MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1515 March, 1995
C.B.C.L. Paper No. 114

A Dynamical Systems Model for Language Change

Partha Niyogi & Robert C. Berwick
This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.

Abstract

Formalizing linguists' intuitions of language change as a dynamical system, we quantify the time course of language change including sudden vs. gradual changes in languages. We apply the computer model to the historical loss of Verb Second from Old French to modern French, showing that otherwise adequate grammatical theories can fail our new evolutionary criterion.

This report describes research done at the Center for Biological and Computational Learning and the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the Center is provided in part by a grant from the National Science Foundation under contract ASC–9217041. Robert C. Berwick was also supported by Vinton-Hayes Fellowship.

1 Introduction

Language scientists 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[1]. Here we formalize these intuitions, to the best of our knowledge for first time, as a concrete, computational, dynamical systems model, investigating its consequences. Specifically, we show that a computational population language change model emerges as a natural consequence of individual language learnability Our computational model establishes the following:

- *Learnability* is a well-known criterion for the adequacy of grammatical theories. Our model provides an *evolutionary* criterion: By comparing the trajectories of dynamical linguistic systems to historically observed trajectories, one can determine the adequacy of linguistic theories or learning algorithms.
- We derive explicit dynamical systems corresponding to parametrized linguistic theories (e.qg. Head First/Final parameter in HPSG or GB grammars) and memoryless language learning algorithms (e.g. gradient ascent in parameter space).
- We illustrate the use of dynamical systems as a research tool by considering the loss of Verb Second position in Old French as compared to Modern French. We demonstrate by computer modeling that one grammatical parameterization in the literature does not seem to permit this historical change, while another does. We can more accurately model the time course of language change. In particular, in contrast to Kroch (1989) and others, who mimic population biology models by imposing an S-shaped logistic change by *assumption*, we explain the time course of language change, and show that it need not be S-shaped. Rather, language-change envelopes are *derivable* from more fundamental properties of dynamical systems; sometimes they are S-shaped, but they can also be nonmonotonic.
- We examine by simulation and traditional phase-space plots the form and stability of possible "diachronic envelopes" given varying conditions of alternative language distributions, language acquisition algorithms, parameterizations, input noise, and sentence distributions systems.

2 The Acquisition-Based Model of Language Change

We first show how a combination of a grammatical theory and a learning paradigm leads directly to a formal

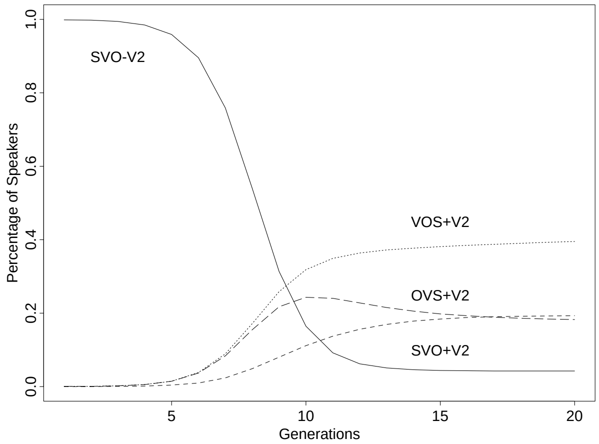

Figure 1: Time evolution of grammars using a greedy learning algorithm. The x-axis is generation time, e.g., units of 20-30 years. The y-axis is the percentage of the population speaking the languages as indicated on the curves, e.g, S(ubject) V(erb) O(bject), with no Verb Second= SVO−V2.

dynamical systems model of language change.

First, informally, consider an adult population speaking a particular language[2]. Individual children attempt to attain their caretaker target grammar. After a finite number of examples, some are successful, but others may misconverge. 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 generations, the linguistic composition evolves as a dynamical system. In the remainder of this paper we formalize this intuition, obtaining detailed figures like the one in 1, showing the evolution of language types over successive generations within a single community. We return to the details later, but let us first formalize our intuitions.

Grammatical theory, Learning Algorithm, 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 Σ in the usual way.

