Part III

Language Change
Chapter 5

Language Change - A Preliminary Model

For the last forty years, synchronic linguistics has been driven by the so called “logical problem of language acquisition” – the problem of how children come to acquire the language of their parents. In part II (Chapters 2, 3, and 4) we discussed in some detail the inherent difficulty of inferring a language on the basis of finite exposure to it. Within the analytic framework of our discussion, we concluded that successful language learning would be possible only in the presence of adequate prior constraints on the class of possible grammars \( \mathcal{H} \) and correspondingly on the class of possible learning algorithms \( \mathcal{A} \).

Linguists in the generative tradition have proposed theories of universal grammar that constrain the range of grammatical hypotheses that children might entertain during language learning. Psycholinguists, developmental psychologists and cognitive scientists have pursued a range of approaches from empirical studies of child language acquisition to formal analyses of the process (Pinker, 1984; Wexler and Culicover, 1980; Crain and Thornton, 1998; Slobin, 1985-97; Gleitman and Landau, 1994; Newport and Aslin (2000); ). These shed light on the possible mechanisms or algorithms of language learning.

Language acquisition may be viewed as the mechanism by which language is transmitted from parent to child — and in fact, from one generation of language users to the next. Perfect language acquisition would imply perfect transmission. Children would acquire perfectly the language of their parents, language would be mirrored perfectly in successive generations and
languages would not change with time. Yet, in Chapter 1, we saw a number of documented cases of such change. Therefore, for languages to change with time, children must do something differently from their parents.

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

In a discussion of the history of English, Lightfoot (1991) clearly points to such a possibility.

As somebody adopts a new parameter setting, say a new verb-object order, the output of that person's grammar often differs from that of other people's. This in turn affects the linguistic environment, which may then be more likely to trigger the new parameter setting in younger people. Thus a chain reaction may be created. (Lightfoot, 1991, p. 162)

In fact, linguists have long been occupied with describing phonological, syntactic, and semantic change, often appealing to an analogy between language change and evolution, but rarely going beyond this. For instance, Lightfoot (1991, chapter 7, pp. 163-65ff.) talks about language change in this way: “Some general properties of language change are shared by other dynamic systems in the natural world ... In population biology and linguistic change there is constant flux ... If one views a language as a totality, as historians often do, one sees a dynamic system”. Indeed, entire books have been devoted to the description of language change using the terminology of population biology: genetic drift, clines, and so forth\(^1\). Other scientists have explicitly made an appeal to dynamical systems in this context; see especially Hawkins and Gell-Mann, 1992. It is only over the last decade, however, that this connection has begun to be seriously and formally explored.

In this chapter, we make explicit the link between learning and evolutionary dynamics. In particular, we show formally that a model of language change emerges as a logical consequence of language acquisition, an argument made informally by Lightfoot (1991). We shall see that Lightfoot’s intuitive

\(^1\) For example, see Nichols (1992), and more recently Mufwene (2001).
5.1 An Acquisition-Based Model of Language Change

How does the combination of a grammatical theory and learning algorithm lead to a model of language change? We begin our treatment by arguing that the problem of language acquisition at the *individual* level leads logically to the problem of language change at the group or *population* level. Consider a population speaking a particular language\(^2\). This is the target language—children are exposed to primary linguistic data (PLD) from this source, typically in the form of sentences uttered by caretakers (adults). The logical problem of language acquisition calls for an explanation for how children acquire this target language from their primary linguistic data — in other words, to come up with an adequate learning theory. Following the development of the previous chapters, we take a learning theory to be simply a mapping from primary linguistic data to the class of phrase structure grammars (computational systems), usually an effective procedure, and so an algorithm. For example, in a typical inductive inference model, given a stream of sentences, an acquisition algorithm would update its grammatical hypothesis with each new sentence according to some computable process. We encountered various formal criteria for learnability all of which require that the algorithm’s output hypothesis converge to the target in some sense as more and more data become available.

Now suppose that we fix an adequate grammatical theory and an adequate acquisition algorithm. There are then essentially two means by which the linguistic composition of the population could change over time. First, if the primary linguistic data presented to the child is altered (due to any number of causes, perhaps to presence of foreign speakers, contact with another population, disfluencies, and the like), the sentences presented to the learner (child) are no longer consistent with a single target grammar. In the face of this input, the learning algorithm might no longer converge to the target grammar. Indeed, it might converge to some other grammar \(g_2\); or it might converge to \(g_2\) with some probability, \(g_3\) with some other probability\(^3\), and

\(^2\)In our analysis this implies that all the adult members of this population have internalized the same grammar (corresponding to the language they speak).

\(^3\)Recall from the previous chapters that convergence is measurable. Therefore it makes sense to talk of the probability of acquiring a particular grammar under any source
so forth. In either case, children attempting to solve the acquisition problem using the same learning algorithm could internalize grammars different from the parental (target) grammar. In this way, in one generation the linguistic composition of the population can change.⁴

Second, even if the PLD comes from a single target grammar, the actual data presented to the learner is truncated, or finite. After a finite sample sequence, children may, with non-zero probability, hypothesize a grammar different from that of their parents. This can again lead to a differing linguistic composition in succeeding generations.

In short, the diachronic model is this: Individual children learn a language based on linguistic input generated from the grammar of their caretaker (the target grammar). After a finite number of examples, some acquire a grammar very similar to that of their caretakers, but others may have acquired something different. The next generation will therefore no longer be linguistically homogeneous. The third generation of children will hear sentences produced by the second—a different distribution—and they, in turn, will attain a different set of grammars. Over successive generations, the linguistic composition evolves as a dynamical system. In this manner, language acquisition and language change become intimately related. Fig. 5.1 provides a pictorial perspective on this state of affairs.

Language acquisition takes a microscopic view of the situation. The object of study is the individual child and one studies how its linguistic (grammatical) hypotheses evolve from example sentence to sentence over its developmental lifetime. If one were to step away and take a macroscopic view of the same phenomena, one could make the object of one's study the entire community or population. If one studied how the linguistic composition of the population were to evolve from generation to generation over evolutionary time scales, then one would end up with models of language change. Such models of language change are driven by models of language learning. Thus every model of language learning at the individual level could have potentially different evolutionary consequences at the population level. This book explores such consequences in some detail.

On this view, language change is a logical consequence of specific as-

⁴Sociological factors affecting language change affect language acquisition in exactly the same way, yet are abstracted away from the formalization of the logical problem of language acquisition. In this same sense, we similarly abstract away such causes here, though they can be brought into the picture as variation in probability distributions and learning algorithms. We shall consider a few such variations in future chapters.
5.1. AN ACQUISITION-BASED MODEL OF LANGUAGE CHANGE

![Diagram showing a microscopic view of language acquisition and a macroscopic view of population and linguistic composition]

Figure 5.1: Language Acquisition takes a microscopic view of the situation. It focuses on the individual language learner and studies how its language (grammar) develops on exposure to primary linguistic data over its critical learning period. Language change and evolution takes a macroscopic view with a focus on the population over generational time scales. One now studies how the linguistic composition of the population evolves from generation to generation.
sumptions about:

1. the grammar hypothesis space—the space of possible grammars that humans might acquire. In the Principles and Parameters framework of modern linguistic theory, this reduces to the choice of a particular parameterization;

2. the language acquisition device—the learning algorithm the child uses to develop grammatical hypotheses on the basis of data;

3. the primary linguistic data—the distribution of sentences that a child is exposed to and that affect its linguistic development.

If we specify (1) through (3) for a particular generation, we should, in principle, be able to compute the linguistic composition for the next generation. In this manner, we can compute the evolving linguistic composition of the population from generation to generation; we arrive at a dynamical system. We now proceed to make this calculation precise. We first review a standard language acquisition framework, and then show how to derive a dynamical system from it. We begin with the simplest possible model—that of two languages in competition with each other.

### 5.2 A Preliminary Model

Our discussion is motivated by a syntactic view of the world where each language is viewed as a set of expressions that are well-formed according to the rules of some underlying grammar. Therefore, languages may be treated formally as subsets of $\Sigma^*$ where $\Sigma$ is a finite alphabet (denoting, for example, the lexical items). Imagine a world with only two possible languages $L_1$ and $L_2$ where each $L_i$ is a subset of $\Sigma^*$ in the usual way. In general, $L_1$ and $L_2$ are not disjoint. Sentences belonging to both $L_1$ and $L_2$ are ambiguous and may be parsed according to the underlying grammar ($g_1$ and $g_2$ respectively) of each language.

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

---

5The general methodology is applicable to phonology just as well. One may view a phonological grammar as defining a set of well formed phonological expressions. The set of such well formed expressions may be defined using a notational system that utilizes a phonological alphabet.
5.2. A PRELIMINARY MODEL

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

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

5.2.1 Learning By Individuals

We begin by examining the acquisition of language by individuals in the population. Language acquisition is the process of developing grammatical hypotheses on the basis of linguistic experience, i.e., exposure to linguistic data during childhood. Within the purview of generative linguistics, this is conceptually regarded as choosing an appropriate grammar from a class of potential grammars $G$ (Universal Grammar or UG) on the basis of primary linguistic data. In this example, we consider the case where there are only two potential grammars — $g_1$ and $g_2$ underlying the languages $L_1$ and $L_2$ respectively.

Consider now a learning procedure (algorithm) to choose a language based on linguistic examples. As we saw in previous chapters, this can, in general, be characterized as a mapping from linguistic data sets to the hypothesis\(^6\) set $\{L_1, L_2\}$. Following the notation developed previously, we let $D_n$ be the set of all potential finite data streams of exactly $n$ sentences each and $D = \bigcup_{i \geq 1} D_i$ to be the set of all finite length data sets. The learning algorithm $A$ is a computable mapping from $D$ to $\{L_1, L_2\}$.

\(^6\)Although we refer to the learner as choosing a language $L_1$ or $L_2$, it is worthwhile to clarify that languages are typically infinite sets of sentences that have finite representations in terms of their grammars. Therefore, learners with their finite brains and finite lifetimes, presumably choose grammars. In general, there may be many grammars that are weakly equivalent (generate the same language) but in our setting in this chapter, there is a one-to-one mapping between the class of grammars $G = \{g_1, g_2\}$ and the class of languages $\mathcal{L} = \{L_1, L_2\}$. Therefore, we identify $G$ with $\mathcal{L}$ and each may be viewed as the hypothesis set in our setting.
Now fix a probability distribution $P$ on $\Sigma^*$ according to which sentences are drawn independently at random and presented to the learner. After $k$ such examples are drawn, the learner’s data set can be denoted by $d_k = \{s_1, s_2, \ldots, s_k\}$ where each $s_i$ is drawn according to the distribution $P$. Clearly $d_k$ is an element of $D_k$. In this setting, it is possible to define the following object

$$p_k = \mathbb{P}[A(d_k) = L_1]$$

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

In this probabilistic setting, it is worthwhile to recall the natural notion of learnability which requires that the learner’s hypothesis must converge to the target as the data goes to infinity. This simply means that if the probability distribution $P$ had support on $L_1$ so that only sentences of $L_1$ occurred in the data sets of the learner (i.e., $L_1$ is the target language), then

$$\lim_{k \to \infty} p_k(A, P = P_1) = 1$$

Similarly, if the probability distribution $P$ had support on $L_2$ so that only sentences of $L_2$ were presented to the learner, then $\lim_{k \to \infty} p_k(A, P = P_2) = 0$. We will evaluate $p_k$ for several different choices of $A$ shortly.

### 5.2.2 Population Dynamics

The previous section discussed how the grammatical hypothesis of the individual learner develops over a series of linguistic examples during a critical learning period (i.e., until linguistic maturation time). The central question in that context is whether or not the learner’s hypothesis gets closer and closer to the target and eventually converges to it as more and more data become available. Of course, convergence to the target only occurs as the number of data goes to infinity — and learners live only finite lives. As a matter of fact, learners do not endlessly update their hypotheses but “mature” after a point and live with their mature hypothesis thereafter. Let us assume that maturation occurs after $K$ examples have been presented to the learner. This assumption is consistent with evidence from developmental psychology suggesting that there is a critical age effect in language learning.
5.2. A PRELIMINARY MODEL

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

We need not have started with a homogeneous adult population. Imagine now that the state of the adult population is denoted by $\alpha_a$ where $\alpha_a$ is the proportion of $L_1$ users in the adult population. Now consider the generation of children. They will receive example sentences from the entire adult population — which in this case is a mixed population. In particular, they will receive examples drawn according to the distribution

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

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

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

In this manner, we see that $\alpha_c$ can be expressed in terms of $\alpha_a$.

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

$$\alpha_{i+1} = p_K(A, \alpha_i P_1 + (1 - \alpha_i) P_2) \quad (5.1)$$

In order to do this, we assumed
1. The population could be isolated into coincident generations.

2. Children receive data drawn from the entire adult population in a manner that reflects the distribution of languages in the adult population.

3. The probability of drawing sentences $P_1$ and $P_2$ do not change with time.

4. The learning algorithm $\mathcal{A}$ constructs a single hypothesis language (grammar) after each example and after maturation ends up with a single language (grammar).

5. Population sizes are infinite.

We will return to a discussion of these assumptions later. Let us now consider some examples where we make specific choices regarding the learning algorithm and derive the evolutionary consequences. In particular, the functional relationship between $\alpha_t$ and $\alpha_{t+1}$ (Eq. 5.1) will be explicitly derived for a number of different algorithms.

### 5.2.3 Some Examples

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

$\mathcal{A}$: Memoryless Learners

A memoryless learner — described in Chapter 4 — is one whose hypothesis at every stage depends only upon the current input sentence and the previous hypothesis it had. There are a wide class of such algorithms and one in particular has received considerable attention in the linguistic parameter setting literature. This is the triggering learning algorithm (TLA) of Gibson and Wexler, 1994 and was extensively analyzed in Chapters 3 and 4. While the algorithm works for any finite parameter space in general, the particular instantiation for the two language case is as follows:

**TLA (Triggering Learning Algorithm)**

- [Initialize] Step 1. Start with an initial hypothesis (either $L_1$ or $L_2$) chosen uniformly at random.
5.2. A PRELIMINARY MODEL

- [Process input sentence] Step 2. Receive a positive example sentence $s_i$ at the $i$th time step.

- [Learnability on error detection] Step 3. If the current grammatical hypothesis parses (generates) $s_i$, then go to Step 2 to receive next example sentence; otherwise, continue.

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

The population at any point in time can be characterized by a single variable ($\alpha_t$ for the $t$th generation) that describes the proportion of $L_1$ users in the population. Now one may ask the question: *If children were TLA learners, then how would the population evolve?*

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

$$a = P_1[L_1 \cap L_2]; \quad 1 - a = P_1[L_1 \setminus L_2]$$
and similarly

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

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

**Theorem 20** The linguistic composition in the $(t+1)$th generation ($\alpha_{t+1}$) is related to the linguistic composition of the $t$th generation ($\alpha_t$) in the following way:

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

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

**Proof** Let the adult population have $\alpha_t$ proportion of $L_1$ users. We need to compute the probability with which the learner acquires $L_1$ after 2 examples. First note that the probability with which a random example belongs to (a)
\( L_1 \setminus L_2 \) (b) \( L_1 \cap L_2 \) (c) \( L_2 \setminus L_1 \) is given by (i) \( \alpha_i(1-a) \) (ii) \( \alpha_i a + (1-\alpha_i)b \) (iii) \( (1-\alpha_i)(1-b) \) respectively. Now with probability \( \frac{1}{2} \), the learner chooses \( L_1 \) as its initial hypothesis. There are two different ways in which it could retain its hypothesis after two examples. These are (i) its hypothesis is \( L_1 \) after both the first and the second example (ii) its hypothesis is \( L_2 \) after the first and \( L_1 \) after the second. Case (i) will occur if both examples lie in \( L_1 \). This happens with probability \( (\alpha_i + (1-\alpha_i)b)^2 \). Case (ii) will occur if the first example is in \( L_2 \setminus L_1 \) and the second example is in \( L_1 \setminus L_2 \). This happens with probability \( ((1-\alpha_i)(1-b))(\alpha_i(1-a)) \).

