Simulating Language Change in the Presence of Non-Idealized Syntax (Corrected)

W. Garrett Mitchener
Mathematics Department
Duke University
Box 90320
Durham, NC 27708
wgm@math.duke.edu

Abstract

Both Middle English and Old French had a syntactic property called *verb-second* or *V2* that disappeared. This paper describes a simulation being developed to shed light on the question of why V2 is stable in some languages, but not others. The simulation, based on a Markov chain, uses *fuzzy grammars* where speakers can use an arbitrary mixture of idealized grammars. Thus, it can mimic the variable syntax observed in Middle English manuscripts. The simulation supports the hypotheses that children use the topic of a sentence for word order acquisition, that acquisition takes into account the ambiguity of grammatical information available from sample sentences, and that speakers prefer to speak with more regularity than they observe in the primary linguistic data.

1 Introduction

The paradox of language change is that on the one hand, children seem to learn the language of their parents very robustly, and yet for example, the English spoken in 800 AD is foreign to speakers of Modern English, and Latin somehow diverged into numerous mutually foreign languages. A number of models and simulations have been studied using historical linguistics and acquisition studies to build on one another (Yang, 2002; Lightfoot, 1999; Niyogi and Berwick, 1996). This paper describes the initial stages of a long term project undertaken in consultation with Anthony Kroch, designed to integrate knowledge from these and other areas of linguistics into a mathematical model of the entire history of English. As a first step, this paper examines the verb-second phenomenon, which has caused some difficulty in other simulations. The history of English and other languages requires simulated populations to have certain long-term behaviors. Assuming that syntax can change without a non-syntactic driving force, these requirements place informative restrictions on the acquisition algorithm. Specifically, the behavior of this simulation suggests that children are aware of the topic of a sentence and use it during acquisition, that children take into account whether or not a sentence can be parsed by multiple hypothetical grammars, and that speakers are aware of variety in their linguistic environment but do not make as much use of it individually.

As discussed in (Yang, 2002) and (Kroch, 1989), both Middle English and Old French had a syntactic rule, typical of Germanic languages, known as *verb-second* or *V2*, in which top-level sentences are re-organized: The finite verb moves to the front, and the topic moves in front of that. These two languages both lost V2 word order. Yang (2002) also states that other Romance languages once had V2 and lost it. However, Middle English is the only Germanic language to have lost V2.

A current hypothesis for how V2 is acquired supposes that children listen for *cue sentences* that cannot be parsed without V2 (Lightfoot, 1999). Specifically, sentences with an initial non-subject topic and finite verb are the cues for V2:

(1) $[_{CP}$ TopicXP $_{C}$V $[_{IP}$ Subject . . .]]

(2) [[On þis gær] *wolde* [þe king Stephne tæcen. . .]]
[[in this year] *wanted* [the king Stephen seize. . .]]
'During this year king Stephen wanted to seize. . . '
(Fischer et al., 2000, p. 130)

This hypothesis suggests that the loss of V2 can be attributed to a decline in cue sentences in speech. Once the change is actuated, feedback from the learning process propels it to completion.

Several questions immediately arise: Can the initial decline happen spontaneously, as a consequence of purely linguistic factors? Specifically, can a purely syntactic force cause the decline of cue sentences, or must it be driven by a phonological or morphological change? Alternatively, given the robustness of child language acquisition, must the initial decline be due to an external event, such as contact or social upheaval? Finally, why did Middle English and Old French lose V2, but not German, Yiddish, or Icelandic? And what can all of this say about the acquisition process?

Yang and Kroch suggest the following hypothesis concerning why some V2 languages, but not all, are unstable. Middle English (specifically, the southern dialects) and Old French had particular features that obscured the evidence for V2 present in the primary linguistic data available for children:

- Both had underlying subject-verb-object (SVO) word order. For a declarative sentence with topicalized subject, an SVO+V2 grammar generates the same surface word order as an SVO grammar without V2. Hence, such sentences are uninformative as to whether children should use V2 or not. According to estimates quoted in (Yang, 2002) and (Lightfoot, 1999), about 70% of sentences in modern V2 languages fall into this category.

- Both allowed sentence-initial adjuncts, which came before the fronted topic and verb.

- Subject pronouns were different from full NP subjects in both languages. In Middle English, subject pronouns had clitic-like properties that caused them to appear to the left of the finite verb, thereby placing the verb in third position. Old French was a pro-drop language, so subject pronouns could be omitted, leaving the verb first.

The Middle English was even more complex due to its regional dialects. The northern dialect was heavily influenced by Scandinavian invaders: Sentence-initial adjuncts were not used, and subject pronouns were treated the same as full NP subjects.

