An evolutionary approach to (logistic-like) language change

Ted Briscoe
Computer Laboratory
University of Cambridge
ejb@cl.cam.ac.uk

Draft – Comments Welcome

Abstract

Niyogi and Berwick have developed a deterministic dynamical model of language change from which they analytically derive logistic, S-shaped spread of a linguistic variant through a speech community given certain assumptions about the language learning procedure, the linguistic environment, and so forth. I will demonstrate that the same assumptions embedded in a stochastic model of language change lead to different and sometimes counterintuitive predictions. I will go on to argue that stochastic models are more appropriate and can support greater demographic and (psycho)linguistic realism, leading to more insightful accounts of the (putative) growth rates of attested changes.

1 Introduction

It has been observed that language changes (often?) spread through a speech community following an S-shaped pattern, beginning slowly, spreading faster, then slowing off before finally extinguishing a competing variant (e.g. Weinreich *et al.* , 1968; Chen, 1972; Bailey, 1973:77; Lass, 1997; Shen, 1997). (This observation even makes it into Crystal's *Cambridge Encyclopedia of Language* along with other statistical chestnuts such as Zipf's law(s).)

Kroch (1990) discusses a number of attested grammatical changes and argues that in each case they can be analysed as cases of competing grammatical subsystems where the rate(s) of change, measured by diverse surface cues in historical texts exemplifying the successful grammatical subsystem, can be fitted to one member of the family of logistic functions which generate such S-shaped curves. Kroch uses the logistic as a tool to demonstrate a single underlying rate of change and thus a single causative factor of competition between (parametrically-defined) grammatical subsystems. Though this work puts the observation on a firmer mathematical foundation and relates it broadly to competition between grammatical subsystems (or perhaps, even more generally, between alternative means of conveying the same meaning) it does not explain why logistic-like growth occurs. (Slide 0 illustrates a logistic curve and the logistic map which generated it and shows Kroch's (1990) presentation of Ellegard's original data, suggesting

S-shaped growth of a number of surface reflexes of a putative single parametric change in the grammar of English.)

It is, of course, not entirely clear that language change always or even ever follows the logistic pattern. (Ogura (1993) has questioned the fit between the logistic curve and Kroch's case study, arguing that the fit is statistically as good for the period up to 1700 even though, the data looks 'less S-curve like' over this window.) There are attested cases, such as the rapid adoption of Hawaiian creole described by Bickerton (e.g. 1984) which are often characterised as an 'instantaneous' or 'rapid' spread of the creole via the first generation of first language learners exposed to the pidgin. There are clearly other logical possibilities: random drift or monotonic change but linear, polynomial, exponential rate of growth; 'variation' rather than change, which doesn't spread thru' the whole/most of the population, etc. (draw some incl. the case of logistic spread converging to stable variation??). I'll assume that S-curves are the norm, but return to the issue of language genesis / creolisation briefly at the end.

Two separate issues are whether competition between variants is taking place at the level of I-language – the (idio)lect) – or E-language – the aggregate output of a speech community, and whether S-shaped spread is the result of lexical diffusion or parametric grammatical competition. The earliest discussions of S-shaped change focus on lexical diffusion or growth through the lexicon (without being explicit about whether this is the E- or I-lexicon). Ogura and Wang (1996) seem to believe that all (S-shaped?) change can be characterised in terms of either E- or I- lexical diffusion and that these can be distinguished in the historical data. Kroch (1990) appears to believe that S-shaped change is caused by syntactic diglossia or competition between parametrically-defined grammatical subsystems within the individual – his evidence comes from the relative frequency of the diverse surface cues in a historical sequence of singly and differently authored texts.

2 The NB Model

Niyogi and Berwick (1997) and Niyogi (2000) (hereafter NB) have developed a model of grammatical change based on a macro-evolutionary deterministic model in which E-languages are treated as dynamical systems, the aggregate output of a population of (adult, stable but possibly different) generative grammars, and evolution of the system corresponds to changes in the distribution of grammars in the population. This distribution changes as each new generation of language learners each acquire a grammar from the data provided by their speech community (i.e. the previous generation of learners who have now acquired an adult grammar). (Slide 1 gives some of the background assumptions.)

The NB model has three main components: a finite set of grammars, UG, from which a learner selects on the basis of triggering data (unembedded / degree-0 sentences); a learning algorithm, LA, used by the learner to choose a grammar, $g \in UG$; and a probability distribution, P, with which triggers are presented to the learner. P is defined in terms of the distribution on triggers within each $g \in UG$ and the proportions of each $g \in UG$ in the current population. A dynamical system can now be defined in which each state of the system is represented by a P for state, s, and the new P' for state $s+1$ can be calculated by an update rule which depends only on P, LA and UG, $P_{pop,s} \longrightarrow^{LA} P_{pop,s+1}$. Crucially, this deterministic update rule relies on the assumption of non-overlapping generations of learners and speakers, and the abstraction to infinite populations. The former assumption makes the analytic calculation of P for each state of the system tractable and the latter abstraction amounts to the assumption that random sampling effects are irrelevant in the calculation of the proportions of learners who converge to specific grammars given P. (see Slide 2)

NB provide a detailed discussion of the derivation of P. The essential point for the following discussion is that the relative frequency of unambiguous triggers which exemplify a unique $g \in G$ is critical for determining which grammar a learner will choose. Intuitively, if two grammars generate languages with largely overlapping triggers (see Slide 3), then given that the learning data is a proper finite subset of the languages, it is more likely that a learner will not sample data distinguishing them, so change (if present) will be slower. NB, in fact, demonstrate that if there is an equal chance of a learner seeing an unambiguous trigger from each variant source grammar exemplified in the linguistic environment, the population will converge to equal proportions of each grammar. On the other hand, if unambiguous triggers generated by g^i are more frequently encountered in the learning data than other unambiguous triggers, we expect learners to converge more often to g^i.

One result which NB demonstrate follows from one instantiation of their model is that the spread of a grammatical variant will be logistic. NB argue that it is a strength of their model that logistic behaviour can be derived analytically from the properties of the update rule, given certain assumptions about UG, LA and P, but is not 'built in' in the first place. To derive the logistic map, NB assume a two grammar / language system in which LA selects between g^1 and g^2 on the basis of 2 triggers drawn from P. If the last trigger is unambiguously from one grammar, then this grammar is selected. If the first trigger is unambiguously from one grammar and the last is ambiguous, then the learner selects a grammar on the basis of the first trigger. Otherwise, a random (unbiased) selection is made. (This LA is equivalent to the Trigger Learning Algorithm (TLA) of Gibson and Wexler (1994) applied to a one parameter system with the critical period for learning set to 2 triggers (hereafter TLA_2 – see Slide 7.)

