we do not attempt in this chapter to provide any review or systematic coverage of the various mathematical approaches that have been taken to the problem of cultural evolution. This would be quite beyond our current scope. For example, Boyd and Richerson (1985) is an important contribution to cultural evolution and could well have been a comparison point for us. We chose to focus on Cavalli-Sforza and Feldman (1981) in part because the mathematical treatment therein was most directly comparable to our approach. Correspondingly, similarities and differences could then be sharply outlined.

between individuals in a population and the evolutionary consequences of such interactions.

9.2 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 is inspired by the transmission of genetic parameters over generations and serves as a point of entry to studying the complex issue of gene-culture co-evolution. In a cultural setting 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 chapter), 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 will acquire each of these traits depends upon its immediate environment – in the standard case of their model (though variations are considered[2]), these traits are acquired from its parents. Thus one can construct Table 9.1.

The first three columns of Table 9.1 are self-explanatory. As one can see, 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

[2]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 (which is briefly discussed at the end of the current chapter)

Paternal Val.	Maternal Val.	$P(ChildVal. = L)$	$P(Types)$	Prob.
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 9.1: 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.

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 the 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:

$$u_{t+1} = Bu_t^2 + Cu_t + D \tag{9.1}$$

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}) 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 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[3] to biological evolution is clear: (1) like gene-types, trait values are

[3]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.

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. In the next section we show how to apply this model to language change.

9.3 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.

9.3.1 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 have discussed many two-language models in previous chapters. In a later section we will also discuss the historical example of syntactic change in English for which this is a 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 L_i possess.

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 different languages with different probabilities.

In the abstract, let us assume that children follow some acquisition algorithm $\mathcal{A}$ 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). Following Chapter 2, we let $\mathcal{D}_k$ be the set of all k-tuples of sentences $(s_1, \ldots, s_k)$ where $s_i \in \Sigma^*$. Each such k-tuple denotes 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:

$$g(\mathcal{A}, P, k) = \sum_{\{w \in \mathcal{D}_k : \mathcal{A}(w) = g_1\}} \prod_{i=1}^{k} P(w_i) \tag{9.2}$$

Recall that each element $w \in \mathcal{D}_k$ is a k-tuple of sentences. In Eq. 9.2 we denote the ith sentence of w by w_i. 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:

Statement 3 *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

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 9.2: 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 $\mathcal{A}$.

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 9.2.

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[4], we see that the population evolves according to Eq. 9.1. Substituting the appropriate $g's$ from Table 9.2 above in place of the $b_i's$ we obtain an evolution that depends upon P_1,P_2,$\mathcal{A}$, and k.

9.3.2 An Alternative Approach

In previous chapters, we have developed a series of models (hereafter, we refer to this class of models as NB[5] 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.

[4]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.

[5]After Niyogi and Berwick (1995, 1997) where the original formulation was articulated and analyzed.

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. This has the form

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

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.

In previous chapters, we have derived 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 populations 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.

9.3.3 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. 9.1. 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

$$s_{t+1} = (f(1) + f(0) - 2f(0.5))\, s_t^2 + (2f(0.5) - 2f(0))\, s_t + f(0) \tag{9.3}$$

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 them. Notice further that the CF update rule is *always* quadratic while the NB update rule is in general not quadratic.
Remarks: It is worthwhile to reflect on the difference in the evolutionary dynamics of CF and NB type models.

1. The evolutionary dynamics of the CF model depends upon the value of f at exactly 3 points. Thus one might have very different update rules in the NB model corresponding to different iterated maps f yet if these different f's agree at $x = 0, \frac{1}{2}$, and 1, the corresponding CF update would be the same.

2. If f is linear, then the NB and CF update rules are exactly the same. If f is nonlinear these update rules potentially differ.

3. The CF update is a quadratic iterated map and has one stable fixed point to which the population converges from all initial conditions. The NB model may have multiple stable fixed points.

4. For some learning algorithms, there may be qualitatively similar evolutionary dynamics for NB and CF models. For example, in the case of the Triggering Learning Algorithm (TLA) discussed below, this is the case. For some other learning algorithms, the qualitative behavior may be quite different. This is the case for the batch and cue based learning algorithms which are bistable in the NB framework but have a single stable fixed point in the CF formulation.

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) 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 derives alternative evolutionary

laws. In a later section we utilize this observation to divide children into classes that depend on their geographical neighborhood. This will allow us to derive a generalization of the NB model for neighborhoods. Before proceeding any further, let us now translate the update rules derived in Niyogi and Berwick (1995,1997) into the appropriate CF models. The update rules are derived for memoryless learning algorithms operating on grammars. We consider an application to English with grammars represented in the Principles and Parameters framework.

9.4 CF Models for Some Simple Learning Algorithms

In this section we consider some simple learning algorithms[6] (like the Triggering Learning Algorithm of Gibson and Wexler (1994), and the batch and cue based learners of previous chapters) and show how their analysis within the NB model can be plugged into Eq. 9.3 to yield the dynamics of linguistic populations under the CF model.

9.4.1 TLA and its Evolution

How will the population evolve if the learning algorithm $\mathcal{A}$ in question is the Triggering Learning Algorithm(or related memoryless learning algorithms in general)? The answer is simple. We know how the TLA driven system evolves in the NB model (from analyses in previous chapters). All we need to do is to plug such an evolution into Eq. 9.3 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.

[6]These algorithms have 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 them by other learning algorithms does not alter the spirit of the major points we wish to make in this chapter 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, a point that we have elaborated at length in previous chapters.

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.

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

$$f(s_t) = \frac{s_t(1-a)}{(1-b)+s_t(b-a)} + \frac{[b-s_t(b-a)]^k[(1-b)+s_t(a+b-2)]}{2[(1-b)+s_t(b-a)]} \tag{9.4}$$

In Eq. 9.4, 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. 9.4 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

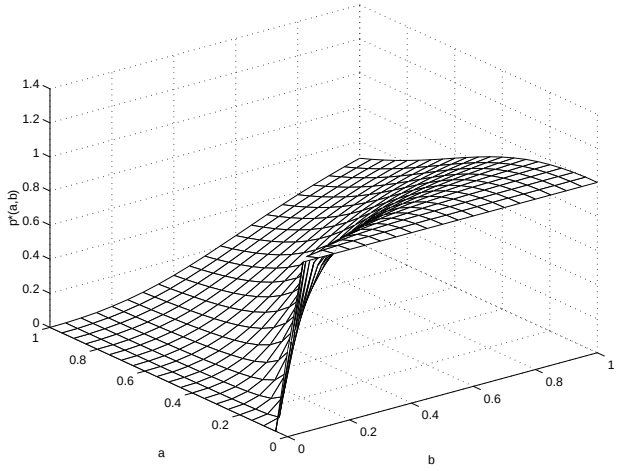

Figure 9.1: The fixed point $p^*(a, b)$ for various choices of a and b for the NB model with $k = 2$.

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. 9.1.

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 $p^* = \frac{1}{2}$. (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. 9.2. 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. 9.3 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

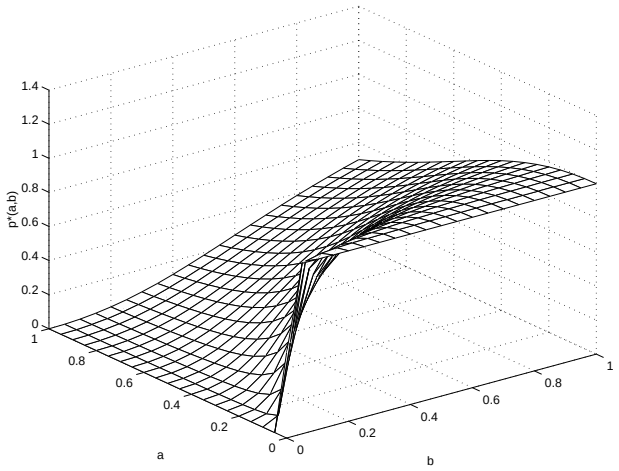

Figure 9.2: The fixed point $p^*(a, b)$ for various choices of a and b for the CF model with $k = 2$.

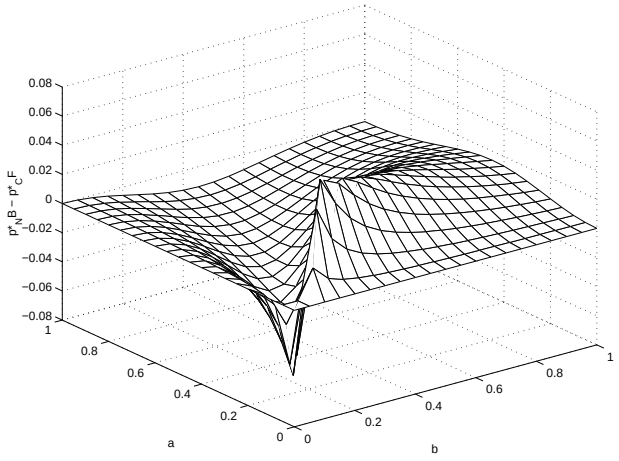

Figure 9.3: The difference in the values of $p * (a, b)$ for the NB model and the CF model $p^*_{NB} - p^*_{CF}$ for various choices of a and with $k = 2$. A flat surface taking a value of zero at all points would indicate that the two were identical. This is not the case.

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.

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 chapter, 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 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 correspond to the case where the learner had 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.

9.4.2 Batch and Cue Based Learners

Consider the batch learner of Chapter 5. The update rule is provided by

$$\alpha_{t+1} = f(\alpha_t) = \sum_{\{(n1,n_2,n_3)|n1\geq n_3;\sum_i n_i=K\}} \binom{K}{n_1 n_2 n_3} p_1(\alpha_t)^{n_1} p_2(\alpha_t)^{n_2} p_3(\alpha_t)^{n_3}$$

where $p_1(\alpha_t) = \alpha_t(1-a); p_3(\alpha_t) = (1-\alpha_t)(1-b); p_2(\alpha_t) = a\alpha_t + b(1-\alpha_t)$. An analyis of this iterated map reveals that for all $0 < a, b < 1$, the system has two stable fixed points $\alpha = \alpha_*$ and $\alpha = 1$ where $\alpha_* < 1$. There is

one unstable fixed point $\alpha' \in (\alpha_*, 1)$ in between. This instantiates a case in which the NB dynamics is bistable. In contrast the CF model will have only one stable fixed point. In fact, in the CF model, we see that 1 is always a fixed point. In this setting, the CF update rule is given by

$$s_{t+1} = (f(0) + 1 - 2f(\frac{1}{2}))s_t^2 + 2(f(\frac{1}{2}) - f(0))s_t + f(0)$$

Putting $s_t = 1$, we see that 1 is a fixed point of the CF system. The derivative at 1 is given by

$$2(f(0) + 1 - 2f(\frac{1}{2})) + 2(f(\frac{1}{2}) - f(0)) = 2(1 - f(\frac{1}{2}))$$

Thus we see that if $f(\frac{1}{2}) > 1$, then $s = 1$ is a stable fixed point, otherwise $s = 1$ is unstable. In contrast, of course, the NB system is always stable at 1.

