Chapter 1

Inferring Leadership Structure From Data on a Syntax Change in English

by W. Garrett Mitchener

Abstract: In a typical human population, some features of the language are bound to be in flux. Variation in each individual's usage rates of optional features reflects language change in progress. Sociolinguistic surveys have determined that some individuals use new features to a greater degree than the population average, that is, they seem to be leading the change. This article describes a mathematical model of the spread of language change inspired by a model from population genetics. It incorporates the premise that some individuals are linguistic leaders and exert more influence on the speech of learning children than others. Using historical data from the rise of *do*-support in English, a maximum likelihood calculation yields an estimate for the influence ratio used in the model. The influence ratio so inferred indicates that 19 of the 200 simulated individuals account for 95% of the total influence, confirming that language change may be driven by a relatively small group of leaders. The model can be improved in any number of ways, but additional features must be selected carefully so as not to produce a computationally intractable inference problem. This project demonstrates how data and techniques from different subfields of linguistics can be combined within a mathematical model to reveal otherwise inaccessible information about language variation and change.

1.1 Introduction

The purpose of this chapter is to introduce a sociolinguistic mathematical model of a structured population of speakers and integrate it with historical data. The project began with a conversation between the author and mathematical biologist Martin Nowak, in which the question at hand was

originally posed as: Given data from a language change, is it possible to infer a population size, and what would that number actually mean? The idea of inferring a population size might suggest trying to estimate how many people lived in medieval England based on the writings of Chaucer, but this is not what that question is meant to ask. Rather, the intent is to formulate a mathematical model with some parameter that represents the number of individuals relevant to some feature of the overall population, such as the size of an average town or friendship network, then to estimate the value of that parameter from data. There are many potentially applicable mathematical models in the population genetics literature in which several genetic variants of a particular species are present, but one of them eventually takes over the entire population. This is called *fixation*. It seems reasonable to interpret variants of a language analogously to genetic variants of a species and to investigate what data about one variant's route to fixation might say about the underlying population. Specifically, the original question of inferring a population size is better posed as follows: Is the population homogeneous, or are some individuals more important than others in driving the change? Is there perhaps a small core of leaders that switch to a new language variant, and the rest of the population simply learns from them and reflects their speech patterns? Inferring a size from language change data would presumably reflect the size of this core. The central task of this chapter is to explain how such a calculation can be carried out, thereby reconstructing the linguistic leadership structure of medieval English society from data on a change in syntax.

The model developed here adapts a genetic model to incorporate sociolinguistic observations. It is then fit to historical data from a syntax change in English. The inferred parameter yields an estimate for the size of the leadership core. This project is part of a growing body of syntheses of linguistic fields and mathematical modeling methods that were generally not combined until the 1990s or so. Some remarks are in order about the challenges and potential of such syntheses.

Mathematics has traditionally been applied with great success to linguistic sub-fields related to formal grammar and the idealized speaker.[1] Statistical studies are essential for understanding language variation and change. Linguists typically use well established tools such as hypothesis tests, VARBRUL, or ANOVA, but these tools have their limits. Recent studies have taken advantage of ever more sophisticated statistical models for

[1]See for example [Chomsky (1965, 1972, 1988); Tesar and Smolensky (2000); Joshi and Schabes (1997); Kroch and Joshi (1985); Stabler (1997); Fong (2005)].

analyzing historical data [Kroch (1989); Warnow (1997)]. This chapter introduces a new kind of tool to this collection and shows how it can deduce unexpected information from well known data.

The use of biological models as bases for linguistic models has been very useful in recent studies of the biological evolution of language,[2] and the study of language change on historical time scales.[3] For example, the logistic sigmoid function, a well known model of the S-curve characteristic of language change, has its roots in the study of population growth in a constrained environment. Many introductory textbooks on differential equations teach logistic growth in conjunction with census data; for example there is a project on the subject in [Edwards and Penney (2008)]. It should be noted that the statistical curve fitting method in that project is somewhat suspect, but since it is within a textbook for a first course on differential equations, there is justification for not choosing a more robust method that might distract students from the primary topic. However, this textbook project highlights a cultural difficulty within mathematics. The mathematical subfield of dynamical systems focuses on the precise and the nonlinear. Statistical inference on the other hand must deal with noisy discrete data, and is frequently limited to linear methods. Combining these two fields correctly can be difficult, and as in the textbook, circumstances often dictate that one field be sacrificed in favor of the other. A better resolution is to use tools from the areas where dynamical systems theory and statistics overlap: Markov chains, and maximum likelihood inference methods.

In addition to mathematical cultural difficulties, this project attempts to combine historical linguistics and sociolinguistics in an unusual way. Sociolinguists have established that social networks contribute to the spread of language changes [Labov (1994, 2001, 2007)]. Present-day investigations can partially identify the relevant social structures from inteviews that reveal details about the friendship networks and speech patterns of many individuals. Such studies that produce a snapshot of the state of a language

[2]See for example [Nowak and Krakauer (1999); Nowak *et al.* (1999b); Trapa and Nowak (2000); Komarova and Nowak (2001a); Nowak *et al.* (1999a, 2000); Plotkin and Nowak (2000); Nowak *et al.* (2001); Komarova and Nowak (2001b); Nowak *et al.* (2002); Nowak and Komarova (2001); Komarova *et al.* (2001); Cangelosi and Parisi (2002); Mitchener (2003a); Mitchener and Nowak (2003); Mitchener (2003b); Mitchener and Nowak (2004); Mitchener (2007)]

[3]See for example [Gibson and Wexler (1994); Niyogi (1998); Niyogi and Berwick (1996, 1997b,a); Niyogi (2006); Kirby (2001); Kirby and Hurford (2002); Briscoe (2000, 2002); Cucker *et al.* (2004); Gold (1967); Pearl and Weinberg (2007); Yang (2002)]

at a single moment in time are called *synchronic*. For example, one might spend a few months recording spontaneous speech by one hundred individuals of varying ages, then estimate the formant frequencies in their vowels or how often they use certain syntactic alternatives. Assuming that adult speech changes very little, the age variation indicates speech patterns going back in time several decades in what is called *apparent time*. The interviews might also include information about socio-economic class and friendships, indicating how one person's speech might influence another's. Unfortunately, the data acquired via interviews takes so much time to gather that the data sets are often far sparser than statisticians would like. Furthermore, interviews and social networking data are generally not available for studying language changes older than the present oldest generation.

Studies of linguistic data across several decades or centuries are called *diachronic* because they compare language use from two or more separated time periods. Corpora consisting of written documents from a across a wide range of time are essential to such studies. Sociolingustic studies sometimes include follow-up interviews years or decades after an initial study, but such projects cannot span the centuries that corpus studies can. Unfortunately, corpora are analogous to fossils or archaeological discoveries in that present-day scientists have no control over the content of ancient documents, or which documents survive to be included in a corpus. Furthermore, the written record contains plenty of linguistic information, but the written language is often distinct from the spoken language in ways that cannot be confirmed centuries later.

The data set of interest for this project is from the change in late Middle English syntax from verb-raising to *do*-support [Ellegård (1953)]. In verb-raising syntax, main verbs are raised from a low position in the syntax tree to various high positions. This means that the main verb raises above the subject, yielding inverted questions, and the main verb raises above negation, so it appears before *not*:

(1.1) Know you what time it is?

(1.2) I know not what time it is.

In *do*-support syntax, the main verb is restricted to a low position in the syntax tree, so when a verb is needed to fill a high position, the auxiliary verb *do* must be inserted:

(1.3) Do you know what time it is?

(1.4) I don't know what time it is.

Affirmative declarative statements have the same surface form under both grammars; the insertion of *do* in this case is actually forbidden under the *do*-support grammar:

(1.5) I know what time it is.

(1.6) *I do know what time it is.

The * indicates an ungrammatical utterance. The second example can be made grammatical by stressing the *do*, which changes the meaning to indicate that the speaker is contradicting a previously made statement. Without that stress, the *do* is ungrammatical. Oddly, the insertion of *do* is also ungrammatical for affirmative subject questions without stress:

(1.7) Who knows what time it is?