Other Germanic languages have some of these factors, but not all. For example, Icelandic has underlying SVO order but does not allow additional adjuncts. It is therefore reasonable to suppose that these confounds increase the probability that natural variation or an external influence might disturb the occurrence rate of cue sentences enough to actuate the loss of V2.

An additional complication, exposed by manuscript data, is that the population seems to progress as a whole. There is no indication that some speakers use a V2 grammar exclusively and the rest never use V2, with the decline in V2 coming from a reduction in the number of exclusively V2 speakers. Instead, manuscripts show highly variable rates of use of unambiguously V2 sentences, suggesting that all individuals used V2 at varying rates, and that the overall rate decreased from generation to generation. Furthermore, children seem to use mixtures of adult grammars during acquisition (Yang, 2002). These features suggest that modeling only idealized adult speech may not be sufficient; rather, the mixed speech of children and adults in a transitional environment is crucial to formulating a model that can be compared to acquisition and manuscript data.

A number of models and simulations of language learning and change have been formulated (Niyogi and Berwick, 1996; Niyogi and Berwick, 1997; Briscoe, 2000; Gibson and Wexler, 1994; Mitchener, 2003; Mitchener and Nowak, 2003; Mitchener and Nowak, 2004; Komarova et al., 2001) based on the simplifying assumption that speakers use one grammar exclusively. Frequently, V2 can never be lost in such simulations, perhaps because the learning algorithm is highly sensitive to noise. For example,

a simple batch learner that accumulates sample sentences and tries to pick a grammar consistent with all of them might end up with a V2 grammar on the basis of a single cue sentence.

The present work is concerned with developing an improved simulation framework for investigating syntactic change. The simulated population consists of individual simulated people called *agents* that can use arbitrary mixtures of idealized grammars called *fuzzy grammars*. Fuzzy grammars enable the simulation to replicate smooth, population-wide transitions from one dominant idealized grammar to another. Fuzzy grammars require a more sophisticated learning algorithm than would be required for an agent to acquire a single idealized grammar: Agents must acquire usage rates for the different idealized grammars rather than a small set of discrete parameter values.

2 Linguistic specifics of the simulation

The change of interest is the loss of V2 in Middle English and Old French, in particular why V2 was unstable in these languages but not in others. Therefore, the idealized grammars allowed in this simulation will be limited to four: All have underlying subject-verb-object word order, and allow sentence-initial adjuncts. The options are V2 or not, and pro-drop or not. Thus, a grammar is specified by a pair of binary parameter values. For simplicity, the pro-drop parameter as in Old French is used rather than trying to model the clitic status of Middle English subject pronouns.

Sentences are limited to a few basic types of declarative statements, following the degree-0 learning hypothesis (Lightfoot, 1999): The sentence may or may not begin with an adjunct, the subject may be either a full noun phrase or a pronoun, and the verb may optionally require an object or a subject. A verb, such as *rain*, that does not require a subject is given an expletive pronoun subject if the grammar is not pro-drop. Additionally, either the adjunct, the subject, or the object may be topicalized. For a V2 grammar, the topicalized constituent appears just before the verb; otherwise it is indicated only by spoken emphasis.

A fuzzy grammar consists of a pair of beta distributions with parameters α and β, following the convention from (Gelman et al., 2004) that the density for $\mathrm{Beta}(\alpha, \beta)$ is

$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, 0 < x < 1. \tag{3}$$

Each beta distribution controls one parameter in the idealized grammar.[1] The special case of $\mathrm{Beta}(1, 1)$ is the uniform distribution, and two such distributions are used as the initial state for the agent's fuzzy grammar. The density for $\mathrm{Beta}(1 + m, 1 + n)$ is a bump with peak at $m/(m + n)$ that grows sharper for larger values of m and n. Thus, it incorporates a natural critical period, as each additional data point changes the mean less and less, while allowing for variation in adult grammars as seen in manuscripts.

To produce a sentence, an agent with fuzzy grammar $(\mathrm{Beta}(\alpha_1, \beta_1), \mathrm{Beta}(\alpha_2, \beta_2))$ constructs an idealized grammar from a pair of random parameter settings, each 0 or 1, selected as follows. The agent picks a random number $Q_j \sim \mathrm{Beta}(\alpha_j, \beta_j)$, then sets parameter j to 1 with probability Q_j and 0 with probability $1 - Q_j$. An equivalent and faster operation is to set parameter j to 1 with probability μ_j and 0 with probability $1 - \mu_j$, where $\mu_j = \alpha_j/(\alpha_j + \beta_j)$ is the mean of $\mathrm{Beta}(\alpha_j, \beta_j)$.

