A Mathematical Model of the Loss of Verb-Second in Middle English

Short title: A Model of the Loss of V2 in Middle English

W. Garrett Mitchener

Duke University Mathematics Department
Box 90320
Room 121 Physics Building
Science Drive
Durham, NC 27708

(In American English)

e-mail: wgm@math.duke.edu

Abstract

Lightfoot (1999) proposes the following explanation for the loss of the verb-second rule in Middle English: There were two regional dialects of Middle English, a northern dialect influenced by Old Norse with a verb-second rule, and a southern dialect with a slightly different word order. Children acquire the verb-second rule based on hearing some critical fraction of cue sentences requiring such a rule. As the dialects experienced increased contact, northern children were less likely to hear enough cue sentences, and consequently acquired a different grammar, resulting in the extinction of the northern dialect.

This hypothesis can be modeled with differential equations. By using dynamical systems methods, the catastrophe in question may be modeled by a mathematical event known as a saddle-node bifurcation. A key part of the model is the function $q(x)$ that gives the probability of learning the northern dialect given that a fraction x of the local population uses it. Other model acquisition algorithms, such as memoryless learner (Niyogi & Berwick 1996), give the mysterious result that verb-second languages should be extremely stable, in contrast to the history of English. This new model provides an explanation for that behavior: Memoryless learners are more sensitive to noise, resulting in a differently shaped q function that does not allow the northern grammar to disappear. This model

demonstrates how dynamical systems theory can be used to study language change and learning models.

1 Introduction

The first step toward understanding syntactic change is to describe the language before, during, and after the transition. Important features of the initial and final grammars must be identified and expressed formally. The rise of new sentence types and the decline of obsolete types must be understood. Once these descriptive questions are answered, the next step is to address more difficult questions: Why did this change occur? Why did it spread? Why did it happen at the time it did rather than earlier or later? Why did a potential change in slightly different circumstances fail to happen? In particular, how well do chance and internal factors explain the change, and to what extent should external factors, such as contact, be invoked to explain the change? This paper illustrates how mathematical models may be used to precisely express and test hypothetical explanations for syntactic change, using the example of the loss of verb-second in Middle English. Specifically, the model is compatible with the hypothesis that the loss of verb-second may be attributed to contact between regional dialects, and that the timing of the change is related to the timing and amount of contact.

The basis for the model is a framework proposed by Lightfoot (1999) for explaining syntactic change: A shift in speech patterns weakens the evidence for the old grammar in the primary linguistic data (PLD)

available to children, and the resulting ambiguous PLD leads children to acquire a new grammar. Their speech further dilutes the PLD for the next generation, resulting in the spread of the new grammar. In the case of Middle English, there is broad agreement in the literature that there were two regional dialects with different verb-second rules, and an increase in contact between them caused the initial shift in speech patterns that ultimately led to the decline of verb-second (Fischer et al. 2000; Kroch 1989; Kroch et al. 2000). The mathematical model is a representation of this process as a continuous dynamical system.

The grammar acquisition process is crucial to the spread of the change. Lightfoot (1999) also proposes a mechanism for the acquisition of verb-second: Children listen for sentences in the PLD that can only be parsed by a grammar with a particular feature, sentences which Lightfoot calls *cues*. Children incorporate that feature into their native language only if the proportion of cue sentences they hear exceeds some threshold. Children learning the northern and southern dialects of Middle English would have been listening for slightly different cues because of the differences in how their native languages treat subject pronouns. At the boundary between the two dialects, the PLD would have been a confusing mixture of the two, and, Lightfoot asserts, the lack of compelling evidence caused children to acquire a grammar without verb-second. Lightfoot's proposed learning process may be expressed easily in mathematical

notation, and the model shows that with reasonable numerical parameter settings it produce the correct result.

Other learning processes have been suggested in the language modeling literature. For example, the memoryless learner (Niyogi & Berwick 1996) is a simple, mathematically convenient learning model; however, it produces the puzzling result that in simulations, verb-second languages are extremely stable and in fact all languages eventually become verb-second. Clearly this does not reflect the current state of the world's languages. The model in this paper shows that grammar acquisition must strike the proper balance between matching the PLD and ignoring noise if it is to correctly predict that Middle English could lose verb-second, and this is precisely where simple memoryless learners fail.