Similarly, with probability \( \frac{1}{2} \), the learner begins with hypothesis \( L_2 \). There are again two different ways in which it could have a hypothesis of \( L_1 \) after two examples. These are given by (i) it flips its hypothesis to \( L_1 \) after the first example and retains \( L_1 \) after the second example. (ii) it retains \( L_2 \) after the first example but switches to \( L_1 \) after the second. The probability with which each of these cases occurs can be easily calculated in a similar manner and putting it all together after some algebra, the update rule is obtained.

A few remarks concerning this dynamical system are in order:

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

Remark 2. The scenario \( a \neq b \) is much more likely to occur in practice — consequently, we are more likely to see logistic change rather than exponential change. Note that logistic change will give rise to the S-shaped pattern that historical linguists have often observed in field studies of language change. See Fig. 5.2 for a graphical display.

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

**Theorem 21.** *Due to the fact that \( a, b \leq 1 \), the dynamical system never enters the chaotic regime.*

Remark 4. We obtain a class of dynamical systems. The quadratic nature of our map comes from the fact that \( K = 2 \). If we choose other values for \( K \) we would get cubic and higher order maps. In general, it is possible to show that for a fixed, finite \( K \), the evolutionary dynamics is given by
Figure 5.2: Evolution of linguistic populations whose speakers differ only in the V2 parameter setting. This reduces to a two language model differing by one linguistic parameter. Note the exponential growth when $a = b$. The different exponential curves are obtained by varying the value $a = b$. When $a$ is not equal to $b$, the system has a qualitatively different (logistic) growth. By varying the values of $a$ and $b$ we get the different logistic curves. It has been the empirical observation of many that language change undergoes an S-shaped trajectory. Here we show how such a trajectory may emerge as a result of the underlying dynamics of individual learners.
Theorem 22 If individual learners in a population of TLA learners have a maturation time $K$, the population evolves according to

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

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

Proof: Recall that the TLA learner may be exactly characterized as a Markov chain — in this case with two states — one (state 1) corresponding to $L_1$ and the other (state 2) to $L_2$. Let this transition matrix be $T$. Then it is easy to see that

$$T_{12} = (1 - \alpha_t)(1 - b) = A; T_{21} = \alpha_t(1 - a) = B$$

Since $T$ is a stochastic matrix, $T_{11} = 1 - T_{12}$ and $T_{22} = 1 - T_{21}$. Let us denote by $T^{(n)}$ the matrix of transitions after $n$ examples have been received. Thus $T^{(n)} = T^n$. Clearly the probability of acquiring $L_1$ after $n$ examples is given by $\frac{1}{2}(T_{11}^{(n)} + T_{21}^{(n)})$. Using the Chapman-Kolmogorov equation, we have

$$T_{11}^{(1)} = 1 - A$$

and

$$T_{11}^{(n)} = (1 - A)T_{11}^{(n-1)} + BT_{12}^{(n-1)}$$

$$= B + (1 - A - B)T_{11}^{(n-1)}$$

By induction, we have

$$T_{11}^{(n)} = \frac{B}{A + B} + \frac{A(1 - A - B)^n}{A + B}$$

Similarly, it is possible to solve for $T_{21}^{(n)}$ as

$$T_{21}^{(n)} = \frac{B}{A + B} - \frac{B(1 - A - B)^n}{A + B}$$

Putting these together, the final expression for the probability of acquiring $L_1$ after $K$ examples is found. Since population size is infinite, this is the proportion of $L_1$ speakers in the next generation and the update rule is thus derived.

The evolutionary modes in this setting may now be investigated. Although the map looks potentially very complex, the dynamics turns out to be surprisingly simple. For any fixed $K$, it is possible to show
5.2. A PRELIMINARY MODEL

1. The function \( f(\alpha) = \frac{B + B(A - B)(1 - A - B)^K}{A + B} \) may be re-expressed by expanding out the term \((1 - A - B)^K\) as \( \sum_{i=0}^{K} \binom{K}{i}(-1)^i(A + B)^i \). When this is done, we see that the expression reduces to

\[
f(\alpha) = \frac{1}{2} + \frac{1}{2} \sum_{i=1}^{K} \binom{K}{i} (A - B)(-(A + B))^{(i-1)}
\]

Thus, the function \( f \) is seen to be a polynomial (in \( \alpha \)) of degree \( K \).

2. There is only one stable fixed point in the interval \([0, 1]\) to which the population converges from all initial conditions.

3. If \( a = b \), the stable point is given by \( \alpha = \frac{1}{2} \). From all initial conditions, populations move to this mix.

4. If \( a \neq b \), then the location of the stable fixed point changes. In particular, if \( a > b \), then the stable fixed point is very close to \( \alpha = 0 \) — the population will mostly speak \( L_2 \) eventually. On the other hand, if \( a < b \), the stable fixed point is very close to \( \alpha = 1 \), the population will mostly speak \( L_1 \) eventually.

To see (2), it is sufficient to show that the map \( \alpha_{t+1} = f(\alpha_t) \) is such that \( f \) is continuous, \( f(0) > 0, f(1) < 0 \), the equation \( x = f(x) \) has only one root in \([0, 1]\) and \( |f'(x)| < 1 \) at this root. The proof of these last two facts is tedious and omitted for our discussion. To see (3), we can substitute \( a = b \) in the update rule. After some algebra, we see that the update rule reduces to

\[
\alpha_{t+1} = \alpha_t(1 - b^K) + \frac{b^K}{2}
\]

This is a linear recurrence relation and it is easily seen that \( \alpha_t \to \frac{1}{2} \) exponentially as \( t \) tends to infinity. To investigate (4) a little more closely, we show in Fig. 5.3 the stable fixed point as \( a \) and \( b \) vary for \( K \) held fixed at \( K = 5 \). Notice that at \( a = b \), the stable fixed point is \( \frac{1}{2} \). When \( a \neq b \), the stable fixed point changes very rapidly to a value close to 0 or 1 depending upon whether \( a > b \) or vice versa. Note, however, that this transition is rapid for small values \( a, b \) while it is more gradual for values of \( a, b \) close to 1. In fact, the gradualness of this transition increases as \( K \) becomes small. For large values of \( K \), the transition becomes sharper and becomes abrupt for \( K = \infty \) as outlined in the next remark.
Figure 5.3: The stable fixed point as a function of $a$ and $b$. Note that if $a \neq b$, the stable fixed point is close to 1 or 0 for most values of $a, b$. The transition of the fixed point from a value close to 1 to a value close to 0 is quite sharp for low values of $a, b$ and gradual for values of $a, b$ close to 1.
5.2. A PRELIMINARY MODEL

Remark 5. If we let the number of examples $K$ become larger and larger (eventually tending to $\infty$) the limiting map is obtained simply as

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

It is easy to see that $f'(\alpha)$ is given by

$$
f'(\alpha) = \frac{(1-b)(1-a)}{[(1-b) + \alpha(b-a)]^2}
$$

For this map, a dynamical systems analysis yields the following results:

1. If $a = b$, then $\alpha_{t+1} = \alpha_t$ — in other words, the initial population mix is preserved for ever.

2. If $a > b$, there are exactly two fixed points. This is immediate from the solutions of $\alpha = f(\alpha)$. $\alpha = 1$ is a stable fixed point while $\alpha = 0$ is an unstable point. This is easily seen by substituting $\alpha = 0$ and $\alpha = 1$ in the expression for $f'(\alpha)$.

3. If $a < b$, the two fixed points switch stability so that $\alpha = 1$ is now unstable while $\alpha = 0$ is the stable fixed point. Again, this is easily seen from the expression for $f'(\alpha)$.

Remark 6. The parameters $a$ and $b$ determine the evolution of the population and its stable modes. As we have mentioned before, they represent respectively the proportion of $L_1$ and $L_2$ sentences respectively that are ambiguous. It is conceivable that one might be able to estimate these parameters from synchronic or diachronic corpora.

A: Batch Error Based Learner

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

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

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

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

$$
\hat{L} = \arg\min_{L_1, L_2} e(L_i)
$$
182 CHAPTER 5. LANGUAGE CHANGE - A PRELIMINARY MODEL

This amounts to the following rule:

1. Group the $K$ example sentences of the data set (PLD) into three classes: (A) those sentences that belong to $L_1$ alone and are not analyzable by the underlying grammar of $L_2$; (B) those sentences that are analyzable by the underlying grammar of $L_1$ but are ambiguous in that they are also analyzable by the underlying grammar of $L_2$, i.e., sentences belonging to $L_1 \cap L_2$; (C) those sentences that belong to $L_2$ alone and are not analyzable by the underlying grammar of $L_1$. Let $n_1, n_2, n_3$ be the number of examples of type A,B,C respectively.

2. Clearly, $n_1 + n_2 + n_3 = K$. Choose $L_1$ if $n_1 > n_3$, choose $L_2$ if $n_3 > n_1$. If $n_1 = n_3$, one might choose either $L_1$ or $L_2$ according to a deterministic or randomized rule.

For illustrative purposes, let us consider a version of this algorithm that chooses $L_1$ if $n_1 \geq n_3$ and $L_2$ otherwise. For this learning algorithm, it is possible to show that the proportion of $L_1$ users in two successive generations ($\alpha_t$ and $\alpha_{t+1}$, respectively) is related by the following update rule.

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

where

$$p_1(\alpha_t) = \alpha_t (1-a); p_2(\alpha_t) = \alpha_t a + (1-\alpha_t) b; p_3(\alpha_t) = (1-\alpha_t)(1-b).$$

In general, the nature of the population dynamics is different in this case from the previous one. An analysis of this iterated map reveals the following

1. If $b = 1$, then $p_3(\alpha) = 0$. Consequently, $n_3$ will always be zero and $n_1 \geq n_3$ with probability one. Therefore, we see that $\alpha_{t+1} = 1$ and remains fixed at this value thereafter.

2. If $b \neq 1$ and $a = 1$, we have $p_1(\alpha) = 0$. Therefore $n_1 = 0$ and the probability that $n_1 \geq n_3$ is the same as the probability that $n_3 = 0$. Therefore the map reduces to

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

For this case, $\alpha = 0$ is not a fixed point while $\alpha = 1$ is a fixed point. The stability of $\alpha = 1$ depends upon the value of $K(1 - b)$. In fact, we
see that $\alpha = 1$ is stable if and only if $b > 1 - \frac{1}{K}$. For smaller values of $b$, this fixed point is unstable and a new stable point arises in the open interval $(0, 1)$.

3. For most other choices of $a, b$ we see that $\alpha = 1$ is stable. There are exactly two other fixed points $\alpha_1 < \alpha_2$ with $\alpha_i \in (0, 1)$. $\alpha_1$ is stable and $\alpha_2$ is unstable.

4. If we let $K$ go gradually to infinity, we see that that $\frac{n_1}{K} \to p_1$ while $\frac{n_2}{K} \to p_3$. Therefore $\alpha_i \to 1$ when $p_1 > p_3$, i.e., $\alpha(1-a) > (1-\alpha)(1-b)$. It is clear that $\alpha = 0$ and $\alpha = 1$ are both stable fixed points while $\alpha = \frac{1-b}{(1-a)(1-b)}$ is an unstable fixed point in between.

5. It is worthwhile to note that the version of the batch learner considered here is asymmetrical in that if $n_1 = n_3$, the algorithm chooses $L_1$. This asymmetry is less important in large $K$ situations where the probability of the event $n_1 = n_3$ is small. One may consider a symmetric version of this algorithm where the learner chooses $L_1$ with probability $\frac{1}{2}$ when $n_1 = n_3$. The evolutionary dynamics corresponding to this algorithm is qualitatively similar to that corresponding to the current (asymmetric) one for $a, b < 1$.

\textbf{A : Cue Based Learner}

A cue based learner examines the data set for cues to a linguistic parameter setting. Let a set $C \subseteq (L_1 \setminus L_2)$ be a set of examples that are cues to the learner that $L_1$ is the target language. If such cues occur often enough in the learner’s data set, the learner will choose $L_1$, otherwise the learner chooses $L_2$. This follows the cue driven approach advocated in Lightfoot (1998) and is often associated with theories of markedness of linguistic parameters.

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

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

the learner chooses $L_1$, otherwise the learner chooses $L_2$.

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

\[
\sum_{K \tau \leq i \leq K} \binom{K}{i} (p \alpha_t)^i (1 - p \alpha_t)^{(K-i)}
\]

and therefore, we get

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

Here, \( \alpha_t \) is the proportion of \( L_1 \) users in the \( t \)th generation. The evolutionary modes of the population clearly depend upon the value of the parameter \( p \).

An analysis reveals the following:

1. For \( p = 0 \), cues are never produced so the only stable point is \( \alpha = 0 \) to which the population converges in exactly one step.

2. For small values of \( p \), \( \alpha = 0 \) is the only fixed point of the population and it is stable.

3. As \( p \) increases a bifurcation occurs when two new fixed points arise. There are three fixed points altogether - \( \alpha = 0 \) which remains stable, \( \alpha = \alpha_1 \) which is unstable and \( \alpha = \alpha_2 > \alpha_1 \) which is stable.

4. For \( p = 1 \), there are two stable fixed point (\( \alpha = 0 \) and \( \alpha = 1 \)) and there is exactly one unstable fixed point in between.

Shown in Fig. 5.4 is the bifurcation diagram as \( p \) is varied for the case where \( K = 50 \) and \( \tau = 0.6 \). Notice that for small values of \( p \), there is only one stable point at \( \alpha = 0 \). As \( p \) increases, a new pair of fixed points arise — one of which is stable and the other is unstable.