To learn from a sentence, an agent first constructs a random idealized grammar as before. If the grammar can parse the sentence, then some of the agent's beta distributions are adjusted to increase the probability that the successful grammar is selected again. If the grammar cannot parse the sentence, then no adjustment is made. To adjust $\mathrm{Beta}(\alpha, \beta)$ to favor 1, the agent increments the first parameter, yielding $\mathrm{Beta}(\alpha + 1, \beta)$. To adjust it to favor 0, the agent increments the second parameter, yielding $\mathrm{Beta}(\alpha, \beta + 1)$.

Within this general framework, many variations are possible. For example, the initial state of an agent, the choice of which beta distributions to update for particular sentences, and the social structure (who speaks to who) may all be varied.

The simulation in (Briscoe, 2002) also makes use of Bayesian learning, but within an algorithm for which learners switch abruptly from one idealized

[1] The beta distribution is the conjugate prior for using Bayesian inference to estimate the probability a biased coin will come up heads: If the prior distribution is $\mathrm{Beta}(\alpha, \beta)$, the posterior after m heads and n tails is $\mathrm{Beta}(\alpha + m, \beta + n)$.

grammar to another as estimated probabilities cross certain thresholds. The smoother algorithm used here is preferable because children do not switch abruptly between grammars (Yang, 2002). Furthermore, this algorithm allows simulations to include children's highly variable speech. Children learning from each other is thought be an important force in certain language changes; for example, a recent change in the Icelandic case system, known as dative sickness, is thought to be spreading through this mechanism.

3 Adaptation for Markov chain analysis

To the learning model outlined so far, we add the following restrictions. The social structure is fixed in a loop: There are n agents, each of which converses with its two neighbors. The parameters α_j and β_j are restricted to be between 1 and N. Thus, the population can be in one of N^{4n} possible states, which is large but finite.

Time is discrete with each time increment representing a single sentence spoken by some agent to a neighbor. The population is represented by a sequence of states $(X_t)_{t\in\mathbf{Z}}$. The population is updated as follows by a transition function $X_{t+1} = \phi(X_t, U_t)$ that is fed the current population state plus a tuple of random numbers U_t. One agent is selected uniformly at random to be the hearer. With probability p_r, that agent dies and is replaced by a baby in an initial state $(\mathrm{Beta}(1,1), \mathrm{Beta}(1,1))$. With probability $1 - p_r$, the agent survives and hears a sentence spoken by a randomly selected neighbor.

Two variations of the learning process are explored here. The first, called LEARN-ALWAYS, serves as a base line: The hearer picks an idealized grammar according to its fuzzy grammar, and tries to parse the sentence. If it succeeds, it updates any one beta distribution selected at random in favor of the parameter that led to a successful parse. If the parse fails, no update is made. This algorithm is similar to Naive Parameter Learning with Batch (Yang, 2002, p. 24), but adapted to learn a fuzzy grammar rather than an idealized grammar, and to update the agent's knowledge of only one syntactic parameter at a time.

The second, called PARAMETER-CRUCIAL, is the same except that the parameter is only updated if it is *crucial* to the parse: The agent tries to parse the sentence with that parameter in the other setting. If the second parse succeeds, then the parameter is not considered crucial and is left unchanged, but if it fails, then the parameter is crucial and the original setting is reinforced. This algorithm builds on LEARN-ALWAYS by restricting learning to sentences that are more or less unambiguous cues for the speaker's setting for one of the syntactic parameters. The theory of cue-based learning assumes that children incorporate particular features into their grammar upon hearing specific sentences that unambiguously require them. This process is thought to be a significant factor in language change (Lightfoot, 1999) as it provides a feedback mechanism: Once a parameter setting begins to decline, cues for it will become less frequent in the population, resulting in further decline in the next generation. A difficulty with the theory of cue-based learning is that it is unclear what exactly "unambiguous" should mean, because realistic language models generally have cases where no single sentence type is unique to a particular grammar or parameter setting (Yang, 2002, p. 34, 39). The definition of a crucial parameter preserves the spirit of cue-based learning while avoiding potential difficulties inherent in the concept of "unambiguous."