One may also consider the cue based learner of Chapter 5. Here the NB dynamics is given by

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

There is a regime where the NB dynamics has only one stable fixed point $\alpha = 0$ and a regime where the NB dynamics has two stable fixed points $\alpha = 0$ and $\alpha = \alpha_* > 0$. Of course, as usual, the CF dynamics has exactly one stable fixed point. Since $f(0) = 0$, we see that the CF dynamics is given by

$$s_{t+1} = (f(1) - 2f(\frac{1}{2}))s_t^2 + 2f(\frac{1}{2})s_t$$

Clearly, we see that $s = 0$ is always a fixed point. To see the stability of this fixed point, we need to differentiate at 0. We obtain

$$2f(\frac{1}{2}) < 1 \Leftrightarrow s = 0 \text{ is stable}$$

Thus when $f(\frac{1}{2}) < \frac{1}{2}$, we see that $s = 0$ is the only stable fixed point. When $f(\frac{1}{2}) > \frac{1}{2}$, we have

$$f(\frac{1}{2}) > \frac{1}{2} > \frac{f(1)}{2}$$

In this regime, the stable fixed point of the CF system is given by $s = \frac{2f(\frac{1}{2})-1}{2f(\frac{1}{2})-f(1)}$.

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

9.4.3 A Historical Example

So far our development has been fairly abstract. To ground the current 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 its 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 chapter, we will collapse the OV/VO and I-final/I-medial distinctions into a single head-complement parameter within the rubric 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) $XP \longrightarrow YP\ X'$ (ii) $X' \longrightarrow X$

2. The **V2** parameter: this denotes the tendency in some languages where the finite verb moves 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. [*PN remark:* implying that the +V2 grammar was therefore lost over time]
>
> (pp. 265-266)

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 chapter. 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.

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 (learned) 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 over the course of this book. 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[7] (with no recursion) that are associated with the +V2

[7]Of course, both L_1 and L_2 have infinite sentences each. Recall that the evolutionary

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 realized (lexical sequences) 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 a 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.*

The ambiguous sentence types are those that have different but valid parses under each of the two grammatical systems. For example, the sentence type SV corresponding to the sentence *John left.* belongs to each of the two languages (extensionally). However, this surface form has two different derivations. Under g_1, the subject *John* is the specifier of the inflectional head. Under g_2, the verb *left* moves from I to C and the +V2 constraint forces the subject to move to occupy the specifier of the C. To an external agent, it is not obvious what the underlying parse is and therefore whether there is underlying verb movement or not.

The set of ambiguous sentence types is simply given by intersecting the

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 psycholinguistic premise is that children learn only on the basis of degree-0 sentences (Gibson and Wexler, 1994; Lightfoot, 1991) and all sentences with recursion are ignored in the learning process. We have adopted this premise for the purposes of this discussion. Therefore only degree-0 sentences are considered in this analysis.

two languages

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

In previous sections, we have considered several variants of both the CF and NB models for two languages in competition. Recall that when the learning algorithm is the TLA, for large k, the qualitative behavior of the two models is similar. In particular, 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. 9.4 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 have not been constructed.

2. The dynamical systems that we have derived and applied to this particular case hold only for the case of memoryless learning algorithms like the TLA. In previous chapters, we discussed a larger variety of learning algorithms and noted that their evolutionary consequences were potentially quite different.

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

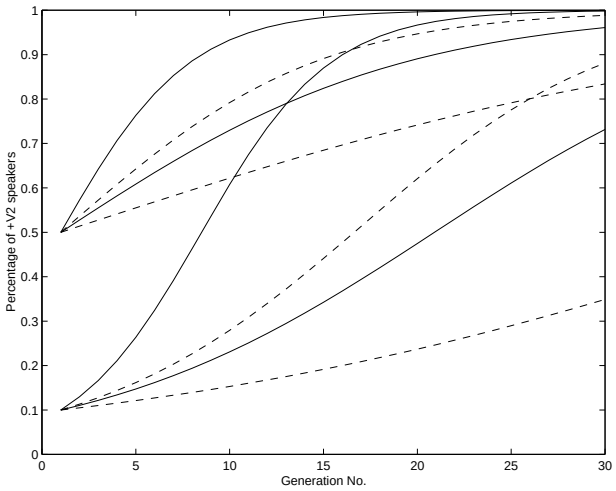

Figure 9.4: Trajectories of V2 growth. Shown in the figure are the evolving trajectories of s_t = proportion of +V2 grammars in the population over successive generations. The solid curves denote the evolutionary trajectories under the NB model; the dotted curves denote the trajectories under the CF 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 in this regime, the NB model has a faster rate of change than the CF model.

the same settings for all other parameters — settings that are the 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.

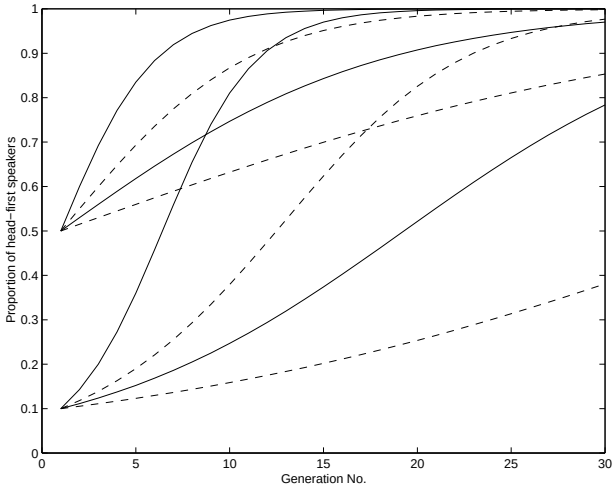

Figure 9.5: Trajectories of head-first growth. Shown in the figure are the evolving trajectories of s_t = proportion of head-first grammars in the population over successive generations. The solid curves denote the evolutionary trajectories under the NB model; the dotted curves denote the trajectories 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.

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. Therefore ambiguous sentences (those parsable with both +V2 and -V2 constraints) constitute a smaller proportion of the total sentence types of such grammars leading to a directional asymmetry in values of a and b in the model framework.

In conclusion, however, it is worthwhile to reiterate again our motivation in working through this particular example of 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 competition (ii) the number of parametric changes that occurred and their description in the context of a grammatical theory (iii) the nature of the learning mechanism that children employ in learning grammars (e.g. monolingual versus bilingual acquisition) and so on. Each of 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.

9.5 A Generalized NB Model for Neighborhood effects

The basic model for vertical transmission of cultural (linguistic) traits by Cavalli-Sforza and Feldman (1981) proceeds by dividing children 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 "neighborhood" effects in linguistic communities.

The key idea here is that in heterogeneous language communities speakers often tend to cluster in linguistically homogeneous neighborhoods. Consequently children growing up in the community might receive data drawn from different distributions depending upon their spatial location within the community at large. 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 neighborhoods. Children receive sentences drawn from different distributions depending upon their location in this neighborhood. 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 neighborhood and all children located in $[s_t, 1]$ are in the L_2 speaking neighborhood. Let us now define the mapping from neighborhood 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. 9.6 are some plausible mappings h that mediate the relation between location of the child in the neighborhood 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.

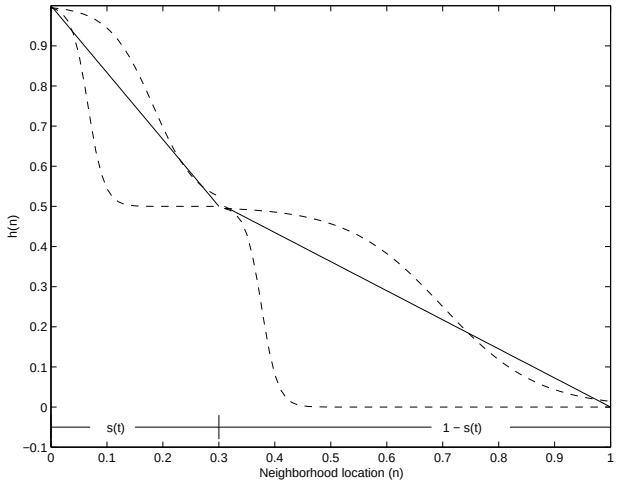

Figure 9.6: Examples of h mappings between the location n and the α type of the children occupying that location. Here the value of s_t (proportion of L_1 speakers) is taken to be 0.3 for illustrative purposes. Therefore the interval $[0, 0.3]$ is the L_1 speaking neighborhood; the interval $[0.3, 1]$ is the L_2 speaking neighborhood. For any location n the value of $h(n)$ represents the proportion of L_1 speakers the child occupying that location is exposed to.

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. 9.5:

$$s_{t+1} = \int_0^1 P_h(\alpha) f(\alpha) d\alpha \tag{9.5}$$

9.5.1 A Specific Choice of Neighborhood 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. 9.6; the solid line of Fig. 9.6):

$$h(0) = 1; h(1) = 0; h(s_t) = \frac{1}{2};$$
$$h(n) = 1 - \frac{1}{2s_t} n \text{ for } n < s_t; h(n) = \frac{1-n}{2(1-s_t)} \text{ for } n > s_t \tag{9.6}$$

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:

$$P_h(\alpha) = 2s_t \text{ if } \alpha > \frac{1}{2}; P_h(\alpha) = 2(1-s_t) \text{ if } \alpha < \frac{1}{2}; P_h(\alpha) = 0 \text{ if } \alpha \notin [0,1]. \tag{9.7}$$

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 some specific choices of learning algorithms. From Eq. 9.4, we see that it is a polynomial of degree k. Putting this into Eq. 9.5, we get the update rule with neighborhood effects to be

$$s_{t+1} = 2(1-s_t)(\int_0^{1/2} f(\alpha) d\alpha) + 2s_t(\int_{1/2}^1 f(\alpha) d\alpha) \tag{9.8}$$

Since α is a dummy variable in the above integral, the effect of the neighborhood 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 CF 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 completely. 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 $a \neq b$.

We see that the particular form of the update rule obtained with such neighborhood effects depends upon the functional form of the 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 neighborhood 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 neighborhoods in *every* generation. Of course, children in a particular neighborhood might acquire different languages. It is implicitly assumed that on maturation, the children (now adults) re-organize themselves into homogeneous neighborhoods. It is this re-organization into homogeneous neighborhoods that prevents the elimination of any one language from the system.

Another (more complete) way to characterize neighborhood 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 neighborhood. It is this function that evolves from generation to generation. Without additional simplifying assumptions, this treatment requires techniques well beyond the scope of this chapter and will be subject of future work. A preliminary account of this approach is provided in the next chapter.

9.6 A Note on Oblique Transmission

A substantial part of Cavalli-Sforza and Feldman (1981) is devoted to the study of oblique and horizontal transmission of cultural traits. In the basic vertical model that we have discussed so far, cultural traits are vertically transmitted from parents to children with varying probabilities. The evolution of such traits is then studied.

In oblique transmission, one considers the effect that members of the parental generation at large have on the transmission of cultural traits. Assume there are two traits L and H as before. Then one models the development of the cultural trait in the individual child in two steps.

1. Stage 1: Children acquire on the basis of preliminary exposure to their parents, a "juvenile" state which is one of L and H. This process is similar to vertical transmission and one may thus characterize the probabilities with which L or H may be attained. These probabilities will depend upon the cultural traits of the parents in the manner that has been discussed previously.

2. Stage 2: Juveniles acquire a "mature" state on the basis of exposure to the rest of the adult population. In this stage, one computes the probability with which trait transitions occur. Thus one characterizes the probability with which a juvenile of type L remains a type L upon maturation and similarly, the probability with which a juvenile of type H remains so after maturation. These trait transition probabilities will now depend upon the frequency of these traits in the entire adult population.

Consider stage 2. What is the probability with which a type L child might change to a type H one? This will surely depend upon the proportion of type H adults in the parental generation. Let $P[L \to H] = \phi_1(\alpha)$ where α is the proportion of type H adults. Similarly, let $P[H \to L] = \phi_2(\alpha)$. If the functional forms, ϕ_1 and ϕ_2 are known, then it is straightforward to compute the dynamics of H types in the population. Indeed, Cavalli-Sforza

and Feldman (1981) consider some simple choices for ϕ_1 and ϕ_2 and conduct an extensive analysis of the dynamics that results.