To see the qualitative nature of the dynamics, we need to investigate the fixed points of the functional map \( \alpha_{t+1} = f(\alpha_t) = g(p \alpha_t) \) where

\[
g(y) = \sum_{i=K \tau}^{K} \binom{K}{i} y^i (1 - y)^{K-i}
\]

By inspection of \( f(\alpha) \), it is easy to see that \( f(0) = 0 \) for all values of \( p \). To show stability, it is sufficient to show that \( |f'(0)| < 1 \). Note that

\[
f'(\alpha) = pg'(p \alpha)
\]
Figure 5.4: The bifurcation diagram for as the parameter $p$ varies. For this example, $K = 50$ and $\tau = 0.6$ Notice how for small values of $p$ there is only one stable point at $\alpha = 0$. As $p$ increases a new pair of fixed points arises — one of which is unstable (dotted) and the other is stable (solid). For any value of $p$ on the $x$-axis, the $y$-axis denotes the values of the fixed points.
Differentiating $g$ term by term, we get (let $K_\tau$ be the smallest integer bigger than $K\tau$)

$$g'(y) = \sum_{k=K_\tau}^{K-1} \binom{K}{k} [ky^{K-1}(1-y)^{K-k} - (K-k)y^k(1-y)^{K-k-1}] + Ky^{K-1}$$

Expanding this out, we see

$$g'(y) = \sum_{k=K_\tau}^{K-1} \frac{K!}{(K-k)!} k y^{K-1}(1-y)^{K-k} - \sum_{k=K_\tau}^{K-1} \frac{K!}{(K-k)!} (K-k)y^k(1-y)^{K-k-1}$$

$$+ Ky^{K-1}$$

Factoring $K$ out of the expression, we have

$$g'(y) = K \left[ \sum_{k=K_\tau}^{K-1} \frac{(K-1)!}{(K-k)!(k-1)!} y^{K-1}(1-y)^{K-k} \right] - K \left[ \sum_{k=K_\tau}^{K-1} \frac{(K-1)!}{k!(K-k-1)!} y^k(1-y)^{K-k-1} - y^{K-1} \right]$$

After cancelling terms, we see

$$g'(y) = K \left( \frac{K-1}{K_\tau-1} \right) y^{K_\tau-1}(1-y)^{K_\tau} \right)$$  \hspace{1cm} (5.2)$$

Clearly, $f'(0) = pg'(0) = 0$. Stability of the fixed point $\alpha = 0$ is shown. We next need to show that there can be at most two fixed points in the interval $(0, 1]$. First note that $f(1) = g(p) < 1$ if $p < 1$.

The fixed points are the points where $f(x)$ crosses the graph of the function $h(x) = x$. Since (i) $f(0) = 0$; (ii) $f'(0) = 0$ and (iii) $f(x)$ is continuous, we see immediately that the number of such crossings in $(0, 1]$ will be even. Let there be $2m$ such crossings in all and let the crossings be indicated by $\alpha_1, \alpha_2, \ldots, \alpha_{2m}$. Further the slope of $f(x)$ will be alternately greater and smaller than $h' = 1$ at each of these points. Thus we have $f'(\alpha_1) > 1$, $f'(\alpha_2) < 1$, $f'(\alpha_3) > 1$, and so on. Let there be 4 or more fixed points. Then clearly, in the interval $(\alpha_1, \alpha_3)$ the graph of $f'$ goes from being $> 1$ to $< 1$ to $> 1$ again. Therefore, there must be a point where $f'' = 0$. By the same logic, in the interval $(\alpha_2, \alpha_4)$ there must be another distinct point
where $f'' = 0$. Therefore there must be at least two distinct points where $f''$ vanishes.

However, it is possible to show that in the entire interval $(0, 1)$, the derivative $f''$ vanishes at most once. To see this, notice that

$$f''(y) = p^2 g''(py)$$

Now $g'(y) = Ay^i(1-y)^j$ where $A, i, j$ can be read off from Eq. 13.5. Therefore

$$g''(y) = A[iy^{i-1}(1-y)^j - jy^i(1-y)^{j-1}] = Ay^{i-1}(1-y)^{j-1}[i - (i + j)y]$$

It is clear that $g''(y)$ has exactly one zero crossing in $(0, 1)$ and therefore $g''(py)$ will have at most one zero crossing in $(0, 1)$. This leads to a contradiction.

A limiting analysis may be conducted by allowing the number of examples $K \to \infty$. It is seen that

$$\frac{k}{K} \to p\alpha_t$$

Therefore, if $p\alpha_t < \tau$, all learners choose $L_2$ and the population evolves to a homogeneous community of $L_2$ speakers in one generation. Conversely, if $p\alpha_t > \tau$, all learners choose $L_1$ and the population evolves to a homogeneous community of $L_1$ speakers in one generation. From this, the following observations may be made

1. If $p < \tau$, then $p\alpha < \tau$ for all $\alpha \in [0, 1]$ and so the only stable fixed point is $\alpha = 0$.

2. If $p > \tau$, then there are two stable fixed points. From all initial conditions $\alpha_0 \in [0, \tau/p)$, the population converges to $\alpha = 0$. From all initial conditions $\alpha_0 \in (\tau/p, 1]$, the population converges to $\alpha = 1$.

We see that the dynamics of the finite sample case (finite $K$) is qualitatively similar to the infinite sample case.

In conclusion we see that a homogeneous population of $L_2$ users will always remain stable and can never change to a population of $L_1$ users. On the other hand, a homogeneous population of $L_1$ users will remain stable only in a certain regime of $p$ values. As $p$ changes the basin of attraction shrinks and after a critical value of $p$ the population switches to a stable mode of $L_2$ speakers. Thus a change from $L_1$ to $L_2$ is possible — a change the other way is never possible. We shall examine later the implications of this for various explanations of directional changes in linguistic grammars.
5.3 Implications and Further Directions

The simple two-language models of the preceding sections are not without significant linguistic applications. In many cases of language change, one finds that there are two variants (dialects, grammars) differing by a significant linguistic parameter that coexist in a population in varying proportions at different points in time. Often, linguistic change leads to the gradual loss of one variant from the population entirely — in many cases following an S-shaped pattern over time. For example, the loss of verb second (V2) from grammars of Old English to that of Modern English is a much studied instance of precisely such a change (Lightfoot, 1998; Kroch and Taylor, 1997). Other examples include the loss of verb second (V2) from Old to Modern French (Clark and Roberts, 1993; Roberts, 1992), the change in subordinate clause word order in Yiddish considered by Santorini (1993) and so on.

As an example of such a change consider the following case of Yiddish reported in Santorini (1993).

5.3.1 An Example from Yiddish

Yiddish underwent some significant linguistic changes from the fifteenth to the nineteenth century A.D. One particular change had to do with the location of the auxiliary verb with respect to the subject and the verb phrase in clauses. For convenience, we reproduce here the discussion about such a change from Chapter 1 of this book. Following Chomsky (1986) one might let the auxiliary verb belong to the functional category INFL (that bears inflectional markers) and thus distinguish between the two basic phrase structure alternatives as in 5(a) and 5(b).

\[
5(a) \ [\text{Spec} \ [\text{VP INF}L]\text{IP} \\
5(b) \ [\text{Spec} \ [\text{INF}L \ \text{VP}]\text{IP}
\]

The inflectional phrase (IP) describes the whole clause (sentence) with an inflectional head (INFL), a verb phrase argument (VP) for this INFL head and a specifier (Spec). The item in specifier position is deemed the subject of the sentence. In modern English, for example, phrases are almost always of type 5(b). Thus the sentence (6)

\[
(6) \ [\text{John} \ [\text{can} \ [\text{read the blackboard} \ [\text{VP}]]]\text{IP}
\]
5.3. IMPLICATIONS AND FURTHER DIRECTIONS

Corresponds to such a type with "John" being in Spec position, "can" being the INFL-head and "read the blackboard" being the verb phrase. If we deem structures like 5(a) to be INFL final and structures like 5(b) to be INFL medial, we find that languages on the whole might be typified according to which of these phrase types is preponderent in the language.\footnote{It is worthwhile to reiterate that while this typological distinction is largely accepted by linguists working in the tradition of Chomsky (1981), there is still considerable debate as to how cleanly languages fall into one of these two types. For example, while Travis (1984) argues that INFL precedes VP in German and Zwart (1991) extends the analysis to Dutch, Schwartz and Vikner (1990) provide considerable evidence arguing otherwise. Part of the complication often arises because the surface forms of sentences might reflect movement processes from some other underlying form in often complicated ways. But this is beyond the scope of this book.}

Interestingly, Yiddish changed from a predominantly INFL final language to a categorically INFL medial one over the course of a transition period from 1400 A.D. to about 1850 A.D. Santorini (1993) has a detailed quantitative analysis of this phenomenon and shown below are two unambiguously INFL final sentences of early Yiddish (taken from Santorini, 1993). Such sentences would be deemed ungrammatical in the modern categorically INFL medial Yiddish.

\textit{ds zi droyf givarn \textit{vern}} (Bovo 39.6, 1507).

that they there-on warned were

\textit{ven der vatr nurt doyts leyan \textit{kan}} (Anshel 11, ca. 1534).

if the father only German read can

To illustrate this point quantitatively, a corpus analysis of Yiddish documents over the ages yields the following statistics shown in Table 5.1. Clauses with simple verbs are analyzed for INFL-medial and INFL-final distributions of phrase structures.

More statistics is available in Santorini (1993) but this simple case illustrates the clear and unmistakable trend in the distribution of phrase types. It is worthwhile to mention here that while Santorini (1993) expresses the statistics within the notational conventions of Chomsky (1986), almost any reasonable grammatical formalism would capture this variation and change with two different grammatical types or forms in competition over time with one gradually yielding to the other over generational time.
Table 5.1: Number of sentences of INFL-medial and INFL-final type in corpora of Yiddish from 1400 to 1950.

<table>
<thead>
<tr>
<th>Time Period</th>
<th>INFL-medial</th>
<th>INFL-final</th>
</tr>
</thead>
<tbody>
<tr>
<td>1400-1489</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>1490-1539</td>
<td>5</td>
<td>37</td>
</tr>
<tr>
<td>1540-1589</td>
<td>13</td>
<td>59</td>
</tr>
<tr>
<td>1590-1639</td>
<td>5</td>
<td>81</td>
</tr>
<tr>
<td>1640-1689</td>
<td>13</td>
<td>33</td>
</tr>
<tr>
<td>1690-1739</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td>1740-1789</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1790-1839</td>
<td>54</td>
<td>3</td>
</tr>
<tr>
<td>1840-1950</td>
<td>90</td>
<td>0</td>
</tr>
</tbody>
</table>

5.3.2 Discussion

We see that the case of Yiddish discussed before falls into a setting where there are two grammatical variants in competition with each other. The grammatical variants are characterized as INFL-medial and INFL-final respectively. In the beginning, all members of the population seemed to be using an INFL-final grammar while by the twentieth century all members were using an INFL-medial grammar. Two aspects of this linguistic change are interesting for our purposes. First, that over a period of time, the population moves from one seemingly stable state to another. Second, both stable states seem to be homogeneous, i.e., while the population passed through mixed modes during intermediate years, the initial and final states of the population seem homogeneous and categorically INFL-final and INFL-medial respectively.

As we have discussed before, the learning of language by children is the key mechanism by which language is transmitted over generational time. Any coherent discussion of why languages change must of necessity take this into account. In fact, one must assign to it a possibly central role in the discourse. While this position has been taken by many historical linguists, only a computational treatment allows this position to be elaborated precisely and exposes the subtleties involved in the analysis.

If we assume that individuals used one or the other variant consistently\(^8\)

\(^8\)This claim is often contentious and sometimes clearly falsifiable empirically. Nonethe-
the dynamical evolution of populations may be characterized using the models described earlier. In this chapter, we considered three different learning models and the evolutionary consequences of each were carefully studied. It is worthwhile to reflect on our general findings.

1. The evolutionary modes of a TLA based learner were characterized by \(a, b\). If \(a = b\), no change was possible. However, if \(a \neq b\), only one stable mode of the population was reached. Whether \(L_1\) or \(L_2\) dominates the population depends upon whether \(a > b\) or not.

2. The evolutionary modes of a batch learner were characterized by \(a, b\). Both \(L_1\) and \(L_2\) were stable modes of the population with each having its own basin of attraction.

3. The evolutionary modes of a cue based learner were characterized by \(p\). For low values of \(p\) only \(L_2\) is stable. For higher values of \(p\) both \(L_1\) and \(L_2\) are stable modes with each having its basin of attraction. A bifurcation between these two evolutionary scenarios occurs at a critical value of \(p\).

How might we explain language change under each of these learning models? First, it is worthwhile to note that under all three models, the stable modes of a population are entirely homogeneous or largely so with one language dominating the other greatly. However, the details are different.

According to (1) the only way change is possible is when \(a \neq b\). Thus a population may be stable with all members speaking \(L_1\) (corresponding to \(a < b\)). If for some reason, in some generation \(a\) becomes greater than \(b\), the stable mode will become unstable. A slight drift out of homogeneity will then move the population entirely to the other language. Both a change from \(L_1\) to \(L_2\) and a change from \(L_2\) to \(L_1\) are possible within this framework.

According to (2) the population will have to jump from one basin of attraction to the other crossing a barrier for the language to change. The only plausible way for this to happen is if language contact as a result of migration brings about a change in the linguistic composition driving the population mix from one basin to the other. Unlike case (1) internally driven change is not possible by drifting of the parameters \(a\) and \(b\) from generation to generation.

less it is a useful exercise to examine the consequences of such a claim because it clarifies the subtle interplay between learning and evolution under various assumptions. See later parts of this book for dropping the assumptions.
According to (3) there is a critical parameter effect and an asymmetry in the evolutionary possibilities. A homogeneous $L_2$ speaking population is the only stable mode for small values of $p$ and therefore no change is possible in this regime. For large values of $p$ both $L_1$ and $L_2$ are stable modes and a change from one to the other requires the population to jump from one basin of attraction to the other. However, parameter drifting might drive a stable $L_1$ speaking population to a stable $L_2$ speaking one if the value of $p$ changes from above the critical threshold to below it. While change in this direction may be internally driven, a change in the other direction is never possible by parameter drifting over generations. Only language change and migration would explain a change in the other direction. This asymmetry is of particular significance in all models of learning and language where a distinction is made between marked and unmarked linguistic parameter settings.

**Major Insights**

There are thus two major insights that have emerged.

1. Different learning algorithms might have different evolutionary consequences. These evolutionary consequences may be tested using historical data. As a result, diachronic facts may be brought to bear as additional evidence in judging the adequacy of different theories of language acquisition.

2. The evolutionary dynamics are typically non-linear and the bifurcations (phase transitions) associated with them may provide a suitable theoretical construct to explain language change. This provides a novel solution to the actuation problem. These results raise the possibility that discontinuous change may result from a continuous drift in the frequencies with which cues, triggers, etc. are produced.

It is worth noting that these insights emerge only as a result of computational analysis following some simplified model construction. The relationship between learning and evolution can be quite subtle and it is hard to see how, for example, one might discover the possibility of bifurcations from verbal arguments alone.

The case of Yiddish described here was provided only as an illustrative example of a situation where the language of the population seemed to be changing along one linguistic dimension. The different models of learning
we have considered may be derived from psycholinguistic considerations of language development — we see clearly in these sections how they have different consequences for language change. In later portions of this book, we will examine similar models in the context of other linguistic changes such as Portuguese, Chinese, French, and English.

5.3.3 Future Directions

The models in this chapter have already brought into sharp focus the interplay between learning and change in linguistic populations. At the same time, one needs to recognize the drastic simplifications that have been made in order to formulate this first coherent model. It is worthwhile to reflect on some of these simplifying assumptions and the possibility of relaxing them in more complex models of this process.