(1.8) *Who does know what time it is?

A previous study of the *do*-support data by Kroch [Kroch (1989)] fit a logistic sigmoid

$$y = \frac{1}{1 + e^{-a(t-t_{1/2})}} \tag{1.9}$$

to the S-curve of the usage rate y of *do*-support over time t. The notation $t_{1/2}$ refers to the fact that when $t = t_{1/2}$, $y = 1/2$. Such a function may be grounded in a logistic population model where the rate of spread of a feature is jointly proportional to the fraction of people who have it and the fraction who do not, as in the logistic differential equation

$$\frac{dy}{dt} = ay(1 - y)$$

to which the function (1.9) is the general solution. The logistic model assumes an infinite, unstructured, homogeneous population. The point of the calculations in [Kroch (1989)] was to infer the rate constant a and demonstrate that the rise of *do*-support in all different kinds of sentences was governed by the same rate constant, although the time offsets $t_{1/2}$ differ. Kroch names this result the *constant rate effect*. Infelicities of the logistic model, such as the fact that it admits populations with a fractional number of people, were not important, nor was the overall population size.

In the interest of inferring a population size from the *do*-support data, we turn to mathematical population genetics. One of the simplest and most flexible tools for working with finite populations is the Moran model

[Nowak (2006)]. The population consists of a finite number of agents, each of which is in one of a fixed number of states. An agent is removed to simulate death, and a new one is created by cloning a randomly selected agent thereby simulating birth. Since births and deaths are paired, the overall population size is fixed.

The mathematical framework at work is the discrete time, finite state space Markov chain: At each time step, the population is in one of a large but finite number of possible states. The dynamics specify that given the current state, the population changes randomly to a new state at the next time step, but the probability of selecting each possible new state is a deterministic function of its current state. It is possible, though often computationally infeasible, to use a vector to represent the probability that the population is in each possible state at a given time step. A stochastic matrix can be used to represent the transition process, and multiplying the distribution vector by the transition matrix gives the distribution vector for the next time step. A computer program with a random number generator can implement the transition process and output a stream of states, that is, a sample trajectory of the model.

The variable influence model in this project starts with the Moran model, but assumes that individuals have different degrees of influence on the speech of others. The cloning step is therefore modified to take influence into account. Initially, most of the agents are in a state representing the old language. A few influential agents start in a different state representing the new language. New agents are more likely to be cloned from the more influential agent, so the new language will spread and is likely to take over. The state of each individual agent must be recorded, which makes for much more complex calculations compared to the original Moran model and the logistic model, in each of which the population state is a single number. The main difficulty is that there are so many possible population states that the vector-and-matrix representation of the Markov chain is computationally infeasible. Instead, the only way to investigate its behavior is to accumulate many sample trajectories and take some sort of average. Therefore, several simplifying assumptions are necessary to formulate a model for which the calculations are feasible. A further complication is determining when enough samples have been collected, so the analysis of the samples will be done two different ways. One is a straightforward average and the other estimates the same average from an approximate density. Based on samples collected over a year of computer time, the two calculations agree, which suggests that we have enough samples. So, despite the numerical

difficulties, it is possible to fit the model to the *do*-support data. The result is an estimate of an influence ratio that indicates the extent to which influence is concentrated in a few individuals.

In the rest of this chapter, we formulate the variable influence model, then test a range of values of the influence ratio to determine which is most harmonious with the Middle English data. The calculations strongly support the conclusion that the population is distinctly skewed, specifically, that a leadership core of around 19 individuals out of the 200 in the simulation account for 95% of the total influence.

1.2 The available data

The available data consists of counts of sentence types from clusters of Middle and Modern English manuscripts and the approximate dates of those clusters [Ellegård (1953); Kroch (1989)]. The sentences of interest are different kinds of questions and negative statements, as these clearly show whether the speaker is generating them with a verb-raising grammar or a *do*-support grammar.

Do-support replaced verb-raising in several stages, affecting some types of sentence before others. The cleanest data is for transitive affirmative questions, as in 'Do you want sugar?' and this subset of the data will be the focus of the remainder of this chapter. See Figure 1.1 for a graph of this data.

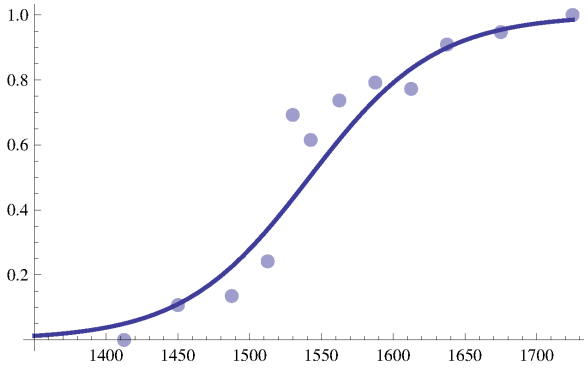

Fig. 1.1 Occurrence rate of *do*-support as a fraction of total sentences, for transitive affirmative questions. The curve is a logistic sigmoid fit to this data via maximum likelihood.

1.3 Formulation of the variable influence model

The simulated population consists of n individual agents, with $Y_i(t)$ representing the type of the i-th agent at time step t, $i \in \{0, 1, \ldots, n-1\}$. We fix $n = 200$. True speech shows context- and individual-dependent variation, but at this stage of the modeling process, a simplification is computationally necessary. We therefore make the simplifying assumption that there are 2 relevant language variants in use, numbered 0 and 1, representing the old verb-raising and new *do*-support grammars, respectively. Thus $Y_i(t) = k$ means that at time t, agent i uses variant k exclusively.

The Markov chain has 2^n possible states because the type of each individual must be tracked separately. This means that if the simulated population is at all large, even 30 individuals, the transition matrix will be too large to compute with directly.

Let $X_k(t)$ be the number of individuals of type k at time t. Thus, if a speaker is selected uniformly at random and asked to produce a sentence, then the sentence is generated by language variant k with probability

$$S_k(t) = \frac{X_k(t)}{n} \tag{1.10}$$

and for each t, $S_0(t) + S_1(t) = 1$.

Let Δt be the real time associated with a unit change it t. The transition function from time step t to $t + 1$ involves examining each agent. With probability $\beta\Delta t$, the agent is replaced, otherwise it remains unchanged, that is $Y_i(t+1) = Y_i(t)$. To replace it, another agent is selected at random and its type is used as $Y_i(t + 1)$. This operation simulates the birth of a new individual who chooses a language variant based on the speech of a single adult. For the calculations in this chapter, Δt is taken to be one year, and $\beta = 1/40$ so that each agent survives for a geometrically distributed random lifetime with a mean of 40 years.

To model a population in which all individuals have equal influence on learning, we would choose the agent to copy in the replacement step uniformly at random from among the whole population. For variable influence, we can assign a score to the i-th slot in the population and choose the agent to copy with probability proportional to that influence score. We will use the function b^i, where b is the *influence ratio*, $0 < b \leq 1$. That is, each individual is a factor b less influential than the next most influential individual. If b is close to 1, then influence is spread through a large part of the population, but if b is even a bit less than 1 then influence is concentrated in a few individuals.

In the initial state, most agents are in state 0 to indicate that the population was dominated by verb-raising initially, but a few agents are in state 1 to trigger the transition to *do*-support. For this chapter, the initial state is that the four most influential agents are in state 1 and the rest are in state 0. The initial time is interpreted as the year 1410, which is about the time of the first data point in the corpus. In the long run, the population will end up in one of two absorbing states, either all state 0 or all state 1. Historically, the English converged to all state 1.

Some sample trajectories are displayed in Figures 1.2 to 1.4. They were hand selected from a small random sample to illustrate the impact of b, and exclude trajectories in which the new language went extinct. For smaller values of b as in Figure 1.2, the usage rate of the new language increases too quickly and overshoots the data. For larger values of b as in Figure 1.4, the population is more influentially uniform, and the trajectory behaves more like a symmetric random walk, almost as likely to go down as up. An intermediate value as in Figure 1.3 fits the data better.