Section 2 gives some background on the word order of Middle English and the differences between the regional dialects. Section 3 describes the mathematical model and its behavior as the two regions mix. In particular, a phenomenon known as a *bifurcation* takes place, resulting in the loss of one language. Section 4 discusses what goes wrong if a memoryless learner is used instead of Lightfoot's cue-based learner. Finally, Section 5 draws conclusions and describes some of the author's ongoing research on mathematical models of language change.

2 Verb-second in Middle English

Middle English had underlying subject-verb-object (SVO) word order, and like most Germanic languages, also had a rule known as verb-second in which top-level sentences are re-organized such that a topic and the finite verb always appear at the front of the sentence. Example (1) comes from the northern dialect, which was heavily influenced by Old Norse in its grammar and vocabulary. This word order will be abbreviated SVO+CPv2.

(1) $[_{CP}\,[_{DP}$ Oþir labur$]_j$ $_V$sal$_i$ $[_{IP}$ t_i þai do $t_j]$]
other labor shall they do
"They must do other labor."
(*The Rule of St. Benet,* Fischer et al. 2000, p.131)

Since embedded sentences do not show this reorganization, the formal description of verb-second word order in this case is that the finite verb is raised to C and a topic, which can be any DP or a sentential adverb, is raised to Spec-CP. Such reorganization is not possible in an embedded sentence because C is already occupied by a complementizer.

Lightfoot proposes that sentences of the form

(2) $[_{CP}\ XP_{Topic}\ V_{Finite}\ [_{IP}\ DP_{Subject}\ \ldots]\ ]$

are the cues for verb-second. Assuming that children have determined that the underlying Middle English word order should be SVO, sentences of the form in (2) have clearly been reorganized from the underlying order, indicating to children that a verb-second rule is required.

Southern Middle English had a slightly different form of verb-second. Pronominal subjects behave differently from full noun phrases, and can appear between the fronted finite verb and the fronted topic, as in (3).

(3) $[_{CP}\,[_{DP}$ alle þese bebodes$]_j$ $_{PRO}$ic $_{V}$habbe$_i$ $[_{IP}$ t_i ihealde t_j fram childhade]]
all these commandments I have kept from childhood
"I have kept all of these commandments from childhood."
(*Vices & Virtues,* Fischer et al. 2000, p. 130)

Furthermore, Old English allowed for verb-second effects in embedded sentences under some circumstances, leading to the suggestion that the finite verb does not raise all the way to C, but stops at some intermediate position between C and I. Fischer et al. (2000) name this position F, so the southern word order will be abbreviated SVO+FPv2+pro.

Since pronominal subjects are common in speech, sentences where the finite verb appears third as in (3) would have been frequent. Northern children at the boundary between the dialects would have heard such sentences and failed to recognize them as cues for verb-second. The resulting shortage of cues then led them to use the modern SVO word order by default.

3 The dynamical system

The proposed mechanism behind the loss of verb-second in Middle English may be expressed mathematically as follows. We first make the simplification of working with two grammars, G_1 and G_2, and assuming that people speak either one or the other, but not both. G_1 is analogous to

the northern SVO+CPv2 dialect, and G_2 may be interpreted as being analogous to either the southern SVO+FPv2+pro or the emerging SVO dialect. Some comments on these simplifying assumptions are called for. The historical situation seems to have involved at least three grammars, namely SVO+CPv2, SVO+FPv2+pro, and SVO. However, a model with two grammars suffices to illustrate how one grammar might displace another, and the resulting dynamical system is much simpler, requiring only two dimensions rather than the four required to express the model for three grammars. Manuscripts also suggest that speakers were diglossic, that is, they used mixtures of verb-second and non-verb-second grammars. The decline in sentences of the forms (1) and (3) seems to be due to a smooth shift among all speakers from using all verb-second to using no verb-second, rather than a decline of exclusively verb-second speakers in favor of exclusively SVO speakers. The model can be reformulated to include diglossia, but the mathematics is significantly more complicated and the overall behavior is essentially the same. So for now, we will ignore diglossia in formulating the model.