1. Multiple Languages: Clearly, there are more than two languages in the world. The space \( G \) represents the hypothesis space that learning algorithms operate on and from which they draw grammatical hypotheses over their learning period. While this space \( G \) is in principle all of universal grammar, in linguistic applications of a more specific nature, one might consider a subset of \( G \) to be the more appropriate object to model and study. For example, in studies of syntax, its acquisition and change, it may be meaningful to ignore the phonological components of UG that might have no bearing on the phenomena at hand. Depending upon the sub-module of syntax under study, other “irrelevant” modules might also be usefully ignored. Thus, the space \( G \) that is formulated in models of language change is really a very low-dimensional projection of the high dimensional space of UG. It is the linguist’s intuition and understanding of the phenomena that provides the appropriate low-dimensional projection. Indeed, linguists might differ on this matter, and the consequences then need to be worked out. Having thus argued that in most useful applications, the space \( G \) will be low dimensional, it might still consist of more than two grammars (languages) and it is important to extend such models to multilingual settings. The extension to \( n \)-language families has already been considered (Niyogi and Berwick, 1997; Yang, 2000). Setting up the models for such cases is easy enough – analytical solutions are harder to come by and one might need to resort to simulations. We consider the \( n \)-language case in the next chapter.
2. **Finite Populations:** One reason we have been able to derive deterministic dynamical maps relating successive generations to each other is the assumption of infinite population size that allows us to take ensemble averages of individual behavior over the entire population. In practice, of course, populations are always finite. If the population sizes are large, then the assumption of infinite sizes may not be too bad. If, on the other hand, population sizes are very small, then one might need to consider the implications of such small sizes more carefully. Let us consider briefly the effect of finite population sizes on the two language models discussed in this paper. Recall that each individual child attains $L_1$ with probability $p_k (A, \alpha_a P_1 + (1 - \alpha_a) P_2)$. From this we concluded that a proportion $p_k$ of the children would end up as $L_1$ users. This statement is exactly true if there were an infinite number of children. Imagine, instead, there were only $N$ children in the population. Each child could end up either as an $L_1$ speaker or as an $L_2$ speaker. In fact, with probability $(p_k)^N$, all children would acquire $L_1$; with probability $(1 - p_k)^N$ all would acquire $L_2$; and different intermediate mixes are possible with probabilities given by the binomial distribution. Thus all evolutionary trajectories are possible, the question is – which ones are likely or probable? The evolution is characterized now as a stochastic process rather than a deterministic dynamical system. The consequences of this can be worked out. The details are beyond the scope of this chapter and are pursued in a later chapter (Chapter 10).

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

Let the state of cohort $t$ be described by a variable $\alpha_t$ (as before, where $\alpha_t$ denotes the proportion of the cohort using the language $L_1$). Consider now the $(t + 1)$th cohort of learning children. Assume that they receive data drawn from the previous three cohorts (who may, for
5.3. **IMPLICATIONS AND FURTHER DIRECTIONS**

example, be characterized as the cohort of young adults, parents, and grandparents respectively) in equal proportions. Then the probability distribution with which data is presented to the \( (t + 1) \)th cohort of learners is given by

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

where we have assumed that all cohorts are equal in size and influence. Given this set up, it is easy to see that \( \alpha_{t+1} \) is now going to be given by

\[
\alpha_{t+1} = p_K(A, P)
\]

and in this manner, \( \alpha_{t+1} \) will depend upon \( \alpha_t, \alpha_{t-1}, \) and \( \alpha_{t-2} \) respectively. The resulting dynamics may be analyzed using the traditional tools.

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

Reality, as always, might be more complicated. Speakers of different linguistic types might reside in different “neighborhoods”. Children born in different neighborhoods might receive data drawn from different probability distributions that reflect the linguistic composition of their neighborhood. For example, one might imagine an \( L_1 \) speaking neighborhood and an \( L_2 \) speaking neighborhood whose population sizes are in the ratio \( \alpha_t : (1-\alpha_t) \). Children born in the \( L_1 \) speaking neighborhood might receive data drawn mostly according to \( P_1 \) while those born in the \( L_2 \) speaking neighborhood might receive data mostly drawn according to \( P_2 \). The evolutionary consequences of such a spatial structure in the population need to be worked out and represents an important direction of future research. Chapters 9 and 10 discuss spatial models.

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

If the target distribution corresponds to a unique grammar, it is certainly reasonable to expect the learner to end up with exactly one language. The case when the target distribution is mixed, i.e., not consistent with a single unique target grammar, natural models of the learning process should allow the possibility of multilingual rather than monolingual acquisition. Thus, for example, in the two language case of this chapter, one might allow the possibility that the learner acquires both languages (in some ratio, perhaps). For the case of English, for example, Kroch and Taylor (1997) argue that learners were effectively bilingual having acquired both dialectal variations in different proportions. Yang (2002) considers such a learning algorithm and explores the evolutionary consequences. We explore the multilingual setting in Chapters 8 and 10.

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

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

In the next few chapters, we will consider a variety of such generalizations and case studies. These will further elucidate the central theme of the current chapter — the nature of the relationship between language learning and language change.
Chapter 6

Language Change - Multiple Languages

6.1 Multiple Languages

In the previous chapter we examined some preliminary models that arise as a result of two languages in competition with each other. We consider now a more general case where \( n \) languages may be potentially present in the population at any point in time. We begin by following our usual logic in deriving dynamical models of language change from models of language acquisition. The basic framework is developed in the next few sections. As we shall see, the \( n \)-language case gives rise to \( n - 1 \) dimensional discrete-time dynamical systems. While a complete analytic understanding is well beyond our current scope, we will conduct two simulation studies later in this chapter to get some insight into the evolutionary trajectories of such systems in a linguistic context. Both these simulation studies are developed in the context of a case of syntactic change observed in the evolution of French from the twelfth century to modern times. Over the course of these simulations, we will see how the dynamical systems framework might allow us engage the issues involved in language change in a concrete, formal, and reasoned fashion.

6.1.1 The Language Acquisition Framework

To formalize the model, we must first state our assumptions about grammatical theories, learning algorithms, and sentence distributions.
1. Denote by $\mathcal{G}$ a family of possible (target) grammars. Each grammar $g \in \mathcal{G}$ defines a language $L(g) \subseteq \Sigma^*$ over some alphabet $\Sigma$ in the usual way. We will particularly consider the case where $\mathcal{G}$ has $n$ grammars denoted by $g_1$ through $g_n$ corresponding to $n$ languages $L_1$ through $L_n$ respectively.

2. Denote by $P$ a probability distribution on $\Sigma^*$ according to which sentences are drawn and presented to the learner. Let speakers of $L_i$ produce sentences according to a distribution $P_i$. Thus $P_i$ has support on $L_i \subseteq \Sigma^*$. Therefore, if the adult population is linguistically homogeneous (with grammar $g_1$) then $P = P_1$. If the adult population is composed of two groups of equal size one of which speaks $L_1$ and the other $L_2$, then $P = \frac{1}{2}P_1 + \frac{1}{2}P_2$.

3. Denote by $\mathcal{A}$ the acquisition algorithm that children use to hypothesize a grammar on the basis of input data. Following our notation developed previously, $\mathcal{A}$ is a computable mapping from the set of all possible finite data streams $\mathcal{D}$ to the set of possible grammatical hypotheses $\mathcal{G}$.

Given this setting we will adopt a notion of probabilistic convergence as our notion of learnability. Thus a grammar $g_i \in \mathcal{G}$ is learnable if on presentation of example sentences in i.i.d. fashion according to $P_i$, the learner converges to the target with probability 1, i.e.,

$$\lim_{n \to \infty} \mathbb{P} [\mathcal{A}(d_n) = g_i] = 1$$

where $d_n \in \mathcal{D}$ is a random data set of $n$ sentences drawn according to $P_i$.

### 6.1.2 From Language Learning to Population Dynamics

The framework for language learning focuses on the behavior of learners attempting to infer grammars on the basis of linguistic data. At any point in time, $n$, (i.e., after encountering $n$ example sentences) the learner may have any of the grammatical hypotheses in $\mathcal{G}$. In particular, let us denote by $p_n(h)$, the probability with which it has the hypothesis $h \in \mathcal{G}$. Now consider what happens when there is a population of such learners. Since an arbitrary learner has a probability $p_n(h)$ of developing hypothesis $h$ (for every $h \in \mathcal{G}$), it follows that a fraction $p_n(h)$ of the population of learners internalize the grammar $h$ after $n$ examples. We therefore have a current state of the learning population after $n$ examples. This state of the population might well be different from the state of the parental population.
6.1. MULTIPLE LANGUAGES

Assume for a moment that after $N$ examples, maturation occurs, i.e., the grammatical hypothesis after $N$ examples crystallizes in the learner’s mind and is retained and used for communication for the rest of its adult life. Then one would arrive at the state of the mature population for the next generation. This new generation now produces sentences for the following generation of learners according to the distribution of grammars in its population. In this manner the process repeats itself and the linguistic composition of the population evolves from generation to generation.

We can now define a discrete time dynamical system by providing its two necessary components:

**A State Space:** a set of system states, $\mathcal{S}$. Here the state space is the space of possible linguistic compositions of the population. Each state is described by a distribution $P_{pop}$ on $\mathcal{G}$ describing the language spoken by the population. At any given point in time, $t$, the system is in exactly one state $s_t \in \mathcal{S}$;

**An Update Rule:** how the system states change from one time step to the next. Typically, this involves specifying a functional mapping, $f$, that maps $s_t \in \mathcal{S}$ to $s_{t+1}$.

As a linguistic example, consider the three parameter syntactic space described in Gibson and Wexler (1994) and examined at length in previous chapters. This system defines eight possible “natural” grammars — thus $\mathcal{G}$ has eight elements. We can picture a distribution on this space as shown in Fig. 6.1. In this particular case, the state space is

$$\mathcal{S} = \{ \mathbf{P} \in \mathbb{R}^8 | \sum_{i=1}^{8} P_i = 1 \}$$

Here we interpret the state as the linguistic composition of the population. For example, a distribution that puts all its weight on grammar $g_1$ and 0 everywhere else indicates a homogeneous population that speaks a language corresponding to grammar $g_1$. Similarly, a distribution that puts a

---

1Maturation seems to be a reasonable hypothesis in this context. After all, it seems even more unreasonable to imagine that learners are forever wandering around in hypothesis space. There is evidence from developmental psychology to suggest a maturational account such that after a certain point children mature and retain their current grammatical hypotheses forever. There are however subtle effects on the adult’s native language due to factors such as second language learning or immersion in a foreign language community as a result of migration. Such effects may be specifically modeled in a systematic manner but we leave aside such considerations for now.
probability mass of 1/2 on \( g_1 \) and 1/2 on \( g_2 \) denotes a population (nonhomogeneous) with half its speakers speaking a language corresponding to \( g_1 \) and half speaking a language corresponding to \( g_2 \).

To see in detail how the update rule may be computed, consider the acquisition algorithm \( \mathcal{A} \). The state at time \( t \) (given by \( P_{\text{pop},t} \)) determines the distribution of speakers of different languages in the parental population as a whole. Therefore, one can obtain the distribution with which sentences from \( \Sigma^* \) will be presented to a typical learner. Recall that the \( i \)th linguistic group in the population, speaking language \( L_i \), produces sentences with distribution \( P_i \) on \( \Sigma^* \). Therefore for any \( \omega \in \Sigma^* \), the probability with which \( \omega \) will be presented to the learner is given by

\[
P(\omega) = \sum_i P_i(\omega) P_{\text{pop},t}(i)
\]

This fixes the distribution \( P \) with which sentences are presented to the learner. Note that we have crucially assumed that there is perfect spatial mixing so that each individual child learner is exposed to the parental population in an unbiased way, i.e., they are exposed to all the different linguistic types in proportion to their numbers in the entire adult population. We will consider departures from this assumption in a later chapter.

The logical problem of language acquisition also assumes some success criterion for attaining the mature target grammar. For our purposes, we take this as being one of two broad possibilities: either (1) the usual scenario of identification in the limit which we shall call the \textit{limiting sample} case; or (2) identification in a fixed, finite time, which we shall call the \textit{finite sample} case.\(^2\)

Consider case (2) first. Here, one draws \( n \) example sentences according to distribution \( P \), and the acquisition algorithm develops hypotheses \( (\mathcal{A}(d_n) \in \mathcal{G}) \). One can, in principle, compute the probability with which the learner will posit hypothesis \( h_i \) after \( n \) examples:

\[
\text{Finite Sample: } \mathbb{P}[\mathcal{A}(d_n) = h_i] = p_n(h_i) \quad (6.1)
\]

Now turn to case (1), the limiting case. Here learnability requires that \( p_n(g_t) \) converge to 1 for the case where a unique target grammar, \( g_t \) exists. However, in general, there need not be a unique target grammar since the

\(^2\)Of course, a variety of other success criteria, e.g., convergence within some epsilon, or polynomial in the size of the target grammar, are possible; each leads to a potentially different language change model. We do not pursue these alternatives here.
6.1. MULTIPLE LANGUAGES

linguistic population can be nonhomogeneous. Even so, recall that since convergence is measurable, the following limiting behavior still exists.

\[ \lim_{n \to \infty} P[\mathcal{A}(d_n) = h_i] = p(h_i) \]  

(6.2)

Let us now turn from the individual child to the population as a whole. For each grammar \( h_i \in \mathcal{G} \), the individual child learner adopts (internalizes) this grammar with probability \( p_n(h_i) \) in the finite sample case or with probability \( p(h_i) \) in the limiting sample case. In a population of such individuals one would therefore expect a proportion \( p_n(h_i) \) or \( p(h_i) \) respectively to have internalized grammar \( h_i \). In other words, the linguistic composition of the next generation is given by \( P_{\text{pop,}t+1}(h_i) = p_n(h_i) \) for the finite sample case and by \( P_{\text{pop,}t+1}(h_i) = p(h_i) \) in the limiting sample case. In this fashion,

\[ P_{\text{pop,}t} \xrightarrow{\mathcal{A}} P_{\text{pop,}t+1} \]

Remarks:

1. In deriving this update rule, we have assumed that population sizes are infinite so that the proportion of \( L_i \) speakers in the next generation is exactly equal to the probability with which the typical child acquires \( L_i \). For small population sizes, the deviation from this may be quite significant. This leads one to specify the linguistic evolution as a stochastic process rather than a deterministic dynamical system and we shall examine the consequences of this in a later chapter.

2. For a Gold-learnable (by the algorithm \( \mathcal{A} \)) family of languages and a limiting sample assumption, homogeneous populations are always stable. This is simply because each child and therefore the entire population always eventually converges to and thus attains the unique target grammar in each generation.

3. The finite sample case, however, is different from the limiting sample case. Suppose we have solved the maturation problem — that is, we know roughly the time, or number of examples \( N \) the learner takes to develop its mature (adult) hypothesis. In that case \( p_N(h) \) is the probability that a child internalizes the grammar \( h \), and \( p_N(h) \) is the percentage of speakers of \( L_h \) in the next generation. Note that under this finite sample analysis, even for a homogeneous population with
all adults speaking a particular language (corresponding to grammar, 
g, say), \( p_N(g) \) will not be 1—that is, there will be a small percentage
of learners who have misconverged. This percentage could blow up
over several generations, and we therefore have potentially unstable
languages.