In general, however, it is clear from the discussion in this and preceding chapters that in the context of language change and evolution, ϕ_1 and ϕ_2 need to be derived from learning- theoretic considerations. Furthermore, ϕ_1 and ϕ_2 are unlikely to be simple in general. Much of the current book may be viewed as a contribution towards the characterization of such quantities for a variety of learning algorithms. For example, if the child learner were using the TLA, the following observations may be made.

1. The behavior of a TLA based learner is characterized by a Markov chain with transition matrix given by $T(\alpha)$ (following the analysis of Chapter 3).

2. Assume in the maturation phase, the learner receives l examples. Then, the probability of a $L_1 \rightarrow L_2$ transition after l examples is simply given by the $(1,2)$ element of $T^l(\alpha)$. Similarly, the probability of a $L_2 \rightarrow L_1$ transition after l examples is given by the $(2,1)$ element of $T^l(\alpha)$. In this manner ϕ_1 and ϕ_2 are easily obtained.

In general, given an NB style analysis, one may then be able to compute the two-stage model outlined above. We do not pursue this in any more detail here.

9.7 Conclusions

In this chapter, 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 change and evolution. Language acquisition serves as the mechanism by which language is transmitted 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 kinds of 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 neighborhoods and have investigated the consequences of such neighborhood 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 chapter, how one might apply the computational framework developed here to test the plausibility of differing accounts.

Chapter 10

Variations and Case Studies

In Chapters 5 through 9, we have developed a variety of models of language change motivated by different potential applications to historically observed real life cases. Over the course of this development, it has become gradually clear that such computational models allow us to bring into sharper focus the issues involved in studying and explaining linguistic diversity and language change.

In order to make initial progress, we have made simplifying assumptions. These assumptions permit a point of entry into the subject and result in valuable first order insights. While some variations of the basic assumptions have been studied in preceding chapters, it is worthwhile to reflect on other important issues that have been inadequately dealt with previously. In this chapter, we pause to consider some of these issues and discuss the nature of the modeling involved in resolving them in some sort of coherent fashion. Each of these represents a direction that is worthy of systematic study in its own right. We hope the preliminary analyses provided below are illustrative.

10.1 Finite Populations

In the analyses of preceding chapters, we have typically assumed that population sizes were infinite. This allowed us to derive deterministic dynamical systems that described the evolution of the population. In general, however, populations are finite and this gives rise to stochastic dynamics as a result.

For large population sizes, the deterministic dynamics and stochastic dynamics have similar behavior, but for small population sizes the two might differ from each other quite substantially. To see this, let us reconsider some

of the two language models of Chapter 5.

10.1.1 Finite Populations

Suppose the space $\mathcal{L}$ of possible languages consists of exactly two languages L_1 and L_2. Suppose learning children follow an algorithm $\mathcal{A}$ to infer a language from the linguistic data they receive over their learning phase. Then following the argument of Chapter 5, one derives the corresponding deterministic population dynamics as

$$\alpha_{t+1} = f(\alpha_t, \mathcal{A}, k)$$

where α_t is the proportion of L_1 users in the tth generation. Recall that f is obtained by computing the probability with which a typical child (following $\mathcal{A}$) acquires L_1 on receiving k linguistic examples. A variety of such maps were derived for different choices of $\mathcal{A}$ and k.

Consider now the situation in which there are a finite number, N, of adults. Each adult in generation t is characterized by a variable $X_i(t)$ (for the ith adult). $X_i(t) \in \{0, 1\}$ takes the value 1 if the ith adult speaks L_1 and the value 0 if he/she speaks L_2. Thus the *linguistic configuration* of the population is denoted by the vector $(X_1(t), X_2(t), \ldots, X_N(t))^T$. The average linguistic behavior of the population is denoted by

$$Y(t) = \frac{1}{N} \sum_{i=1}^{N} X_i(t)$$

$Y(t)$ is the fraction of the population that uses L_1. Note that $Y(t) \in \{\frac{0}{N}, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N}{N}\}$. Further, for very large N, we have $Y(t) \approx E[X_i(t)] = \alpha_t$, i.e., the fraction of L_1 users is close to what it would have been if the population size were infinite.

10.1.2 Stochastic Dynamics

Let us now characterize the evolution of $Y(t)$ over time. Consider the next generation of N children who are exposed to linguistic data from the previous generation.

A typical child of the next generation receives data drawn at random from the previous generation which contains a proportion $Y(t)$ of L_1 users. What is the probability with which such a child will attain L_1? This is simply given by $f(Y(t), \mathcal{A}, k)$. Thus for the ith child we see that $X_i(t+1)$ is

a random variable that takes the value 1 with probability $f(Y(t), \mathcal{A}, k)$ and the value 0 with probability $1 - f(Y(t), \mathcal{A}, k)$.

The average linguistic behavior is correspondingly given by

$$Y(t+1) = \frac{1}{N} \sum_{i=1}^{N} X_i(t+1)$$

Thus $Y(t+1)$ is a random variable that takes values in $\{\frac{0}{N}, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N}{N}\}$. Further, the probability distribution of $Y(t+1)$ is determined entirely by the value of $Y(t)$. The evolution of $Y(t)$ can be characterized by a Markov chain with $N+1$ states. Each state is identified with an element of $\{\frac{0}{N}, \ldots, \frac{N}{N}\}$. The transition matrix of this chain is given by $(1 \leq i, j \leq N+1)$

$$T_{ij} = \mathbb{P}[Y(t+1) = \frac{j-1}{N} \mid Y(t) = \frac{i-1}{N}] = \binom{N}{j-1} f_{i-1}^{j-1} (1 - f_{i-1})^{N-(j-1)} \tag{10.1}$$

where

$$f_i = f(\frac{i}{N}, \mathcal{A}, k)$$

Thus we have stochastic evolution of the average linguistic behavior from generation to generation: $Y(t) \to Y(t+1)$. It is now possible to study the effect of N on this evolutionary behavior

10.1.3 Evolutionary Behavior as a function of N

First, let us recall that for infinite N, we have for any t

$$Y(t) = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} X_i(t) = \alpha_t = E[X_i(t)] = f(Y(t-1), \mathcal{A}, k)$$

Now let us consider the evolution for finite N.

Large N

For large N, one may invoke the central limit theorem and conclude that $Y(t)$ is normally distributed for it is the average of a large number of i.i.d. variables. The mean and variance of $Y(t)$ are given by $\mu = E[X_i(t)]$ and $\sigma^2 = \frac{1}{N}\mu(1-\mu)$ respectively.

Thus, we have

$$Y(t) \approx f(Y(t-1)) + \eta(f(Y(t)), N)$$

where η is a normally distributed random variable with zero mean and variance given by $\frac{1}{N}f(Y(t))(1-f(Y(t)))$. Since $f(Y(t)) \in [0,1]$, we have the variance of η upper bounded by $\frac{1}{4N}$. For notational convenience, we have dropped the explicit dependence of f on $\mathcal{A}$ and k.

In other words, the evolution of average linguistic behavior is a noisy version of the deterministic dynamics where the variance of the noise is bounded by $\frac{1}{4N}$. Clearly, the larger N is, the more closely the dynamics will resemble the behavior of the infinite system. To see this, let us just consider as an example, the dynamics of the TLA learner with $k = 2$. Recall that this was provided by the following quadratic map (Chapter 5)

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

where $A = \frac{1}{2}((1-b)^2 - (1-a)^2); B = b(1-b) + (1-a); C = \frac{b^2}{2}$. When $a = b$, the system has exponential change to a stable point of $\alpha^* = \frac{1}{2}$. When $a > b$ or $a < b$, the system has logistic growth to a fixed point given by a valid solution to the quadratic equation.

Shown in Fig. 10.1 are the evolutionary trajectories for $a = b$ and $a < b$ respectively. Note how the trajectories are a noisy version of the deterministic ones. The noisiness decreases with N so that for large N, the trajectories start resembling the original deterministic dynamics quite closely.

Small N

For small values of N, a normal approximation is no longer valid and one will have to understand the behavior of the Markov chain given by Eq. 10.1. Note that if the following is true,

$$\forall i, 0 < f_i < 1$$

the Markov chain is ergodic and has a stationary distribution given by a solution to

$$\mathbf{x}^T T = \mathbf{x}^T$$

Thus although the dynamics has no fixed points or stable attractors anymore, the stationary distribution characterizes how often the population is in the state $Y(t) = \frac{i}{N}$ (for each i) over time. $\mathbf{x}$ is an $N+1$ dimensional vector whose jth element is proportional (equal, if suitably scaled) to the probability that $Y(t) = \frac{j-1}{N}$ in the limit. Therefore, if the stationary distribution has a sharp peak at $\mathbf{x}(j)$, then it corresponds to $Y(t)$ taking the value $\frac{j-1}{N}$ most often.

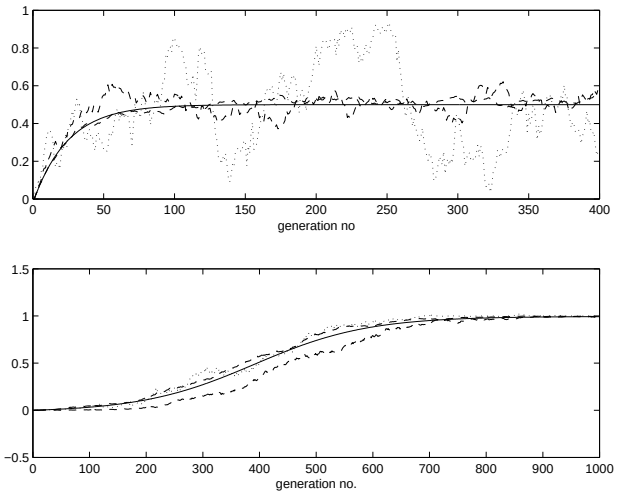

Figure 10.1: Evolutionary trajectories for the quadratic map. The top panel shows the case $a = b = 0.2$ when the deterministic system moves exponentially to a stable attractor of $\alpha^* = \frac{1}{2}$. This is shown by the solid curve. The superimposed dotted curves show how three random trajectories for finite $N(= 10000, 800, 100)$. The bottom panel shows the case $a = 0.01; b = 0.02$. Again, the solid curve is the infinite population trajectory and the dotted curves are three random trajectories for $N = 20000, 5000, 1000$ respectively. 1000 generations have been simulated. In both cases, the "noisiness" of the trajectories decrease with N.

Thus, we should expect the stationary distribution of the chain to have peaks at or close to the stable points of the original map f. Let us examine this for some choices of $\mathcal{A}$ and correspondingly f.

The Triggering Learning Algorithm

For finite k the map f has been derived for the TLA in Chapter 5. Let us consider the case $a = b$ as before. In general, for an arbitrary k, the map f is given by

$$\alpha_{t+1} = f(\alpha_t) = \alpha_t(1 - b^k) + \frac{b^k}{2}$$

This is a linear map and $\alpha_t \to \frac{1}{2}$. One can compute T for various choices of k and b. The numerical simulations reported below were conducted for $k = 9$ and $b = 0.7$. No particular significance should be attributed to the choice of these parameter values.

Fig. 10.2 shows the stationary distribution as a function of the size of the population N. Notice that for large N, the distribution is peaked around $\frac{1}{2}$ as expected. As N decreases the distribution starts flattening out. Note that for each choice of N, the stationary distribution is a probability distribution with support on the set $\{\frac{0}{N}, \frac{1}{N}, \ldots, \frac{N}{N}\}$. Since the support set is different for each N, the scale on the y-axis will be different for each of the distributions. However the spread can be inspected visually.