To model learning, we assume that the sentences accepted by G_1 form a superset of those accepted by G_2. Sentences not accepted by G_2 are therefore cues for G_1. Children hear n sentences total, and if m or more of them are cues, they acquire G_1 else they acquire G_2. Cue sentences are

produced frequently by speakers of G_1 at a rate $p_1 \approx 30\%$, and rarely by speakers of G_2 at a rate $p_2 \approx 5\%$. The choice of 30% is based on a figure cited in (Lightfoot, 1999). The choice of 5% is arbitrary, and was made to represent a reasonably large amount of noise in the PLD, due for example to exceptional phrases such as "Never before has such-and-such been attempted." The behavior of the model is essentially unchanged for a range of values of p_1 and p_2; the only requirement is that p_2 should be much smaller than p_1. Mathematically, acquisition is modeled by the function $q(x)$ as defined in equation (4), which represents the probability a child will acquire G_1 given that a fraction x of the surrounding population uses G_1 and a fraction $1-x$ use G_2:

$$q(x) = \sum_{j=m}^{n} \binom{n}{j} \gamma^j (1-\gamma)^{n-j} \quad \text{where } \gamma = p_1 x + p_2 (1-x) \tag{4}$$

The number γ is the probability that a cue sentence is spoken if a speaker is selected at random and asked to produce a sentence. With the choices $n = 100$ and $m = 20$, the function $q(x)$ has the shape shown in Figure 1.

@@ Insert Figure 1 Here

@@ Insert Figure 2 Here

The model population is divided into two regions, north and south. The state of the population consists of two functions of time, $x_N(t)$ and $x_S(t)$,

representing the fraction of the population in the two regions that speaks G_1 at time t. Both are between 0 and 1. A pair of differential equations defines how they change in time:

$$\begin{aligned} \dot{x}_N &= q(x_N) - x_N + \alpha(x_S - x_N) \\ \dot{x}_S &= q(x_S) - x_S + \beta(x_N - x_S) \end{aligned} \tag{5}$$

The dot represents the derivative with respect to time. For this formulation, the units of time have been scaled so that the birth and death rates are both 1. The $q(x)$ terms represent the fraction of births that yield a child who learns G_1. The $-x$ terms represent the death of speakers of G_1. The numbers α and β represent migration or mixing rates between the two regions. A system of differential equations such as (5) is known as a *dynamical system,* and is studied with techniques such as linear stability analysis and Lyapunov functions (Strogatz, 1994). A picture called a *phase portrait* describes the behavior of the system. See Figure 2. The arrows are called a *vector field* and represent the direction in which population states flow, as given by equation (5). There are nine singularities called *fixed points* in the vector field where $\dot{x}_N$ and $\dot{x}_S$ are both zero; these are denoted by dots and represent equilibrium states of the population, that is, states for which there is no tendency to flow. Some are stable, and the population will return to such a state if disturbed. These stable fixed points are called *sinks* and are drawn as black dots. The others are unstable, and come in two types. A *source* is the reverse of a

sink, and a population near but not exactly on top of a source will move away. Sources are denoted by white dots. A *saddle* repels most nearby populations, but they follow a path that initially brings them near the saddle, then swerve. Saddles are denoted by circles with crosses. Dotted lines (called *stable manifolds*) separate trajectories that swerve in different directions upon approaching a saddle point. Dashed lines (called *unstable manifolds*) are drawn along trajectories that flow most directly away from a saddle and attract those swerving trajectories. The locations and stabilities of the nine fixed points are determined by the shape of the $q(x)$ graph, specifically, where it crosses the line $q(x) = x$, and by the mixing parameters α and β.

In Figure 2, there are four sinks, one in each corner, representing the extreme states of the population where each region uses one grammar exclusively. This is the scenario when there is no mixing between the regions. The bottom right sink is analogous to the situation before the extinction of northern Middle English, where the population is split: The northern region speaks exclusively G_1, and the southern speaks exclusively G_2. Any population using a similar mixture of the two grammars will flow along the vector field and converge to the sink in the lower right. It will remain there indefinitely even in the presence of small perturbations. The sink in the lower left represents a population in which

both regions speak exclusively G_2, and attracts populations where G_1 is used by fewer than about 60% of northerners.