2. Denote by P, the distribution with which sentences of Σ^* are presented to the individual learner (child). More specifically, let P_i be the distribution with which sentences of the ith grammar $g_i \in \mathcal{G}$ are presented if there is a speaker of g_i in the adult population. Thus, if the adult population is linguistically homogeneous (with grammar g_1) then $P = P_1$. If the adult population speaks 50 percent $L(g_1)$ and 50 percent $L(g_2)$ then $P = \frac{1}{2}P_1 + \frac{1}{2}P_2$.

3. Denote by $\mathcal{A}$ the learning algorithm that children use to hypothesize a grammar on the basis of input data.

[1] One notable exception is Kroch, 1989, whose account we explore below.

[2] In our framework, this implies that the adult members of this population have internalized the same grammar.

If d_n is a presentation sequence of n randomly drawn examples, then learnability (Gold, 1967) requires (for every target grammar g_t),

$$Prob[\mathcal{A}(d_n) = g_t] \longrightarrow_{n\rightarrow\infty} 1$$

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

A State Space ($\mathcal{S}$): a set of system states. 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.[3]

An Update Rule: how the system states change from one time step to the next. Typically, this involves specifying a function, f, that maps $s_t \in S$ to s_{t+1}[4]

In our case the update rule can be derived directly from learning algorithm $\mathcal{A}$ because learning can change the distribution of languages spoken from one generation to the next. For example, given $P_{pop,t}$, we see that any any $\omega \in \Sigma^*$ is presented with probability $P(\omega) = \sum_i P_i(\omega)P_{pop,t}(i)$.

The learning algorithm $\mathcal{A}$ uses the linguistic data (n examples, indicated by d_n) and conjectures hypotheses ($\mathcal{A}(d_n) \in \mathcal{G}$). One can, in principle, compute this probability[5] with which the learner will develop an arbitrary hypothesis, h_i, after n examples:

$$\textbf{Finite Sample: } Prob[\mathcal{A}(d_n) = h_i] = p_n(h_i) \quad (1)$$

Learnability requires $p_n(g_t)$ to go to 1, for the unique target grammar, g_t, if such a grammar exists. In general, there is no unique target grammar since we have nonhomogeneous linguistic populations. However, the following limiting behavior can still exist:

$$\textbf{Limiting Sample: } \lim_{n\rightarrow\infty} Prob[\mathcal{A}(d_n) = h_i] = p_i \quad (2)$$

Thus, with probability $p_n(h_i)$,[6] an arbitrary child will have internalized grammar h_i. Thus, in the next generation, a proportion $p_n(h_i)$ of the population has grammar h_i, i.e., the linguistic composition of the next generation is given by $P_{pop,t+1}(h_i) = p_i$(or $p_n(h_i)$). In this fashion, we have an update rule,

$$P_{pop,t} \longrightarrow^{\mathcal{A}} P_{pop,t+1}$$

[3]As usual, one needs to be able to define a σ-algebra on the space of grammars, and so on. This is unproblematic for the cases considered in this paper because the set of grammars is finite.

[4]In general, this mapping could be fairly complicated. For example, it could depend on previous states, future states, and so forth; for reasons of space we do not consider all possibilities here. For reference, see Strogatz (1993).

[5]The finite sample situation is always well defined; see Niyogi, 1994.

[6]Or p_i, depending upon whether one wishes to carry out a finite sample, or a limiting sample analysis for learning within one generation.

Generality of the approach. Note that such a dynamical system exists for every choice of $\mathcal{A}$, $\mathcal{G}$, and P_i (relative to the constraints mentioned earlier). In short then,

$$(\mathcal{G}, \mathcal{A}, \{P_i\}) \longrightarrow \mathcal{D}(\text{ dynamical system})$$

Importantly, this formulation does *not* assume any particular linguistic theory, learning algorithm, or distribution over sentences.