The true trajectory for each N is given by the evolution of the corresponding Markov chain. The initial condition was taken to be $x = 0$, i.e., all members of the population speaking L_2. From this initial condition, the deterministic system moves exponentially to $\frac{1}{2}$. Shown in Fig. 10.3 are trajectories for the deterministic and the Markov chain simulations respectively. Note how the population average moves to $\frac{1}{2}$ in a noisy manner with the degree of noise increasing as N becomes smaller. In each of the cases considered in figs 10.2 and 10.3 respectively, the stationary distribution is peaked at $\frac{1}{2}$ which means that eventually the population is mostly divided equally between L_1 and L_2 speakers. An interesting inversion occurs when N is very small. Consider, for example, the situation in Fig. 10.4 where the stationary distribution is plotted for $N = 25$ and $N = 23$ respectively. At $N = 24$, the distribution is still peaked at $\frac{1}{2}$. However, at $N = 23$, the distribution suddenly inverts and now has two peaks near 0 and 1 respectively. This means that *most of the time* the population consists of mostly L_1 speakers or mostly L_2 speakers but it flip-flops between these two situations. Thus, we see an inversion — for "large" N, the population moves to

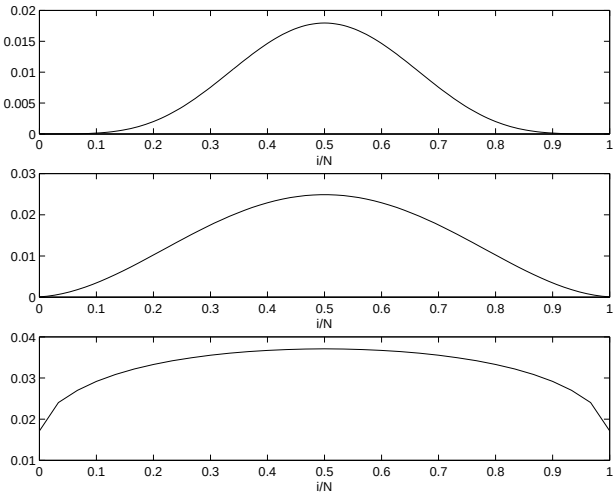

Figure 10.2: Stationary distributions for three choices of N. The peakedness of the distribution decreases from top panel ($N = 150$) through the middle panel ($N = 75$) to the bottom panel ($N = 30$).

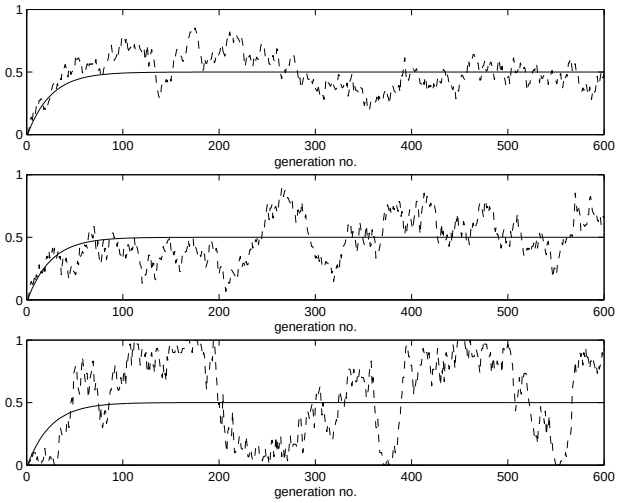

Figure 10.3: Trajectories for the random evolution of the population with the detailed Markov analysis for population sizes given by $N = 150$ (top panel), $N = 75$ (middle panel), and $N = 30$ (bottom panel) respectively.

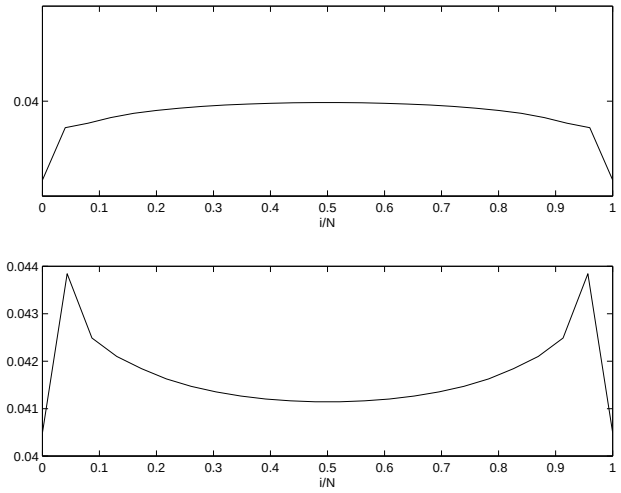

Figure 10.4: Stationary distributions for $N = 25$ (top panel) and $N = 23$ (bottom panel) respectively.

a heterogeneous linguistic state. For "small" N, the population at any point in time is almost always homogeneous. In this sense, the behavior at small N may be qualitatively different from the behavior at large N.

Cue Learner; finite k

Consider now the cue learner described in Chapter 5. In the two language setting, there is a probability p with which speakers of L_1 might produce cues for the next generation. The map f for the infinite population case was derived to be

$$f(\alpha_t) = \sum_{i=l}^{k} \binom{k}{i} (p\alpha_t)^i (1 - p\alpha_t)^{k-i}$$

It was seen that for a fixed k and l, as p varied, there was a bifurcation so that for large values of p there are two stable fixed points. One of these is $\alpha_1^* = 0$ and the other is $\alpha_2^* < 1$. In between is an unstable fixed point. Therefore, in this setting the infinite population may converge to either of the two stable points depending upon the initial conditions.

In sharp contrast, the finite population will *always* converge eventually to 0. To see this, let us first make some simple observations about the transition matrix of the Markov chain characterizing the behavior of the finite population. Note that

$$P(Y(t+1) = \frac{j}{N}|Y(t) = 0) = \binom{N}{j} f_0^j (1 - f_0)^{N-j}$$

But $f_0 = f(0, \mathcal{A}, K) = 0$. Hence we have $P(Y(t+1) = j|Y(t) = 0) = 0$ for every $j \neq 0$. Thus clearly 0 is an absorbing state. Since for every other $i > 0$, we have $0 < f_i < 1$, it is easily seen that (i) there are no other absorbing states (ii) every other state has a non zero probability of transiting to 0.

By the standard theory of Markov chains reviewed in Chapters 3 and 4, it is immediate that the chain will eventually settle to 0, i.e., the finite population will eventually converge to a situation where all members speak L_2. The other stable point of the infinite population analysis is never discovered.

Summary and Further Directions

Thus we see that the behavior of the finite population may be quite different from the infinite population in subtle ways. In general the following statements may be made

1. If there are no extremal fixed points, i.e., neither 0 nor 1 is a fixed point, then the behavior of the finite population is similar to the infinite one for large N and may differ for small N. This is exemplified by the example of the TLA studied in this section.

2. Extremal fixed points correspond to absorbing states of the finite population. When this occurs, no other stable attractors of the deterministic system are discovered in the finite population setting.

What we have presented so far is only the tip of the iceberg as far as understanding the effects of finite population sizes is concerned. We hope, however, to have provided some sense of the techniques involved in studying this issue in greater detail. The general question reduces to understanding the relation between the deterministic dynamical system (infinite population) and the corresponding stochastic process (finite population). If the deterministic system displays bifurcations or potentially chaos, it is quite unclear how the corresponding stochastic process would behave. The analysis in this section was restricted to the two language case where we obtained

some preliminary insights. For the more general case when $|\mathcal{L}| = n$, one obtains a Markov chain with n^N different states. An effective analysis of such chains presents considerable combinatorial challenges. Such questions are beyond the scope of this book and remain to be investigated systematically. The interested reader is referred to Komarova and Nowak (2003) for more detailed analysis for the case of finite populations in language. A large literature exists on genetic drift in finite populations (see Kimura (1983)) where similar questions are studied in the context of biological evolution.

10.2 Spatial Effects

In most of our models so far, we have assumed that all children receive their primary linguistic data from the same source distribution. This source distribution depends upon the mixture of different linguistic types in the parental population. In other words, we assumed a social connectivity pattern that is "perfectly mixed" in the sense that the different linguistic agents move freely in the society and influence each other equally.

Reality, as always, is more complicated. For example, the community may be spatially segregated into different "neighborhoods" with each such neighborhood having a different mix of linguistic types. Children born in these different neighborhoods are therefore exposed to different source distributions and end up having different linguistic experiences. This leads naturally to spatially distributed models and a brief discussion was conducted in Chapter 9.

Let us reconsider some of these spatial effects. We begin by representing the geographical extent of the population by the square interval $[0,1] \times [0,1]$. To keep matters simple, let us again work within the confines of the two-language setting where $\mathcal{L}$, the space of possible languages, consists of exactly two languages L_1 and L_2. Then the linguistic distribution of the population in space and time may be represented by a function

$$g_t(x,y) : [0,1] \times [0,1] \to [0,1]$$

Thus $g_t(x,y) \in [0,1]$ is the proportion of L_1 users at location (x,y) and time t.

10.2.1 Spatial Variation and Dialect Formation

Consider a typical child born at location (x,y). Assume that it is exposed to linguistic data from mature language users at that location. If it uses a

learning algorithm $\mathcal{A}$ and operates on the basis of k examples, then one might compute the probability with which such a child would acquire L_1. Following the usual analysis of previous chapters, we conclude that this probability is equal to $f(g_t(x,y), \mathcal{A}, k)$ and therefore

$$g_{t+1}(x,y) = f(g_t(x,y), \mathcal{A}, k)$$

In this setting, at each location (x,y), one has dynamics given by the deterministic map f. In previous chapters (particularly, Chapter 5 for two language models) we showed how to compute $f(\cdot, \mathcal{A}, k)$ for a variety of different learning algorithms. It is worthwhile to reflect on the implications of this spatial model.

Imagine that f has two stable fixed points $\alpha = 0$ and $\alpha = 1$ respectively. We encountered several maps having such a property. This suggests that populations would converge to a unique language (either L_1 or L_2) depending upon the initial conditions. Now imagine a spatially distributed population with the initial conditions provided by $g_0(x,y)$. At each location (x,y) the population would converge to a fixed language. However, the language that "emerges" would depend upon the initial condition. At all points (x,y) where the value $g_0(x,y)$ lies in the basin of attraction of $\alpha = 0$, we see that eventually L_2 would be spoken by everyone. Correspondingly, at all points (x,y) where the value $g_0(x,y)$ lies in the basin of attraction of $\alpha = 1$, the population would eventually converge on L_1. Thus linguistic differentiation would emerge in much the same way as species get formed. Local dialects would appear.

Figure 10.5 illustrates this. There are six panels in the figure. Each panel shows the distribution of L_1 speakers (across space) at a particular point in time. We assume that the learners use the cue based strategy of Chapter 5. If the cue frequency p is within a certain interval, there are two stable fixed points $\alpha = 0$ and $\alpha \approx 1$ with an unstable fixed point given by $\alpha = \alpha^*$. In our simulations, p was chosen so that $\alpha^* = 0.34$.

The top two panels denote the initial state of linguistic differentiation in the square region for two initial conditions that are only very slightly different from each other. At any point (x,y) in the region, the darkness is proportional to the percentage of L_2 speakers. The initial condition is mostly gray indicating that a mixture of both kinds of speakers exists in all regions. Further, in both cases, there is a region in the upper left hand corner where the mixture is slightly different from the rest of the region. In the upper left panel, the initial condition is such that the initial percentage of L_1 speakers is

0.33 and 0.35 respectively (in the upper left subregion and rest of the square). In the upper right panel, the initial percentage of L_1 speakers in these same regions is 0.31 and 0.33 respectively. Thus the upper left hand corner has a slightly lower initial proportion of L_1 speakers (higher proportion of L_2 speakers) than the rest. The two initial conditions corresponding to the upper two panels are numerically different but qualitatively the same. In fact, the two pictures are hardly distinguishable by eye.