These modifications result in a finite-state Markov chain with several useful properties. It is *irreducible*, which means that there is a strictly positive probability of eventually getting from any initial state to any other target state. To see this, observe that there is a tiny but strictly positive probability that in the next several transitions, all the agents will die and the following sentence exchanges will happen just right to bring the population to the target state. This Markov chain is also *aperiodic*, which means that at any time t far enough into the future, there is a strictly positive probability that the chain will have returned to its original state. Aperiodicity is a consequence of irreducibility and the fact that there is a strictly positive probability that the chain does not change states from one time step to the next. That happens when a hearer fails to parse a sentence, for example. An irreducible aperiodic Markov chain always has a *stationary distribution.* This is a probability distribution on its states, normally denoted π, such that the probability that $X_t = x$ converges to

$\pi(x)$ as $t \to \infty$ no matter what the initial state X_0 is. Furthermore, the transition function preserves π, which means that if X is distributed according to π, then so is $\phi(X, U)$. The stationary distribution represents the long term behavior of the Markov chain.

Agents have a natural partial ordering $\succeq$ defined by

$$\begin{aligned}(\mathrm{Beta}(\alpha_1, \beta_1), \mathrm{Beta}(\alpha_2, \beta_2)) \\ \succeq (\mathrm{Beta}(\alpha_1', \beta_1'), \mathrm{Beta}(\alpha_2', \beta_2')) \\ \text{if and only if} \\ \alpha_1 \geq \alpha_1', \beta_1 \leq \beta_1', \alpha_2 \geq \alpha_2', \text{ and } \beta_2 \leq \beta_2'. \quad (4)\end{aligned}$$

This ordering means that the left-hand agent is slanted more toward 1 in both parameters. Not all pairs of agent states are comparable, but there are unique maximum and minimum agent states under this partial ordering,

$$\begin{aligned}A_{\max} &= (\mathrm{Beta}(N, 1), \mathrm{Beta}(N, 1)), \\ A_{\min} &= (\mathrm{Beta}(1, N), \mathrm{Beta}(1, N)),\end{aligned}$$

such that all agent states A satisfy $A_{\max} \succeq A \succeq A_{\min}$. Let us consider two population states X and Y and denote the agents in X by A_j and the agents in Y by B_j, where $1 \leq j \leq n$. The population states may also be partially ordered, as we can define $X \succeq Y$ to mean all corresponding agents satisfy $A_j \succeq B_j$. There are also maximum and minimum population states $X_{\max}$ and $X_{\min}$ defined by setting all agent states to $A_{\max}$ and $A_{\min}$, respectively.

A Markov chain is *monotonic* if the set of states has a partial ordering with maximum and minimum elements and a transition function that respects that ordering. There is a perfect sampling algorithm called *monotonic coupling from the past (MCFTP)* that generates samples from the stationary distribution π of a monotonic Markov chain without requiring certain properties of it that are difficult to compute (Propp and Wilson, 1996). The partial ordering $\succeq$ on population states was constructed so that this algorithm could be used. The transition function ϕ mostly respects this partial ordering, that is, if $X \succeq Y$, then with high probability $\phi(X, U) \succeq \phi(Y, U)$. This monotonicity property is why ϕ was defined to change only one agent per time step, and why the learning algorithms change that agent's knowledge of at most one parameter per time step. However, ϕ does not quite respect $\succeq$, because one can construct X, Y, and U such that $X \succeq Y$ but $\phi(X, U)$ and $\phi(Y, U)$ are not comparable. So, MCFTP does not necessarily produce correctly distributed samples. However, it turns out to be a reasonable heuristic, and until further theory can be developed and applied to this problem, it is the best that can be done.

The MCFTP algorithm works as follows. We suppose that $(U_t)_{t \in \mathbf{Z}}$ is a sequence of tuples of random numbers, and that $(X_t)_{t \in \mathbf{Z}}$ is a sequence of random states such that each X_t is distributed according to π and $X_{t+1} = \phi(X_t, U_t)$. We will determine X_0 and return it as the random sample from the distribution π. To determine X_0, we start at time $T < 0$ with a list of all possible states, and compute their futures using ϕ and the sequence of U_t. If ϕ has been chosen properly, many of these paths will converge, and with any luck, at time 0 they will all be in the same state. If this happens, then we have found a time T such that no matter what X_T is, there is only one possible value for X_0, and that random state is distributed according to π as desired. Otherwise, we continue, starting twice as far back at time $2T$, and so on. This procedure is generally impractical if the number of possible states is large. However, if the Markov chain is monotonic, we can take the shortcut of only looking at the two paths starting at $X_{\max}$ and $X_{\min}$ at time T. If these agree at time 0, then all other paths are squeezed in between and must agree as well.

4 Tweaking

Since this simulation is intended to be used to study the loss of V2, certain long term behavior is desirable. Of the four idealized grammars available in this simulation, three ought to be fairly stable, since there are languages of these types that have retained these properties for a long time: SVO (French, English), SVO+V2 (Icelandic), and SVO+pro-drop (Spanish). The fourth, SVO+V2+pro-drop, ought to be unstable and give way to SVO+pro-drop, since it approximates Old French before it changed. In any case, the population ought to spend most of its time in states where most of the agents use one of the four grammars predominantly, and neighboring agents should have similar fuzzy grammars.