4. The formulation is very general. Any triple \( \langle \mathcal{A}, \mathcal{G}, \{P_i\} \rangle \) yields a
dynamical system.\(^3\) In short:

\[
\langle \mathcal{G}, \mathcal{A}, \{P_i\} \rangle \rightarrow \text{ (dynamical system)}
\]

Note that the formulation does not assume any particular linguistic
theory — the elements of \( \mathcal{G} \) could be specified in any suitable formal-
ism from Optimality Theory to Government and Binding to Connect-
ionism. Nor have we assumed any particular learning algorithm or
distribution with which sentences are drawn. Of course, we have im-
plicitly assumed a learning model, i.e., positive examples are drawn in
i.i.d. fashion and presented to the learner. Our dynamical systems for-
malization follows as a logical consequence of this learning framework.
One can conceivably imagine other learning frameworks—these would
potentially give rise to other kinds of dynamical systems—but we do
not formalize them here.

In previous chapters we examined the problem of learnability within
parametric systems. In particular, we showed how the behavior of any mem-
oryless learning algorithm can be modeled as a Markov chain. This analysis
allows us to solve equations 6.1 and 6.2 and thereby obtain the update equa-
tions for the associated dynamical system. Let us now show how to derive
such models in detail. We first provide the particular \( \langle \mathcal{G}, \mathcal{A}, \{P_i\} \rangle \) triple,
and then give the update rule.

A learning system triple.

1. \( \mathcal{G} \): Assume there are \( n \) binary valued linguistic parameters—this leads
to a space \( \mathcal{G} \) with \( 2^n \) different grammars.

2. \( \mathcal{A} \): Let us imagine that the child learner follows some memoryless
(incremental) algorithm to set parameters. For the most part, we

\(^3\)Note that this probability could evolve with generations as well. That will complete
all the logical possibilities. However, for simplicity, we assume that this does not happen.
6.1. MULTIPLE LANGUAGES

will assume that the algorithm is the “Triggering Learning Algorithm” (TLA) or one of the variants discussed in previous chapters.

3. \{P_i\}: Let speakers of the ith language, \(L_i\), in the population produce sentences according to the distribution \(P_i\). For the most part we will assume in our simulations that this distribution is uniform on degree-0 (unembedded) sentences.

**The update rule.** We can now compute the update rule associated with this triple. Suppose the state of the parental population is \(P_\text{pop}\) on \(G\). Then one can obtain the distribution \(P\) on the sentences of \(\Sigma^*\) according to which sentences will be presented to the learner. Once such a distribution is obtained, then given the Markov equivalence established earlier, we can compute the transition matrix \(T\) according to which the learner updates its hypotheses with each new sentence. From \(T\) one can finally compute the following quantities, one for the “finite sample” case and one for the “limiting sample” case:

\[
P[\text{Learner’s hypothesis} = h_i \in G \text{ after } m \text{ examples}] = \{\frac{1}{2^n}1^T_{2^n}T^m\}[i] \tag{6.3}
\]

In the above equation, \(1_{2^n}\) represents the \(2^n\) dimensional column vector with all its components taking the value 1 and \(1^T_{2^n}\) is simply its transpose. The \(i, j\) element of the matrix \(T^m\) contains the probability with which the learner moves from an initial hypothesis of \(h_i\) to a hypothesis of \(h_j\) after exactly \(m\) examples. We have assumed that the learner chooses an initial hypothesis uniformly at random from the \(2^n\) different hypotheses in \(G\). \(\frac{1}{2^n}1^T_{2^n}T^m\) is therefore a row vector and \(\{\frac{1}{2^n}1^T_{2^n}T^m\}[i]\) denotes its \(i\)th component.

Similarly, making use of the limiting distributions of Markov chains (Resnick, 1992) one can obtain the following (where \(ONE\) is a \(2^n \times 2^n\) matrix with all ones).

\[
P[\text{Learner’s hypothesis} = h_i \text{ “in the limit”}] = \{\frac{1}{2^n}1^T_{2^n}(I - T + ONE)^{-1}\}[i] \tag{6.4}
\]
These expressions allow us to compute the linguistic composition of the population from one generation to the next according to our analysis of the previous section.

*Remark.* The limiting distribution case is more complex than the finite sample case and requires some careful explanation. There are two possibilities. If there is just a single target grammar, then, by definition, the learners all identify the target correctly in the limit, and there is no further change in the linguistic composition from generation to generation. This case is essentially uninteresting. If there are two or more target grammars, then recalling our analysis of learnability from the previous chapters, there can be no absorbing states in the Markov chain corresponding to the parametric grammar family. In this situation, a single learner will oscillate between some set of states in the limit. In this sense, learners will not converge to any single, correct target grammar. However, there is a sense in which we can characterize limiting behavior for learners: although a given learner will visit each of these states infinitely often in the limit, it will visit some more often than others. The exact proportion of the time the learner will be in a particular state is given by equation 6.4 above. Therefore, since we know the fraction of the time the learner spends in each grammatical state in the limit, we assume that this is the probability with which it internalizes the grammar corresponding to that state in the Markov chain. An alternative interpretation is worth emphasizing. The limiting distribution may be viewed as characterizing the probability with which the learner will use each of the grammars after seeing a large (infinite) number of examples. Hence, one might interpret the final state of the learner as having internalized multiple grammars. For more discussion on multilingual acquisition, see later chapters.

One can summarize the basic computational framework for modeling language change as follows:

1. Let \( \pi_1 \) be the initial population mix, i.e., the percentage of different language speakers in the community. Assuming that the \( i \)th group of speakers produces sentences with probability \( P_i \), we obtain the probability \( P \) with which sentences in \( \Sigma^* \) occur for the next generation of learners.

2. From \( P \) we obtain the transition matrix \( T \) for the Markov learning model and using equations 6.3 and 6.4, we derive the distribution of the linguistic composition \( \pi_2 \) for the next generation.

3. The second generation now has a population mix of \( \pi_2 \). We repeat
6.2. Example 1: A Three Parameter System

Recall that every choice of \( \mathcal{G}, \mathcal{A}, \{P_i\} \) gives rise to a unique dynamical system. We begin by making specific choices for these three elements:

1. \( \mathcal{G} \): This is a 3-parameter syntactic subsystem described in Gibson and Wexler (1994). Thus \( \mathcal{G} \) has exactly 8 grammars, generating languages from \( L_1 \) through \( L_8 \) described in previous chapters.

2. \( \mathcal{A} \): The memoryless algorithms we consider are the TLA and the variants obtained by dropping either or both of the single-valued and greediness constraints.
3. \( \{P_i\} \): For the most part, we assume sentences are produced according to a uniform distribution on the degree-0 sentences of the relevant language, i.e., \( P_i \) is uniform on (degree-0 sentences of) \( L_i \).

Ideally, a complete investigation of diachronic possibilities would involve varying \( G \), \( A \), and \( P \) and characterizing the resulting dynamical systems by their phase space plots. Rather than explore this entire space, we first consider only systems evolving from homogeneous initial populations, under four basic variants of the learning algorithm \( A \) while holding \( G \) and \( \{P_i\} \) fixed. This will give us an initial grasp of how linguistic populations may change. Indeed, linguistic change has been studied before; even the dynamical system metaphor itself has been invoked. Our computational paradigm allows us to go further than these previous qualitative descriptions; for example, (1) we can precisely calculate what the rates of change will be; (2) we can determine what diachronic population curve changes will look like, without stipulating in advance that they must be S-shaped (sigmoid) or not, and without curve fitting to a pre-defined functional form.

6.2.1 Homogeneous Initial Populations

First we consider the case of a homogeneous population—no noise or confounding factors like foreign target languages. How stable are the languages in the 3-parameter system in this case? To determine this, we begin with a finite-sample analysis with \( n = 128 \) example sentences (recall by the analysis of previous chapters that learners converge to all target languages in the 3-parameter system with high probability after hearing this many sentences). Some small proportion of the children misconverge; the goal is to see whether this small proportion can drive language change—and if so, in what direction. To give the reader some idea of the possible outcomes, let us consider the four possible variations in the learning algorithm (±Single-step, ±Greedy) holding fixed the sentence distributions and size of the learning sample.

**Variation 1:** \( A = TLA \) (+Single Step, +Greedy); \( P_i = \text{Uniform} \);
Finite Sample = 128

Suppose the learning algorithm is the triggering learning algorithm (TLA). The table below shows the language mix after 30 generations. Languages are numbered from 1 to 8. Recall that +V2 refers to a language that has the verb second property, and −V2 to one that does not.
6.2. **Example 1: A Three Parameter System**

![Diagram](image)

Figure 6.1: A simple illustration of the state space for the 3-parameter syntactic case. There are 8 grammars. A probability distribution on these 8 grammars, as shown above, can be interpreted as the linguistic composition of the population. Thus, a fraction $P_1$ of the population have internalized grammar, $g_1$, and so on.

<table>
<thead>
<tr>
<th>Initial Language</th>
<th>Change to Language?</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(-V2)$ 1</td>
<td>2 (0.85), 6 (0.1)</td>
</tr>
<tr>
<td>$(+V2)$ 2</td>
<td>2 (0.98); stable</td>
</tr>
<tr>
<td>$(-V2)$ 3</td>
<td>6 (0.48), 8 (0.38)</td>
</tr>
<tr>
<td>$(+V2)$ 4</td>
<td>4 (0.86); stable</td>
</tr>
<tr>
<td>$(-V2)$ 5</td>
<td>2 (0.97)</td>
</tr>
<tr>
<td>$(+V2)$ 6</td>
<td>6 (0.92); stable</td>
</tr>
<tr>
<td>$(-V2)$ 7</td>
<td>2 (0.54), 4 (0.35)</td>
</tr>
<tr>
<td>$(+V2)$ 8</td>
<td>8 (0.97); stable</td>
</tr>
</tbody>
</table>

Table 6.1: Language change driven by misconvergence from a homogeneous initial linguistic population. A finite-sample analysis was conducted allowing each child learner 128 examples to internalize its grammar. After 30 generations, initial populations drifted (or not, as shown in the table) to different final linguistic compositions. For example, the third row shows that a homogeneous initial population consisting entirely of $L_3$ speakers evolves over many generations to one with mostly $L_6$ (48 percent) and $L_8$ (38 percent) speakers and a smattering of other speakers.
Observations. Some striking patterns regarding the resulting population mixes may be noted. We collect together some observations from preliminary simulations.

1. *All the +V2 languages are relatively stable*, i.e., the linguistic composition did not vary significantly over 30 generations. This means that every succeeding generation mostly acquired the target parameter settings—a smattering acquired alternative settings—but no significant parameter drifts were observed over time.

2. *In contrast, populations speaking −V2 languages all drift to +V2 languages.* Thus a population speaking \( L_1 \) winds up speaking mostly \( L_2 \) (85%). A population speaking language \( L_7 \) gradually shifts to a population with 54 percent speaking \( L_2 \) and 35 percent speaking \( L_4 \) (with a smattering of other speakers) and apparently remains basically stable in this mix thereafter. Note that the relative stability of +V2 languages and the tendency of −V2 languages to drift to +V2 is exactly contrary to evidence in the linguistic literature. Lightfoot (1991), for example, claims that the tendency to lose V2 dominates the reverse tendency in the world’s languages. Certainly, both English and French lost the V2 parameter setting—an empirically observed phenomenon that needs to be explained. Immediately then, we see that our dynamical system does not evolve in the expected manner. The reason could be due to any of the assumptions behind the model: the parameter space, the learning algorithm, the initial conditions, or the distributional assumptions about sentences presented to learners. Exactly which is in error remains to be seen, but nonetheless our example shows concretely how assumptions about a grammatical theory and learning theory can make evolutionary, diachronic predictions—in this case, incorrect predictions that falsify the assumptions.

3. *The rates at which the linguistic composition change vary significantly from language to language.* Consider for example the change of \( L_1 \) to \( L_2 \). Figure 6.2 shows the gradual decrease in speakers of \( L_1 \) over successive generations along with the increase in \( L_2 \) speakers. We see that over the first 6 or seven generations very little change occurs, but over the next six or seven generations the population changes at a much faster rate. Note that in this particular case the two languages differ only in the V2 parameter, so the curves essentially plot the gain of V2. In contrast, consider Fig. 6.3 which shows the decrease of \( L_5 \).
speakers and the shift to \( L_2 \). Here we note a sudden change: over a space of just 4 generations, the population shifts completely. Analysis of the time course of language change has been given some attention in linguistic analyses of diachronic syntax change, and we return to this issue later.

4. We see that in many cases a homogeneous population splits up into different linguistic groups, and seems to remain stable in that mix. In other words, certain combinations of language speakers seem to asymptote towards equilibrium (at least through 30 generations). For example, a population of \( L_7 \) speakers shifts over 5–6 generations to one with 54 percent speaking \( L_2 \) and 35 percent speaking \( L_4 \) and remains that way with no shifts in the distribution of speakers. Of course, we do not know for certain whether this is really a stable mixture. It could be that the population mix could suddenly shift after another 100 generations. What we really need to do is characterize the stable points of these dynamical systems. Other linguistic mixes can be inherently unstable; they might drift systematically to stable situations, or might shift dramatically (as with language \( L_1 \)).

5. It seems that the observed instability and drifts are to a large extent an artifact of the learning algorithm. Remember that the TLA suffers from the problem of local maxima.\(^4\) We note that those languages whose acquisition is not impeded by local maxima (the +V2 languages) are stable over time. Languages that have local maxima are unstable; in particular they drift to the local maxima over time. Consider \( L_7 \). If this is the target language, then there are two local maxima (\( L_2 \) and \( L_4 \)) and these are precisely the states to which the system drifts over time. The same is true for languages \( L_5 \) and \( L_3 \). In this respect, the behavior of \( L_1 \) is quite unusual since it actually does not have any local maxima, yet it tends to flip the V2 parameter over time.

Now let us consider a different learning algorithm from the TLA that does not suffer from local maxima problems, to see whether this changes the dynamical system results.

\(^4\)We regard local maxima of a language \( L_i \) to be alternative absorbing states (sinks) in the Markov chain for that target language. This formulation differs slightly from the conception of local maxima in Gibson and Wexler (1994), a matter discussed at some length in Niyogi and Berwick (1993). Thus according to our definition \( L_4 \) is not a local maxima for \( L_5 \) and consequently no shift is observed.
Figure 6.2: Percentage of a population (denoted as a proportion) speaking languages $L_1$ (-V2) and $L_2$ (+V2), denoted on the Y-axis, as the population evolves over some number of generations, measured on the X-axis. The plot has been shown only up to 20 generations, as the proportions of $L_1$ and $L_2$ speakers do not vary significantly thereafter. Note that this curve is “S” shaped. Kroch (1989) imposes such a shape using models from population biology, while we derive this shape as an emergent property of our dynamical model. $L_1$ and $L_2$ differ only in the V2 parameter setting. The initial condition is a homogeneous $L_1$ speaking population.
Figure 6.3: Percentage of the population speaking languages \( L_5 \) (SVO\,-\,V2) and \( L_2 \) (VOS\,+\,V2) as the population evolves over a number of generations. Note that a complete shift from \( L_5 \) to \( L_2 \) occurs over just 4 generations.
Table 6.2: Language change driven by misconvergence. A finite-sample analysis was conducted allowing each child learner (following the TLA with single-value dropped) 128 examples to internalize its grammar. Initial populations were linguistically homogeneous, and they drifted to different linguistic compositions. The major language groups after 30 generations have been listed in this table. Note how all initially homogeneous populations tend to the same composition.