The average shapes of trajectories are shown in Figures 1.5 to 1.7. These display quartiles of samples of many trajectories, and approximately indicate the distribution of the population as a function of time for several different values of b.

To understand why the trajectories have the shape that they do, consider first the beginning of the change, where only a few of the most influential agents are in state 1. Each time step replaces approximately βn agents, and many of the new ones will be clones of influential agents and therefore type 1. This yields an approximately linear growth in the usage rate of the new grammar, but slightly concave-up. The curvature happens because as more influential agents are replaced, the fraction of new agents of type 1 each step will increase. It is often very slight and does not match the distinct initial upward curve of the usual sigmoid trajectory of language changes, which suggests that some modifications should be made to the model in future experiments.

The downward curve at the top of the simulated trajectories is at least qualitatively in agreement with the downward curve typical of language changes. This curvature happens because once most of the population has switched to type 1, a large fraction of the agents that get replaced were already type 1, so the net change of the usage rate of the new grammar is slower.

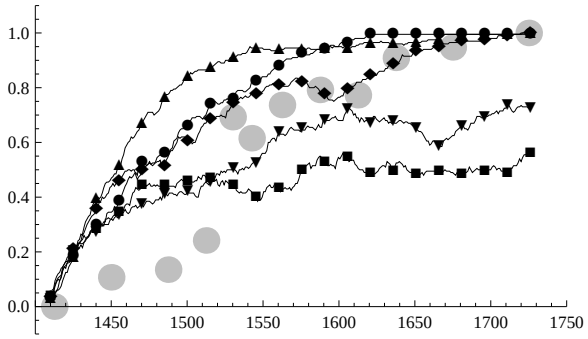

Fig. 1.2 Trajectories, shown as the fraction $S_1(t)$ of the population in state 1 as a function of time. For these runs, $b = 0.8$. Different runs are marked by different symbols. Big gray dots mark the *do*-support usage rate from the corpus.

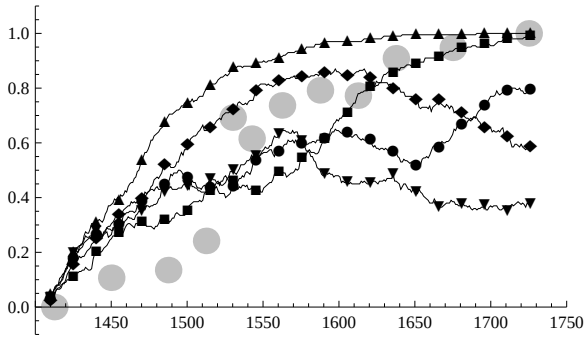

Fig. 1.3 Trajectories, shown as the fraction $S_1(t)$ of the population in state 1 as a function of time, for $b = 0.85$.

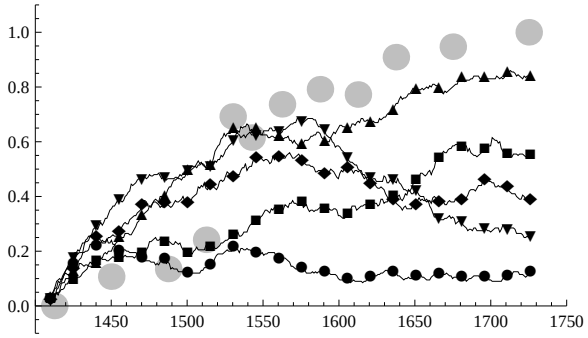

Fig. 1.4 Trajectories, shown as the fraction $S_1(t)$ of the population in state 1 as a function of time, for $b = 0.9$.

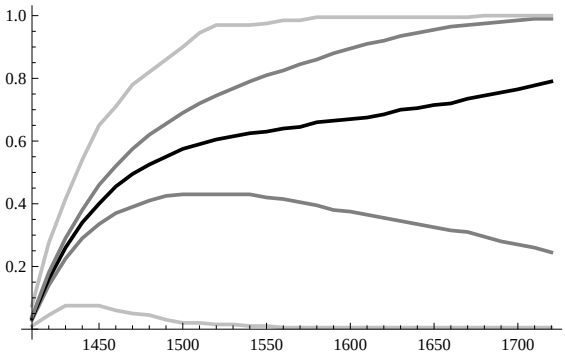

Fig. 1.5 An ensemble envelope of $S_1(t)$ as a function of time. The curves show the minimum, first quartile, median, third quartile, and maximum at each time step over 5000 sample trajectories, for $b = 0.8$.

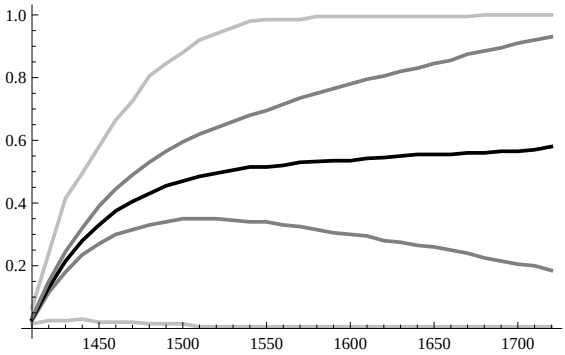

Fig. 1.6 An ensemble envelope of $S_1(t)$ as a function of time, for $b = 0.85$.

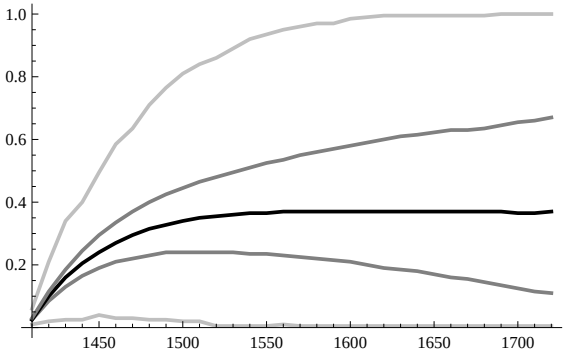

Fig. 1.7 An ensemble envelope of $S_1(t)$ as a function of time, for $b = 0.9$.

1.4 Fitting the *do*-support data

1.4.1 *The maximum likelihood method*

To represent the data, let $t_1, t_2, \ldots$ be the times at which clusters of manuscripts are available, and define $m_k(j)$ be the number of sentences of type k found in the manuscripts at time t_j.

Any model of linguistic dynamics can be tuned to the manuscript count data through maximum likelihood. The idea comes from Bayesian inference [Gelman *et al.* (2004)]. We are really interested in the value of a parameter b. Probability theory is a mathematical way to express partial information about unknown quantities. So we will treat B as a random variable for the parameter b, and our confidence that B has a particular value is represented by a probability distribution $\mathbb{P}(B \in db) = \mathrm{p}(b)\, db$. Until we obtain data, our information about B is some *prior distribution* that incorporates any assumptions we might need to bring to the model. We assume $0 \leq B \leq 1$ but any value in this range is equally likely, so the prior is the uninformative uniform distribution on the interval $[0, 1]$, that is $\mathrm{p}(b) = \mathbf{1}(0 \leq b \leq 1)$.

The addition of data, also treated as a random variable, causes us to have more confidence in some values than others. This modified knowledge is represented by a *posterior distribution* $\mathrm{p}(b \mid m)$.[4] Bayes's formula gives

$$\mathrm{p}(b \mid m) = \frac{\mathrm{p}(m \mid b)\, \mathrm{p}(b)}{\mathrm{p}(m)}$$

The distribution $\mathrm{p}(m \mid b)$ is called the *likelihood*, meaning the probability of observing the data m given a particular value of the b. Since we are taking the prior distribution to be the uniform distribution, and since $\mathrm{p}(m)$ is the same no matter what b is, we can treat these as unknown constants and use the form

$$\mathrm{p}(b \mid m) \propto \mathrm{p}(m \mid b).$$

The obvious choice of b is one in which we have high confidence after examining the data, that is, a b for which the posterior is high. Since the posterior differs from the likelihood by a constant factor, we can choose b to be the value that maximizes the likelihood. The point of this is that the likelihood can be computed based on the model of which b is a parameter.