The deterministic update rule is defined in terms of the consequent probabilities of LA selecting g^1 or g^2 given P. If these probabilities are not equal then the population will converge logistically to the grammar better represented in triggering data over time. If they are equal then the population will stabilise with equal proportions of g^1 and g^2 grammars, and not change thereafter. The critical assumption for the analytic derivation of logistic behaviour lies not in the specific assumptions about UG, LA or P, but rather in D, the model of a dynamical system that NB adopt. (This is not to say that UG, P and LA are not important – Robert Clark (1996) demonstrates via simulation that logistic change is the exception rather than norm in the NB model, and NB only derive this behaviour analytically for the specific case of selecting between two grammars using TLA_2. NB also show graphs displaying logistic and sometimes exponential spread of a variant grammar mostly through whole population, based on simulation of this dynamical model with a wider range of LAs and more complex multi-parameter UG fragments.)

NB characterise the states of the system in terms of the proportion of *average* or arbitrary learners exposed to P who converge to g^1 (equivalenty g^2). This is a macro-evolutionary deterministic model in which what is modelled is the gross statistical behaviour of learners (and thus of linguistic systems), rather than the behaviour of individual learners within the population (e.g. Renshaw, 1991). The macro model effectively builds in the assumption that variation in input samples across individual learners is irrelevant or, equivalently, that the population is infinite. However, for TLA_2 and other models of LA based on very small samples this looks like a big and unrealistic assumption which together with D makes the whole model questionable.

Yang (2000) in related work proposes a different model of LA in which the learner converges to a stable weighting of one or more $g \in G$ in an attempt to model sociolinguistic variation and take account of Kroch's observation that singly-authored historical texts show evidence of logistic-like spread of a variant *within* individual speakers. Yang sketches how logistic spread might follow from his LA by showing that the weighting of a preferred $g \in G$ will increase over time logistically with respect to a dispreferred one. This result is derived from a more plausible LA than that of NB, and does not rest on TLA_2 but it does require the same assumptions about UG (that the set $g \in UG$ is finite) and D (non-overlapping generations and an infinite population).

3 The Stochastic NB Model

If we replace D with a stochastic micro-evolutionary model, D' in which there is a finite population of non-overlapping generations, and we model the behaviour of each individual learner while keeping assumptions about P, LA and UG identical, we find different behaviour – at least until population sizes (or better networks of linguistic interaction over which change is measured) become large. The differences are most obvious when we consider the case where each learner has an equal chance of being exposed to an unambiguous trigger from g^1 or g^2. In the NB deterministic model this results in stasis, but in a stochastic version of their model stasis is highly improbable.

For simplicity assume a starting point in which there are equal numbers of g^1 and g^2 grammars in the population, $\frac{1}{2}$ of triggers from g^1 and from g^2 can distinguish the two grammars (i.e. are unambiguous with respect to the source grammar which generated them), and P is a uniform distribution (so triggers are equiprobable). The probability that a learner selecting a grammar based on 2 triggers will select randomly, because the two triggers are ambiguous, is $\frac{1}{4}$, because for each independently drawn observation from P the chance of seeing an ambiguous trigger is $\frac{1}{2}$. Therefore, the learner will select g^1 (equiv. g^1) on the basis of data with probability $\frac{3}{8}$ ($P = 0.375$).

For stasis we require exactly half of the learners to acquire g^1. Suppose there are 100 learners; what is the probability that exactly half will select g^1 in the first generation? The data provided to each learner is stochastically independent so this is equivalent to asking how probable is it that in 100 tosses of an unbiased coin exactly 50 will come up heads, and is given by the binomial theorem: $P = 0.0795$ (e.g. McColl, 1995). Therefore, it is very improbable that the distribution P will remain unaltered, and unbiased between g^1 and g^2, for the next generation of learners. This result is in marked contrast from that of NB and follows directly from modelling the fact that each individual learner will be exposed to a different (random) sample of triggers.

To see how likely it is that, given a biased distribution of unambiguous triggers, P, on g^1 and g^2q, the dominant grammar will spread logistically through the population given D', we need to consider the shape of the skewed binomial distribution arising from the bias. For example, if we minimally modify the example above by assuming that $\frac{3}{4}$ of the adult population speak g^1, the probability that a learner will acquire g^1 given 2 triggers is now $\frac{11}{16}$ ($P = 0.687$). (Note that it is not $\frac{12}{16}$ ($P = 0.75$) because of the possibility of selection according to the initial unbiased setting when the triggering data seen is ambiguous.) Consequently, the probability that more than 75 learners will acquire g^1 is only $P = 0.070$, though the probability that more than 50 will acquire g^1 is $P > 0.999$. In fact, the distribution peaks at 69 learners predicting not logistic growth but rather a probable slight decline in the number of g^1 speakers in the next generation. In the limit, if the whole population speak g^1, the probability that a learner will select g^1 is $\frac{7}{8}$ ($P = 0.875$) because there remains a $\frac{1}{8}$ chance that a learner will see 2 ambiguous triggers and select g^2 randomly.

It might be objected that these results follows primarily from choosing UG and P with a high proportion of ambiguous triggers or small trigger samples (as in TLA_2), so that learners frequently select grammars randomly (though UG and P here are in this respect similar to several of the more realistic examples NB consider, derived from Gibson and Wexler, 1994). If we assume, that g^1 and g^2 are as highly differentiated as possible and share no triggers, then a learner will select between them with probability directly correlated with the proportions of g^1 and g^2 speakers in the adult population. In the case of equal proportions, the probability that exactly half the population of learners will acquire g^1 (equivalently g^2) is still given by the binomial distribution, and thus remains low. The binomial distributions for each generation of learners will now peak at exactly the point predicted by the proportions of adult grammars in the current generation, but this still only allows us to predict that +/–13 learners around this peak will acquire g^1 with $P > 0.99$ for a population of 100 learners. (See Slide 4 for a summary of the stochastic NB model, D', examples of relevant binomial distributions etc.)

In fact, the appropriate mathematical model for describing D' is very similar to that for genetic drift in finite populations (e.g. Maynard Smith, 1998:15f; Renshaw, 1991). General results derived for such models suggest that we should expect to see (random) oscillations in the proportion of g^1 speakers in the population with (temporary) fixation on 0% or 100% within (every) $2N$ generations for a population of size N (with $s.d.N$). However, it is not possible to predict the individual or overall direction of these oscillations because they are caused by variation in a sequence of a series of sampling events (learning). Simulating D' allows us to tighten up these general predictions and explore the consequences of varying the number, n, of triggers to the LA (e.g. TLA$_n$, see Slide 7), the proportions of unambiguous triggers for each source grammar, and the initial ratio of g^1 and g^2 speakers.