**Table 6.2:** Language change driven by misconvergence.

<table>
<thead>
<tr>
<th>Initial Language</th>
<th>Change to Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-V2 \ 1$</td>
<td>2 (0.41), 4 (0.19), 6 (0.18), 8 (0.13)</td>
</tr>
<tr>
<td>$+V2 \ 2$</td>
<td>2 (0.42), 4 (0.19), 6 (0.17), 8 (0.12)</td>
</tr>
<tr>
<td>$-V2 \ 3$</td>
<td>2 (0.40), 4 (0.19), 6 (0.18), 8 (0.13)</td>
</tr>
<tr>
<td>$+V2 \ 4$</td>
<td>2 (0.41), 4 (0.19), 6 (0.18), 8 (0.13)</td>
</tr>
<tr>
<td>$-V2 \ 5$</td>
<td>2 (0.40), 4 (0.19), 6 (0.18), 8 (0.13)</td>
</tr>
<tr>
<td>$+V2 \ 6$</td>
<td>2 (0.40), 4 (0.19), 6 (0.18), 8 (0.13)</td>
</tr>
<tr>
<td>$-V2 \ 7$</td>
<td>2 (0.40), 4 (0.19), 6 (0.18), 8 (0.13)</td>
</tr>
<tr>
<td>$+V2 \ 8$</td>
<td>2 (0.40), 4 (0.19), 6 (0.18), 8 (0.13)</td>
</tr>
</tbody>
</table>

**Variation 2:** $A = +$Greedy, $-Single\ value; P_t = Uniform; Finite\ Sample = 128$

Consider a simple variant of the TLA obtained by dropping the single valued constraint. This implies that the learner is no longer constrained to change just one parameter at a time: on being presented with a sentence it cannot analyze, it chooses any of the alternative grammars and attempts to analyze the sentence with it. Greediness is retained; thus the learner retains its original hypothesis if the new one is also not able to analyze the sentence. Given this new learning algorithm, and retaining all the other original assumptions, Table 6.2 shows the distribution of speakers after 30 generations.

**Observations.** In this situation there are no local maxima, and the evolutionary pattern takes on a very different nature. There are two distinct observations to be made.

1. *All homogeneous populations eventually drift to a strikingly similar population mix, irrespective of what language they start from.* What is unique about this mix? Clearly it is a kind of stable fixed point? Is
it the only one, however? What does it depend upon as a function of linguistic and other parameters? Further simulations and theoretical analyses are needed to resolve these questions; we leave them as open at this point.

2. All homogeneous populations drift to a population mix of only + V2 languages. Thus, the V2 parameter is gradually set over succeeding generations by all people in the community (irrespective of which language they speak). In other words, as before, there is a tendency to gain V2 rather than lose V2, contrary to the empirical facts.

As an example, Fig. 6.4 shows the changing percentage of the population speaking the different languages starting off from a homogeneous population speaking L5. As before, learners who have not converged to the target in 128 examples are the driving force for change here. Note again the time evolution of the grammars. For about 5 generations there is only a slight decrease in the percentage of speakers of L5. Then the linguistic patterns switch rapidly over the next 7 generations to a relatively stable mix. Note the S-shaped nature of some (but not all) of the trajectories.

Variations 3 & 4: –Greedy, ±Single Value constraint; \( P_i = \text{Uniform; Finite Sample} = 128 \)

Having dropped the single value constraint, we consider the next obvious variation in the learning algorithm: dropping greediness while varying the single value constraint. Again, our goal is to see whether this makes any difference in the resulting dynamical system. This gives rise to two different learning algorithms: (1) allow the learning algorithm to pick any new grammar at most one parameter value away from its current hypothesis (retaining the single-value constraint, but without greediness, that is, the new grammar does not have to be able to parse the current input sentence); (2) allow the learning algorithm to pick any new grammar at each step (no matter how far away from its current hypothesis).

In both cases, the population mix after 30 generations is the same irrespective of the initial language of the homogeneous population. These results are shown in Table 6.3.

Observations:
Figure 6.4: Time evolution of grammars using a greedy learning algorithm with no single value constraint in place.

<table>
<thead>
<tr>
<th>Initial Language</th>
<th>Change to Language?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Any Language</td>
<td>1 (0.11), 2 (0.16), 3 (0.10), 4 (0.14)</td>
</tr>
<tr>
<td>(Homogeneous)</td>
<td>5 (0.12), 6 (0.14), 7 (0.10), 8 (0.13)</td>
</tr>
</tbody>
</table>

Table 6.3: Language change driven by misconvergence, using two different acquisition algorithms that do not obey a local gradient-ascent rule (a greediness constraint). A finite-sample analysis was conducted with the learning algorithm following a random-step algorithm or else a single-step algorithm, along with 128 examples to internalize its grammar. Initial populations were linguistically homogeneous, and they drifted to different linguistic compositions. The major language groups after 30 generations have been listed in this table. Note that all initially homogeneous populations converge to the same final composition.
6.2. EXAMPLE 1: A THREE PARAMETER SYSTEM

1. Both algorithms yield dynamical systems that arrive at the same population mix after 30 generations. The path by which they arrive at this mix is, however, not the same (see figure 6.5).

2. The final population mix contains all languages in significant proportion. This is in distinct contrast to the previous situations, where we saw that –V2 languages were eliminated over time.

6.2.2 Modeling Diachronic Trajectories

With some basic intuitions in hand as to how diachronic systems may evolve given different learning algorithms, we turn next to the question of population trajectories. One aspect of diachronic evolutionary trajectories that has attracted repeated attention from historical linguists is the so called S shape that they often take.

For example, Bailey (1973) proposed a "wave" model of linguistic change: linguistic replacements follow an S-shaped curve over time. In Bailey's own words (taken from Kroch, 1989):

A given change begins quite gradually; after reaching a certain point (say, twenty percent), it picks up momentum and proceeds at a much faster rate; and finally tails off slowly before reaching completion. The result is an S-curve: the statistical differences among isolects in the middle relative times of the change will be greater than the statistical differences among the early and late isolects.

While we see already that some evolutionary trajectories in our computational model may have a "linguistically classical" S-shape, the smoothness of such S-shaped trajectories may vary considerably. Crucially, however, our current formalization allows us to be much more precise than this. Unlike the previous work in diachronic linguistics that we are familiar with, we can explore systematically the space of possible trajectories and thus examine the factors that affect their evolutionary time course, without assuming an a priori S-shape.

The idea that linguistic changes follow (and ought to follow) an S-curve has also been proposed by Osgood and Sebeok (1954) and Weinreich, Labov, and Herzog (1968). More specific logistic forms have been advanced by Altmann et al (1983) and Kroch (1989). Here, the idea of a logistic functional form is borrowed from population biology where it is demonstrable that the
Figure 6.5: Time evolution of linguistic composition for the situations where the learning algorithm is -Greedy, +Single Value constraint (dotted line), and -Greedy, -Single Value (solid line). Only the percentage of people speaking $L_1$ (-V2) and $L_2$ (+V2) are shown. The initial population is homogeneous and speaks $L_1$. The percentage of $L_1$ speakers gradually decreases to about 11 percent. The percentage of $L_2$ speakers rises to about 16 percent from 0 percent. The two dynamical systems converge to the same population mix; however, their trajectories are not the exactly the same—the rates of change are a little different, as shown in this plot.
6.2. EXAMPLE 1: A THREE PARAMETER SYSTEM

logistic governs the replacement of organisms and of genetic alleles that differ in Darwinian fitness. However, Kroch (1989) concedes that “unlike in the population biology case, no mechanism of change has been proposed from which the logistic form can be deduced.”

Crucially, in our case, we suggest a specific mechanism of change: an acquisition-based model where the combination of grammatical theory, learning algorithms, and distributional assumptions on sentences drive change. The specific form might or might not be S-shaped, and might have varying rates of change. Furthermore, as we saw over the last two chapters, many different equations may all be consistent with an S-shaped trajectory and these equations may be derived from different learning algorithms.

Among the other factors that affect evolutionary trajectories are maturation time—the number of sentences available to the learner before it internalizes its adult grammar—and the probability distributions according to which such sentences are presented to the learner. We examine these in turn.

The Effect of Maturation Time or Sample Size

One obvious factor influencing the evolutionary trajectories is the maturational time, i.e., the number \(N\) of sentences the child is allowed to hear before forming its mature hypothesis. This was fixed at 128 in all the systems shown so far (based in part on our explicit computation for the Markov convergence time in this situation). Figure 6.6 shows the effect of varying \(N\) on the evolutionary trajectories. As usual, we plot only a subspace of the population. In particular, we plot the percentage of \(L_2\) speakers in the population with each succeeding generation. The initial composition of the population was taken to be homogeneous (with all adult language users speaking \(L_1\)).

Observations.

\(^5\)Of course, we do not mean to say that we can simulate any possible trajectory—that would make the formalism empty. Rather, we are exploring the initial space of possible trajectories, given some example initial conditions that have been already advanced in the literature. Because the mathematics for dynamical systems is in general quite complex, at present we cannot make general statements of the form, “under these particular initial conditions the trajectory will be sigmoidal, and under these other conditions it will not be.” We have conducted only very preliminary investigations demonstrating that potentially at least, reasonable, distinct initial conditions can lead to demonstrably different trajectories.
1. The initial rate of change of the population is highest when the maturation time is smallest, i.e., the learner is allowed the least amount of time to develop its mature hypothesis. This is not surprising. If the learner were allowed access to a lot of examples to make its mature hypothesis, most learners would reach the target grammar. Very few would misconverge, and the linguistic composition would change little over the next generation. On the other hand, if the learner were allowed very few examples to develop its hypothesis, many would misconverge, possibly causing great change over one generation.

2. The “stable” linguistic compositions seem to depend upon maturation time. For example, if learners are allowed only 8 examples, the percentage of $L_2$ speakers rises quickly to about 0.26. On the other hand, if learners are allowed 128 examples, the percentage of $L_2$ speakers eventually rises to about 0.41.

3. Note that the trajectories do not have an S-shaped curve. The proportion of $L_2$ speakers rises beyond its stable value before falling again and reaching the equilibrium position. Part of this is surely the consequence of plotting one variable of what is essentially a multivariable dynamical system. There are two further remarks that are worth making in this context. First, recall the many different examples of two-language models that we considered in the previous chapter. For some parameter values, the change observed was S-shaped, for others it was not. Therefore, the fact that linguistic trajectories in general need not be S-shaped ought not to come as a surprise in the $n$-dimensional setting. Second, when linguists study historical change they focus usually on one or two most salient (fastest changing) linguistic parameters. The S-shaped curves referred to in the historical linguistics literature are usually plots of one parameter in what is always a multidimensional evolutionary system. In this sense, our one-dimensional plots derived from multidimensional systems takes on an added touch of realism.

4. The maturation time is related to the order of the dynamical system. In particular, following our discussion on learning models in previous chapters, it is easy to see that if $N$ is the maturation time, then the dynamical systems correspond to $N$th degree polynomial maps.
### 6.2. Example 1: A Three Parameter System

![Graph showing time evolution of linguistic composition](image)

**Figure 6.6:** Time evolution of linguistic composition when varying maturation time (sample size). The learning algorithm used is the +Greedy, −Single value. Only the percentage of people speaking \( L_2 \) (+V2) is shown. The initial population is homogeneous and speaks \( L_1 \). The maturation time was varied through 8, 16, 32, 64, 128, and 256, giving rise to the six curves shown. The curve with the highest initial rate of change corresponds to 8 examples for maturation time. The initial rate of change decreases as the maturation time \( N \) increases. The value at which these curves asymptote also seems to vary with the maturation time, and increases monotonically with it.
CHAPTER 6. LANGUAGE CHANGE - N LANGUAGES

The Effect of Sentence Distributions \( (P_i) \)

Another important factor influencing evolutionary trajectories is the distribution \( P_i \) with which sentences of the \( i \)th language, \( L_i \), are presented to the learner. In a certain sense, the grammatical space and the learning algorithm jointly determine the order of the dynamical system. On the other hand, sentence distributions are much like the parameters of the dynamical system (see Sec. 6.2.3). Clearly the sentence distributions affect rates of convergence within one generation. Further, by putting greater weight on certain word forms rather than others, they might influence systemic evolution in certain directions. We have already encountered the possibly subtle effects that sentence distributions might have in the two language models of Chapter 5. We saw that not only did the numerical value of the fixed points change with sentence distributions, there may be bifurcations leading to a qualitative change in the dynamics altogether.

To illustrate the idea in the multidimensional setting, consider an example focusing on the interaction between \( L_1 \) and \( L_2 \) speakers in the community as the sentence distributions with which those speakers produce sentences change. Recall that so far in this chapter, we have assumed that all speakers produce sentences with uniform distributions on degree-0 sentences of their respective languages. Now we consider alternative distributions parameterized by a value \( p \):

1. Let \( L_{1,2} = L_1 \cap L_2 \).

2. \( P_1 \) : Speakers of \( L_1 \) produce sentences so that all degree-0 sentences of \( L_{1,2} \) are equally likely and their total probability is \( p \). Further, sentences of \( L_1 \setminus L_{1,2} \) are also equally likely, but their total probability is \( 1 - p \).

3. \( P_2 \) : Speakers of \( L_2 \) produce sentences so that all degree-0 sentences of \( L_{1,2} \) are equally likely and their total probability is \( p \). Further, sentences of \( L_2 \setminus L_{1,2} \) are also equally likely, but their total probability is \( 1 - p \).

4. Other \( P_i \)'s are all uniform over degree-0 sentences.

The parameter \( p \) determines the weight on the sentence patterns in common between the languages \( L_1 \) and \( L_2 \). Figure 6.7 shows the evolution of the \( L_2 \) speakers as \( p \) varies. Here the learning algorithm is +Greedy, +Single
6.2. EXAMPLE 1: A THREE PARAMETER SYSTEM

value (TLA, or local gradient ascent) and the initial population is homogeneous, 100% $L_1$; 0% $L_2$. Note that the system moves in different ways as $p$ varies. When $p$ is very small (0.05), that is, sentences common to $L_1$ and $L_2$ occur infrequently, we find that in the long run the percentage of $L_2$ speakers does not increase; the population stays put with mostly $L_1$ speakers. However, as $p$ grows, more strings of $L_2$ occur, and the dynamical system changes so that the long-term percentage of $L_1$ speakers decreases and that of $L_2$ speakers increases. When $p$ reaches 0.75 the initial population evolves into a completely $L_2$ speaking community. After this, as $p$ increases further, we notice (see $p = 0.95$) that the $L_2$ speakers increase but can never rise to 100 percent of the population; there is still a residual $L_1$ speaking component. This is to be expected, because for such high values of $p$, many strings common to $L_1$ and $L_2$ occur frequently. This means that a learner could sometimes converge to $L_1$ just as well as $L_2$, and some learners indeed begin to do so, increasing the number of the $L_1$ speakers.