3 Language Change in Parametric Systems

We next instantiate our abstract system by modeling some specific cases. Suppose we have a "parameterized" grammatical theory, such as HPSG or GB, with n boolean-valued parameters and a space $\mathcal{G}$ with 2^n different languages (in this case, equivalently, grammars). Further take the assumptions of Berwick and Niyogi (1994), regarding sentence distributions and learning: P_i is uniform on unembedded sentences generated by g_i and $\mathcal{A}$ is single step, gradient ascent. To derive the relevant update rule we need the following theorem and corollaries, given here without proof (see Niyogi, 1994):

Theorem 1 *Any memoryless incremental algorithm that attempts to set the values of the parameters on the basis of example sentences, can be modeled exactly by a Markov Chain. This Markov chain has 2^n states with state corresponding to a particular grammar. The transition probabilities depend upon the distribution P with which sentences occur, and the learning algorithm $\mathcal{A}$ (which is essentially a recursive function from data to hypotheses).*

Corollary 1 *The probability that the learner internalizes hypothesis h_i after m examples (solution to equation 1) is given by,*

$$Prob[\textit{ Learner's hypothesis } = h_i \in \mathcal{G} \textit{ after } m \textit{ examples}]$$

$$= \{\tfrac{1}{2^n}(1,\ldots,1)'T^m\}[i]$$

Similarly, making use of limiting distributions of Markov chains (see Resnick, 1992) one can obtain the following:

Corollary 2 *The probability that the learner internalizes hypothesis h_i "in the limit" (solution to equation 2) is given by*

$$Prob[\textit{ Learner's hypothesis } = h_i \textit{ "in the limit"}]$$

$$= (1,\ldots,1)'(I - T + ONE)^{-1}$$

where ONE is a $\frac{1}{2^n} \times \frac{1}{2^n}$ matrix with all ones.

This yields our required dynamical system for parameter-based theories:

1. Let $\mathbf{\Pi_1}$ be the initial population mix. Assume P_i's as above. Compute P according from $\pi_{\mathbf{1}}$, and P_i's.
2. Compute T (transition matrix) according to the theorem.
3. Use the corollaries to the theorems to obtain the update rule, to get the population mix $\mathbf{\Pi_2}$.
4. Repeat for the next generation.

4 Example 1: A Three Parameter System

Let us consider a specific example to illustrate the derivation of the previous section: the 3-parameter syntactic subsystem describe in Gibson and Wexler (1994) and Niyogi and Berwick (1994). Specifically, posit 3 Boolean parameters, Specifier first/final; Head first/final; Verb second allowed or not, leading to 8 possible grammars/languages (English and French, SVO−Verb second; Bengali and Hindi, SOV−Verb second; German and Dutch, SOV+Verb second; and so forth). The learning algorithm is single-step gradient ascent. For the moment, take P_i to be a uniform distribution on unembedded sentences in the language. Let us consider some results we obtain by simulating the resulting dynamical systems by computer. Our key results are these:

1. All +Verb second populations remain stable over time. Nonverb second populations tend to *gain* Verb second over time (e.g., English-type languages change to a more German type) contrary to historically observed phenomena (loss of Verb second in both French and English) and linguistic intuition (Lightfoot, 1991). This evolutionary behavior suggests that either the grammatical theory or the learning algorithm are incorrect, or both.

2. Rates of change can vary from gradual S-shaped curves (fig. 2) to more sudden changes (fig. 3).

3. Diachronic envelopes are often logistic, but not always. Note that in some alternative models of language change, the logistic shape has sometimes been *assumed* as a starting point, see, e.g., Kroch (1982, 1989). However, Kroch concedes that "unlike in the population biology case, no mechanism of change has been proposed from which the logistic form can be deduced". On the contrary, we propose that the logistic form is derivative, in that it sometimes arises from more fundamental assumptions about the grammatical theory, acquisition algorithm, and sentence distributions. Sometimes a logistic form is not even observed, as in fig. 3.

4. In many cases the homogeneous population splits into stable linguistic groups.

A variant of the learning algorithm (non-single step, gradient ascent) yields figure 1 shown at the beginning of this paper. Here again, populations tend to gain Verb-Second over time.

Next, see fig. 4 for the effect of maturation time on evolutionary trajectories.

Finally, so we have assumed that the P_i's were uniform. Fig. 5 shows the evolution of the L_2 (V O S +V2) speakers as p varies.