Slides 5–6 illustrate the behaviour of the TLA in D' given these varying conditions. In each case, the Y-axis shows the percentage of g^1 speakers and the X-axis time measured in generations. The red lines show the behaviour of the TLA with 2 triggers, green dashes 4 triggers, blue short dashes 6 triggers, and pink dots 20 triggers, respectively.

For Slide 5 all runs begin with initially equal proportions of g^1 and g^2 speakers with an equal chance of seeing an unambiguous g^1 or g^2 trigger drawn from either source grammar. This is the 'stasis' case discussed above. In the stochastic model, rather than observing stasis at equal proportions of g^1 and g^2 speakers, we either get oscillation around this 'attractor' ($n = 2$) or a tendency to (temporarily) fixate on one or other grammar as a result of the positive feedback inherent in the calculation of P for successive states of the model ($n > 2$). Decreasing the 'overlap' between g^1 and g^2 (i.e. increasing the overall proportion of unambiguous triggers) makes the system less stable because it decreases the influence of unbiased random guessing and increases the positive feedback dynamics.

(Increasing the size of the population tenfold dampens the positive feedback dynamics and consequent degree of oscillation, so there is a greater tendency for fixation on one or other grammar, but oscillation still occurs with low number of triggers, as unbiased guessing during learning dominates the overall behaviour of the system.

While populations of 1000 or more are quite plausible in the context of attested language change, they are not always so – there were 13 indigeneous members of the population of Pitcairn when the 8 Bounty mutineers arrived, and the consequences linguistic and otherwise were dramatic (Romaine, 1988)! More importantly, though populations might typically run to the thousands, networks of strong and regular linguistic interaction are probably limited to group sizes of around 100-150 (Dunbar, 1993; Milroy, 1992; Nettle, 1999).)

Slide 6 shows runs where initially there is a single g^1 speaker in a population of 100 but there is a greater proportion of unambiguous triggers for g^1 than g^2. In the top graph a g^1 unambiguous trigger is twice as likely as a g^2 one. In the bottom graph, a g^1 unambiguous trigger has only a 1/20 greater likelihood. In the top case, g^1 grows exponentially in the population but only reliably stays at fixation for $n = 6$ triggers. For higher numbers of triggers, the population stays robustly fixated on $g = 2$. In the bottom case, $g = 1$ only grows (exponentially) when $n = 2$ and doesn't reach fixation.

None of this behaviour fits a logistic curve well. The quasi-random drift observed in most cases isn't attested in (major) language change to my knowledge. Furthermore, it is clear that we only get spread of a minority variant under conditions where unbiased guessing dominates the system (i.e. low numbers of triggers, lower probability of unambiguous triggers). Under these conditions, the TLA is a very poor model of language learning, because unbiased guessing amounts to mislearning half the time, on average. The NB model with LA set to TLA_n for finite n, 'explains' change in terms of mislearning.

4 Desiderata for the Language Learning Algorithm

The TLA (Gibson and Wexler, 1994) (see Slide 7) is an implausible model of language learning because it predicts that the child will take a memoryless walk through grammar space starting from a random complete grammar and may not converge to $g^t \in UG$ even when exposed to a fair sample of triggering data from g^t (e.g. Brent, 1996; Briscoe, 1999, 2000a; Niyogi and Berwick, 1996). Here I outline a framework for thinking about grammatical learning in the context of language change and some desiderata for the LA drawing on and extending Briscoe (2000b). Then I briefly present one model of the LA which instantiates the framework and satisfies these desiderata, simplifying Briscoe (1999) and Villavicencio (2000).

4.1 The Framework

The framework modifies fairly standard learnability conditions to take account of the insight that E-language is a dynamical system, that speech communities may contain mixed populations of variant grammars which may even change during the learning period, and that learners take time to learn and are not 'input matchers' (e.g. Lightfoot, 1999; Briscoe, 2000b) but typically select between competing alternatives often with a bias in favour of one, and that in situations of genuine diglossia learners will acquire multiple grammars (Kroch, 1989, Yang, 2000). (See Slide 8.)

In common with many accounts of language learning we assume that grammatical acquisition is based on exposure to a finite number of triggers during the (critical) learning period. We also define a trigger relatively uncontroversially as a pairing of a surface form (SF) and logical form (LF) so that the task of the learner is to parametrically select $g \in UG$ so that the SF-LF mapping exemplified in triggers can be expressed correctly. However, these triggers are drawn from an unknown (and possibly non-stationary) mix of source grammars defined by $P_{pop,S_t - S_{t+n}}$ over the course of the learning period as discussed in sections 1 and 2 above.

The most important desideratum of LA is learnability or accuracy, which leads to stasis or language maintenance in conditions of linguistic homogeneity. When a (fair) trigger sample is drawn from a single source grammar, then LA should acquire that grammar, or one representing the same SF-LF mapping, with very high probability. In this special case, the framework is very close to the standard learnability framework (e.g. Niyogi, 1999).

In the case where the sample is from mixed sources, if triggers represent parametrically different ways of realising the same LF then the parameters expressed by the more frequent trigger will be acquired provided the alternative is K times less probable. Data selectivity incorporates Lightfoot's (1999) insight that input is not matched and internalized grammars typically don't allow (parametric) variation, but is in fact a more fundamental requirement of any successful model of LA. Without it, LA will not be able to cope with noise in the input caused by trigger miscategorizations due, for example, to the indeterminacy of parameter expression (Clark, 1992, Briscoe, 2000c) nor with the scenario envisaged by Bickerton (e.g. 1984) and carefully documented by Newport (1999) in which conflicting and inconsistent input is 'regularised' and fashioned into a single consistent generative grammar.

Inductive bias over and above the hard constraints in UG yields soft constraints or preferences on grammars within the hypothesis space. Such preferences can arise from general principles of learning such as Occam's Razor in the form of a simplicity metric over the representational framework employed (e.g. Mitchell, 1997), from 'functional' considerations such as the relative parsability of different triggers (e.g. Kirby, 1998), or from prehistorical contingent properties of grammars genetically assimilated during the period of adaptation for the language faculty (Briscoe, 2000a). For example, in a parametric LA the learner might select between alternative grammars compatible with the triggering data (so far) by selecting the one involving the least number of parameters or the one requiring the fewest non-default settings of parameters (Briscoe, 2000b). Such biases can be formalised in a Bayesian / Minimum Description Length framework in terms of a prior probability distribution on grammars (Briscoe, 1999, 2000c).

Inductive bias predicts that language learning will result from the interplay of such biases with triggering data (e.g. Kiparsky, 1996), but the Bayesian formulation ensures that the data will win out if a parameter is robustly specified. (Cosmides and Tooby (1996) and others have argued that a Bayesian perspective on learning provides a good account of many aspects of human learning / reasoning.)