Yet from such qualitatively similar initial conditions, as time unfolds, the spatial distribution of languages evolves in very different ways. In the panels on the left, we see that strong linguistic differentiation emerges so that the two different areas become linguistically homogeneous but with two different languages. Different regions with different languages have emerged. On the other hand, in the panels on the right, we see that in the long run, there is no difference between the linguistic composition of the two regions. Both regions are in the basin of attraction of the same language (L_2) and eventually this language is spoken throughout with no spatial segregation at all. No dialect formation emerges.

Thus the initial conditions determine which language comes to be spoken in a particular region. Very slight differences in initial conditions might lead to different languages being spoken and linguistic diversity might thus arise because of different initial conditions in different regions of the world.

10.2.2 A General Spatial Model

A more general account of spatial dynamics is provided by the following model. We develop this within the two-language framework as before. Let $X \subset \mathbb{R} \times \mathbb{R}$ be a spatial[1] region. At location $x \in X$ one may let $g_t(x) \in [0,1]$ denote the proportion of L_1 users in the population at time t. . Thus the spatial distribution of linguistic types at time t is given by the function $g_t : X \to \mathbb{R}$. We wish to characterize the evolution of g_t over time.

We denote spatial influences by the *influence function*

$$I : X \times X \to R$$

where $I(z,x)$ denotes the influence of speakers at location $x \in X$ on learners at location $z \in X$. For normalization purposes, we require that $\int_X I(z,x)dx =$

[1] More generally we may identify X with any parameters of social, ethnic, economic or other factors. For example each $x \in X$ may denote a particular socio-economic group and $g_t(x)$ then denotes the linguistic composition of that group. The influence function I characterizes the linguistic influence of one group on another.

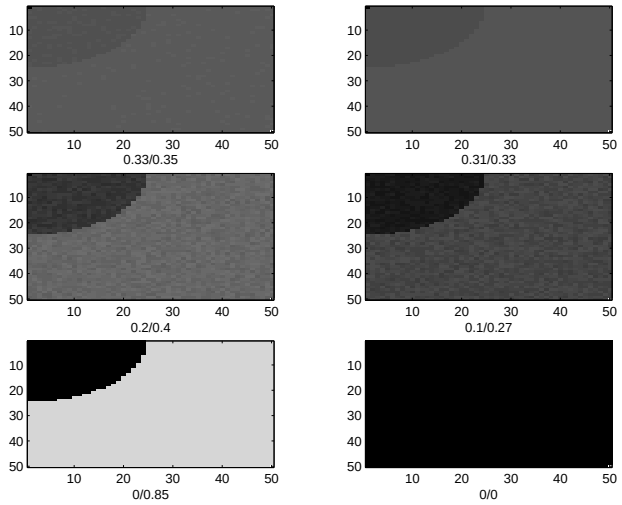

Figure 10.5: The evolution of linguistic diversity in space and time. Each panel represents the linguistic diversity in space at a point in time. The darkness of the (x, y) location is proportional to the percentage of L_2 speakers in the population. The three left panels (from top to bottom) represent snapshots of the linguistic diversity at three points in the evolutionary process from an initial condition (top), after 10 generations (middle) and after 20 generations (bottom). The three right panels (from top to bottom) represent snapshots of the linguistic diversity from a different initial condition (top), after 10 generations (middle) and after 20 generations (bottom). In each panel, there are two subregions — the upper left hand corner and the rest — with different proportions of L_2 speakers. These are indicated by the numbers x/y below each panel where x and y are the percentage of L_1 speakers in the two subregions respectively. Although the two initial conditions represented by top two panels are qualitatively the same, the numerical difference has put them in different basins of attraction. As a result the population evolves in different ways. In the left panels, regional linguistic differentiation occurs, in the right panels, it does not.

1. Thus for any fixed z, the quantity $I(z,x)$ may be interpreted as a probability density function denoting the likelihood of drawing a random example from a speaker at location x. If X were identified with social factors, then the influence function I characterizes the linguistic influence of one social group on another leading to a mathematization of socio-linguistic networks in the spirit of Milroy and Milroy (1985). More recently, there has been a profusion of interest in the structure and formation of networks in a variety of areas from biology, sociology, to the internet (see, for example, Strogatz, 2001; Newman et al, 2003 for popular and accessible treatments).

Consider a child learner born at location z. If this learner develops a language based on inputs drawn from speakers in its location, the evolution of language at that location would be given by $f(g_t(z))$ following the arguments of the previous section. However, if the child is exposed to data from different locations according to I, then one may define the intermediate function $h_t : X \to \mathbb{R}$ where

$$h_t(z) = \int_X I(z,x) g_t(x) dx$$

Here $h_t(z)$ denotes the overall probability with which a learner at z might encounter a speaker of L_1 at time t. Thus the evolution of language is characterized by by $f(h_t(z))$. Therefore, we get

$$g_{t+1}(z) = f(h_t(z)) = f \circ (L_I g_t)(z) \tag{10.2}$$

where $\circ$ denotes composition and L_I is a linear operator given by $L_I f = \int_X I(z,x) f(x) dx$.

The evolution of g_t may be studied for different choices of I and f. A proper investigation is beyond the scope of the current book but let us provide some preliminary insights. Let $f : [0,1] \to [0,1]$ denote a monotonic map whose iteration leads to dynamics with two stable attractors $\alpha_1 < \alpha_2$ and an unstable fixed point $\alpha^* \in (\alpha_1, \alpha_2)$. The basin of attraction of α_1 is given by $[0, \alpha^*)$ and that of α_2 is given by $(\alpha^*, 1]$. We have encountered such maps before. Then if spatial evolution of language is provided by Eq. 10.2, the following results are true.

Proposition 2 *If g_t is the constant function, it remains so for all time. The evolution of the constant is characterized by f.*

Proof: Let $g_t(z) = \alpha_t$. Then it is immediate that $h_t(z) = \int_X I(z,x) \alpha_t dx = \alpha_t$. Therefore $g_{t+1}(z) = f(\alpha_t) = \alpha_{t+1}$. Thus g_t is always a constant function. ■

In other words, proposition 2 states that if there is no spatial diversity to begin with, such diversity will not arise. Recall that the dialect formation model of the previous section required some initial diversity distributed around the boundary between two basins of attraction of two different attractors. This result is therefore consistent with the earlier results.

Proposition 3 *If at all locations z, we have $g_t(z) > \alpha^*$ then spatial diversity will be eliminated, i.e., $\forall z(g_t(z) \to \alpha_2)$.*

Proof: Let $\alpha_t = \inf_{z \in X} g_t(z)$. If X is compact, we know that $\alpha_t > \alpha^*$. We also know that for every z, $h_t(z) = \int_X I(z,x) g_t(x) dx \geq \int_X I(z,x) \alpha_t dx = \alpha_t$. Since f is monotone, we have $f(h_t(z)) \geq f(\alpha_t)$. Therefore $\alpha_{t+1} \geq f(\alpha_t)$. Iterating k times, we get $1 \geq \alpha_{t+k} \geq f^k(\alpha_t)$. Since $\lim_{k=\infty} f^k(\alpha_t) = \alpha_2$, we have that $\lim_{k=\infty} \alpha_{t+k} \geq \alpha_2$.

Now let $\beta_t = \sup_{z \in X} g_t(z)$. We see that $h_t(z) = \int_X I(z,x) g_t(x) \leq \int_X I(z,x) \beta_t dx \leq \beta_t$. Therefore $f(h_t(z)) \leq f(\beta_t)$ from which we conclude $\beta_{t+1} \leq f(\beta_t)$. Iterating k times, we get $\lim_{k=\infty} \beta_{t+k} \leq \lim_{k=\infty} f^k(\beta_t) = \alpha_2$. Since $\alpha_t \leq \beta_t$, the result follows. ■

Even if there is diversity to begin with, if all the initial conditions lie in the basin of attraction of one attractor, then spatial diversity will eventually be eliminated as the linguistic composition moves towards that attractor at all locations. If all initial conditions are not within the same basin of attraction, then more complicated behavior may arise. In the dialect formation model of the previous section, we assumed that $I(z,x) = \delta(z-x)$ where δ is the Dirac delta function. In other words different regions were linguistically isolated from each other. This led to each region developing its own language depending upon the initial conditions of that region. For other kinds of spatial interactions, other evolutionary possibilities emerge.

Proposition 4 *Even though f has no cycles, the spatial evolution could display oscillations because of spatial interactions.*

Proof: We construct an example of cyclic behavior in the population. Let X be divided into two disjoint regions X_1 and X_2 such that $X_1 \cup X_2 = X$. Let $g_t(x) = \frac{1}{2} + \epsilon$ for all $x \in X_1$ and $g_t(x) = \frac{1}{2} - \epsilon$ for all $x \in X_2$. We will choose ϵ appropriately later.

Now let I be as follows. For all $z \in X_1$, we have $\int_{X_2} I(z,x) dx = \beta > \frac{1}{2}$. Similarly, for all $z \in X_2$, we have $\int_{X_1} I(z,x) dx = 1 - \beta$. Let us characterize the evolution of spatial diversity for this choice of I.

We see that for $z \in X_1$, we have $h_t(z) = \int_X I(z,x)g_t(x) = \beta(\frac{1}{2} - \epsilon) + (1-\beta)(\frac{1}{2} + \epsilon) = \frac{1}{2} + (1-2\beta)\epsilon$. Therefore, we get

$$g_{t+1}(z) = f(\frac{1}{2} - (2\beta - 1)\epsilon)$$

Similarly, for $z \in X_2$, we have $g_{t+1}(z) = f(\frac{1}{2} + (2\beta - 1)\epsilon)$.

Consider $f : [0,1] \to [0,1]$ to be monotone increasing with $f(0) = 0; f(\frac{1}{2}) = \frac{1}{2}; f(1) = 1; \forall 0 \leq \epsilon \leq \frac{1}{2}, f(\frac{1}{2} + \epsilon) + f(\frac{1}{2} - \epsilon) = 1$ Further let the dynamics arising from the iteration of f be such that there are two stable attractors (0 and 1) and one unstable fixed point ($\frac{1}{2}$) in the middle. Then f has no cycles.

For such an f we have $f'(\frac{1}{2}) > 1$. Therefore, it is possible to find ϵ^* and β such that

$$f(\frac{1}{2} - (2\beta - 1)\epsilon^*) = \frac{1}{2} - \epsilon^*$$

For this choice of ϵ^* we see that g_t cycles between $\frac{1}{2} + \epsilon^*$ and $\frac{1}{2} - \epsilon^*$ in a period 2 cycle at all locations. ■

In the construction of the previous proposition there are essentially two regions X_1 and X_2 where each region influences the other more than itself. This leads to a cycling behavior as each region follows the other. By having n regions, one can get period n cycles. One may eventually get chaos though we do not explore that possibility here.

In summary, we see that the spatial distribution of language and the interactions between different regions may have considerable effect on the dynamics depending upon both the initial conditions and the nature of the influence function I. It is worth noting in conclusion that the formulation is general enough to cover many different kinds of variation. Thus X need not represent spatial diversity but may be used to model ethnic, social, economic, or other forms of diversity that may influence the language learning process at any point.