4.1 Nonhomogeneous Populations

Note that instead of starting with homogeneous populations, one could consider any nonhomogeneous initial condition, e.g. a mixture of English and German speakers. Each such initial condition results in a grammatical trajectory as shown in fig. 6. One typically characterizes dynamical systems by their phase-space plots. These contain all the trajectories corresponding to different initial conditions, exhibited in fig. 7.

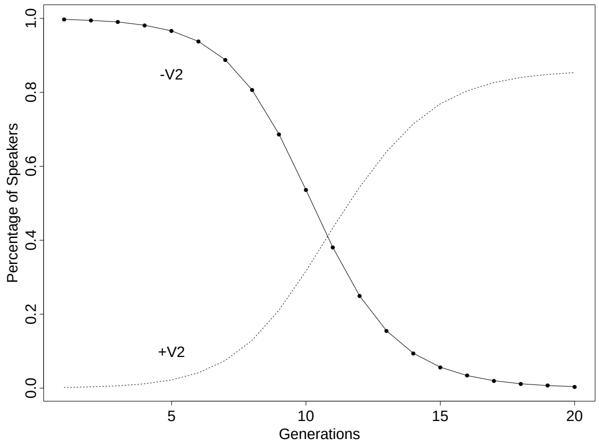

Figure 2: Percentage of the population speaking languages of the basic forms V(erb) O(bject) S(ubject) with and without Verb second. The evolution has been shown upto 20 generations, as the proportions do not vary significantly thereafter. Notice the "S" shaped nature of the curve (Kroch, 1989, imposes such a shape by fiat using models from population biology, while we derive this form as an emergent property of our dynamical model, given varying starting conditions). Also notice the region of maximum change as the Verb second parameter is slowly set by increasing proportion of the population, with no external influence.

Finally, the following theorem characterizes stable nonhomogeneous populations:

Theorem 2 (Finite Case) *A fixed point (stable point) of the grammatical dynamical system (obtained by a memoryless learner operating on the 3 parameter space with k examples to choose its mature hypothesis) is a solution of the following equation:*

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

If the learner were given infinite time to choose its hypothesis, then the fixed point is given by

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

where ONE is the 8×8 matrix with all its entries equal to 1.

Proof (Sketch): Both equations are obtained simply by setting $\Pi(t+1) = \Pi(t)$. ∎

Remark: Strogatz (1993) suggests that higher dimensional nonlinear mappings are likely to be chaotic. Since our systems fall into such a class, this possible chaotic behavior needs to be investigated further; we leave this for future publications.

5 The Case of Modern French

We briefly consider a different parametric system (studied by Clark and Roberts, 1993) as a test of our model's

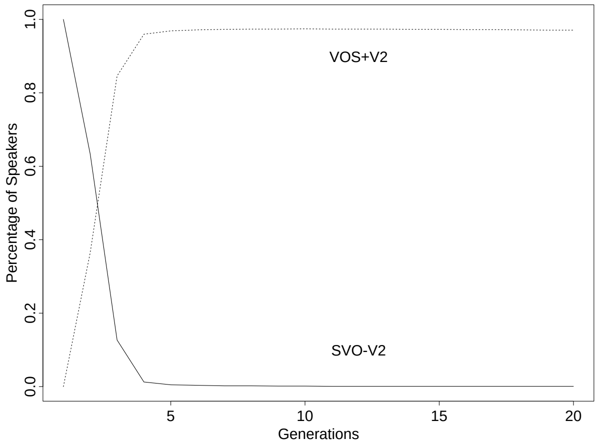

Figure 3: Percentage of the population speaking languages S V O −Verb second (English) and V O S (+Verb second) as it evolves over the number of generations. Notice the sudden shift over a space of 3-4 generations.

ability to impose a diachronic criterion on grammatical theories. The historical context in which we study this is the evolution of Modern French from Old French, in particular, the loss of Verb second.