@@ Insert Figure 3 Here

Now consider what happens when the mixing parameters are stepped up to allow more interaction between the regions. See Figure 3. As α and β increase, the vector field changes, and the fixed points shift. Initially, the population can remain in a split state with most northerners speaking G_1 and most southerners speaking G_2. Eventually, the lower right sink collides with a nearby saddle in an event called a *saddle-node bifurcation.* The term *bifurcation* refers to the fact that two features of the phase portrait have collided and annihilated one another, and the name *saddle-node* refers to the fact that one of the features was a saddle and the other was a nodal sink (as opposed to a spiral sink, which does not occur in this model). After the bifurcation, there is no longer a stable split state for populations to converge to, so they are attracted to the sink in the bottom left corner, resulting in the extinction of G_1.

@@ Insert Figure 4 Here

We may also set the mixing parameters into continuous motion, as in Figure 4. The population is initially placed in a split state, with exclusively G_1 in the north and exclusively G_2 in the south. As the

mixing parameters slowly increase, the population tracks the bottom right sink as it shifts, maintaining a split state. When the bifurcation occurs, that sink vanishes, and the population flows to the bottom left sink and G_1 disappears. Thus, the timing of the loss of G_1 is determined by the timing and strength of the mixing between the two regions.

4 Comparison to memoryless learner

Niyogi and Berwick (1996) studied a simple learning algorithm called *memoryless learner,* which searches a universal grammar (UG) consisting of a finite set of grammars $U = \{G_1, G_2, \ldots, G_n\}$ as follows. It starts with a randomly selected hypothesis grammar $H \in U$. Given a sentence from the environment, if H can parse the sentence, the learner stays there, otherwise it switches to another randomly selected hypothesis, possibly one it has already visited, hence the term *memoryless*. The process ends after a fixed number of sentences, and the hypothesis at that point is the output of the algorithm. In one of their simulations, also discussed in (Lightfoot 1999), a model UG consisting of eight grammars determined by three binary parameters is studied under memoryless learning. Oddly, all verb-second languages are stable in this simulation and non-verb-second languages tend to extinction in favor of their verb-second counterparts. (A similar phenomenon was observed by Briscoe (2000) under certain circumstances in his more complex simulation.)

@@ Insert Figure 5 Here

The model (5) yields a mathematical explanation for this unexpected behavior. If we replace the cue-based learning algorithm with memorlyess learning on $U = \{G_1 = \text{SVO} + \text{v2}, G_2 = \text{SVO}\}$, then the function $q(x)$ changes shape dramatically, as in Figure 5. A single cue sentence is enough to cause a memoryless learner to choose SVO+v2 over SVO, and it will never have reason to switch hypotheses again. Because of this hypersensitivity, memoryless learners are unlikely to acquire G_2 even if the presence of G_1 in the population is minimal. This skews $q(x)$ and causes the phase portrait of (5) to appear as in Figure 6, where the only sink is in the upper right and represents a population where both regions speak exclusively G_1. There is no way for the bifurcation from Section 3 to take place, and there is no way for G_1 to disappear.

@@ Insert Figure 6 Here

In summary, the dynamical system presented here shows that memoryless learning is overly sensitive to noise, and cue-based learning provides a more historically accurate alternative.

5 Conclusion and future work

The dynamical system model presented in this paper shows that mathematical modeling techniques can help linguists to express

hypotheses about language change precisely, and to then use mathematics to understand how these models behave. Specifically, we have seen how the loss of a regional dialect of Middle English may be understood as a consequence of a saddle-node bifurcation. The model is compatible with the hypothesis that contact between regional dialects caused the loss of verb-second in Middle English. Furthermore, Lightfoot's cue-base acquisition algorithm provides a mechanism by which such a change might spread, but only when contact is sufficiently high. The time course of the change is directly tied to the strength and timing of the contact.