Data sensitivity models, rather crudely, Kroch's (1990) insight that grammatical variants compete in I-language/idiolect. We approximate this in terms of a frequency based threshold K, and a (Labovian) conditioning context for sociolinguistically-motivated variation. (The difference between noise or inconsistency and such variation is that the former lacks consistent conditioning contexts, and not necessarily that it is rarer.) When parametric variation is 'robustly' exemplified in triggering data, LA should acquire multiple grammars.

4.2 Bayesian Incremental Parameter Setting

I now present a simplified version of Bayesian incremental parameter setting (BIPS) described in Briscoe (1999, 2000c), which I argue meets the desiderata above (though data sensitivity needs more work). (See Slide 9.) The presentation is deliberately made as similar to that of the TLA as possible. The critical difference is that BIPS sets parameters incrementally according to an incremental version of Bayes law (in which, roughly, the prior for the next trigger is defined by the posterior from the previous). Therefore, the learner never makes (random) guesses (though in the absence of enough triggering data a parameter can take on a default (unmarked) value). Furthermore, BIPS is sensitive to the frequency with which triggers expressing alternate parameter values are exemplified in the data, implementing selectivity (and potentially sensitivity).

4.3 BIPS integrated with the GCG/TDFS model

Briscoe (1999) demonstrates that the BIPS model integrated with Generalized Categorial Grammar (GCG) embedded in the Typed Default Feature Structure (TDFS) default inheritance framework can reliably learn a target grammar from finite preclassified but noisy triggers from a hypothesis space of around 300 distinct grammars exemplifying typological constituent order differences. Villavicencio (2000) demonstrates that the same model, combined with a variant of Siskind's algorithm for preclassifying triggers (Waldron, 2000), can acquire English clause and verbal argument structure from a sample of about 1000 caretaker utterances to a single child (drawn from the Sachs/CHILDES corpus).

The integration of BIPS with a concrete model of UG introduces a further source of inductive bias in that not all parametric variation is independent. Thus, the setting of parameter$_i$ to one specific value may *then* require the setting of dependent parameter$_j$, while the opposite setting for i does not. This, in effect, orders the hypotheses $g \in UG$ considered by the learner. Other attempts to define concrete parametric models have the same property (Clark, 1992; Dresher, 1999).

4.4 BIPS in the NB Stochastic Model

BIPS with a $P = 0.8$ prior bias in favour of the parameter setting yielding g^1 displays logistic-like spread of g^1 through the population under a wide variety of conditions. Slide 10 shows the case where initially there is one g^1 speaker out of a population of 100 speakers and there is a 1/20 higher chance of seeing an unambiguous trigger for g^1 over g^2. Depending on the number, n, of triggers sampled g^1 either spreads exponentially ($n = 2$), logistically ($n = 4 - 6$) or doesn't spread at all (e.g. $n = 20$), reflecting the interplay of the prior bias and the 'robustness' of the triggering data defined, as before, by $P_{pop,s}$.

The behaviour of BIPS compared to TLA is considerably less 'random' because it derives as accurate an estimate of the probability of each parameter setting (equiv. g^1 / g^2 here) as is possible from the trigger sample and it incorporates inductive bias. When the 'balance' of data and inductive bias is near perfect, logistic-like monotonic change is highly likely. When the balance is perfect, the model will behave randomly but this is now so finely specified that we are very unlikely to observe the quasi-random oscillations across generations. (Furthermore, given inductive bias, attractors will now always be 0/1??.)

In essence, logistic-like spread occurs with the BIPS model because data selectivity yields an increase in the probability of triggers generated by the favoured grammar (variant). If data-sensitivity lead to the acquisition of a both grammars and their use by a speaker in proportions that precisely mirrored the acquisition data, then this effect would be nullified. Therefore, in the case where the threshold K is not reached (diglossia), we need to posit an alternative source of the change in the distribution, $P_{pop,s}$ – for example, Nettle (1999) demonstrates with stochastic models that sociolinguistic factors can act as amplifiers of variation, predicting e.g. that a variant will spread via more frequent usage, post acquisition.

5 The OG Stochastic Model

In a stochastic model, it is possible to incorporate more demographic realism in the form of, for example, overlapping generations, though it makes mathematical analysis of the model harder because there are now potentially learner-learner as well as adult-learner interactions making P harder to calculate for each state of the system (see Briscoe, 2000b). Nevertheless, overlapping generations may be important to understanding language changes which involve dramatic demographic change and is a likely (alternative) source of logistic-like spread of a variant through a speech community.

In a model like that described in Briscoe (2000b) with approximately 100 speakers in which speakers are removed after 10 'generations' and six new learners are added every 'generation' (and learn in one generation), repeating the exact same run described in the previous section, and shown in Slide 10, results in the same qualitative behaviour but now the growth of g^1 is 'bumpier' and slower taking about 3 times as long to spread through the population – see Slide 11.

6 Language Genesis and the Subset Principle

The results for BIPS illustrated above do scale up to more complex simulations of multiparametric systems undergoing change in which the learning period is spread out over several time steps of the system. However, rather than show more graphs of logistic-like curves, I'll finish by illustrating the application of the BIPS+GCG/TDFS model to language genesis. (For more details see Briscoe, 2000c.)

Language genesis appears to pose a problem for an essentially selectionist (parametric) model of language learning of the type presented here. Bickerton (e.g. 1984) has argued that the abrupt pidgin-creole transition is a consequence of the first and subsequent (overlapping) generations of children exposed to inconsistent pidgin input each without exception acquiring a superset creole grammar. This implies a rate of change correlated with the (typically high and rising) birthrate and proportion of child language learners in the plantation community. Roberts (1998), using a large database of Hawaiian pidgin and creole utterances, has revised Bickerton's original claim slightly, by arguing that some aspects of the transition in Hawaii took two full generations to emerge. Nevertheless, this careful empirical work by-and-large confirms Bickerton's original claims that the creole emerges very abruptly and embodies a much richer grammatical system than the pidgin. One whose properties are not directly exemplified in the super- or sub-stratum languages to which learners might be exposed, and are very similar to the properties of other creoles which emerged at geographically and historically unrelated points. (My simulations, nevertheless, show logistic-like spread through the *whole* population of children and adults with convergence on the creole only when the original generation of adults have all died.)

If there is an element of 'invention' in creolisation how could this arise? The account that I will pursue here is that in some respects the primary linguistic data that creole learners are exposed to is so uninformative that they retain their prior default-valued parameter settings. However, this is not enough to ensure that a superset grammatical system will emerge if the data never exemplifies, however indirectly, a particular grammatical phenomenon. When exposed exclusively to a subset language, the BIPS+GCG/TDFS model reliably acquires subset languages and does not go 'beyond the evidence' (mislearn) to predict a superset language of the subset language learners have been exposed to (as is required by the desideratum of learnability).