Loss of Verb-Second (from Clark and Roberts, 1993)

Mod.	*Puis entendirent-ils un coup de tonerre.
	then they heard a clap of thunder.
Old	Lors oirent ils venir un escoiz de tonoire.
	then they heard come a clap of thunder

Recall that simulations in the previous section indicated an (historically incorrect) tendency to gain Verb second over time. We now consider Clark and Roberts' (1993) alternative 5-parameter grammatical theory. These parameters include: (1) Null subjects or not; (2) Verb second; and three other binary parameters that we need not detail here, yielding 32 possible languages (grammars). It has been generally argued that in the middle French period, word forms like Adv(erb) V(erb) S(ubject) decreased in frequency, while others like Adv S V increased; eventually bringing about a loss of Verb second. We can now test this hypothesis with the model, varying initial conditions about population mixtures, foreign speakers, etc.

Starting from just Old French, our model shows that, even without foreign intrusion, eventually speakers of Old French die out altogether, and within 20 generations, 15 percent of the speakers have lost Verb second completely; see fig. 8. However, note that this is not sufficient to attain Modern French, and the change is too slow. In order to more closely duplicate the historically observed trajectory, we consider an initial condition consisting more like that actually found: a mix of Old French and data from Modern French (reproducing the intrusion of foreign speakers and reproducing data similar to that obtained from the Middle French period, see Clark and Roberts, 1993 for justification).

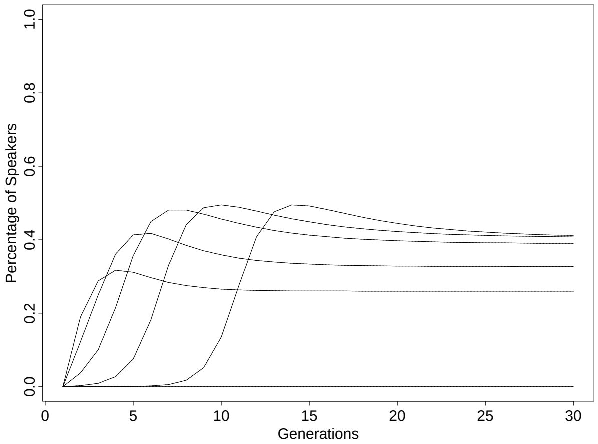

Figure 4: Time evolution of linguistic composition for the situations where the learning algorithm used is gradient ascent. Only the percentage of people speaking V(erb) O(bject) S(ubject) (+Verb second) is shown. The initial population is homogeneous and speaks V O S (−V2). The maturational time (number of sentences the child hears before internalizing a grammar) is varied through 8, 16, 32, 64, 128, 256, giving rise to six curves. The curve with the highest initial rate of change corresponds to the situation where only 8 examples were allowed to the learner to develop its mature hypothesis. 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.

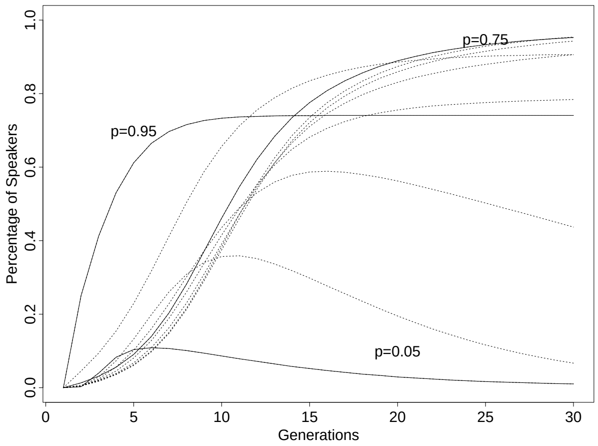

Figure 5: The evolution of V(erb) O(bject) S(ubject) +Verb second speakers in a community given different sentence distributions, P_i's. The P_i's were perturbed (with parameter p denoting the extent of the perturbation) around a uniform distribution. The algorithm used was single-step, gradient ascent. The initial population was homogeneous, with all members speaking a V(erb) O(bject) S(ubject) −Verb second type language. Curves for $p = 0.05, 0.75$, and 0.95 have been plotted as solid lines. If we wanted the population to completely *lose* the Verb second parameter, the optimal choice of p is 0.75 (not 1 as expected).