[4] In classical frequentist statistics, one might assume that B has a normal distribution with mean μ and variance σ^2 and express this information as a confidence interval. That is, $\mathrm{p}(b \mid m) \propto e^{-(b-\mu)^2/\sigma^2}$. The Bayesian approach allows the posterior to be more general.

By computing it for many values of b, one can then sketch the posterior distribution up to a scale factor and select the maximum.

Given estimates $s_k(t)$ of the population-wide usage rate of variant k at time t, the likelihood comes from the binomial distribution

$$\mathrm{p}\left(m \mid s\right) = \prod_j \binom{m_0(j) + m_1(j)}{m_1(j)} s_0(t_j)^{m_0(j)} s_1(t_j)^{m_1(j)} \tag{1.11}$$

A straightforward calculation gives an upper bound on the likelihood of the transitive affirmative *do*-support data. For each time point t_j, there is a value of $\hat{s}_j$ that maximizes

$$\hat{s}_j^{m_0(j)} (1 - \hat{s}_j)^{m_1(j)}$$

Putting those values of $\hat{s}_j$ in for $s_0(t_j)$, using $s_1(t_j) = 1 - \hat{s}_j$, and taking the product yields the upper bound. It should be noted that a model can only achieve that bound by over-fitting the data. For the transitive affirmative question data, the upper bound is 5.48×10^{-10}. Let ρ be the natural logarithm of this upper bound, so $\rho = -21.3248$. For reference, the likelihood achieved by the logistic curve in Figure 1.1 is 2.04×10^{-17}.

Logistic dynamics yield a fairly simple explicit formula for the likelihood in terms of two unknown parameters. The curve in Figure 1.1 was drawn by assuming the form $s_1(t) = 1/(1 + \exp(-a(t - t_{1/2})))$ and solving for a and $t_{1/2}$.

In contrast, it is not possible to use an explicit formula for the likelihood with the variable influence Markov chain. Instead, the likelihood must be computed by conditioning on the population's complete history. Let $\mathcal{H}$ be the set of all possible histories $y_i(t)$ of the population. If y is given, then the type counts $x_k(t)$ and the overall usage rates $s_k(t)$ are known in terms of y. Thus

$$\begin{aligned} \mathrm{p}\left(m \mid b\right) &= \sum_{y \in \mathcal{H}} \mathrm{p}\left(m \mid s\right) \mathrm{p}\left(y \mid b\right) \\ &= \mathbb{E}\left(\mathrm{p}\left(m \mid S\right)\right) \end{aligned} \tag{1.12}$$

Unfortunately, the summation over $\mathcal{H}$ is computationally infeasible. For any reasonable population size, such as the modest $n = 200$ used in this project, there are too many possible histories. A Monte Carlo method that averages over a random sample $\mathcal{S}(b)$ of possible histories generated with a particular value of b is the obvious alternative:

$$\mathrm{p}\left(m \mid b\right) \approx \frac{1}{|\mathcal{S}(b)|} \sum_{y \in \mathcal{S}(b)} \mathrm{p}\left(m \mid s\right) \tag{1.13}$$

An important property of (1.11) is that if either language variant goes extinct too early in the trajectory, then for some j, $s_1(t_j)$ will be one or zero, thereby zeroing out the entire product. The syntactic change is known to have taken place, and the old language persisted for some time. Therefore any sample trajectory in which that change is impossible will be discarded, that is, we condition on the fact that the Markov chain must be absorbed into the state of all 1s but not before the old syntax disappears from the written record.

It should be noted that this calculation is different from what is normally meant by the terms *Markov chain Monte Carlo*, in which the goal is to concoct a Markov chain whose stationary distribution matches some desired distribution and to then sample from it. Rather, for this model the Markov chain itself is the random process of primary interest. We are not interested in a stationary distribution but in trajectories themselves, starting from a particular starting point and moving toward absorption.

1.4.2 *The Monte Carlo calculation*

Although the average (1.13) is computationally feasible, it turns out to require a huge sample size to achieve acceptable results. The core difficulty is that the trajectories y for which $\mathrm{p}\,(m \mid s)$ are largest are relatively uncommon, but the corresponding values of $\mathrm{p}\,(m \mid s)$ are several orders of magnitude larger than the likelihood values contributed by bulk of the samples. In other words, the average (1.13) is dominated by rare events. Figure 1.8 shows a histogram of $\ln \mathrm{p}\,(m \mid S)$ for $500,000$ samples using $b = 0.85$. That is, the horizontal scale is logarithmic. For reference, the smallest positive number representable in the standard 64-bit floating point format is about 2×10^{-308} or about e^{-708}. To avoid hardware underflow errors, the *logarithm* of the likelihood has to be computed all along. An important reason to condition on the fact that the change took place is that none of the likelihood samples can be zero, for which the logarithm would be undefined.

The author wrote a computer program and ran it sporadically on a shared computer cluster over the course of a year to generate $170,000,000$ samples of the log-likelihood for each of 12 different values of b. On this cluster, the program takes approximately 9.5 hours to produce 500,000 sample runs for each of the 12 values of b, which means that the full data set required about 3200 hours or about 134 days of cluster computing time. These samples were then processed in several different ways as described in the following subsections.

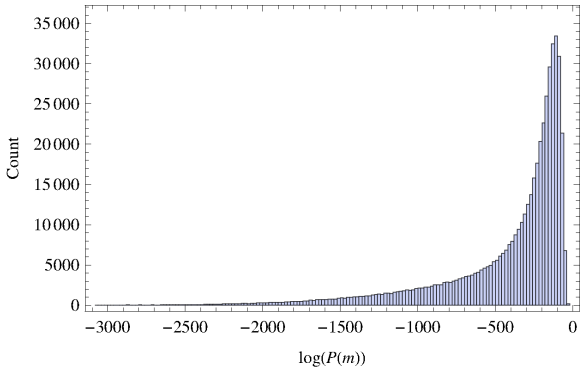

Fig. 1.8 Histogram of 500,000 samples of $\ln \mathbb{P}(m \mid S)$ with $b = 0.85$.

1.4.3 *Direct average*

The obvious approximation to the likelihood is a direct average of the samples as in (1.13). The safest way to compute it is to use a program like Maple or Mathematica to read the log-likelihood samples, apply $\exp(\cdot)$ using arbitrary precision arithmetic rather than hardware floating point arithmetic, and take the average. Likelihood ought to be a smooth function of b, but it takes many millions of samples to produce a reasonably clean plot. The results are shown in Figure 1.9. As is typical of Monte Carlo methods, the accuracy of the result depends on the square-root of the sample size, so the error bars are fairly large even with $170,000,000$ samples. Nevertheless, the likelihood increases dramatically as b decreases from 1 (evenly distributed influence) to $b = 0.85$, which indicates that influence is concentrated in a few members of the population.