In preliminary experiments, the set of possible

sentences did not contain expletive subject pronouns, sentence initial adverbs, or any indication of spoken stress. Thus, the simulated SVO language was a subset of all the others, and SVO+pro-drop was a subset of SVO+V2+pro-drop. Consequently, the PARAMETER-CRUCIAL learning algorithm was unable to learn either of these languages because the non-V2 setting was never crucial: Any sentence that could be parsed without V2 could also be parsed with it. In later experiments, the sentences and grammars were modified to include expletive pronouns, thereby ensuring that SVO is not a subset of SVO+pro-drop or SVO+V2+pro-drop. In addition, marks were added to sentences to indicate spoken stress on the topic. In the simulated V2 languages, topics are always fronted, so such stress can only appear on the initial constituent, but in the simulated non-V2 languages it can appear on any constituent. This modification ensures that no language within the simulation is a subset of any of the others.

The addition of spoken stress is theoretically plausible for several reasons. First, the acquisition of word order and case marking requires children to infer the subject and object of sample sentences, meaning that such thematic information is available from context. It is therefore reasonable to assume that the thematic context also allows for inference of the topic. Second, Chinese allows topics to be dropped where permitted by discourse, a feature also observed in the speech of children learning English. These considerations, along with the fact that the simulation works much better with topic markings than without, suggests that spoken emphasis on the topic provides positive evidence that children use to determine that a language is not V2.

It turns out that the maximum value N allowed for α_j and β_j must be rather large. If it is too small, the population tends to converge to a *saturated* state where all the agents are approximately $\hat{A} = (\mathrm{Beta}(N, N), \mathrm{Beta}(N, N))$. This state represents an even mixture of all four grammars and is clearly unrealistic. To see why this happens, imagine a fixed linguistic environment and an isolated agent learning from this environment with no birth-and-death process. This process is a Markov chain with a single absorbing state $\hat{A}$, meaning that once the learner reaches state $\hat{A}$ it cannot change to any other state: Every learning step requires increasing one of the numerical parameters in the agent's state, and if they are all maximal, then no further change can take place. Starting from any initial state, the agent will eventually reach the absorbing state. The number of states for an agent must be finite for practical and theoretical reasons, but by making N very large, the time it takes for an agent to reach $\hat{A}$ becomes far greater than its life span under the birth-and-death process, thereby avoiding the saturation problem. With $p_r = 0.001$, it turns out that 5000 is an appropriate value for N, and effectively no agents come close to saturation.

After some preliminary runs, the LEARN-ALWAYS algorithm seemed to produce extremely incoherent populations with no global or local consensus on a dominant grammar. Furthermore, MCFTP was taking an extremely long time under the PARAMETER-CRUCIAL algorithm. An additional modification was put in place to encourage agents toward using predominantly one grammar. The best results were obtained by modifying the speaking algorithm so that agents prefer to speak more toward an extreme than the linguistic data would indicate. For example, if the data suggests that they should use V2 with a high probability of 0.7, then they use V2 with some higher probability, say, 0.8. If the data suggests a low value, say 0.3, then they use an even lower value, say 0.2. The original algorithm used the mean μ_j of beta distribution $\mathrm{Beta}(\alpha_j, \beta_j)$ as the probability of using 1 for parameter j. The biased speech algorithm uses $f(\mu_j)$ instead, where f is a sigmoid function

$$f(\mu) = \frac{1}{1 + \exp(2k - 4k\mu)} \tag{5}$$

that satisfies $f(1/2) = 1/2$ and $f'(1/2) = k$. The numerical parameter k can be varied to exaggerate the effect. This modification leads to some increase in coherence with the LEARN-ALWAYS algorithm; it has minimal effect on the samples obtained with the PARAMETER-CRUCIAL algorithm, however MCFTP becomes significantly faster.

The biased speech algorithm can be viewed as a smoother form of the thresholding operation used in (Briscoe, 2002), discussed earlier. An alternative interpretation is that the acquisition process may involve biased estimates of the usage frequencies of syntactic constructions. Language acquisition re-

quires children to impose regularity on sample data, leading to creoles and regularization of vocabulary, for instance (Bickerton, 1981; Kirby, 2001). This addition to the simulation is therefore psychologically plausible.