10.3 Multilingual Learners

In studying the relationship between learning and evolution, we have always assumed that the language acquisition process leads the learner to a single grammar. This is acquired on the basis of the primary linguistic data that it is exposed to over the learning period. Formally, the learning algorithm is treated as a map

$$\mathcal{A} : \text{ Data } \to \mathcal{H}$$

where $\mathcal{H}$ is a class of possible natural language grammars. Much of the interesting and nontrivial population dynamics arises when the primary linguistic data comes from a mixture of sources, e.g., from two different grammatical systems that are present in the adult generation.

An interesting and important issue is whether and when the learning child actually becomes bilingual (or more generally multilingual). In this case, the child does not acquire a single grammatical system but instead acquires multiple distinct ones. For example, if there are two linguistic systems (variants) L_1 and L_2 in the adult population, the learning child might actually acquire both. We consider this possibility in this section.

The issue of acquiring multiple linguistic systems raises a host of subsidiary questions. For one, how might the learning child even know that there are two (or n) different linguistic systems in its linguistic environment? The literature on language acquisition is often imprecise on this question. One approach to multilingual acquisition is to claim that the child categorizes individuals into linguistic groups and figures out (somehow) that there are n different systems to acquire. Thereafter, acquisition proceeds in parallel as if there were n targets and the linguistic data is streamed based on the individual generating it. For example, a child might conclude from the social dynamics that the grandparents speak one language, the parents another, and the neighbors a third different language. The child then proceeds to learn three different languages treating the data differently depending upon who produces it. This is usually the case when the lexical items used by the different language users are quite distinct. The languages are thus kept apart and multilingual acquisition reduces to three parallel monolingual acquisition systems. The previously studied monolingual models apply immediately. This is classical multilingualism.

However, the case of greater interest is when the linguistic systems interact. This is especially so if they share the same or similar lexical items but have different grammatical rules. In this setting, it is as if the child neither acquires a single system, nor multiple separate systems but rather a "mixture" of competing systems that it uses simultaneously. For example, in their insightful study of language contact and change, Kroch et al (1997,2000) argue strongly that over the middle English period, there were southern and northern dialects that came into contact and speakers had multiple conflicting systems that they used simultaneously. The most natural model for this version of multilingual behavior is to treat linguistic knowledge as a distribution over competing grammatical systems. Thus the learning algorithm

is a map

$$\mathcal{A}: \text{ Data } \rightarrow \{\mu \mid \mu \text{ is a probability measure on } \mathcal{H}\}.$$

If $\mathcal{H}$ consists of exactly n elements, i.e., there are n linguistic types, then probability measures on $\mathcal{H}$ can be identified with elements of the $n-1$-dimensional simplex $\Delta^{(n-1)}$. Thus any measure on $\mathcal{H}$ is given by $(\alpha_1, \ldots, \alpha_n)$ where $\sum_{i=1}^{n} \alpha_i = 1$ and $\alpha_i \geq 0$. Therefore the learning algorithm may be characterized as

$$\mathcal{A}: \text{ Data } \rightarrow \Delta^{(n-1)}.$$

We consider this for the case $n = 2$ in what follows. It is hoped that the analysis is illuminating and generalizations to the general n case can be easily imagined.

10.3.1 Bilingualism Modeled as a Lambda Factor

Following the treatment of Kroch et al (1997), let us consider an explicit model of bilingualism. Individuals are now bilingual with two internal grammars (say g_1 and g_2) that they alternate between in production and comprehension. A parameter $\lambda \in [0, 1]$ determines the frequency of usage between the grammars. Thus, for example, a speaker/hearer with $\lambda = 0.2$ would use g_1 20 percent of the time and use g_2 80 percent of the time.

The Bilingual Individual and the Population

Imagine a population of bilingual adults with the ith adult using the two grammars in the ratio λ_i. Each adult therefore has his/her characteristic λ and one can characterize the variation in the adult population by a distribution over values of λ. Accordingly, let $P(\lambda)$ denote a probability density function having support in $[0, 1]$ that characterizes the distribution of λ values in the adult population.

Consider a particular adult having a value λ. The probability with which this adult produces an arbitrary sentence $s \in \Sigma^*$ is given by

$$P_\lambda(s) = \lambda P_1(s) + (1 - \lambda) P_2(s) \tag{10.3}$$

where $P_1(s)$ is the probability with which s is generated while using grammar g_1 and $P_2(s)$ is the probability with which s is generated while using grammar g_2.

Eq. 10.3 characterizes the probability with which sentences are produced (and consequently received by listeners/children) by a single adult speaker. However, in reality the true source is not a single adult speaker but the entire population of adults with adults having λ values given by the distribution $P(\lambda)$. Although it seems like this might complicate matters considerably, the analysis turns out to be quite simple.

To see this, consider the probability with which an arbitrary sentence $s \in \Sigma^*$ is produced. This is given by

$$P(s) = \int_0^1 P_\lambda(s) P(\lambda) d\lambda \tag{10.4}$$

Putting in the value of $P_\lambda(s)$ from Eq. 10.3, into Eq. 10.4, we have

$$P(s) = P_1(s) \int_0^1 \lambda P(\lambda) d\lambda + P_2(s) \int_0^1 (1-\lambda) P(\lambda) d\lambda$$

Clearly,

$$P(s) = E[\lambda] P_1(s) + (1 - E[\lambda]) P_2(s)$$

where $E[\lambda] = \int \lambda P(\lambda) d\lambda$ is the mean value of λ in the population and clearly lies between 0 and 1. Thus having a mixed source with distribution $P(\lambda)$ is *equivalent* to having a single source switching between the two grammars with $E[\lambda]$. Therefore, from the point of view of the child, the variation (of λ's) in the adult population can be replaced by a single adult whose λ-value is given by the population mean. This will considerably simplify matters in the analysis of change.

The Child as a Lambda Learner

Children are exposed to sentences and from this exposure, they acquire (estimate) a value of λ that they then employ as adult users of language to switch between the two grammars in sentence production. According to this point of view, and in contrast to other models of language acquisition considered in this context, child learners are *not* explicitly monolingual. They do not attain a single unique grammar but rather acquire both. This reflects the fact that a single grammar is not adequate to account for the conflicting data they receive but rather two grammars are attained and they are in competition with the parameter λ deciding the rate at which the two grammars are employed. Crucially, this parameter λ is *estimated* by children on the basis of the primary linguistic data they receive.

There is also variation in the population of children. Let us assume that all children use the *same* algorithm (procedure; denoted by $\mathcal{A}$) to determine a suitable value of λ. Let us also assume that each child in the community receives its primary linguistic data from the bilingual adult population that has been characterized in the earlier section. However, each child receives a *different* random draw of sentences from the *same* mixed source of adults. Because each child receives a potentially different data set, it attains a different value of λ if the primary linguistic data is finite. Thus, one has a population of children who receive different sets of sentences and therefore attain different values of λ leading to a variation in λ-values in the population of children.

Under these assumptions, one might now attempt to characterize the distribution of λ-values in the population of children. By doing so, one will have determined the relationship between the linguistic composition of children and that of adults, i.e., the relationship between the linguistic compositions of two successive generations. The precise nature of such a relationship is determined by the λ-estimation procedure the canonical child employs and characterizes the evolutionary consequences (over generational time scales) of λ-estimation (over developmental time scales).

We now outline three candidate λ-estimation procedures a child could potentially employ. In the next section, we will determine their evolutionary consequences.

Imagine a child receives n random sentences from the adult population. Sentences can be of three types:

1. sentences that are analyzable by g_1 but not by g_2, i.e., those that belong to $L_1 \setminus L_2$.

2. sentences that are analyzable by both g_1 and g_2, i.e., those that belong to $L_1 \cap L_2$.

3. sentences that are analyzable by g_2 but not by g_1, i.e., those that belong to $L_2 \setminus L_1$.

Of the n sentences the child receives, let n_1 be of type 1, n_2 be of type 2, and n_3 be of type 3. Clearly, $n = n_1 + n_2 + n_3$.

Now, three candidate algorithms for λ-estimation the child might utilize are:

1. $\mathcal{A}_1 : \hat{\lambda} = \frac{n_1}{n_1+n_3}$. Thus the child ignores all ambiguous sentences and uses only unambiguous triggers in its count.

2. $\mathcal{A}_2 : \hat{\lambda} = \frac{n_1}{n_1+n_2+n_3}$. The child interprets ambiguous sentences (type 2) as having been produced by g_2.

3. $\mathcal{A}_3 : \hat{\lambda} = \frac{n_1+n_2}{n_1+n_2+n_3}$. The child interprets ambiguous sentences (type 3) as having been produced by g_1.

Two remarks are in order. Algorithms $\mathcal{A}_2$ and $\mathcal{A}_3$ are mirror images in that g_2 and g_1 respectively are given preferential treatment in the analysis of ambiguous sentences. If g_2 and g_1 represent first and second language grammars respectively, this implies a tendency of the child to analyze ambiguous sentences exclusively using first or second language respectively. Alternatively, one may consider g_2 or g_1 (respectively) to be a default or *marked* state according to which ambiguous sentences are interpreted. Various interpretations may thus be given to the asymmetry of estimation.

Evolutionary Consequences

Let u_t be the mean value of λ's in the population of adults, i.e, $u_t = E[\lambda]$. Similarly, let u_{t+1} be the mean value of λ's in the population of children (after maturation; therefore the population of the next generation of adults at time $t+1$). It is possible to derive the functional relationship between u_{t+1} and u_t under the assumptions described in earlier sections and we provide and analyze such relationships below.

First, some preliminaries. Recall that 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 theirs. Further, let

$$a = P_1[L_1 \cap L_2]$$

and

$$b = P_2[L_1 \cap L_2]$$

Thus a, b are weights that P_1 and P_2 respectively put on the ambiguous sentences that are analyzable under both grammars. It turns out that these are the only parameters of P_1 and P_2 that enter the functional relationship between u_{t+1} and u_t for the three λ-estimation procedures described above.

Theorem 23 *Let u_t be the mean value of λ in generation t and u_{t+1} be the mean value of λ in generation $t+1$. Assume that individual learners*

estimate λ according to each of the three algorithms described earlier. Then u_t and u_{t+1} are related by the following for each of the three procedures

$$\mathcal{A}_1 : u_{t+1} = \frac{u_t(1-a)}{u_t(1-a) + (1-u_t)(1-b)}$$

$$\mathcal{A}_2 : u_{t+1} = u_t(1-a)$$

$$\mathcal{A}_3 : u_{t+1} = u_t(1-b) + b$$

Proof: Consider $\mathcal{A}_2$. For this setting,

$$u_{t+1} = E[\frac{n_1}{n_1+n_2+n_3}] = \frac{E[n_1]}{n} = u_t(1-a)$$

Next consider $\mathcal{A}_3$. For this setting,

$$u_{t+1} = E[\frac{n_1+n_2}{n_1+n_2+n_3}] = \frac{E[n_1]+E[n_2]}{n} = u_t(1-a) + u_t a + (1-u_t)b$$

Finally, consider $\mathcal{A}_1$.

$$u_{t+1} = E[\frac{n_1}{n_1+n_3}] = \sum \binom{n}{n_1 n_2 n_3} \alpha^{n_1}\beta^{n_2}\gamma^{n_3}\frac{n_1}{n_1+n_3}$$

where $\alpha = (1-a)u_t; \beta = au_t + b(1-u_t); \gamma = (1-b)(1-u_t)$ respectively. Now

$$\sum \binom{n}{n_1 n_2 n_3} \alpha^{n_1}\beta^{n_2}\gamma^{n_3}\frac{n_1}{n_1+n_3} = \sum_{k=0}^{n}\sum_{n_1=0}^{k}\binom{k}{n_1}\binom{n}{k}\alpha^{n_1}\beta^{n-k}\gamma^{k-n_1}\frac{n_1}{k}$$

$$= \sum_{k=0}^{n}\binom{n}{k}\beta^{n-k}(1-\beta)^k\sum_{n_1=0}^{k}\binom{k}{n_1}(\frac{\alpha}{1-\beta})^{n_1}(\frac{\gamma}{1-\beta})^{k-n_1}\frac{n_1}{k}$$

$$= \sum_{k=0}^{n}\binom{n}{k}\beta^{n-k}(1-\beta)^k(\frac{\alpha}{1-\beta}) = \frac{\alpha}{1-\beta} = \frac{u_t(1-a)}{u_t(1-a)+(1-u_t)(1-b)}$$

■

Let us examine the consequences of this result.