This example shows us that if we wanted a homogeneous $L_1$ speaking population to move to a homogeneous $L_2$ speaking population, by choosing our distributions appropriately, we could drive the grammatical dynamical system in the appropriate direction. It suggests another important application of the dynamical system approach: one can work backwards, and examine the conditions needed to generate a change of a certain kind. By checking whether such conditions could have possibly existed historically, we can falsify a grammatical theory or a learning paradigm. Note that this example showed the effect of sentence distributions, and how to alter them to obtain desired evolutionary envelopes. One could, in principle, alter the grammatical theory or the learning algorithm in the same fashion—leading to a tool to aid the search for an adequate linguistic theory.\(^6\)

6.2.3 Nonhomogeneous Populations: Phase-Space Plots

For our three-parameter system, we have been able to characterize the update rules for the dynamical systems corresponding to a variety of learning algorithms. Each dynamical system has a specific update procedure according to which the states evolve from some homogeneous initial population. A more complete characterization of the dynamical system would be achieved by obtaining phase-space plots of this system. Such phase-space plots are

\(^6\)Again, we stress that we obviously do not want so weak a theory that we can arrive at \textit{any} possible initial conditions simply by carrying out reasonable changes to the sentence distributions. This may, of course, be possible; we have not yet examined the general case.
Figure 6.7: The evolution of $L_2$ speakers in the community for various values of $p$ (a parameter related to the sentence distributions $P_i$, see text). The algorithm used was the TLA, the initial population was homogeneous, speaking only $L_1$. The curves for $p = 0.05, 0.75,$ and $0.95$ have been plotted as solid lines.
pictures of the state-space $S$ filled with trajectories obtained by letting the system evolve from various initial points (states) in the state space.

**Phase-Space Plots: Grammatical Trajectories**

We have described earlier the relationship between the state of the population in one generation and the next. In our case, let $\Pi$ denote an 8-dimensional vector variable (state variable). Specifically, $\Pi = (\pi_1, \ldots, \pi_8)'$ (with $\pi_i \geq 0$ and $\sum_{i=1}^{8} \pi_i = 1$, i.e., $\Pi \in \Delta(7)$ where $\Delta(7)$ is the 7-dimensional simplex) as we discussed before. The following schema reiterates the chain of dependencies involved in the update rule governing system evolution. The state of the population at time $t$ (in generations), allows us to compute the transition matrix $T$ for the Markov chain associated with the memoryless learner. Now, depending upon whether we want (1) an asymptotic analysis or (2) a finite sample analysis, we compute (1) the limiting behavior of $T^m$ as $m$ (the number of examples) goes to infinity (for an asymptotic analysis), or (2) the value of $T^N$ (where $N$ is the number of examples after which maturation occurs). This allows us to compute the next state of the population. Thus $\Pi(t + 1) = g(\Pi(t))$ where $g$ is a complex non-linear relation.

$$\Pi(t) \Longrightarrow P \; \Sigma^* \Longrightarrow T \Longrightarrow T^m \Longrightarrow \Pi(t + 1)$$

If we choose a certain initial condition $\Pi_1$, the system will evolve according to the above relation and one can obtain a trajectory of $\Pi$ in the 8 dimensional space over time. Each initial condition yields a unique trajectory and one can then plot these trajectories obtaining a phase-space plot. Each such trajectory corresponds to a curve in the seven dimensional hyperplane (simplex) given by $\sum_{i=1}^{8} \pi_i = 1$. One cannot directly display such a high dimensional object, but we plot in Fig. 6.8 the projection of a particular trajectory onto a two dimensional subspace given by $(\pi_1(t), \pi_2(t))$ (the proportion of speakers of $L_1$ and $L_2$) at different points in time.

As mentioned earlier, with a different initial condition we get a different grammatical trajectory. The complete state space picture is thus filled with all the different trajectories corresponding to different initial conditions. They all seem to be converging to the same fixed point. Fig. 6.9 shows this.

**Stability Issues**

The phase-space plots show that many initial conditions yield trajectories that seem to converge to a single point in the state space. In the dynamical
Figure 6.8: Subspace of a phase-space plot. The plot shows $(\pi_1(t), \pi_2(t))$ as $t$ varies, i.e., the proportion of speakers speaking languages $L_1$ and $L_2$ in the population. The initial state of the population was homogeneous (speaking language $L_1$). The algorithm used was +Greedy−Single value.
Figure 6.9: Subspace of a Phase-space plot. The plot shows \((\pi_1(t), \pi_2(t))\) as \(t\) varies for different nonhomogeneous initial population conditions. The algorithm used was +Greedy −Single value.
systems terminology, this corresponds to a stable fixed point of the system—a population mix that stays at the same composition. Many natural questions arise at this stage. What are the conditions for stability? How many fixed points are there in a given linguistic system? How can we solve for them? These are interesting questions but detailed answers are not within the scope of the current chapter. We will provide partial answers in certain contexts over the course of the rest of this book. For the time being, in lieu of a more complete analysis, let us first consider at least the equations that allow one to characterize the stable population mixes.

First, some notational preliminaries. Consider a memoryless learning algorithm. As before, let \( P_i \) be the distribution on the sentences of the \( i \)th language \( L_i \). From \( P_i \), we can construct \( T_i \), the transition matrix whose elements are given by the explicit procedure documented in previous chapters. The matrix \( T_i \) characterizes the Markov development of a memoryless learner when the target language is \( L_i \) (and sentences from the target are produced according to \( P_i \)). Note that fixing the \( P_i \)'s fixes the \( T_i \)'s and thus the \( P_i \)'s are a different sort of “parameter” that characterize how the dynamical system evolves.\(^7\)

Now let the state of the parent population at time \( t \) be \( \Pi(t) \). Therefore the probability distribution according to which the learner is exposed to example sentences is given by

\[
P = \sum_{i=1}^{8} \pi_i(t) P_i
\] (6.5)

Recall that the memoryless learner chooses hypotheses depending only upon its previously held grammatical hypothesis and the newly available example sentence. In this sense, one may usefully characterize the memoryless learner as a computable mapping

\[
\mathcal{A} : \mathcal{H} \cup \Sigma^* \rightarrow \mathcal{H}
\]

where \( \mathcal{A}(h, s) \) determines what the algorithm will hypothesize next if it currently held the hypothesis \( h \in \mathcal{H} \) and sentence \( s \in \Sigma^* \) was presented to it.

\(^7\)There are thus two distinct kinds of parameters in our model: first, parameters that define the \( 2^n \) languages and define the state-space of the system; and second, the \( P_i \)'s that characterize the way in which the system evolves and are therefore the parameters of the complete grammatical dynamical system.
6.2. EXAMPLE 1: A THREE PARAMETER SYSTEM

The transition matrix $T$ under the influence of examples presented according to $P$ may then be easily derived. Recall that

$$T_{ij} = \mathbb{P}[h_i \rightarrow h_j] = \sum_{s \in A} P(s)$$

where $A = \{ s \in \Sigma^* | \mathcal{A}(h_i, s) = h_j \}$ is the set of all sentences on which the learning algorithm would change its hypothesis\(^8\) from $h_i$ to $h_j$. However, using Eq. 6.5, we see that $T_{ij}$ is simply

$$= \sum_{s \in A} \sum_{k=1}^{8} \pi_k(t) P_k(s) = \sum_{k=1}^{8} \pi_k(t) \sum_{s \in A} P_k(s) = \sum_{k=1}^{8} \pi_k(t) T_k(i, j)$$

where $T_k(i, j)$ refers to the $i, j$ element of the matrix $T_k$. The matrix $T_k$ is the transition matrix of the Markov chain characterizing the behavior of the learner when the target language is $L_k$ and therefore example sentences are drawn according to $P_k$.

Since $T$ can be expressed in terms of $\pi_i$'s and $T_i$'s, we have the following statements.

**Statement 1 (Finite Case)** A fixed point of the grammatical dynamical system (derived from a $\pm$ Greedy $\pm$ Single value learner operating on the 8 parameter space with $k$ examples to choose its final hypothesis) is a solution of the following equation:

$$\Pi' = (\pi_1, \ldots, \pi_8) = (1/8, \ldots, 1/8)' \left( \sum_{i=1}^{8} \pi_i T_i \right)^k$$

**Note:** This equation is obtained simply by setting $\Pi(t+1) = \Pi(t)$. Note however, that this is an example of a nonlinear multidimensional iterated function map. The complete analysis of such dynamical systems is nontrivial and beyond the scope of the current chapter.

Similarly, for the limiting (asymptotic) case, the following holds:

---

\(^8\) Strictly speaking, the memoryless algorithms we have been considering in this example are all variants of the TLA. Such algorithms are not deterministic but randomized for which the transition matrix formulae were derived in earlier chapters on learning. While the derivation presented here holds only for deterministic memoryless algorithms an extension to the randomized case is easily achieved. We omit such a derivation for expository purposes though the reader may check the validity of the extension.
Statement 2 (Limiting or Asymptotic Analysis) A fixed point of the grammatical dynamical system (derived from a ±Greedy ±Single value learner operating on the 8 parameter space with infinite examples to choose its mature hypothesis) is a solution of the following equation:

$$\Pi' = (\pi_1, \ldots, \pi_8) = (1, \ldots, 1)\left(I - \sum_{i=1}^{8} \pi_i T_i + ONE\right)^{-1}$$

where $ONE$ is the $8 \times 8$ matrix with all its entries equal to 1.

Note: Again this is trivially obtained by setting $\Pi(t + 1) = \Pi(t)$. The expression on the right provides an analytical expression for the update equation in the asymptotic case. See Resnick (1992) for details. All the caveats mentioned before in the Finite case statement apply here as well.

Remark. We have barely scratched the surface as far as the theoretical characterization of these grammatical dynamical systems is concerned. The main purpose of this chapter is to continue the argument that these dynamical systems exist as a logical consequence of assumptions about a grammatical space and an acquisition theory. We have demonstrated only some preliminary simulations with these higher dimensional systems. From a theoretical perspective, it would be much more valuable to have complete characterizations of such systems. Because the systems described above are multidimensional and non-linear, the possibility exists for multiple stable and unstable equilibria, as well as more complicated behavior such as cycles, bifurcations, and ultimately chaos. Keep in mind, however, that some of these regimes may not be entered because the matrices $T_i$ are constrained to be stochastic matrices. This is analogous to the one-dimensional models we examined in the previous chapter where the corresponding dynamical systems did not enter the chaotic regime because the parameters $a, b$ were bounded.

6.3 Example 2: Syntactic Change in French — Revisiting Clark and Roberts (1993)

So far, our examples have been based on a 3-parameter linguistic theory for which we derived several different dynamical systems. Our goal was to concretely instantiate our philosophical arguments, sketching the factors
that influence evolutionary trajectories. In this section, we briefly consider a different parametric linguistic system studied by Clark and Roberts, 1993. The historical context in which Clark and Roberts advanced their linguistic proposal is the evolution of Modern French from Old French. Their parameters are intended to capture some, but of course not all, of this change. They too use a learning algorithm—in their case, a genetic algorithm—to account for historical change but do not analyze their model from the dynamical systems viewpoint. Here we adopt their parameterization, with all its strengths and weaknesses, but consider an alternative learning paradigm and the dynamical systems approach.

Extensive simulations in the earlier section reveal that while the learnability problem of the 3-parameter space can be solved by stochastic hill climbing algorithms, the long term evolution of these algorithms have a behavior that is at variance with the diachronic change actually observed in historical linguistics. In particular, we saw how there was a tendency to gain rather than lose the V2 parameter setting. While this could well be an artifact of the class of learning algorithms considered, a more likely explanation is that loss of V2 (observed in many of the world’s languages like French, English, and so forth) is due to an interaction of parameters and triggers other than those considered in the previous section. We investigate this possibility and begin by first reviewing Clark and Roberts’ alternative parametric theory.

\subsection{The Parametric Subspace and Data}

We now consider a syntactic space requiring 5 (boolean-valued) parameters. We do not attempt to describe these parameters. The interested reader should consult Haegeman (1991) as well as Clark and Roberts (1993) for details.

1. $p_1$: Case assignment under agreement ($p_1 = 1$) or not ($p_1 = 0$).

2. $p_2$: Case assignment under government ($p_2 = 1$) or not ($p_2 = 0$).

   Relevant triggers for this parameter include “Adv V S”, “S V O”.


4. $p_4$: Null Subject. Here relevant triggers would include “wh V S O”.

5. $p_5$: Verb-second V2. Triggers include “Adv V S”, and “S V O”.


Table 6.4: A list of the key sentences that serve as triggers to drive learning in
the five parameter system of Clark and Roberts (1993). Parameter settings
are indicated in brackets. An asterisk denotes that the relevant sentence is
parsable under both settings of the relevant parameter. An X or Y denotes
a placeholder that may be occupied by any phrase. wh refers to a question
word like who, where and so on. V refers to verb, S to subject, O to object,
pro to pronoun, s to subject clitic and adv to adverb.

These 5 parameters define a space with 32 grammars. Each grammar in
this parametrized system can be represented by a string of 5 bits depending
upon the values of \( p_1, \ldots, p_5 \). For instance, the first bit position corresponds
to case assignment under agreement. We can now look at the surface strings
(sentences) generated by each such grammar. For the purpose of explaining
how Old French changed to Modern French, Clark and Roberts consider a
list of key sentences provided in Table 6.4. The parameter settings required
to generate each sentence are provided in brackets; an asterisk is a “doesn’t
matter” value and an “X” means any phrase.

The parameter settings provided in brackets represent the grammars that
generate the sentence. For example, the sentence form “adv V S” (corre-
sponding to *quickly ran John* — an incorrect word order in English) is
generated by all grammars that have case assignment under government (the
second element of the array set to 1, \( p_2 = 1 \)) and verb second movement
(\( p_5 = 1 \)). The other parameters can be set to any value. Clearly there are
8 different grammars that can generate (alternatively parse) this sentence.
Similarly there are 16 grammars that generate the form S V O (8 corre-
sponding to parameter settings of \([1**1] \) and 8 corresponding to parameter
settings of [1***0]) and 4 grammars that generate ((s V Y).

Remark. Note that the set of sentences Clark and Roberts consider is only a subset of the total number of degree-0 sentences generated by the 32 grammars in question. In order to directly compare their model with ours, we have not attempted to expand the data set or fill out the space any further. As a result, the grammars do not all have unique extensional properties, i.e., some pairs of grammars generate the same set of sentences and are weakly equivalent in this setting.

6.3.2 The Case of Diachronic Syntactic Change in French

Let us continue with the analysis of Clark and Roberts within this parameter space. Historical data, primarily in the form of written texts, suggest that the language spoken in France underwent a parametric change from the twelfth century A.D. to modern times. In particular, it has been observed that both V2 and prodrop are lost. Examples illustrating this change are provided below.

Loss of null subjects: pro-drop

(Old French; +pro drop)
Si firent (pro) grant joie la nuit
‘thus (they) made great joy the night’

(Modern French; −pro drop)
(*) Ainsi s'amusaient bien cette nuit
‘thus (they) had fun that night’

Loss of V2

(Old French; +V2)
Lors oirent ils venir un escoiz de tonoire
‘then they heard come a clap of thunder’

(Modern French; −V2)
(*) Puis entendirent-ils un coup de tonerre.
‘then they heard a clap of thunder’

Clark and Roberts (1993) observe that it has been argued that this transition was brought about by the introduction of new word orders during the fifteenth and sixteenth centuries resulting in generations of children acquiring slightly different grammars and eventually culminating in the grammar
of modern French. A brief reconstruction of the historical process into three
basic periods (after Clark and Roberts, 1993) is as follows.

**Old French; setting [11011]** The language spoken in the twelfth and
thirteenth centuries had verb-second movement and null subjects, both of
which were dropped by the twentieth century. Of particular interest in
the analysis of Clark and Roberts are some of the sentence types that are
generated by the parameter settings corresponding to Old French (Table 6.5).