Inductive bias is enough to predict the rapid development of, for example, SVO constituent order from essentially randomly ordered pidgin triggering data – see Slide 12. (Four interaction cycles correspond to a single generation in this model. Both learners exhibit inductive bias in terms of dependencies between the setting of parameters but only the 'default' learner has inductive bias in favour of (SVO) parameter settings.) However, to predict acquisition of a superset language in the BIPS+GCG/TDFS model it is necessary to appeal to (limited and indirect) super- or sub-stratum influence to trigger acquisition of complex categories allowing for clausal postmodification, unbounded extraction, and so forth. If such categories are reliably expressed somewhere in the triggering data for each learner, even with inconsistent ordering, then the BIPS+GCG/TDFS model will 'switch on' a generic unordered form of these categories, and predict their ordering behaviour via inductive bias. In the case of Hawaii, this is possible, but Bickerton (1984) claims creolisation has occurred in 'marooned' communities in which learners would have had no access to non-pidgin input and Kegl *et al* (1999) argue that the first learners of Nicaraguan Sign Language were such a community. Quite probably, the interplay of pidgin data with inductive bias is more complex than my simulations allow, and/or language learning is more inventive than a strict interpretation of the Subset Principle (Berwick, 1985) suggests?

7 Conclusions

See Slide 14 for main focussed conclusions – below some meanderings:

For some types of language change the idealisation of the dynamical model, D, to infinite populations may not be harmful; for example, (E-lexical) diffusion through American English within the last 50 years might be such a case. However, even then we would need to be clear that there is an analytic and thus predictive advantage to deterministic modelling for *realistic* versions of UG, LA and P, and this has not been demonstrated as yet.

In all cases where evolution of a linguistic system is likely to have taken place in small relatively isolated speech communities – for example, modelling of prehistoric development or of a process like creolisation, where the relevant populations are likely to have been at most in the low hundreds – abstracting away from sampling issues is dangerous.

Furthermore, the specific behaviour which we want to derive, such as logistic change in the system, may simply follow directly from more realistic demographic assumptions than are possible with deterministic models. For example, population movement, birthrate, the proportion of language learners in the population and the resultant linguistic mix of the population are critical factors in understanding creolisation (very fast language change) / language genesis. In larger populations, spatial distribution and resultant networks of interaction are also likely to be very important (see Nettle, 1999 and Niyogi, 2000 for examples of stochastic and deterministic models which take account of such factors).

The model must satisfy certain constraints of realism which I have argued derive primarily from properties of the language learning algorithm assumed – most notably language maintenance or stasis in a homogeneous linguistic environment. Given language maintenance via the accuracy or learnability of the learning algorithm, one likely first language acquisition mechanism for spread of a minority variant is via (some form of) inductive bias. However, it is also likely that there are sociolinguistic and other amplifiers of variation too (e.g. Nettle, 1999:18) less closely connected to LA and also likely that these competing pressures interact to create a complex and many peaked fitness landscape for language change. Under these conditions, languages are best characterised and modelled not just as dynamical systems but as *complex adaptive systems* (Briscoe, 2000b).

In order to really sort out which of these possible causes of spread of a variant were relevant to an attested change would require more data about the change than is likely to be forthcoming from historical texts. However, thorough studies of contemporary changes, like the emergence and subsequent development of Nicaraguan Sign Language (Kegl *et al.* , 1999), might yet yield the requisite detail.

References

Bickerton, D. (1984) 'The language bioprogram hypothesis', *The Behavioral and Brain Sciences, vol.7.2,* 173–222.

Brent, M. (1996) 'Advances in the computational study of language acquisition', *Cognition, vol.61,* 1–38.

Briscoe, E.J. (1999) 'The Acquisition of Grammar in an Evolving Population of Language Agents' in (ed) Muggleton, S. (ed.), *Linkoping Electronic AI Journal, 21 (Special Issue: Proc. of Machine Intelligence, 16),* http://www.etaij.org.

Briscoe, E.J. (2000a, in press) 'Grammatical Acquisition: Inductive Bias and Coevolution of Language and the Language Acquisition Device', *Language, vol.76.2.*

Briscoe, E.J. (2000b, in press) 'Evolutionary perspectives on diachronic syntax' in Susan Pintzuk, George Tsoulas and Anthony Warner (ed.), *Diachronic Syntax: Models and Mechanisms,* Oxford: Oxford University Press.

Briscoe, E.J. (2000c, in press) 'Grammatical acquisition and linguistic selection' in Briscoe, E.J. (ed.), *Language Acquisition and Linguistic Evolution: Formal and Computational Approaches,* Cambridge University Press, Cambridge.

Clark, Robert A.J. (1996) *Internal and External Factors Affecting language Change: A Computational Model,* MSc Dissertation, University of Edinburgh.

Clark, Robin (1992) 'The selection of syntactic knowledge', *Language Acquisition, vol.2.2,* 83–149.

Cosmides, L. and Tooby, J. (1996) 'Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgement under uncertainty', *Cognition, vol.58,* 1–73.

Dunbar, R. (1993) 'Coevolution of neocortical size, group size and language in humans', *Behavioral and Brain Sciences, vol.16,* 681–735.

Gibson, E. and Wexler, K. (1994) 'Triggers', *Linguistic Inquiry, vol.25.3,* 407–454.

Kiparsky, P. (1996) 'The shift to head-initial VP in Germanic' in Thrainsson, H., Epstein, S. and Peters, S. (ed.), *Studies in Comparative Germanic Syntax, Vol II,* Kluwer, Dordrecht.

Kirby, S. (1998) 'Fitness and the selective adaptation of language' in Hurford, J., Studdert-Kennedy, M., and Knight, C. (ed.), *Approaches to the Evolution of Language,* Cambridge University Press, Cambridge, pp. 359–383.

Kroch, A. (1990) 'Reflexes of grammar in patterns of language change', *Language Variation and Change, vol.1,* 199–244.

Lass, R. (1997) *Historical Linguistics and Language Change,* Cambridge University Press, Cambridge.

Lightfoot, D. (1999) *The Development of Language: Acquisition, Change, and Evolution,* Blackwell, Oxford.

Maynard-Smith, J. (1998) *Evolutionary Genetics,* Oxford University Press, Oxford, 2nd ed.,.

McColl, J.H. (1995) *Probability,* Edward Arnold, London.

Milroy, J. (1992) *Sociolinguistic Variation and Language Change,* ??.

Mitchell, T. (1997) *Machine Learning,* McGraw Hill.

Nettle, D. (1999) *Linguistic Diversity,* Oxford University Press, Oxford.