5 Results

In all of the following results, the bound on α_j and β_j is $N = 5000$, the sigmoid slope is $k = 2$, the probability that an agent is replaced when selected is $p_r = 0.001$, and there are 40 agents in the population configured in a loop where each agent talks to its two neighbors. See Figure 1 for a key to the notation used in the figures.

First, let us consider the base line LEARN-ALWAYS algorithm. Typical sample populations, such as the one shown in Figure 2, tend to have large transition areas between regions of coherence.[2]

A sample run using the PARAMETER-CRUCIAL learning algorithm is shown in Figure 3. This population is quite coherent, with neighbors generally favoring similar grammars, and most speakers using non-V2 languages. Remember that the picture represents the internal data of each agent, and that their speech is biased to be more regular than their experience. There is a region of SVO+V2 spanning the second row, and a region of SVO+pro-drop on the fourth row with some SVO+V2+pro-drop speakers. Another sample dominated by V2 with larger regions of SVO+V2+pro-drop is shown in Figure 4. A third sample dominated by non-pro-drop speakers is shown in Figure 5. The MCFTP algorithm starts with a population of all $A_{\max}$ and one of $A_{\min}$ and returns a sample that is a possible future of both; hence, both V2 and pro-drop may be lost and gained under this simulation.

In addition to sampling from the stationary distribution π of a Markov chain, MCFTP estimates the chain's *mixing time*, which is how large t must be for the distribution of X_t to be ε-close to π (in total variation distance). The mixing time is roughly how long the chain must run before it "forgets" its initial state. Since this Markov chain is not quite monotonic, the following should be considered a heuristic back-of-the-napkin calculation for the order of magnitude of the time it takes for a linguistic environment to forget its initial state. Figures 3 and 4 require 29 and 30 doubling steps in MCFTP, which indicates a mixing time of around 2^{28} steps of the Markov chain. Each agent has a probability p_r of dying and being replaced if it is selected. Therefore, the probability of an agent living to age m is $(1-p_r)^m p_r$, with a mean of $(1-p_r)/p_r$. For $p_r = 0.001$, this gives an average life span of 999 listening interactions. Each agent is selected to listen or be replaced with probability $1/40$, so the average lifespan is approximately $40,000$ steps of the Markov chain, which is between 2^{15} and 2^{16}. Hence, the mixing time is on the order of $2^{28-16} = 4096$ times the lifespan of an individual agent. In real life, taking a lifespan to be 40 years, that corresponds to at least $160,000$ years. Furthermore, this is an underestimate, because true human language is far more complex and should have an even longer mixing time. Thus, this simulation suggests that the linguistic transitions we observe in real life taking place over a few decades are essentially transient behavior.

6 Discussion and conclusion

With reasonable parameter settings, populations in this simulation are able to both gain and lose V2, an improvement over other simulations, including earlier versions of this one, that tend to always converge to SVO+V2+pro-drop. Furthermore, such changes can happen spontaneously, without an externally imposed catastrophe. The simulation does not give reasonable results unless learners can tell which component of a sentence is the topic. Preliminary results suggest that the PARAMETER-CRUCIAL learning algorithm gives more realistic results than the LEARN-ALWAYS algorithm, supporting the hypothesis that much of language acquisition is based on cue sentences that are in some sense unambiguous indicators of the grammar that generates them. Timing properties of the simulation suggest that it takes many generations for a population to effectively forget its original state, suggesting that further research should focus on the simulation's transient behavior

[2] The version of this document in the proceedings of the ACL workshop was written based in part on a picture generated by a computer program that turned out to have a bug in its implementation of LEARN-ALWAYS. The corrected picture is much more reasonable than the original, and is included in this corrected version. However, in the interest of changing this document as little as possible, I am postponing more extensive discussion of LEARN-ALWAYS and will include it in a future paper.

rather than on its stationary distribution.

In future research, this simulation will be extended to include other possible grammars, particularly approximations of Middle English and Icelandic. That should be an appropriate level of detail for studying the loss of V2. For studying the rise of V2, the simulation should also include V1 grammars as in Celtic languages, where the finite verb raises but the topic remains in place. According to Kroch (personal communication) V2 is thought to arise from V1 languages rather than directly from SOV or SVO languages, so the learning algorithm should be tuned so that V1 languages are more likely to become V2 than non-V1 languages.

The learning algorithms described here do not include any bias in favor of unmarked grammatical features, a property that is thought to be necessary for the acquisition of subset languages. One could easily add such a bias by starting newborns with non-uniform prior information, such as $\mathrm{Beta}(1, 20)$ for example. It is generally accepted that V2 is marked based on derivational economy.[3] Pro-drop is more complicated, as there is no consensus on which setting is marked.[4] The correct biases are not obvious, and determining them requires further research.