1. The three different λ-estimation techniques lead to three different dynamical systems. Each dynamical system characterizes the evolution of the mean λ value of the population from generation to generation. Crucially, nowhere have we *assumed* a particular form for the evolution of this mean λ-value. Instead, we have postulated different mechanisms by which children might estimate a value of λ from the data they receive. By taking ensemble averages over the entire population of children, we have then been able to *derive* the evolution of mean-λ as a logical consequence. The three different λ-estimation procedures essentially differ on how they treat ambiguous sentences, i.e., those that are parsed under both grammatical hypotheses in question. This results in the different evolutionary properties. One can now compare these predicted evolutionary trajectories to the ones that have been actually observed in history.

2. The three dynamical systems have different evolutionary properties. The evolution under λ-estimation method $\mathcal{A}_1$, is characterized by equations that are familiar to us from Chapter 5. Recall that such a system has the following behavior: (a) if $a = b$, then $u_{t+1} = u_t$, i.e, there is no change over generational time; (b) if $a > b$, then $u_t \longrightarrow 0$ over generational time; (c) if $a < b$, then $u_t \longrightarrow 1$ over generational time.

3. Under λ-estimation method $\mathcal{A}_2$, it is easy to check that $u_t \longrightarrow 0$ (for all $a > 0$).

4. Under λ-estimation method $\mathcal{A}_3$, it is similarly easy to check that $u_t \longrightarrow 1$ (for all $b > 0$).

10.3.2 Further Remarks

Memoryless Bilingual Learner

The learning algorithm considered in the previous sections operates in a batch mode where it needs to keep a count of the different kinds of sentences (whether $s \in L_1 \setminus L_2$ etc.) it has received. A different algorithm that is more faithful to the memoryless online principles discussed in Chapter 3 was developed and considered by Yang (2000). A brief exposition follows:

We consider the case where $|\mathcal{H}| = 2$. Thus there are two grammars g_1 and g_2 in $\mathcal{H}$ characterizing two different languages L_1 and L_2 respectively. Before seeing any examples the learner has a uniform prior on $\mathcal{H}$. This may be characterized by the number $p_1(0) = p_2(0) = \frac{1}{2}$. Thus $p_1(0)$ denotes the

"weight" for g_1 and $p_2(0) = 1 - p_1(0)$ denotes the weight for g_2. The learner receives an example stream $s_1, s_2, \ldots$, one at a time. After the nth example has been received, let the distribution be characterized by $(p_1(n), p_2(n))$ where $p_1(n) + p_2(n) = 1$. A new sentence s_{n+1} is now heard by the learner. Then

1. with probability $p_i(n)$ the learner picks g_i.

2. **if** (s_{n+1} understood by g_i),
   $\mathrm{p}_i(n+1) = \quad (1\text{-}\gamma)p_i(n) + \gamma$
   $\mathrm{p}_j(n+1) = \quad (1\text{-}\gamma)p_j(n); j \neq i$
   else
   $\mathrm{p}_i(n+1) = \quad (1\text{-}\gamma)p_i(n)$
   $\mathrm{p}_j(n+1) = \quad (1\text{-}\gamma)p_j(n) + \gamma; j \neq i$

In this algorithm p_1 plays the role of λ in the previous discussion. Instead of estimating λ using frequency counts, the algorithm estimates it using a classical linear reward penalty scheme (Bush and Mosteller, 1955). Suppose maturation occurs after k sentences have been received. Then, the variable of interest is $p_1(k)$. This characterizes the probability with which the mature learner will use the grammatical system g_1 in future interactions. One might say that the learner estimates λ to be $\hat{\lambda} = p_1(k)$.

In the population setting, different children will receive different random draws of sentences and arrive at different estimates of λ. There is variation in λ values of the children. As before, it is of interest to to characterize the distribution and in particular the average value of $\hat{\lambda} = p_1(k)$ in the population of child learners. We show how to do this.

Note that in general $p_1(n+1)$ is related to $p_1(n)$ in the following way:

$$p_1(n+1) = (1-\gamma)p_1(n) + \gamma X_n$$

where X_n is a random variable taking values in $\{0, 1\}$. Denote the probability of $X_n = 1$ by A_n. Therefore, by taking expectations we have

$$E[p_1(n+1)|p_1(n)] = (1-\gamma)p_1(n) + \gamma A_n \tag{10.5}$$

Now let us compute A_n. As usual, we assume that u_t is the mean value of the $\lambda's$ in the adult population (generation t). Therefore the following probabilities are immediate. The probability that a random sentence belongs to $L_1 \setminus L_2$ is equal to $u_t(1-a)$. The probability that a random sentence belongs to L_1 is equal to $u_t + (1-u_t)b$. In order to compute A_n, note that $X_n = 1$ when either of two disjoint events occur:

1. The learner chooses g_1 (with probability $p_1(n)$) and the sentence $s_{n+1} \in L_1$. This entire event occurs with probability

$$(u_t + (1 - u_t)b)\, p_1(n)$$

2. The learner chooses g_2 (with probability $(1 - p_1(n))$) and the sentence $s_{n+1} \in L_1 \setminus L_2$. This entire event occurs with probability

$$(u_t(1 - a))\, (1 - p_1(n))$$

Thus we have

$$A_n = (u_t + (1 - u_t)b)\, p_1(n) + (u_t(1 - a))\, (1 - p_1(n))$$

Putting this into Eq. 10.5 we have

$$E[p_1(n + 1)|p_1(n)] = \alpha p_1(n) + \beta$$

where

$$\alpha = 1 - \gamma + \gamma(au_t + b(1 - u_t))$$

and

$$\beta = u_t(1 - a)\gamma$$

Taking expectations with respect to $p_1(n)$ we get the following recurrence relation

$$E[p_1(n + 1)] = \alpha E[p_1(n)] + \beta$$

Since $E[p_1(0)] = \frac{1}{2}$, we have

$$u_{t+1} = E[p_1(k)] = \frac{1}{2}\alpha^k + \frac{1 - \alpha^k}{1 - \alpha}\beta$$

For the limiting case of $k = \infty$, we have

$$u_{t+1} = \frac{\beta}{1 - \alpha} = \frac{u_t(1 - a)}{(1 - a)u_t + (1 - b)(1 - u_t)}$$

Note that this is *exactly* the same update rule as obtained in the previous section.

Final Remarks

While the bilingual learners were elaborated for the two language setting, natural generalizations to the n language case exist. For example, for the λ-estimation procedures discussed earlier, one might consider algorithms that count *cues*. Elements of $L_i \setminus L_j$ count as cues and one might count occurrences of members of each of these sets in the primary linguistic data and combine them usefully to obtain a total estimate $\hat{\lambda} = (\lambda_1, \ldots, \lambda_n) \in \Delta^{(n-1)}$. Many such estimates are possible and we do not explore their evolutionary consequences here. Similarly, the linear reward scheme of Yang (2000) has a natural extension to the n language case. This general extension is considered in Yang (2000) and we direct the interested reader to that exposition.

On a concluding note, it is worthwhile to reflect on the difference between bilingual and monolingual learning when batch learning algorithms are used.

Consider first the monolingual setting. The learning algorithm chooses L_1 if $n_1 > n_3$. Recall that n_1 is the number of sentences belonging to $L_1 \setminus L_2$ that occur in the data set and n_3 is the corresponding number belonging to $L_2 \setminus L_1$. The update rule for this case is given by

$$\alpha_{t+1} = \sum_{n_1 > n_3} \binom{n}{n_1 n_2 n_3} p_1(t)^{n_1} p_2(t)^{n_2} p_3(t)^{n_3}$$

where $p_1(t) = \alpha_t(1-a); p_3(t) = (1-\alpha_t)(1-b); p_2(t) = 1 - p_1(t) - p_3(t)$. Each individual in the population uses exactly *one* language and α_t is the proportion of them that use L_1 in generation t. The dynamics of this map has *two* stable equilibria, $\alpha^* \approx 0$ and $\alpha^* \approx 1$. Thus, if a community is homogeneous with most members speaking L_2, it is hard to imagine how one might move to one consisting of mostly L_1 speakers. The population will have to jump from one basin of attraction into another. This may only come about because of large scale language contact or migratory effects.

In contrast, consider the bifurcations that occur with the bilingual learner. Each individual in the population uses *both* languages in the ratio $\lambda_i(t)$ (for the ith individual in the tth generation). One is interested in the evolution of the mean $u_t = E[\lambda_i(t)]$ (where the expectation is taken with respect to individuals.) The update rule is provided by

$$u_{t+1} = \frac{u_t(1-a)}{u_t(1-a) + (1-u_t)(1-b)}$$

In this setting, there is only *one* stable equilibrium although both $\alpha^* = 0$ and $\alpha^* = 1$ are fixed points. Which of these is stable depends upon the

relationship of a and b with respect to each other. Thus a switch in usage frequencies from $a > b$ to $b > a$ would cause a switch in the stability of the population. It is easy to imagine how language change might come about.

It is interesting to note, however, that with *memoryless* algorithms the update rule (whether the monolingual TLA or the scheme outlined earlier (Yang, 2000)) gives rise to similar bifurcation diagrams where language change is easier to imagine.

10.3.3 A Bilingual Model for French

Let us briefly consider the application of this point of view to re-analyze the case of syntactic change in French during the period from the fourteenth century to the seventeenth century (A.D.). Our first computational analysis of this case was conducted in Chapter 6 within a 5 parameter monolingual framework of Clark and Roberts, 1993. Here we redo the analysis within a bilingual framework. Our analysis continues to draw heavily from the work of Ian Roberts (linguistic work in Roberts, 1993 and computational work in Clark and Roberts, 1993) and is influenced by a more recent treatment in Yang, 2002.

Linguistic Background

Let us recall the syntactic setting. The discussion that follows is conducted within the Principles and Parameters tradition (Chomsky, 1981) of linguistic theory. There were two dominant parametric changes that occurred in French syntax over the period under consideration. First, there was loss of subject (pro) drop. In Old French, (like modern Italian), a pronominal subject could be dropped as the following examples show.

Loss of null subjects

1. * Ainsi s'amusaient bien cette nuit. (ModF)
   thus (they) had fun that night

2. Si firent grant joie la nuit. (OF)
   thus (they) made great joy the night

Second, there was loss of verb-second phenomena (V2). Old French was a V2 language so that V could raise to C (with the specifier typically filled) and occupy therefore the second position in the linear order of the constituents. This is no longer true as the following examples show.

Loss of V2

1. * Puis entendirect-ils un coup de tonnerre. (ModF)
   then heard-they a clap of thunder

2. Lors oirent ils venir un escoiz de tonoire (OF)
   then heard they come a clap of thunder

Thus the situation is simply summarized as follows. In the beginning there was a relatively stable and homogeneous grammatical system that was +V2 and had null subjects (pro drop). At the end, there was again a relatively stable and homogeneous grammatical system that had lost both V2 and pro drop. In the middle there was variation with multiple grammatical variants co-existing in the population.

Computational Analysis

We make the following assumptions.