Newport, E. (1999) 'Reduced input in the acquisition of signed languages: contributions to the study of creolization' in DeGraff, M. (ed.), *Language Creation and Language Change,* MIT Press, Cambridge, Ma., pp. 161–178.

Niyogi, P. (1999) *The Informational Complexity of Learning from Examples,* Kluwer, Drodrecht.

Niyogi, P. (2000, in press) 'Theories of Cultural Change and their Application to Language Evolution' in Briscoe, E.J. (ed.), *Language Acquisition and Linguistic Evolution: Formal and Computational Approaches,* Cambridge University Press, Cambridge.

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

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

Ogura, M. and Wang, S. (1996) 'Evolutionary theory and lexical diffusion' in Fisiak, Jacek and Marcin Krygier (ed.), *Advances in English Historical Linguistics,* ??.

Renshaw, E. (1991) *Modelling Biological Populations in Space and Time,* Cambridge University Press, Cambridge.

Roberts, S. (1998) 'The role of diffusion in the genesis of Hawaiian creole', *Language, vol.74.1,* 1–39.

Romaine, S. (1988) *Pidgin and Creole Languages,* Longman, London.

Shen, Z. (1997) 'Exploring the dynamic aspect of sound change', *Journal of Chinese Linguistics, vol.??,* ??.

Villavicencio, A. (2000) 'Learning a GCG using BIPS??', *Proceedings of the 3rd. Computational Linguistics in the UK (CLUK),* ITRI, Brighton.

Yang, C. (2000, submitted) 'Internal and External Forces in Language Change', *Language Variation and Change,*

(Penultimate drafts of my in press papers are available at:
http://www.cl.cam.ac.uk/users/ejb/papers.html)

Grammatical Change

Grammatical acquisition is parametric
– v2-on/off; head-initial/final;...

Grammatical change is the result of parameter 'resetting / reanalysis' across generations during grammatical acquisition
– v2-on $\rightarrow$ v2-off

I-language, idiolect change is 'immediate' (Lightfoot, etc)

E-language, spread of change across community is S-shaped (Kroch, etc)

An idiolect/I-language is a well-formed stringset with mappings to LFs defined by a generative grammar

The language of a speech community/E-language is a dynamical system – the aggregate output of a changing population of speakers (generative grammars)

Derive 'logistic' spread from a dynamical model (Niyogi and Berwick)
[logistic curve + Ellegard's data]

The Niyogi/Berwick Model

1) A class of grammars, UG
2) A learning algorithm, LA to select $g \in UG$ from data, t_n
3) A probability distribution, P, on triggers

A dynamical system consists of a sequence of states through time: $s, s+1, \ldots$

States are defined by the proportions of speakers (generative grammars) in the population: a distribution $P_{pop,s}$ on $g \in UG$ at s

The distribution P on triggers can be calculated from $P_{pop,s}$:

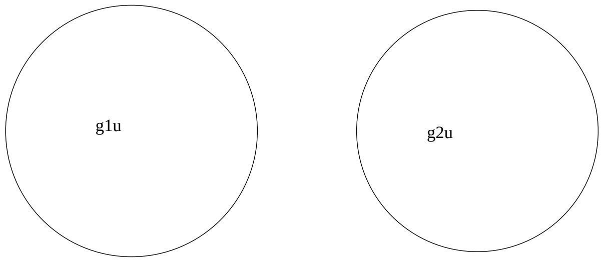
$$P(t_i) = \sum_{g^j \in UG} P^j(t_i) P_{pop,s}(g^j)$$

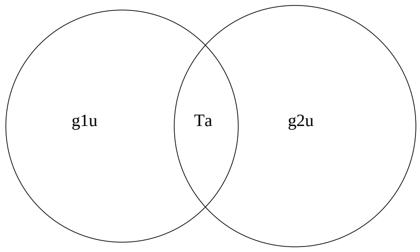
An update rule defines $P_{pop,s+1}$: $P_{pop,s} \longrightarrow^{LA} \; P_{pop,s+1}$

Non-overlapping generations – $P_{pop,s}$ defines triggering data for $S_{pop,s+1}$ of learners
Infinite population – deterministic update rule models behaviour of an *average* learner

One parameter, LP = TLA, 2 triggers, update rule = logistic map provided that $P^1(T_u(L(g^1))) \neq P^2(T_u(L(g^2)))$
If $P^1(T_u(L(g^1))) = P^2(T_u(L(g^2)))$, stasis

Triggering Data for (Related) Languages





g_1: *-V2; Head-first; Spec-first*

$L_1 = \{$ S V, S V O, S V O1 O2, S Aux V, S Aux V O, S Aux V O1 O2, Adv S V, Adv S V O, Adv S V O1 O2, Adv S Auz V, Adv S Aux V O, Adv S Aux V O1 O2 $\}$

g_2: *+V2; Head-first; Spec-first*

$L_2 = \{$ S V, S V O, O V S, S V O1 O2, O1 V S O2, O2 V S O1, S Aux V, S Aux V O, O Aux S V, S Aux V O1 O2, O1 Aux S V O2, O2 Aux S V O1, Adv S V, Adv V S O, Adv V S O1 O2, Adv Aux S V, Adv Aux S V O, Adv Aux S V O1 O2$\}$

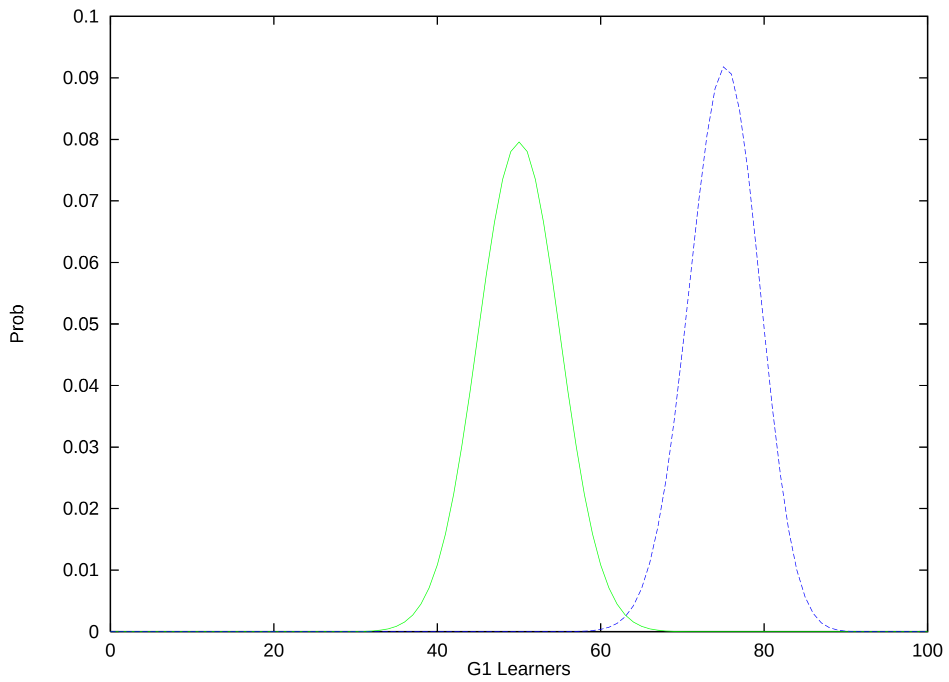
$L_1 \cap L_2 = \{$ S V, S V O, S V O1 O2, S Aux V, S Aux V O, S Aux V O1 O2$\}$