The model makes a number of simplifying assumptions that should be relaxed in future work. In particular, manuscript data suggests that speakers of Middle English used varying mixtures of verb-second and non-verb-second grammars. The model can be extended to allow for such speakers at the expense of additional complexity: Regions must now be represented as densities where $u(z,t)$ is the probability at time t that a speaker selected at random from the region uses G_1 a fraction z of the time. The result is an infinite dimensional differential equation, which could potentially have much more complex behavior than the two dimensional model discussed here. However, for reasonable choices of the learning process, the means of the regional densities obey the same dynamics as seen in Section 3, so this simplified model still gives useful results.

An alternative question is to study in more detail what learning processes generate the correct behavior in the infinite dimensional model. Currently, much less is known about how children acquire usage frequencies than how they might acquire particular syntactic structures, and the infinite dimensional model might shed some light on this subject.

The current model splits the population into two compartments, which is a fairly crude approximation to the spatial structure of medieval England. The model could be improved by adding additional spatial structure, for example, a network of discrete communities or a continuous population density. A network of discrete communities requires a higher dimensional dynamical system. A continuous population density requires a system of partial differential equations, which are generally much more difficult to understand than dynamical systems.

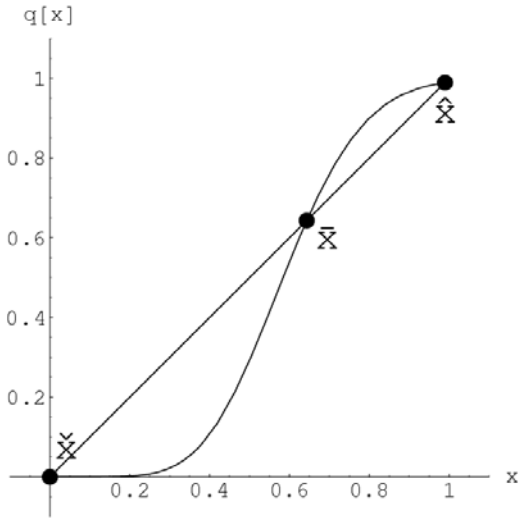
The dynamical system model is deterministic, so it requires an external event (the increase in contact between regional dialects) to initiate the syntactic change. It does not allow for the possibility that the loss of verb-second might have happened purely spontaneously. To remedy this situation, the model may be altered to include random events by reformulating it as a set of stochastic differential equations. Such an improvement would allow further investigation into how much of the change should be attributed to contact and how much to random chance,

but at the expense of substantially increasing the mathematical complexity of the model.

Other future work includes detailed simulations of individual agents that may speak many more possible grammars. The plan is to use the minimalist framework to construct grammars, and use ideas from (Yang 2002) as the basis of a learning algorithm. The population will include social and spatial structure as well as simulated literacy. Eventually, the results of the simulation will be compared to manuscript data from the Pennsylvania Parsed Corpus of Middle English, and simplified mathematical models will be constructed to better understand the essential details.