Further extensions will include more complex population structure and literacy, with the goal of eventually comparing the results of the simulation to data from the Pennsylvania Parsed Corpus of Middle English.

References

Derek Bickerton. 1981. *Roots of Language*. Karoma Publishers, Inc., Ann Arbor.

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

E. J. Briscoe. 2002. Grammatical acquisition and linguistic selection. In E. J. Briscoe, editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge University Press.

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

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis*. Chapman & Hall/CRC, second edition.

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

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

Natalia L. Komarova, Partha Niyogi, and Martin A. Nowak. 2001. The evolutionary dynamics of grammar acquisition. *Journal of Theoretical Biology*, 209(1):43–59.

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

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

W. Garrett Mitchener and Martin A. Nowak. 2003. Competitive exclusion and coexistence of universal grammars. *Bulletin of Mathematical Biology*, 65(1):67–93, January.

W. Garrett Mitchener and Martin A. Nowak. 2004. Chaos and language. *Proceedings of the Royal Society of London, Biological Sciences*, 271(1540):701–704, April. DOI 10.1098/rspb.2003.2643.

W. Garrett Mitchener. 2003. Bifurcation analysis of the fully symmetric language dynamical equation. *Journal of Mathematical Biology*, 46:265–285, March.

Partha Niyogi and Robert C. Berwick. 1996. A language learning model for finite parameter spaces. *Cognition*, 61:161–193.

Partha Niyogi and Robert C. Berwick. 1997. A dynamical systems model for language change. *Complex Systems*, 11:161–204.

James Gary Propp and David Bruce Wilson. 1996. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(2):223–252.

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

[3] Although Hawaiian Creole English and other creoles front topic and *wh*-word rather than leaving them *in situ*, so it is unclear to what degree movement is marked (Bickerton, 1981).

[4] On one hand, English-speaking children go through a period of topic-drop before learning that subject pronouns are obligatory, suggesting some form of pro-drop is the default (Yang, 2002). On the other hand, creoles are thought to represent completely unmarked grammars, and they are generally not pro-drop (Bickerton, 1981).

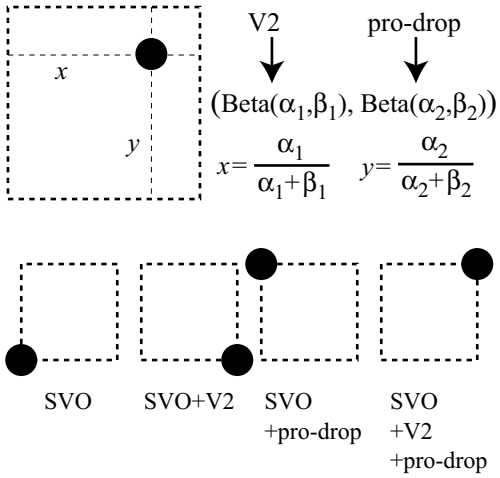

Figure 1: Key to illustrations. Each agent is drawn as a box, with a dot indicating its fuzzy grammar. The means of its beta distributions are used as the coordinates of the dot. The distribution for the V2 parameter is used for the horizontal component, and the distribution for the pro-drop parameter is used for the vertical component. Agents using predominantly one of the four possible idealized grammars have their dot in one of the corners as shown.

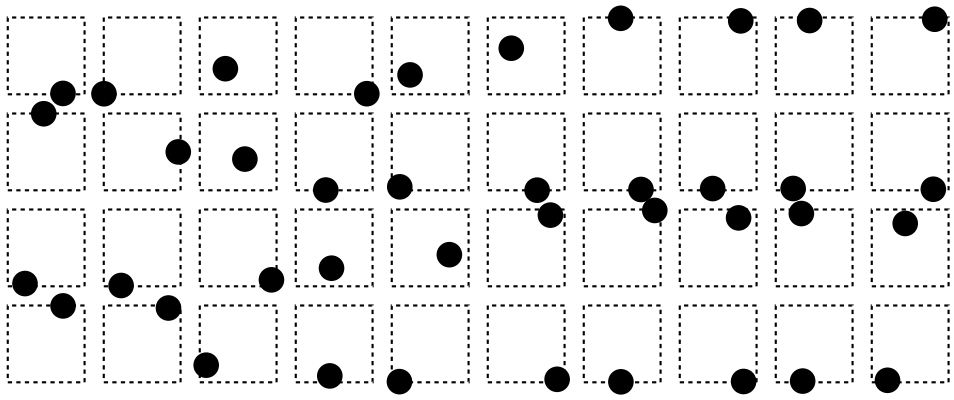

Figure 2: A population of 40 under the LEARN-ALWAYS algorithm. Each agent speaks to its neighbors, and the population should be read left to right and bottom to top. The rightmost agent in each row is neighbors with the leftmost agent in the next row up. The bottom left agent is neighbors with the top right agent. The corresponding picture in the ACL proceedings was generated by a computer program with a bug. This picture is generated by the corrected program.