The Stochastic NB Model

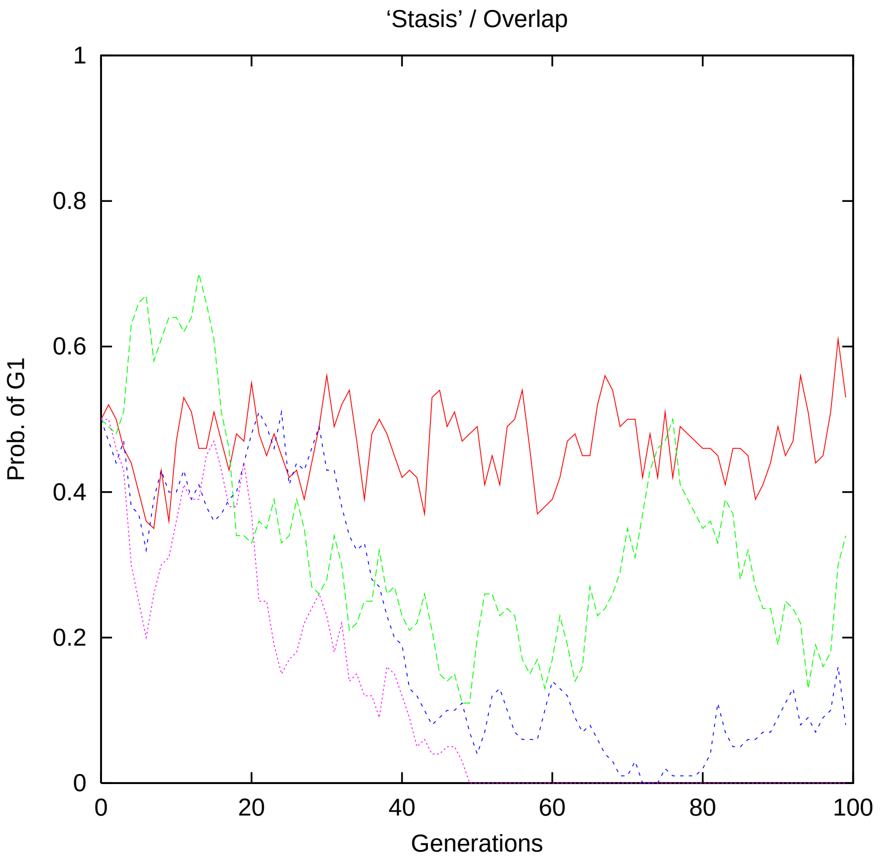
Finite population of 100 speakers (generative grammars)
Non-overlapping generations
One parameter – $g^1/g^2 \in UG$
LA = TLA
$P(t_i) = \Sigma_{g^j \in UG} P^j(t_i) P_{pop,s}(g^j)$

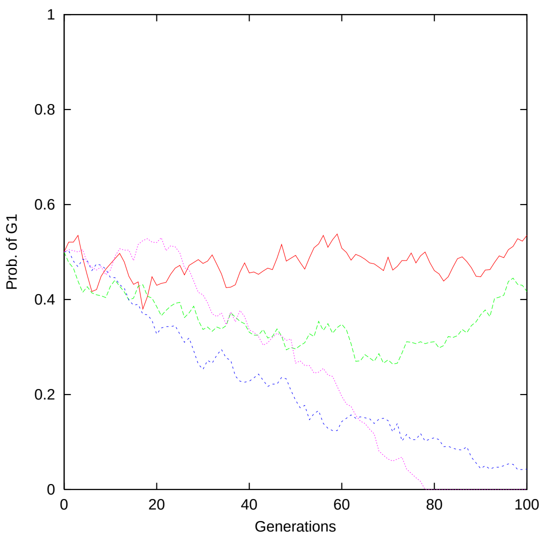
Each learner randomly samples triggers from P n times
Probability of any given learner selecting g^1 given n triggers from P is a Bernoulli trial: $P(TLA(t_n iid.P_{pop,s}) \rightarrow \; g^1)$

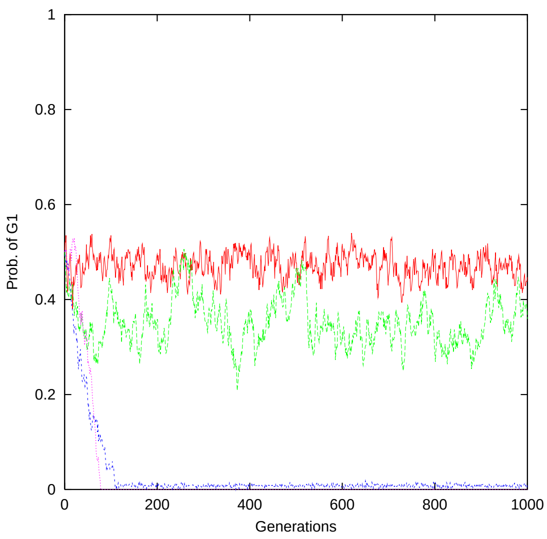
Assume $P(TLA(t_n iid.P_{pop,s}) \rightarrow \; g^1) = 0.5$ (0.75)
Probability that exactly half (50) learners will acquire g^1 is given by the binomial theorem: $P = 0.0795$ (0.0000001):