1. Each speaker is potentially bilingual/multilingual with multiple grammatical systems that provide the basis for linguistic use.

2. Similarly, each child potentially acquires multiple grammatical systems based on its linguistic experience. In particular, in periods when there is linguistic variation in the adult population and the data received is not consistent with a single grammar, the child will accordingly acquire multiple systems.

For illustrative purposes, we will focus on the competition between two grammatical systems. The two grammars are denoted by g_+ and g_- respectively. The corresponding sets of surface expressions (sentences) are denoted by L_+ and L_- respectively. When using the grammatical system g_+ speakers produce sentences with a probability distribution P_+ (over L_+) and similarly when using g_- speakers produce sentences with a probability distribution P_- (over L_-).

For example, if g_+ were a head-first grammar without verb second movement and no prodrop (like modern French), then L_+ consists of elements like (a) **SVO** (subject-verb-object; like the modern English *Mary sees the children* or the modern French *Marie voit les enfants*) (b) **XSVO** (like the English *After dinner, John read the newspaper.*) and so on. In general, in our analysis, various choices may be made for g_+ and g_- and the evolutionary consequences may then be examined.

Since speaker/learners are potentially bilingual, each speaker has a *grammatical mix factor* $\lambda \in [0,1]$ that characterizes how often the speaker uses g_+ as opposed to g_-. In general, a speaker with mix factor λ produces sentences with a probability distribution given by $\lambda P_+ + (1-\lambda)P_-$. Note that this distribution is over the set $L_+ \cup L_-$. Thus there may be *internal* variation within each speaker and the expressions produced by such a speaker are not consistent with a single unique grammar. One may view the space of possible grammatical systems as $\mathcal{G} = \{h | h = \lambda g_1 + (1-\lambda) g_2\}$ where $\mathcal{G}$ is a space of formal convex combinations denoting multiple grammatical systems.

As noted in earlier sections, there is also *external* variation in the adult population. Thus different individuals have potentially different λ values and one can therefore imagine the distribution of λ values in the adult population. A summary statistic for the average linguistic behavior of the population as a whole may be provided by the mean value of λ which we denote by $E[\lambda]$. One is usually interested in the evolution of this quantity over generational time.

Bifurcations and Syntactic Change

We have discussed evolutionary dynamics for a variety of learning algorithms. Two algorithms are of particular interest. One is algorithm $\mathcal{A}_1$ of Section 10.3.1. The other is the memoryless λ-estimation procedure of Section 10.3.2. Both are algorithms by which the individual child might estimate a λ value on the basis of linguistic data. Recall that both learning algorithms resulted in the same evolutionary dynamics for $E[\lambda]$ over time. This dynamical system is given by

$$x_{t+1} = \frac{a x_t}{a x_t + b(1-x_t)} \tag{10.6}$$

where

1. $x_t = E[\lambda_t]$ is the average value of λ in the parental generation (generation t).
2. $a = \sum_{s \in L_+ \setminus L_-} P_+(s)$ is the probability with which a speaker when using g_+ produces a trigger for that grammatical system.
3. $b = \sum_{s \in L_- \setminus L_+} P_-(s)$ is the probability with which a speaker when using g_- produces a trigger for that grammatical system.

Recall the dynamical behavior of Eq. 10.6.

1. If $a > b$ then $x = 1$ is the only stable point. From all initial conditions, the population will converge over evolutionary time to a homogeneous population of g_+ users.

2. If $a < b$ then $x = 0$ is the only stable point. From all initial conditions, the population converges to g_-.

3. If $a = b$ then $x_{t+1} = x_t$ for all t. There is no change.

We may interpret the facts of language change in terms of this bifurcation. Thus, on this account, one would suggest that a homogeneous stable population of g_+ users ($x = 1$) could become unstable if the frequencies of sentences changed so that a became less than b while before it was the other way around. Under this condition, we see that the introduction of even the slightest variation in the population would cause the language of the community to move to one of g_- users, i.e., large scale language change as a result of a bifurcation. It is also interesting to note that while syntactic diglossia is permitted within the grammatical and acquisition framework, it is usually eliminated over time unless a is exactly equal to b in such models.

Looking more closely at the grammatical theories and the data, we find that *if there was no pro-drop*, a +V2 grammar tends to be quite stable in conmparison to a -V2 grammar if this is the only parametric difference between the two grammars. Following the analysis in Roberts (1993) and Yang (2000), we may take the two grammars to be:

1. g_+: the +V2 grammar has expressions like **SVO** (subject-verb-object; with verb typically in C and subject in spec-C) and **VS** patterns like **XVSO**, **OVS** and so on.

2. g_-: the -V2 grammar (like Modern French) has expressions like (a) **SVO** (subject-verb-object; with subject in spec-IP) (b) **XSVO** (in general, V> 2 patterns).

Following our analysis above, we see that **SVO** patterns do not count as triggers. The proportion of **XSVO** (trigger for -V2) and **XVSO** (trigger for +V2) patterns in the speech of g_- and g_+ users respectively will determine the evolution of the population. Preliminary statistics (following Yang, 2000) based on the speech of modern $-$V2 (like English and French) and +V2 (like German and Dutch) languages suggest that $a = 0.3$ while $b = 0.2$. Consequently, the +V2 grammar would remain stable.

Let us now consider a +V2 (with pro drop) grammar in competition with a -V2 (with prodrop) grammar. Then we have the following patterns[2]: (Note that +V2 grammars with pro drop will not generate **VO** expressions presumably because subj is in spec-CP and this needs to be filled.)

1. g_+: **SVO; XVSO; XVO**
2. g_-: **SVO; VO; XVO; XSVO**

Let us do a simple calculation. Assume that with probability p the subject is a pronoun rather than a full NP. Further, with probability d, the pronoun is dropped in a pro-drop (null subject) language. If $d = 1$ then the prodrop is obligatory. If $d = 0$ then the language does not allow prodrop. Then we see that

$$\begin{aligned} P_+(SVO) &= 0.7 \\ P_+(XVSO) &= 0.3((1-p) + p(1-d)) \\ P_+(XVO) &= 0.3pd \end{aligned}$$

To clarify the logic of the calculations, let us consider the probability with which **XVSO** would be produced by a speaker of g_+. With probability 0.3 a baseform of **XVSO** would be produced. Now we need to calculate the probability with which this is overtly expressed, i.e., the subject is not dropped. There are two cases: (i) the subject position is filled by a full NP (with probability $1-p$) in which case it cannot be dropped (ii) the subject position is filled with a pronoun (probability p) but this pronoun is

[2]These patterns are provided for illustrative purposes. In reality, of course, there are many more expressions that are generated by each of the two grammars but these may be deemed irrelevant to the discussion at hand. Consequently, the probabilities provided may be treated as normalized after discarding these irrelevant distributions. More precisely, we are assuming the following. L_+ and L_- each contain a potentially infinite number of expressions. We restrict our discussion to a set $A \subset \Sigma^*$ of expressions where $A = \{\mathbf{SVO}, \mathbf{XVSO}, \mathbf{XVO}, \mathbf{VO}, \mathbf{XSVO}\}$. Then for any element $a \in A$, when we put in values for $P_+(a)$ in our calculations, we actually use the value $\frac{P_+(a)}{P_+(A \cap L_+)}$ and similarly for P_-. There are two other potentially important considerations that we have eliminated from our current discussion. First, we are restricting ourselves to verb-medial grammars. It has been proposed that +V2 systems tend to be more stable in verb-final grammatical systems than in verb-medial ones. We do not explore this issue in part because of our supposition that French was verb-medial throughout. Second, there was a point when subject pronouns started behaving as clitics and it is quite possible that this behavior affected the interpretation of surface expression during the language learning phase and altered the relevant primary linguistic data to weaken V2. We do not consider this issue any further.

not dropped (probability $1-d$). Multiplying this out and adding the two cases, we obtain $P_+(XVSO) = 0.3((1-p)+p(1-d))$. Probability values for the other expressions are obtained similarly.

Now consider probabilities with which g_- speakers produce their expressions. It is simply seen that

$$\begin{aligned} P_-(SVO) &= 0.8(1-pd) \\ P_-(VO) &= 0.8pd \\ P_-(XSVO) &= 0.2(1-pd) \\ P_-(XVO) &= 0.2pd \end{aligned}$$

Given P_- and P_+, we can calculate a and b to be

$$a = 0.3(1-pd)$$

and

$$b = 0.8pd + 0.2(1-pd)$$

From this we see that for $a > b$ we need

$$0.3(1-pd) > 0.8pd + 0.2(1-pd)$$

or

$$pd < \frac{1}{9}$$

Thus we see that if $d = 0$, i.e., the language has no prodrop, then the dynamics is in the regime $a > b$ and correspondingly, the +V2 grammar is stable as our first analysis showed already. On the other hand, if $d > \frac{1}{9p}$ then a bifurcation occurs and the +V2 grammar becomes unstable. One might then ask, how come a +V2 and pro drop grammar (as old French putatively was) remained stable in the first place? According to this analysis it must be because p was small (so that the product $pd < \frac{1}{9}$). Now notice that if this were the state of affairs, then the only way in which change would come about is if p increased to cross the threshold. While this is happening it is crucial that the null subject is not being lost, i.e., d is not decreasing. By this analysis +V2 is lost before the null subject is lost. If the null subject were lost first, then $d = 0$ and the dynamics would always be in the regime $pd < \frac{1}{9}$ and the +V2 parameter would always remain stable. On this account, +V2 is lost before the null subject was lost. Further, +V2 was lost because of the increase in p, i.e., the use of pronominal subjects in the speech of the times.

The above analysis is empirically anecdotal as we have plugged in plausible numbers for the probabilities. The point of the exercise was to show again how a bifurcation may be at the root of language change and how the conditions for change depend in a subtle way on the frequencies with which expressions are produced. In this case, the product pd is seen to determine the stability of the language. Further, we obtain a linguistic prediction, that +V2 must have been lost before the null subject parameter was lost. The loss of +V2 must have been triggered by the increase in the use of pronominal subjects.

10.4 Conclusions

We have considered three separate issues that arise in the accurate modeling of the evolution of linguistic populations.

First, there is the question of the finiteness of population sizes. In general, assuming an infinite population size allows one to derive deterministic dynamics by taking suitable ensemble averages over the population. When populations are finite, the contingencies of individual decisions do not converge to their theoretical mean behavior. As a result, the population dynamics is now stochastic. For large or medium population sizes, this stochastic behavior is like a noisy version of the original deterministic dynamics. For small population sizes, the stochastic behavior could be systematically different.

Second, we considered the effects of spatial organization of the population into linguistic neighborhoods. The most interesting result from such a consideration is a proposal about dialect formation. Since there may be multiple attractors for the population dynamics, the actual stable language that emerges in a given region will depend upon the initial conditions of the population in that region. Assuming that humans are distributed with different initial conditions in different regions of the world, one might then see the formation of different dialects with each dialect corresponding to a attractor of the linguistic system.

Third, we examined the population dynamics for multilingual learners. In a two language setting, we considered two different bilingual learning procedures and derived the corresponding dynamical equations. In those linguistic settings where multilingual behavior is important to take into account, models of this sort will need to be developed. A case study of syntactic change in French was then conducted within this paradigm. From this anal-

ysis, we concluded that a bifurcation could have lead to the loss of +V2. We conjecture that +V2 must have been lost before null subject was lost and further that it must have been the increase in the use of pronominal subjects in the speech of the times that triggered the loss of +V2. These predictions are empirically verifiable.

Each of these investigations is incomplete but suggestive. It shows how a relevant aspect of language evolution might be systematically modeled and the kinds of results one might potentially obtain. We leave a more extensive coverage of these issues to future research.