The confidence intervals shown in Figure 1.9 are drawn by computing a sample standard deviation $\bar{s}$ for the set of likelihood samples, then plotting $\pm 2\bar{s}/\sqrt{|\mathcal{S}|}$, assuming there is sufficient data to invoke the central limit theorem. Interpreting the confidence intervals, there is enough data to assert that the maximum is at no less than 0.9 and most likely at 0.85

1.4.4 *Fitting the density*

It is useful to process the samples a second way to confirm that enough data has been collected. An alternative to the raw average is to fit a curve to the log-likelihood histogram and use an integral to compute the likelihood. To make this process numerically simpler, the log-likelihood

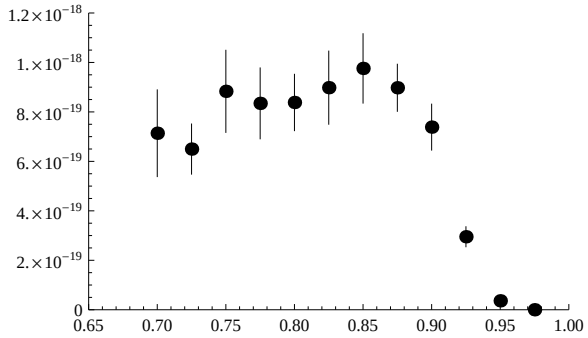

Fig. 1.9 Estimates of $\mathrm{p}\,(m \mid b)$ from a direct average. Whiskers indicate a 95% confidence interval.

samples are transformed by subtracting them from the log-upper-bound ρ, thereby yielding all positive values that can theoretically go all the way down to 0. That is, given a random history Y and the corresponding X and S, set $Z = \rho - \ln \mathrm{p}\,(m \mid S)$. A histogram of Z can be obtained from Figure 1.8 by reflecting it about the vertical axis and shifting it horizontally. Let $\mathrm{p}\,(z \mid b)$ be the density function for Z and let $\bar{f}(z; b)$ be an estimate of $\mathrm{p}\,(z \mid b)$ obtained through a curve fit. Then

$$\begin{aligned} \mathrm{p}\,(m \mid b) &= \mathbb{E}\,(\exp Z) \\ &= \int_0^\infty e^{\rho - z}\, \mathrm{p}\,(z \mid b)\, dz \\ &\approx \int_0^\infty e^{\rho - z} \bar{f}(z; b) dz \end{aligned} \tag{1.14}$$

This method does not escape from all the difficulties of the direct average method. The integral is very sensitive to the density near $z = 0$ because that is where most of e^{-z} is concentrated, but that is precisely where there is the least data and the most uncertainty. The curve fit effectively smooths the histogram in that area.

The algorithm used to fit the density for each value of b is as follows. Sample histories are transformed into samples for Z, which are then grouped into bins of width 1. Empty bins are discarded. Each bin B_n contains numbers $z_1, z_2, \ldots \in [n, n+1)$ which are mapped to the point

$$(u_n, v_n) = \left(\frac{1}{|B_n|} \sum_i z_i, \frac{|B_n|}{|\mathcal{S}|} \right) \tag{1.15}$$

where $|\mathcal{S}|$ is the total number of samples in all the bins. The left-most bins usually contain fewer points than the others because high likelihood

estimates are rare. The average of the numbers in B_n is used for the horizontal coordinate of the point rather than the midpoint of the bin because it better reflects the off-center numbers in those left-most bins, which yields more stable likelihood estimates.

Although the shape of the density of Z consists of a smooth hill and a long tail, it does not seem to be well represented by the commonly occurring gamma or extreme-value distributions. Instead, a fairly general form was chosen for the fit function based on trial and error and asymptotic considerations. A polynomial $a_0+a_1\lambda+a_2\lambda^2$ is fit to the log-log points $(\ln u_n, \ln v_n)$ with $u_n \leq 40$ using the method of least squares, with each point weighted by the number of samples in the corresponding bin B_n. The transformation to Z ensures that all the u_n are positive so $\ln u_n$ is defined. The weighting is important because it continues the trend of points not quite at the extreme left where more data is available, while not ignoring the points derived from sparser data at the extreme left. This process yields a fit to a function of the form

$$\begin{aligned} \bar{f}(z) &= e^{a_0+a_1\lambda+a_2\lambda^2} \text{ where } \lambda = \ln z \\ &= c_0 z^{c_1} e^{-c_2(\ln z)^2} \end{aligned} \tag{1.16}$$

with

$$c_0 = e^{a_0} > 0, c_1 = a_1 > 0, c_2 = -a_2 > 0,$$

This form takes advantage of the fact that the graph of $(\ln u_n, \ln v_n)$ is fairly smooth, and that we need only fit the left side of the hill. It does not match the tail of the density, but it does not need to. Only the shape of the density for small z matters. See Figure 1.10.

A seemingly better fit to the log-log points can be found by including higher powers of λ. However, for some b values, doing so yields negative coefficients on the $(\ln z)^3$ terms and therefore a singularity as $z \to 0$. The correct asymptotic behavior at 0 requires that the coefficients on $(\ln z)^2$ be negative and those on $(\ln z)^3$ be positive. Then, as $z \to 0$, $-(\ln z)^2 \to -\infty$ and $(\ln z)^3 \to -\infty$, so $\bar{f}(z) \to 0$.

Given those approximate densities, the integral approximations of the likelihoods are shown in Figure 1.12. The fitting procedure gives 95% confidence intervals for the parameters, which are mapped into confidence intervals for the likelihood estimates. The results are essentially the same as from using the raw average, as in Figure 1.9, with the maximum at $b = 0.85$.

This process is particularly sensitive to the value of ρ. The calculations were carried out once with an incorrect ρ, which resulted in a very noisy

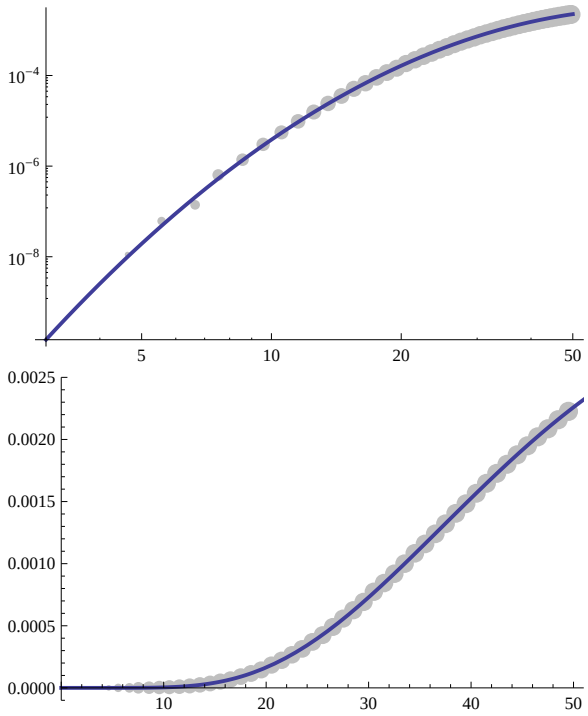

Fig. 1.10 Top: Log-log plot of (u_n, v_n) and $\bar{f}$ fit to the points with $u_n \leq 40$, for $b = 0.85$. Bottom: Same functions on normal scale. Areas of dots are proportional to $\ln(|B_n| + 1)$.

likelihood graph with much larger error bars. Therefore, the form (1.16) is probably not optimal. However, the error bars with the correct ρ are very small, and the overall shape agrees well with the likelihood estimates from the direct average.

Monte Carlo method maximum likelihood method

1.5 Results and discussion

To begin, we should compare the likelihood estimates from the two methods. See Figure 1.13. Both methods indicate that the maximum satisfies $0.8 \leq b \leq 0.85$. They are largely in agreement, suggesting that $170,000,000$ is nearly enough samples to estimate $\mathrm{p}\,(m \mid b)$ across the range of interest.

Let us be overly conservative and consider $b = 0.9$. Recall that within the simulation, the influence score of individual i is b^i. The ratio of the

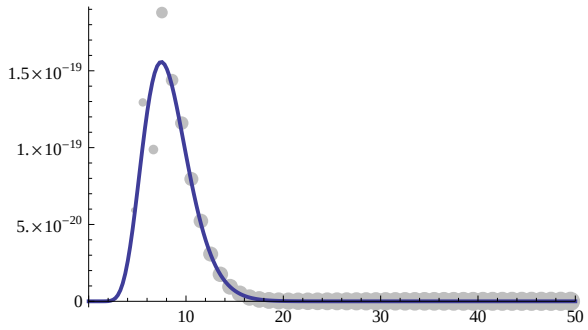

Fig. 1.11 Plot of $(u_n, e^{\rho - u_n} v_n)$ and $e^{\rho - z} \bar{f}(z)$ for $b = 0.85$.