Note that from this data set it appears that the parameters corresponding
to Case agreement and nominative clitics remain ambiguous. In particular,
Old French is in a subset-superset relation with another language (generated
by the parameter settings of 11111). In this case, possibly some kind of
subset principle (Berwick, 1985) could be used by the learner; otherwise it is
not clear how the data would allow the learner to converge to the Old French
grammar in the first place. None of the ±Greedy, ±Single value algorithms
would converge uniquely to the grammar of Old French.

The string (X)VS occurs with frequency 58% and SV(X) occurs with
34% in Old French texts. It is argued that this frequency of (X)VS is high
enough to cause the V2 parameter to trigger to +V2.

**Middle French** In Middle French, the data is not consistent with any of
the 32 target grammars. This corresponds therefore to a heterogeneous
population speaking a mixture of the different grammars giving rise to all
the different data types. Analysis of texts from that period reveal that some
old forms (like Adv V S) decreased in frequency and new forms (like Adv S
V) increased. It is argued in Clark and Roberts that such a frequency shift
causes "erosion" of V2, brings about parameter instability and ultimately
convergence to the grammar of Modern French. In this transition period
(i.e. when Middle French was spoken/written) the data is of the form shown
in Table 6.6.

<table>
<thead>
<tr>
<th>Old French</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>adv V S –</td>
<td>[*1**1]</td>
</tr>
<tr>
<td>S V O –</td>
<td>[*1**1]</td>
</tr>
<tr>
<td>wh V S O</td>
<td>[*1**1]</td>
</tr>
<tr>
<td>X (pro) V O</td>
<td>[*1**1]</td>
</tr>
</tbody>
</table>

Table 6.5: Sentence types corresponding to the parameter settings of Old
6.3. EXAMPLE 2: SYNTACTIC CHANGE IN FRENCH — REVISITING CLARK AND ROBERTS (1993)

<table>
<thead>
<tr>
<th>Middle French</th>
</tr>
</thead>
<tbody>
<tr>
<td>adv V S</td>
</tr>
<tr>
<td>SVO</td>
</tr>
<tr>
<td>wh V S O</td>
</tr>
<tr>
<td>wh V s O</td>
</tr>
<tr>
<td>X (pro)V O</td>
</tr>
<tr>
<td>X V s</td>
</tr>
<tr>
<td>X s V</td>
</tr>
<tr>
<td>X S V</td>
</tr>
<tr>
<td>(s)VY</td>
</tr>
</tbody>
</table>

Table 6.6: A listing of sentence types in the Middle French period.

Thus, we have old sentence patterns like Adv V S (though it decreases in frequency and becomes only 10%), SVO, X (pro)V O and whVSO. The new sentence patterns which emerge at this stage are adv S V (increases in frequency to become 60%), X subjclitic V, V subjclitic (pro)V Y (null subjects), whV subjclitic O.

Modern French [10100] By the eighteenth century, French had lost both the V2 parameter setting as well as the null subject parameter setting. The sentence patterns consistent with Modern French parameter settings are SVO [*1**1] or [1***0], X S V [1***0], V s O [**1**]. Note that this data, though consistent with Modern French, will not trigger all the parameter settings. In this sense, Modern French (just like Old French) is not uniquely learnable from data. However, as before, we shall not concern ourselves overly with this, for the relevant parameters (V2 and null subject) are uniquely set by the data here.

6.3.3 Some Dynamical System Simulations

Using a TLA like learning algorithm as a starting point, we can again obtain a dynamical system characterization for the linguistic population in this parametric space. For illustrative purposes, we show the results of two simulations conducted with such dynamical systems.
Homogeneous Populations [Initial—Old French]

Let us conduct a simulation on this new parameter space using the Triggering Learning Algorithm. Recall that the relevant Markov chain in this case has 32 states. Our goal will be to see if misconvergence by learners can be a sufficient reason to drive a population from speakers of Old French to speakers of Modern French. Consequently, we start the simulation with a homogeneous population speaking Old French (parameter setting = 11011).

Just as before, we can observe the linguistic composition of the population over several generations. It is observed that in one generation, 15 percent of the children converge to grammar 01011; 18 percent to grammar 01111; 33 percent to grammar 11011 (target) and 26 percent to grammar 11111 with very few having converged to other grammars. Thereafter, the population consists mostly of speakers of these 4 languages, with one important difference: 15 percent of the speakers eventually lose V2. In particular, they have acquired the grammar 11110. Fig. 6.10 shows the percentage of the population speaking each of the 4 languages mentioned above as the population evolves over 20 generations. Notice that in the space of a few generations, the speakers of 11011, and 01011 have dropped out altogether. Most of the population now speaks language 1111 (46 percent) and 01111 (27 percent). Fifteen percent of the population speaks 11110 and there is a smattering of other speakers. The population remains roughly stable in this configuration thereafter.

Observations:

1. On examining the four languages to which the system converges after one generation, we notice that they share the same settings for the principles [Case assignment under government], [pro drop], and [V2]. These correspond to the three parameters which are uniquely set by data from Old French. The other two parameters can take on any value. Consequently 4 languages are generated all of which satisfy the data from Old French.

2. Recall our earlier remark that due to insufficient data, there were equivalent grammars in the parameter system. It turns out that in this particular case, the grammars (01011) and (11011) are identical as far as their extensional properties are concerned; as are the grammars (11111) and (01111). Thus, extensionally, there are only two languages.
Figure 6.10: Evolution of speakers of different languages in a population starting off with speakers only of Old French.
3. There is a subset relation between the two sets described in (2). The grammar \((11011)\) is in a subset relation with \((11111)\). This explains why after a few generations most of the population switches to either \((11111)\) or \((01111)\) (the superset grammars).

4. An interesting feature of the simulation is that 15 percent of the population eventually acquires the grammar \((11110)\), i.e., they have lost the V2 parameter setting. This is the only sign of instability of V2 that is apparent from our simulations so far (for greedy algorithms which are psychologically preferred). Recall that for such algorithms, the V2 parameter was very stable in the previously conducted simulations using a three parameter system. This suggests that an explanation for loss of V2 may lie in the structure of UG and the interaction of the parameters that make up UG.

**Heterogenous Populations (Mixtures)**

The earlier section showed that with no new (foreign) sentence patterns the grammatical system starting out with only Old French speakers showed some tendency to lose V2. However, the grammatical trajectory did not terminate in Modern French. In order to more closely duplicate this historically observed trajectory, we examine alternative initial conditions. We start our simulations with an initial condition which is a mixture of two sources; data from Old French and data from New French (reproducing in this sense, data similar to that obtained from the Middle French period). Thus children in the next generation observe new surface forms. Most of the surface forms observed in Middle French are covered by this mixture.

**Observations:**

1. On performing the simulations using the TLA as a learning algorithm on this parameter space, an interesting pattern is observed. Suppose the learner is exposed to sentences where 90 percent are generated by the grammar of Old French \((11011)\) and 10 percent by the grammar of Modern French \((10100)\). We find that within one generation, 22 percent of the learners have converged to the grammar \((11110)\) and 78 percent to the grammar \((11111)\). Thus the learners set each of the parameter values to 1 except the V2 parameter setting. Now Modern French is a non-V2 language; and 10 percent of data from Modern French is sufficient to cause 22 percent of the speakers to lose V2. This is the behavior over one generation. The new population (consisting of
6.3. **EXAMPLE 2: SYNTACTIC CHANGE IN FRENCH — REVISITING CLARK AND ROBERTS (1993)**

78 percent speaking grammar (11111) and 22 percent speaking grammar (11110) remains stable for ever.

2. Fig. 6.11 shows the proportion of speakers who have lost V2 after one generation, as a function of the proportion of sentences from the Modern French Source. The shape of the curve is interesting. For small values of the proportion of the Modern French source, the slope of the curve is greater than 1. Thus there is a greater tendency of speakers to lose V2 than to retain it. Thus 10 percent of novel sentences from the Modern French source causes 20 percent of the population to lose V2; similarly 20 percent of novel sentences from the Modern French source causes 40 percent of the speakers to lose V2. This effect wears off later. This seems to capture computationally the intuitive notion of many linguists that a small change in inputs provided to children could drive the system towards larger change.

3. Unfortunately, there are several shortcomings of this particular simulation. First, we notice that mixing Old and Modern French sources does not cause the desired (historically observed) grammatical trajectory from Old to Modern French (corresponding in our system to movement from state (11011) to state (10100) in the Markov Chain). Although we find that a small injection of sentences from Modern French causes a larger percentage of the population to lose V2 and gain subject clitics (which are historically observed phenomena), nevertheless, the entire population retains the null subject setting and case assignment under government. It should be mentioned that Clark and Roberts argue that the change in case assignment under government is the driving force which allows alternate parse-trees to be formed and causes the parametric loss of V2 and null subject. In this sense, it is a more fundamental change.

4. If the dynamical system is allowed to evolve, it ends up in either of the two states (11111) or (11110). This is essentially due to the subset relations these states (languages) have with other languages in the system. Another complication in the system is the equivalence of several different grammars (with respect to their surface extensions) e.g. given the data we are considering, the grammars (01011) and (11011) (Old French) generate the same sentences. This leads to multiplicity of paths, convergence to more than one target grammar and general inelegance of the state-space description.
Figure 6.11: Tendency to lose V2 as a result of new word orders introduced by the Modern French source in the population dynamics resulting from a TLA based Markov learner.
6.3. EXAMPLE 2: SYNTACTIC CHANGE IN FRENCH — REVISITING CLARK AND ROBERTS (199325)

General Insights and Future Directions

We have considered two different scenarios for the plausible change of a population of Old French speakers to one of Modern French speakers. One scenario embodies the hypothesis that the change is internally driven with misconverging speakers changing the linguistic composition over generational time. Our simulations with initial conditions corresponding to homogeneous populations were directed at exploring this hypothesis. While some signs of instability of the V2 parameter are observed, only about 15 percent of the population loses this parameter over time. A second scenario follows the hypothesis that the change is driven by language contact with two different linguistic types in contact with each other and the population drifting gradually from one to the other. Again, we see that small injections of Modern French speakers in the population may have an effect that is disproportionate to their numbers in the first few generations. This effect however decreases over generational time so that eventually only 22 percent have lost V2. This population mix remains stable thereafter.

Thus neither of the two simulations are able to replicate the change from Old to Modern French. These simulations are used as an aid in our reasoning for the consequences of the two above mentioned scenarios are difficult to deduce by verbal arguments alone. Furthermore, the computational framework compels the historical linguist to articulate the explanations for change in a precise manner. The very fact that these two precisely articulated scenarios fail to reproduce the change demonstrate the power of this framework to falsify potential explanations.

One may now consider several possible reasons for the failure to explain the complete change from Old to Modern French. First, using more data and filling out the state-space might yield greater insight. Second, TLA-like hill climbing algorithms do not pay attention to the subset principle explicitly. Given the number of subset-superset relations among grammars in the 5 parameter space, it would be interesting to explicitly program this into the learning algorithm and observe the evolution thereafter. Third, there are often cases when several different grammars generate the same sentences or at least equally well fit the data. Algorithms which operate only on surface strings are unable then to distinguish between these grammars. As a result one finds convergence to all of the weakly equivalent grammars with different probabilities in our stochastic setting. We saw an example of this for convergence to the four states earlier. Clark and Roberts (1993) suggest an elegance criterion by looking at the parse-trees to decide between
these grammars. This difference between strong generative capacity and weak generative capacity can easily be incorporated into the model and its consequences examined more thoroughly. Transition probabilities, now, will not depend upon the surface properties of the grammars alone, but also upon the elegance of derivation for each surface string. Finally, rather than the evolution of the population, one could look at the evolution of the distribution of sentence types. One can also obtain bounds on the frequencies with which the new data in the Middle French Period must occur so that the correct drift is observed.

We do not explore any of these directions here but list them as examples of meaningful questions that arise largely due to the application of computational thinking to the historical problem at hand. As one explores these questions further, a more nuanced understanding of the reasons behind the change will doubtless emerge.

6.4 Conclusions

In this chapter, we have continued our exploration of the relationship between language learning and language change. The central argument of this book has been that any specification of (i) linguistic theory as articulated by models of generative grammar (broadly construed) and (ii) learning theory as articulated by models of how those grammars are attained by learning children — leads logically to models of grammatical evolution and diachronic change. These models of grammatical evolution therefore represent the evolutionary consequences of models of linguistic theory and language learning.

From a programmatic perspective, this argument has two important consequences. First, it allows us to take a formal, analytic view of historical linguistics. Most accounts of language change have tended to be descriptive in nature. In contrast, we place the study of historical or diachronic linguistics in a formal framework. In this sense, our conception of historical linguistics is closest in spirit to evolutionary theory and population biology.

Second, this approach allows us to formally pose a diachronic criterion for the adequacy of grammatical theories. A significant body of work in learning theory, has already sharpened the learnability criterion for grammatical theories—in other words, the class of grammars \( G \) must be learnable by some psychologically plausible algorithm from primary linguistic data. Now we can go one step further. The class of grammars \( G \) (along with a proposed learning algorithm \( A \)) can be reduced to a dynamical system
6.4. CONCLUSIONS

whose evolution must be consistent with that of the true evolution of human languages (as reconstructed from historical data).

In Chapter 5, we considered one-parameter models of language change. In this chapter, we considered the general \( n \)-language setting where \( n \) different linguistic types are potentially present in the population at all times. To concretely demonstrate that the grammatical dynamical systems need not be impossibly difficult to compute (or simulate), we explicitly showed how to transform parameterized theories, and memoryless learning algorithms to dynamical systems.

The specific simulations in this chapter have been conducted with the goal of generating some insight into syntactic change and we considered in some detail the case of syntactic change in French from the twelfth to the twentieth century. While the simulations conducted here are too incomplete to have any long term linguistic implications, we hope, it certainly forms a starting point for research in this direction. Some interesting insights obtained along the way merit reiteration. We list these below.

1. We were able to shed some light on the time course (the S-shape) of evolution. In particular, we saw how this was a derivative of more fundamental assumptions about initial population conditions, sentence distributions, and learning algorithms.

2. We were able to formally develop notions of system stability. Thus, certain linguistic parameters could change with time, others might remain stable. This can now be measured, and the conditions for stability or change can be investigated. In Chapter 5, we saw analytically how the stability of a system might change as a result of drift in frequencies. These correspond to bifurcations and provide an important theoretical construct for explaining language change and evolution.

3. We were able to demonstrate how one could manipulate the system (by changing the algorithm, or the sentence distributions, or maturational time) to allow evolution in certain directions. These logical possibilities suggest the kinds of modifications needed in the linguistic theory for greater explanatory adequacy.

4. In the case study of French, we saw that the V2 parameter was more stable in the 3-parameter case, than it was in the 5 parameter case. This suggests that the explanation for the loss of V2 (actually observed in history) may reside in the nature of the parameters available to the
learning child, i.e., the structure of $G$ (though we must suggest great caution before drawing strong conclusions on the basis of this study).

Now that the basic framework has been developed in some generality, in the next few chapters we will apply this kind of thinking to particular cases of language change studied in historical linguistics. This exercise will allow us to make issues concrete by grounding our models in real cases. More generally, it will also allow us to better understand the role of such models in reasoning about evolutionary and historical phenomena in linguistic populations.