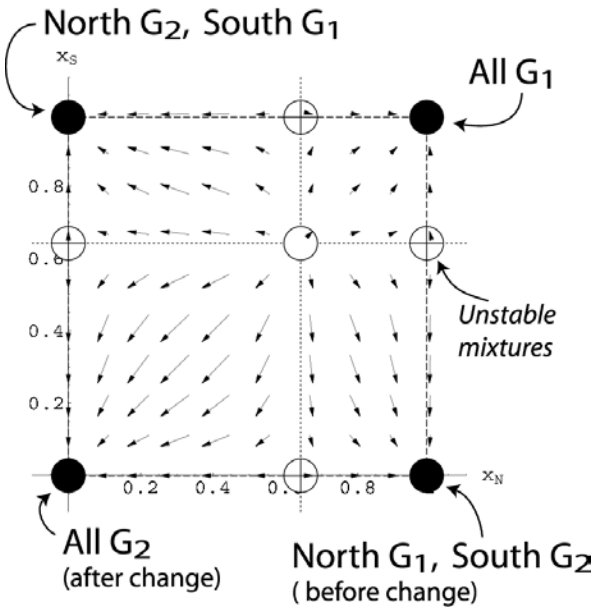
Figures

Figure 1



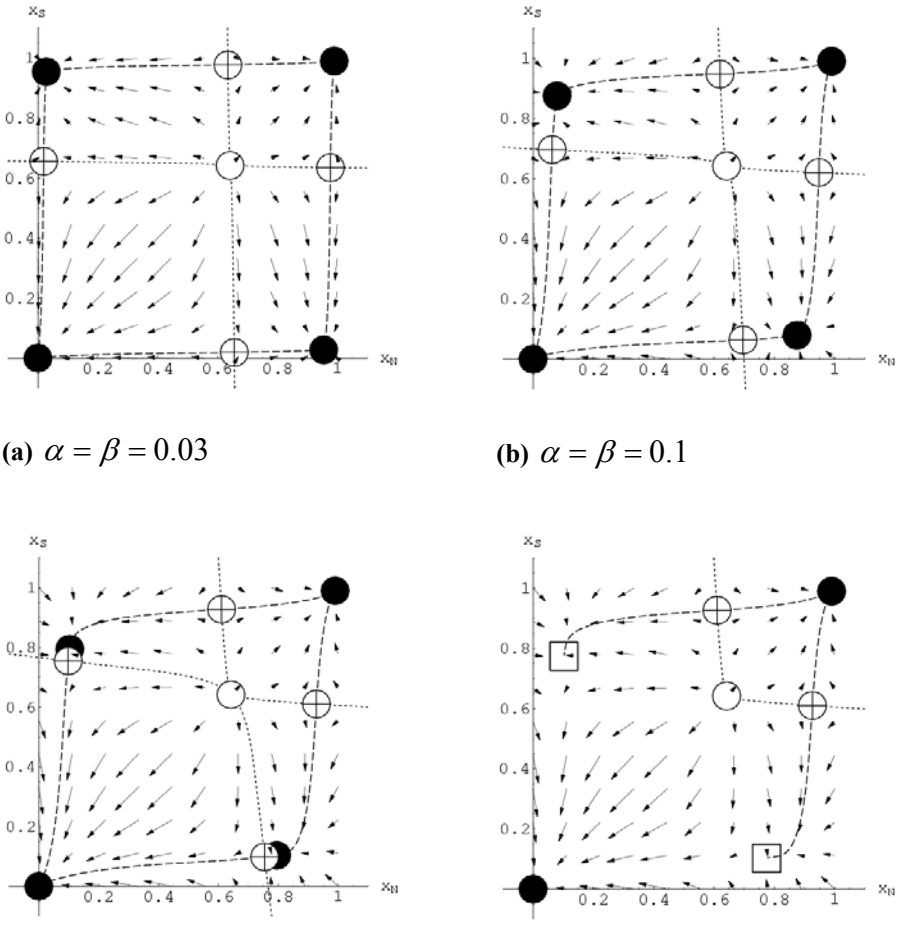
A graph of the learning function $q(x)$ for Lightfoot's cue-based learning algorithm. The line $q(x) = x$ and the three points of intersection are drawn for reference.

Figure 2



Phase portrait for the dynamical system, with no migration between regions, that is $\alpha = \beta = 0$.