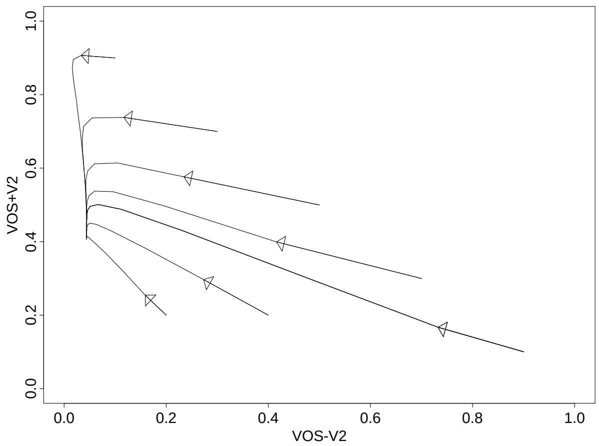

Figure 7: Subspace of a Phase-space plot. The plot shows the number of speakers of V(erb) O(bject) S(ubject) (−Verb second and +Verb second) as t varies. The learning algorithm was single step, gradient ascent. The different curves correspond to grammatical trajectories for different initial conditions.

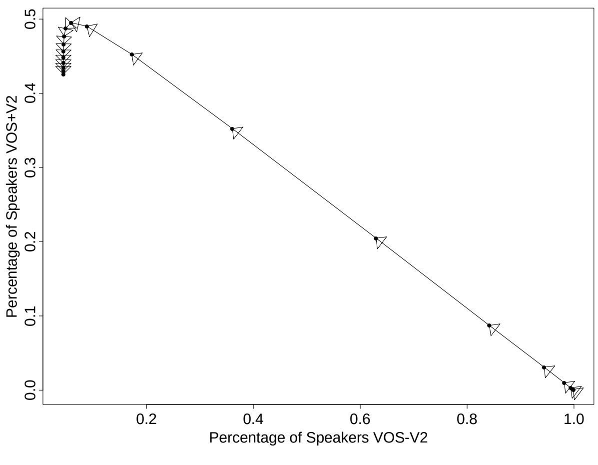

Figure 6: Subspace of a Phase-space plot. The plot shows the number of speakers of V(erb) O(bject) S(ubject) (−Verb second and +Verb second) as t varies. The learning algorithm was single step, gradient ascent.

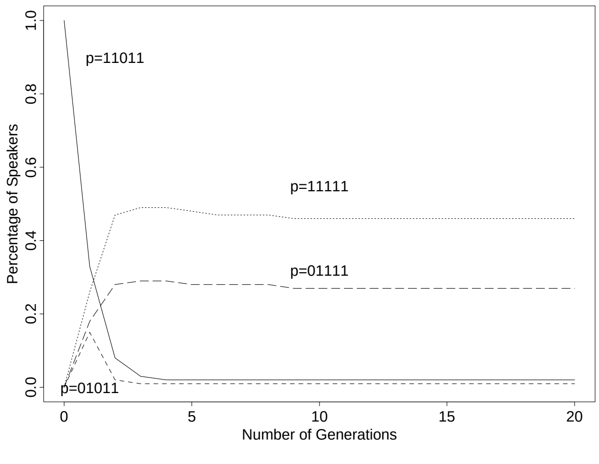

Figure 8: Evolution of speakers of different languages in a population starting off with speakers only of Old French. The "p" settings may be ignored here.

Given this new initial condition, fig. 9 shows the proportion of speakers losing Verb second after *one* generation as a function of the proportion of sentences from the "foreign" Modern French source. Surprisingly small proportions of Modern French cause a disproportionate number of speakers to lose Verb second, corresponding closely to the historically observed rapid change.

6 Conclusions

A learning theory (paradigm) attempts to account for how children (the individual child) solve the problem of language acquisition. By considering a population of such *individual* "child" learners, we arrive at a model of *emergent*, global, population language behavior. Consequently, whenever a linguist proposes a new grammatical or learning theory, they are also implicitly proposing a particular theory of language change, one whose consequences need to be examined. In particular, we saw the gain of Verb second in the 3-parameter case did not match historically observed patterns, but the 5-parameter system did. In this way the dynamical systems model supports the 5-parameter linguistic system to explain some changes in French. We have also greatly sharpened the informal notions of the time course of linguistic change, and grammatical stability. Such evolutionary systems are, we believe, useful for testing grammatical theories and explicitly modeling historical language change.