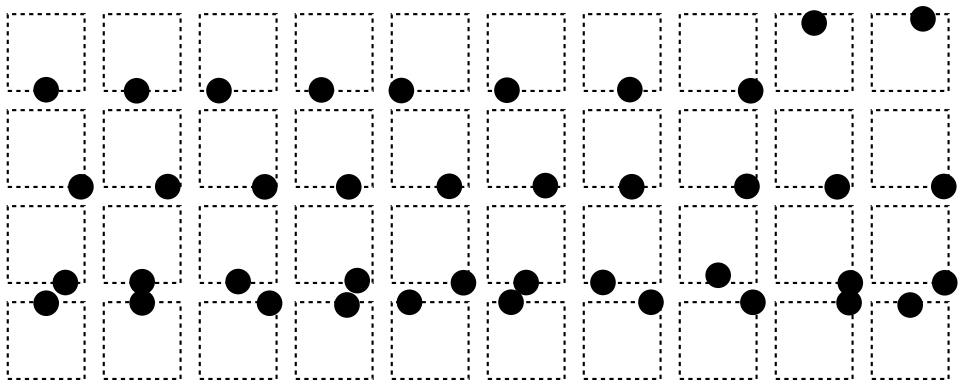

Figure 3: A population of 40 under the PARAMETER-CRUCIAL algorithm. Each agent speaks to its neighbors, and the population should be read left to right and bottom to top.

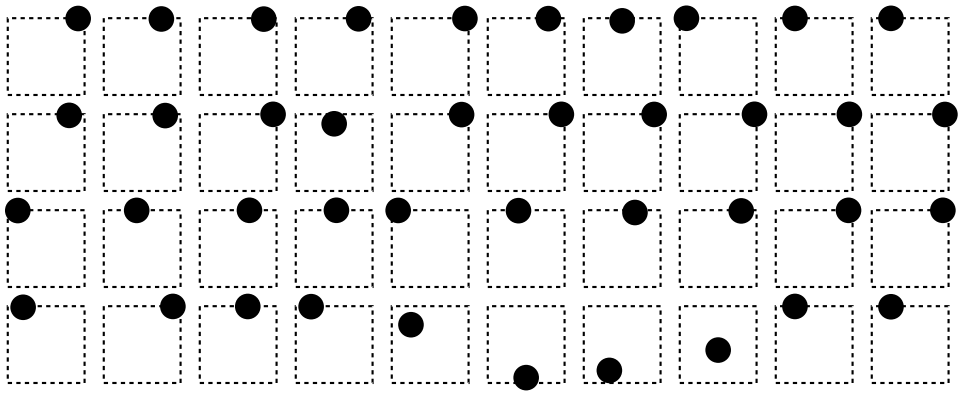

Figure 4: A population of 40 under the PARAMETER-CRUCIAL algorithm. Each agent speaks to its neighbors, and the population should be read left to right and bottom to top.

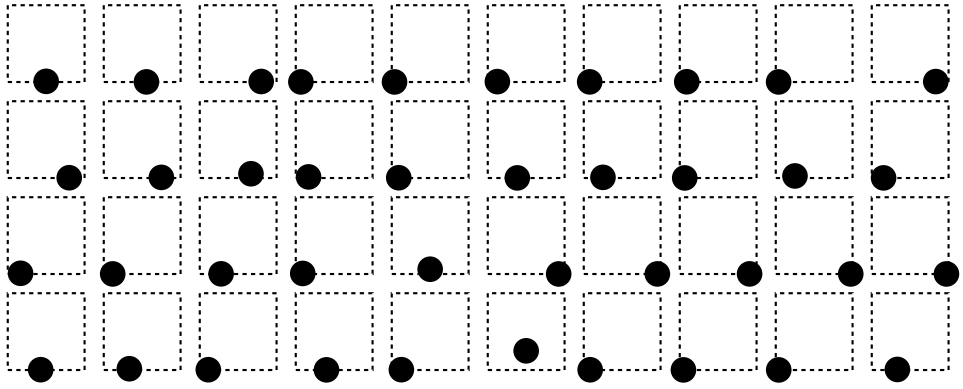

Figure 5: A population of 40 under the PARAMETER-CRUCIAL algorithm. Each agent speaks to its neighbors, and the population should be read left to right and bottom to top.