Figure 3

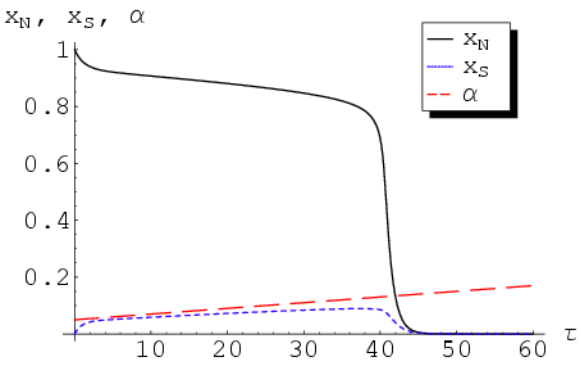


(a) $\alpha = \beta = 0.03$ **(b)** $\alpha = \beta = 0.1$

(c) $\alpha = \beta = 0.15$ **(d)** $\alpha = \beta = 0.152985\ldots$

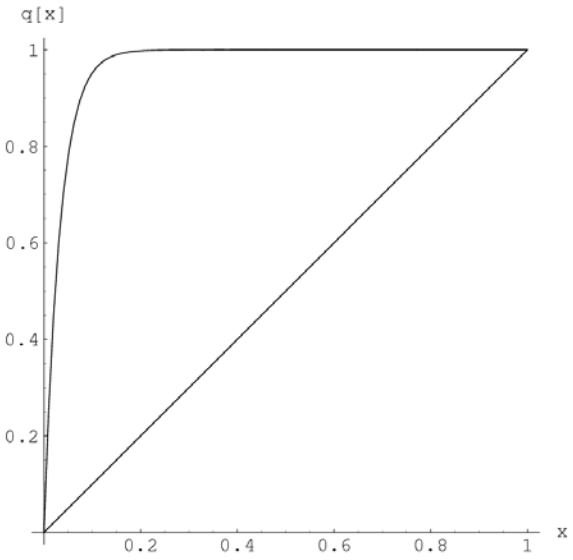
Phase portraits for different values of the mixing parameters. As mixing increases, the sinks in the upper left and lower right corners collide in a pair of saddle-node bifurcations. The square in picture (d) shows where the collision takes place.

Figure 4



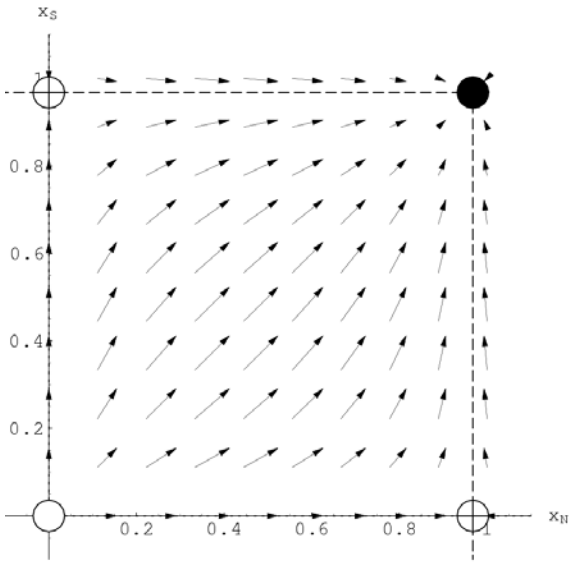
Time traces of the population when the mixing parameters increase smoothly: $\alpha = \beta =$ a linear function of time τ. The time axis is in rescaled units, not years. As the mixing parameters increase, the bifurcation takes place and x_N converges rapidly to 0.

Figure 5



A graph of $q(x)$ for the memoryless learner. The line $q(x) = x$ is included for reference.

Figure 6



Phase portrait for the memoryless learner and no mixing between regions.

All populations tend to the sink in the upper right.

References

Fischer, Olga, Ans van Kemenade, Willem Koopman & Wim van der Wurff. 2000. *The Syntax of Early English.* Cambridge University Press.

Kroch, Anthony. 1989. "Reflexes of grammar in patterns of language change." *Language Variation and Change* 1:3. 199-244.

Kroch, Anthony, Ann Taylor & Donald Ringe. 2000. "The Middle English verb-second constraint: A case study in language contact and language change." *Textual Parameters in Older Language.* ed. by Susan Herring, Pieter van Reenen & Lene Schløsler. John Benjamins Publishing Company: Philadelphia, Penn.

Lightfoot, David. 1999. The Development of Language: Acquisition, Changes and Evolution. Blackwell Publishers, Malden Mass.

Niyogi, P. and Robert C. Berwick. 1996. "A Language Learning Model for Finite Parameter Spaces." *Cognition* 61. 161-193.

Strogatz, Steven H. 1994. *Nonlinear Dynamics and Chaos.* Perseus Books: Reading, Mass.

Yang, Charles D. 2002. *Knowledge and Learning in Natural Language.* Oxford University Press.