References

[Altmann1982] G. et al Altmann. 1982. A law of change in language. In B. Brainard, editor, *Historical Linguistics*, pages 104–115, Studienverlag Dr. N. Brockmeyer. Bochum, FRG.

[Clark and Roberts1993] R. Clark and I. Roberts. 1993. A computational model of language learnability and language change. *Linguistic Inquiry*, 24(2):299–345.

[Gibson and Wexler1994] E. Gibson and K. Wexler. 1994. Triggers. Linguistic Inquiry, To appear.

[Haegeman1991] L. Haegeman. 1991. *Introduction to Government and Binding Theory*. Blackwell: Cambridge, USA.

[Kroch1982] A. S. Kroch. 1982. Grammatical theory and the quantitative study of syntactic change. In *Paper presented at NWAVE 11, Georgetown University*.

[Kroch1989] A. S. Kroch. 1989. Function and grammar in the history of english: Periphrastic "do.". In Ralph Fasold, editor, *Language change and variation*. Amsterdam:Benjamins. 133-172.

[Kroch1990] Anthony S. Kroch. 1990. Reflexes of grammar in patterns of language change. *Language Variation and Change*, pages 199–243.

[Lightfoot1991] D. Lightfoot. 1991. *How to Set Parameters*. MIT Press, Cambridge, MA.

[Niyogi1994] P. Niyogi. 1994. *The Informational Complexity of Learning From Examples*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

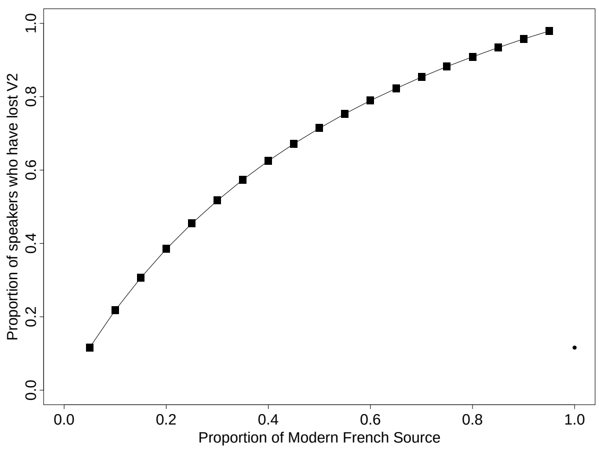

Figure 9: Tendency to lose Verb second as a result of new word orders introduced by Modern French sources in the dynamical systems model.

[Resnick1992] S. Resnick. 1992. *Adventures in Stochastic Processes*. Birkhauser.

[Strogatz1993] S. Strogatz. 1993. *Nonlinear Dynamics and Chaos*. Addison-Wesley.

References

[1] R. Clark and I. Roberts. A computational model of language learnability and language change. *Linguistic Inquiry*, 24(2):299–345, 1993.

[2] G. Altmann et al. A law of change in language. In B. Brainard, editor, *Historical Linguistics*, pages 104–115, Studienverlag Dr. N. Brockmeyer., 1982. Bochum, FRG.

[3] E. Gibson and K. Wexler. Triggers. *Linguistic Inquiry*, 25, 1994.

[4] L. Haegeman. *Introduction to Government and Binding Theory*. Blackwell: Cambridge, USA, 1991.

[5] A. S. Kroch. Grammatical theory and the quantitative study of syntactic change. In *Paper presented at NWAVE 11, Georgetown Universtiy*, 1982.

[6] A. S. Kroch. Function and gramar in the history of english: Periphrastic "do.". In Ralph Fasold, editor, *Language change and variation*. Amsterdam:Benjamins. 133-172, 1989.

[7] Anthony S. Kroch. Reflexes of grammar in patterns of language change. *Language Variation and Change*, pages 199–243, 1990.

[8] D. Lightfoot. *How to Set Parameters*. MIT Press, Cambridge, MA, 1991.

[9] S. Resnick. *Adventures in Stochastic Processes*. Birkhauser, 1992.

[10] S. Strogatz. *Nonlinear Dynamics and Chaos*. Addison-Wesley, 1993.