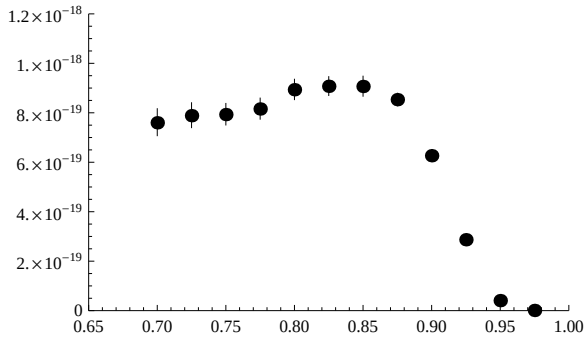

Fig. 1.12 Estimates of $\mathrm{p}(m \mid b)$ from integrating against a fitted density. Whiskers indicate a 95% confidence interval.

net influence of the first m individuals to the total influence across the population indicates the degree to which influence is concentrated among the most influential individuals. This ratio is plotted in Figure 1.14. The 29 most influential individuals account for 95% of the total influence for $b = 0.9$. For $b = 0.875$, the 23 most influential account for 95%. The b that maximizes the likelihood appears to be at most $b = 0.85$, for which the 19 most influential individuals account for 95%.

Even though more samples would help to pin down the correct value of b, the collected data is definitely more consistent with a variable-influence population than a flat population: A leadership core of approximately 19 people account for most of the total influence. This suggests that if it were possible to survey a large number of people and somehow determine who was most influential on each of their speech patterns, we should expect 19

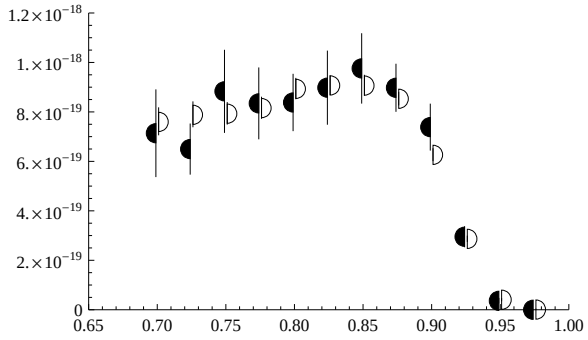

Fig. 1.13 Estimates of $\mathrm{p}\,(m \mid b)$. For each value of b, the black half dot is the estimate from the direct average and the white half dot is the estimate from integrating against a fitted density. Whiskers indicate a 95% confidence interval.

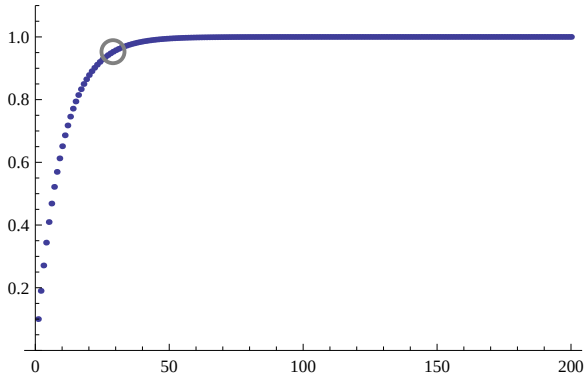

Fig. 1.14 Ratio of the net influence of the first h individuals to the total influence across the population, as a function of h, for $b = 0.9$. The gray circle is centered at $h = 29$, which accounts for 95%.

or so linguistic leaders for each community within the overall population.

1.6 Future directions

This project focuses on inferring a single macroscopic feature of medieval English society from data about syntax. However, the results suggest many improvements, further questions, and other applications of this mathematical modeling technique.

There is plenty of room for improvement in the variable influence model,

but at a cost. Many of its features as presented here are fixed at some reasonable value so that b is the only parameter that has to be inferred. The inescapable difficulty is that adding detail will add variables, which must then either be set arbitrarily, like the choice of $\beta = 1/40$, or inferred from samples of the Markov chain. There is no reason to suppose that the unknown variables can be inferred independently in seeking to maximize the likelihood. That is, if we seek to maximize the likelihood allowing some new unknown c to vary, the maximum might occur at some $b_{\max}$ and $c_{\max}$ where $b_{\max}$ is not 0.85 as found here (although it ought to be close). To maximize the likelihood will require many samples at a much larger set of values of (b, c), which will require a lot of computer time. Likelihood provides a metric for how important a particular variable is to the model. If adding the variable increases the likelihood significantly without overfitting the data, then it is important and the increase in likelihood quantifies how much. Otherwise, it can probably be omitted, and computer time can be better spent on some other feature. On a practical note, the sampling program could be written more efficiently, but any improvements will probably not be sufficient to allow for inferring more than two or three interdependent variables in a reasonable amount of time. New mathematical tools for dealing with Markov chains such as this model are therefore needed.

To give a specific comparison, the logistic sigmoid in Figure 1.1 fits the *do*-support data better than the variable influence model, in that the likelihood of the data given the sigmoid model (2.04×10^{-17}) is higher than the likelihood given the variable influence model with the best choice of b (10^{-18}). This difference is to be expected because the logistic model has two variables, a and $t_{1/2}$, but the variable influence model has only one, b. The difference should not be interpreted as a failure of the variable influence model, because the two models give different information. Specifically, the logistic model does not give any information about heterogeneity of the population. Based on Figure 1.3, the main drawback to the variable influence model is that its trajectories do not match the slow growth of the change near its beginning, so refinement should focus first on this feature to increase the likelihood.

With these concerns in mind, the question naturally arises of whether the extra effort required by models of this kind is worthwhile. Alternatively, one could try to run analyses based around a series of hypothesis tests. For example, we might consider as a null hypothesis that children learn from all adults in their neighborhood equally, then ask whether a certain data set allows us to reject this hypothesis at some confidence level. This approach

has the advantage of being computational easier, and the statistics are potentially conventional and well understood. However, such a project gives no indication of the degree to which individuals with varying influence might drive language change. A hypothesis test may give the result that certain models are statistically distinct, but it gives no information about what the difference means, or whether it might be statistically significant but subordinate to some other stronger force. A VARBRUL-based analysis might initially seem to be a reasonable alternative. It gives relative strengths of various factors on the probability of an overall binary outcome, but it can only be used if the factors in question are also binary. Estimating the size of a leadership core is not possible with VARBRUL, for example.

In contrast, the variable influence model includes a continuous parameter b that indicates the strength of the effect. In the calculations, it ranges from no effect at $b = 1$ to concentrating influence in 19 individuals at $b = 0.85$. The likelihood calculation allows us to estimate how probable each value of b is given the data, so we are essentially testing a whole range of hypotheses rather than two as in a standard hypothesis test.

An obvious improvement would be a more realistic representation of language usage and the learning process. The variable influence model currently assumes that children exactly copy one other individual's speech, when they should learn from several, including adults and peers. Furthermore, an individual's state should include more possibilities than using one language variant or another exclusively. Recognizing that the discrete categorial tools of formal grammars and idealized speakers are insufficient for representing the intricate variations of language, there is increasing interest in using probability in conjunction with traditional formalisms to understand and represent language [Bod *et al.* (2003); Shannon and Weaver (1949); Mitchener (2005)]. These features could be incorporated into the current model and would likely be worth the computational cost.

An additional improvement would be in the interpretation of the manuscript data. The likelihood formula (1.11) implicitly models the creation of the corpus by selecting individuals uniformly at random and asking for a sentence, which is rather naïve. The corpus contains collections of manuscripts written at estimated times by relatively few speakers in a variety of genres, and these are the ones that happen to have survived the centuries and been cataloged by linguists. There is clearly room for an improved model of corpus formation, but it, too, would introduce additional parameters that must be fixed or inferred.

The present model does not try to account for how leaders arise. Stud-

ies reported in [Labov (2001, 1994)] suggest that leadership in language change is determined more by personality, attitude, gender, and friendship networks than any conventional notion of economic or political power. Furthermore, is not clear how to properly scale the results of Section 1.5 to larger populations. Simply increasing the population size n is not likely to affect b significantly because the less influential bulk of the population effectively copies the proportions of the leadership core. Larger populations will have more complex social structure, in which some people are very influential but over distinct subsets of the overall population. The ordered influence structure used here was simple and gave reasonable results, but it would be more realistic to represent the population as graph. Agents would be vertices, edges would indicate linguistic influence, and new agents could be incorporated through some form of preferential attachment process.[5] For example, each new agent might be linked or not to each existing agent with probability determined by the number of links the existing agent already has. Another alternative would be to retain the flat structure but use a function other than b^i for the influence of the i-th individual. However, each of these potential improvements might make it more computationally demanding to fit the model to the corpus data.

Labov and his collaborators have accumulated considerable data on phonetic changes in cities, including information about specific informants who seem to be leading these changes. It should be possible to fit the variable influence model to that phonetic data, in which case its conclusions about leadership structure can be compared to the collected sociological information. Such a project would provide an additional means of verifying this method of statistical analysis.

The model developed in this chapter began as a population genetics model, although the application is sociological and no explicit use is made of natural selection. However, it should be possible to adapt the variable influence model for use in studying biological evolution. Consider for example a model of the evolution of imitation posed by Boyd and Richerson [**?**]. Their underlying model was a set of individual agents who choose a behavior based either on their observation of the environment, or by copying a randomly selected individual when their observation is inconclusive. The mathematics is greatly simplified by assuming that all agents are essentially interchangeable, and by focusing on the dynamics of average properties of the population. Leadership structure breaks the assumption

[5] See [**??**] and forthcoming articles by Swarup.

of interchangeability, and may rule out the reduction of the dynamics to average properties. It would be informative to revisit the Boyd-Richerson model with leadership structure added in, and see which of their results still hold and which are modified.

1.7 Conclusion

In conclusion, corpus data concerning the rise of *do*-support at the expense of verb-raising, in conjunction with an agent-based population model, is consistent with the hypothesis that influence is distributed unevenly through the population. The maximum likelihood method, computed two different ways from sample runs of the variable influence Markov chain model, yields an estimate of the influence ratio. That estimate asserts that approximately 19 out of the 200 people within the simulation account for 95% of the total linguistic influence. This project provides an important mathematical tool for combining sociolinguistics with historical methods and sophisticated mathematical models, but there is plenty of room for improvement.

The author gratefully acknowledges that this project was supported by a grant from the National Science Foundation (DMS #0734783).

Appendix: Probability and notation

The notation for probability distributions can be confusing, particularly when mixing continuous and discrete distributions and when conditional probability is involved. I provide this appendix to assist readers who may not be as familiar with some of the concepts and my preferred notation.

Whenever possible, a capital letter (as in X) is used for a random variable and the corresponding lower case letter (as in x) is used for a non-random value that it might take. For instance, the density for X would be written in terms of x, and a calculation involving random samples would be written in terms of X.

The notation

$$\mathbb{P}\left(X \in dx\right) = f(x)dx$$

indicates that X has a *continuous distribution* with *probability density func-*

tion f, so that

$$\mathbb{P}\left(a < X < b\right) = \int_a^b f(x)dx$$

For a random variable N with a *discrete distribution*, we can write

$$\mathbb{P}\left(N = n\right) = g(n)$$

to indicate that g is the *probability mass function* for N. A mass function can also be expressed using delta measures. The symbol $\delta_z(x)$ is special notation for using an integral $\int$ to pick out discrete values of a function:

$$\int \phi(x)\delta_z(x)dx = \phi(z)$$

so the mass function for N can also be expressed as

$$\mathbb{P}\left(N \in dx\right) = \sum_n g(x)\delta_n(x)dx.$$

The bar notation indicates conditioning. That is, $X \mid Y$, read "X given Y," means that we modify the distribution of X by assuming Y is known. The basic property of conditioning is that

$$\mathbb{P}\left(X \in dx \text{ and } Y \in dy\right) = \mathbb{P}\left(X \in dx \mid Y = y\right)\mathbb{P}\left(Y \in dy\right).$$

Since you can also condition on X,

$$\mathbb{P}\left(X \in dx \text{ and } Y \in dy\right) = \mathbb{P}\left(Y \in dy \mid X = x\right)\mathbb{P}\left(X \in dx\right).$$

Combining the two gives Bayes's formula,

$$\mathbb{P}\left(X \in dx \mid Y = y\right) = \frac{\mathbb{P}\left(Y \in dy \mid X = x\right)\mathbb{P}\left(X \in dx\right)}{\mathbb{P}\left(Y \in dy\right)}$$

which expresses the distribution of X given Y in terms of the distribution of Y given X. (The dy's effectively cancel; the details involve the Radon-Nikodym derivative and are well beyond the scope of this chapter.)

Since it's usually more convenient to work with densities (or to use delta measures to pretend that discrete distributions have densities), the notation $\mathrm{p}\left(x\right)$ is often used to indicate the density of the random variable X tied by convention to the same letter in lower case,

$$\mathbb{P}\left(X \in dx\right) = \mathrm{p}\left(x\right)dx.$$

Conditional densities are expressed as

$$\mathbb{P}\left(X \in dx \mid Y = y\right) = \mathrm{p}\left(x \mid y\right)dx.$$

The conditioning formula becomes

$$\mathrm{p}\left(x, y\right) = \mathrm{p}\left(x \mid y\right) \mathrm{p}\left(y\right)$$

and Bayes's formula becomes

$$\mathrm{p}\left(x \mid y\right) = \frac{\mathrm{p}\left(y \mid x\right) \mathrm{p}\left(x\right)}{\mathrm{p}\left(y\right)}.$$

Since the $\mathrm{p}\left(\cdot\right)$ notation is compact and expressive, it is normally overloaded to indicate the mass function of a discrete random variable as well, as in $\mathrm{p}\left(n\right)$. The interested reader must use context to determine the correct rigorous interpretation for $\mathrm{p}\left(\cdot\right)$.

The *expected value* or *mean* or *first moment* of a continuous random variable X is

$$\mathbb{E}\left(X\right) = \int x \, \mathbb{P}\left(X \in dx\right) = \int x \, \mathrm{p}\left(x\right) dx.$$

For a discrete random variable N, the formula is the same except that the integral becomes a sum:

$$\mathbb{E}\left(N\right) = \int x \, \mathbb{P}\left(N \in dx\right) = \int x \sum_n \mathrm{p}\left(n\right) \delta_n(x) dx = \sum_n n \, \mathrm{p}\left(n\right).$$

The notation for an *indicator function* is

$$\mathbf{1}\left(\text{condition}\right) = \begin{cases} 1 & \text{if the condition is true} \\ 0 & \text{if the condition is false} \end{cases}$$

Bibliography

Apolloni, A., Kumar, V. A., Marathe, M. V. and Swarup, S. (2009). Computational epidemiology in a connected world, *Computer* **42**, pp. 83–86, doi: 10.1109/MC.2009.386.

Bod, R., Hay, J. and Jannedy, S. (eds.) (2003). *Probabilistic Linguistics* (MIT Press, Cambridge, MA).

Boyd, R. and Richerson, P. J. (1988). An evolutionary model of social learning: The effects of spatial and temporal variation, in T. R. Zentall and J. Bennett G. Galef (eds.), *Social Learning* (Lawrence Erlbaum Associates, Inc., Hillsdale, New Jersey), pp. 29–48.

Briscoe, E. J. (2000). Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device, *Language* **76**, 2, pp. 245–296.

Briscoe, E. J. (ed.) (2002). *Linguistic Evolution through Language Acquisition: Formal and Computational Models* (Cambridge University Press).

Cangelosi, A. and Parisi, D. (eds.) (2002). *Simulating the Evolution of Language* (Springer-Verlag).

Chomsky, N. (1965). *Aspects of the Theory of Syntax* (MIT Press, Cambridge, MA).

Chomsky, N. (1972). *Language and Mind* (Harcourt Brace Jovanovich, New York).

Chomsky, N. (1988). *Language and Problems of Knowledge* (MIT Press).

Cucker, F., Smale, S. and Zhou, D.-X. (2004). Modeling language evolution, *Foundations of Computational Mathematics* **4**, 3, pp. 315–343.

di Sciullo, A. M. (ed.) (2005). *UG and External Systems: Language, brain and computation*, no. 75 in Linguistics Today (John Benjamins).

Edwards, C. H. and Penney, D. E. (2008). *Differential Equations and Boundary Value Problems: Computing and modeling*, 4th edn. (Pearson Prentice Hall).

Ellegård, A. (1953). *The Auxiliary do: The Establishment and Regulation of Its Use in English*, *Gothenburg Studies in English*, Vol. II (Almqvist and Wiksell).

Fong, S. (2005). Computation with probes and goals: A parsing perspective, in [di Sciullo (2005)], pp. 311–333.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd edn. (Chapman & Hall/CRC).

Gibson, E. and Wexler, K. (1994). Triggers, *Linguistic Inquiry* **25**, pp. 407–454.

Gold, E. M. (1967). Language identification in the limit, *Information and Control* **10**, pp. 447–474.

Joshi, A. and Schabes, Y. (1997). Tree-Adjoining grammars, in *Handbook of Formal Languages 3: Beyond Words*, chap. 2, no. 3 in Handbook of Formal Languages (Springer-Verlag), ISBN 978-3-540-60649-9, pp. 69–120, doi: 10.1.1.30.502.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity, *IEEE Transactions on Evolutionary Computation* **5**, 2, pp. 102–110.

Kirby, S. and Hurford, J. R. (2002). The emergence of structure: An overview of the iterated learning model, in [Cangelosi and Parisi (2002)], pp. 121–148.

Komarova, N. L., Niyogi, P. and Nowak, M. A. (2001). The evolutionary dynamics of grammar acquisition, *Journal of Theoretical Biology* **209**, 1, pp. 43–59.

Komarova, N. L. and Nowak, M. A. (2001a). The evolutionary dynamics of the lexical matrix, *Bulletin of Mathematical Biology* **63**, 3, pp. 451–485.

Komarova, N. L. and Nowak, M. A. (2001b). Natural selection of the critical period for language acquisition, *Proceedings of the Royal Society of London, Series B* **268**, pp. 1189–1196.

Kroch, A. (1989). Reflexes of grammar in patterns of language change, *Language Variation and Change* **1**, pp. 199–244.

Kroch, A. S. and Joshi, A. K. (1985). The linguistic relevance of tree adjoining grammar, Tech. Rep. MS-CIS-85-16, University of Pennsylvania, URL `http://repository.upenn.edu/cis_reports/671/`.

Labov, W. (1994). *Principles of Linguistic Change: Internal Factors*, Vol. 1 (Blackwell, Cambridge, MA).

Labov, W. (2001). *Principles of Linguistic Change: Social Factors*, Vol. 2 (Blackwell, Cambridge, MA).

Labov, W. (2007). Transmission and diffusion, *Language* **83**, 2, pp. 344–387.

Mitchener, W. G. (2003a). Bifurcation analysis of the fully symmetric language dynamical equation, *Journal of Mathematical Biology* **46**, pp. 265–285, doi: 10.1007/s00285-002-0172-8.

Mitchener, W. G. (2003b). *A Mathematical Model of Human Languages: The interaction of game dynamics and learning processes*, Ph.D. thesis, Princeton University.

Mitchener, W. G. (2005). Simulating language change in the presence of non-idealized speech, in *Proceedings of the Second Workshop on Psychocomputational Models of Human Language Acquisition* (Association for Computational Linguistics), pp. 10–19.

Mitchener, W. G. (2007). Game dynamics with learning and evolution of universal grammar, *Bulletin of Mathematical Biology* **69**, 3, pp. 1093–1118, doi:10.1007/s11538-006-9165-x.

Mitchener, W. G. and Nowak, M. A. (2003). Competitive exclusion and coexistence of universal grammars, *Bulletin of Mathematical Biology* **65**, 1, pp.

67–93, doi:10.1006/bulm.2002.0322.

Mitchener, W. G. and Nowak, M. A. (2004). Chaos and language, *Proceedings of the Royal Society of London, Biological Sciences* **271**, 1540, pp. 701–704, doi:10.1098/rspb.2003.2643.

Niyogi, P. (1998). *The Informational Complexity of Learning* (Kluwer Academic Publishers, Boston).

Niyogi, P. (2006). *The Computational Nature of Language Learning and Evolution* (MIT Press, Boston).

Niyogi, P. and Berwick, R. C. (1996). A language learning model for finite parameter spaces, *Cognition* **61**, pp. 161–193.

Niyogi, P. and Berwick, R. C. (1997a). A dynamical systems model for language change, *Complex Systems* **11**, pp. 161–204, URL `ftp://publications.ai.mit.edu/ai-publications/1500-1999/AIM-1%515.ps.Z`.

Niyogi, P. and Berwick, R. C. (1997b). Evolutionary consequences of language learning, *Linguistics and Philosophy* **20**, pp. 697–719.

Nowak, M. A. (2006). *Evolutionary Dynamics: Exploring the equations of life* (Harvard University Press).

Nowak, M. A. and Komarova, N. L. (2001). Towards an evolutionary theory of language, *Trends in Cognitive Sciences* **5**, 7, pp. 288–295.

Nowak, M. A., Komarova, N. L. and Niyogi, P. (2001). Evolution of universal grammar, *Science* **291**, 5501, pp. 114–118.

Nowak, M. A., Komarova, N. L. and Niyogi, P. (2002). Computational and evolutionary aspects of language, *Nature* **417**, 6889, pp. 611–617.

Nowak, M. A. and Krakauer, D. C. (1999). The evolution of language, *Proceedings of the National Academy of Sciences, USA* **96**, pp. 8028–8033.

Nowak, M. A., Krakauer, D. C. and Dress, A. (1999a). An error limit for the evolution of language, *Proceedings of the Royal Society of London, Series B* **266**, pp. 2131–2136.

Nowak, M. A., Plotkin, J. and Jansen, V. A. A. (2000). Evolution of syntactic communication, *Nature* **404**, 6777, pp. 495–498.

Nowak, M. A., Plotkin, J. and Krakauer, D. C. (1999b). The evolutionary language game, *Journal of Theoretical Biology* **200**, pp. 147–162.

Pearl, L. and Weinberg, A. (2007). Input filtering in syntactic acquisition: Answers from language change modeling, *Language Learning and Development* **3**, 1, pp. 43–72.

Plotkin, J. and Nowak, M. A. (2000). Language evolution and information theory, *Journal of Theoretical Biology* **205**, pp. 147–159.

Retoré, C. (ed.) (1997). *Logical Aspects of Computational Linguistics*, *Lecture Notes in Computer Science*, Vol. 1328 (Springer-Verlag), ISBN 978-3-540-63700-4.

Ritt, N., Schendl, H., Dalton-Puffer, C. and Kastovsky, D. (eds.) (2006). *Medieval English and its Heritage: Structure, meaning and mechanisms of change*, *Studies in English Medieval Language and Literature*, Vol. 16 (Peter Lang, Frankfurt), proceedings of the 13th International Conference on English Historical Linguistics.

Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Information*

(University of Illinois Press, Urbana, Illinois).

Stabler, E. (1997). Derivational minimalism, in [Retoré (1997)], pp. 68–95, URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.31%27`.

Swarup, S. and Gasser, L. (2007). Unifying evolutionary and network dynamics, *Phys. Rev. E* **75**, 6, p. 066114, doi:10.1103/PhysRevE.75.066114.

Tesar, B. and Smolensky, P. (2000). *Learnability in Optimality Theory* (MIT Press).

Trapa, P. E. and Nowak, M. A. (2000). Nash equilibria for an evolutionary language game, *Journal of Mathematical Biology* **41**, pp. 172–1888.

Warnow, T. (1997). Mathematical approaches to comparative linguistics, *Proceedings of the National Academy of Sciences, USA* **94**, pp. 6585–6590.

Yang, C. D. (2002). *Knowledge and Learning in Natural Language* (Oxford University Press, Oxford).

Index