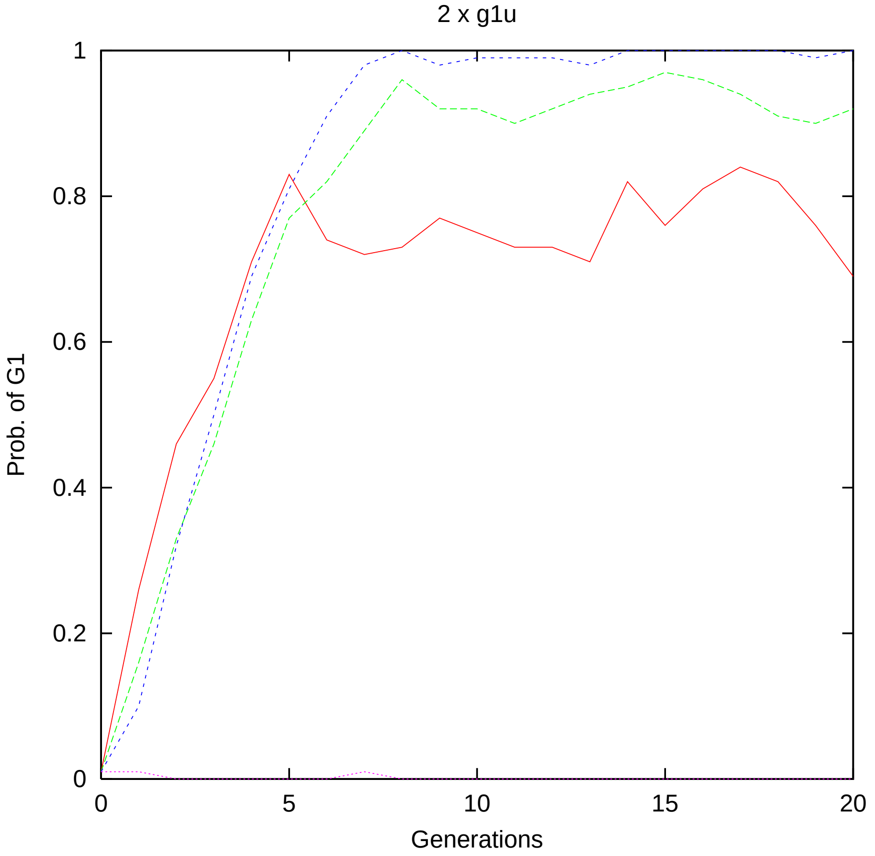
Stochastic NB Model – g1u=g2u

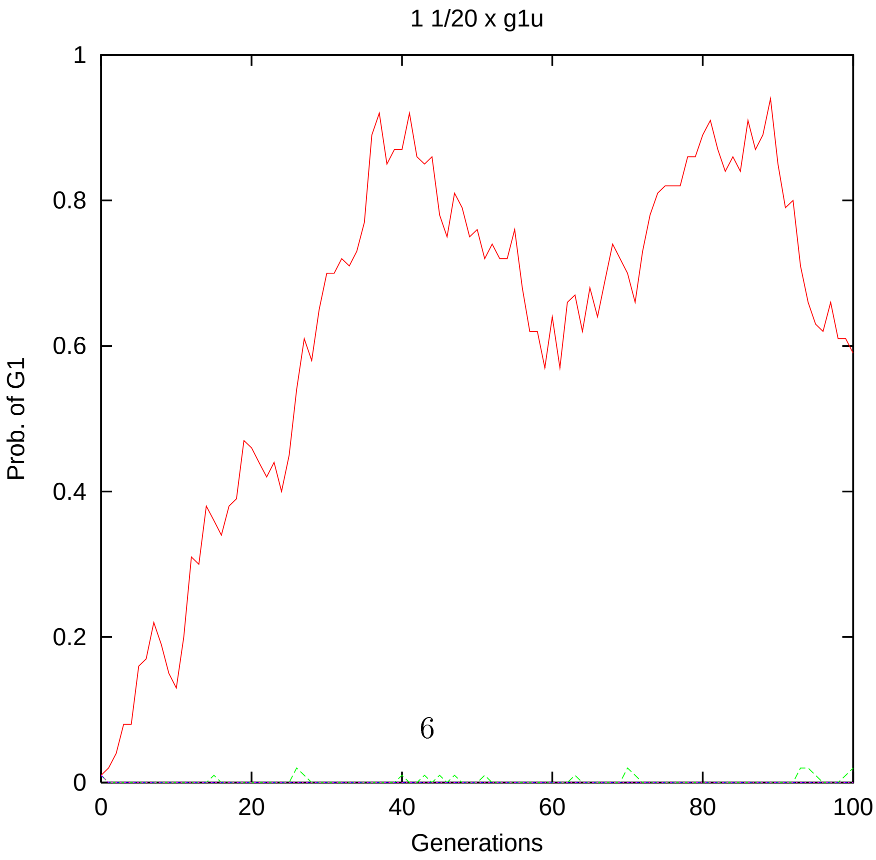






Stochastic NB Model – g1u>g2u





The trigger learning algorithm

TLA (Gibson and Wexler)

1) Randomly select $g \in UG$ (i.e. randomly set parameters of UG);
2) Randomly sample a trigger (degree-0/1, SF-LF pair), t_i from $P(T(L(g^t))$;
3) If the current grammar parses/generates t_i goto-2, else:
4) Select one random parameter, flip its setting, retain the new setting iff that change allows t_i to be parsed/generated, goto-2.
5) Stop if for trigger, $t_i, i = n$ (i.e. at end of critical period).

Local, incremental, greedy and memoryless search through grammar space from random starting point

Do children start from 'random' complete grammars?
Do children (re)(re)visit previous complete grammars?
Do children make a randomly guess the values of unexpressed parameters?

The Language Learning Algorithm

- **Finite Triggering Data** – finite number n of triggers, t, during the (critical) learning period: $t =$ SF:LF degree 0/1 sentence pair sampled from $T = P_{pop,S_t - S_{t+n}}$ (i.e. non-stationary mixed sources, $L(g^1), L(g^2) \ldots$)
- **Learnability** – $P(LA(UG, t_n) \rightarrow \; g^t) > 1 - \epsilon$ where t_n is the nth random sample drawn from $T = P(L(g^{t'}))$ and $L(g^{t'}) = L(g^t)$ (i.e. a single source learnt accurately)
- **Selectivity** – $P(LA(UG, t_n) \rightarrow \; g^t) > 1 - \epsilon$ where $P(t_i \in L(g^t)) > K.P(t_j \neg \in L(g^t))$ and $t_i \wedge t_j \rightarrow \; LF_k$ (i.e. better represented alternative learnt if above some threshold K)
- **Inductive Bias** – $g = argmax_{g \in UG} P(g)P(t_n \mid g)$ (i.e. there is an ordering on $g \in G$, deriving from Occam's Razor / Simplicity (MDL), Parsability, Genetic assimilation)
- **Sensitivity** – $P(LA(UG, t_n) \rightarrow \; g^t \wedge g^{t'}) > 1 - \epsilon$ where $\neg P(t_i \in L(g^t)) > K.P(t_j \in L(g^{t'}))$ and $t_i \wedge t_j \rightarrow \; LF_k$ and $P(t_j \mid C) > P(t_i \mid C)$ (i.e. both alternatives learnt if threshold, K not reached and there is a conditioning context for the rarer alternate)

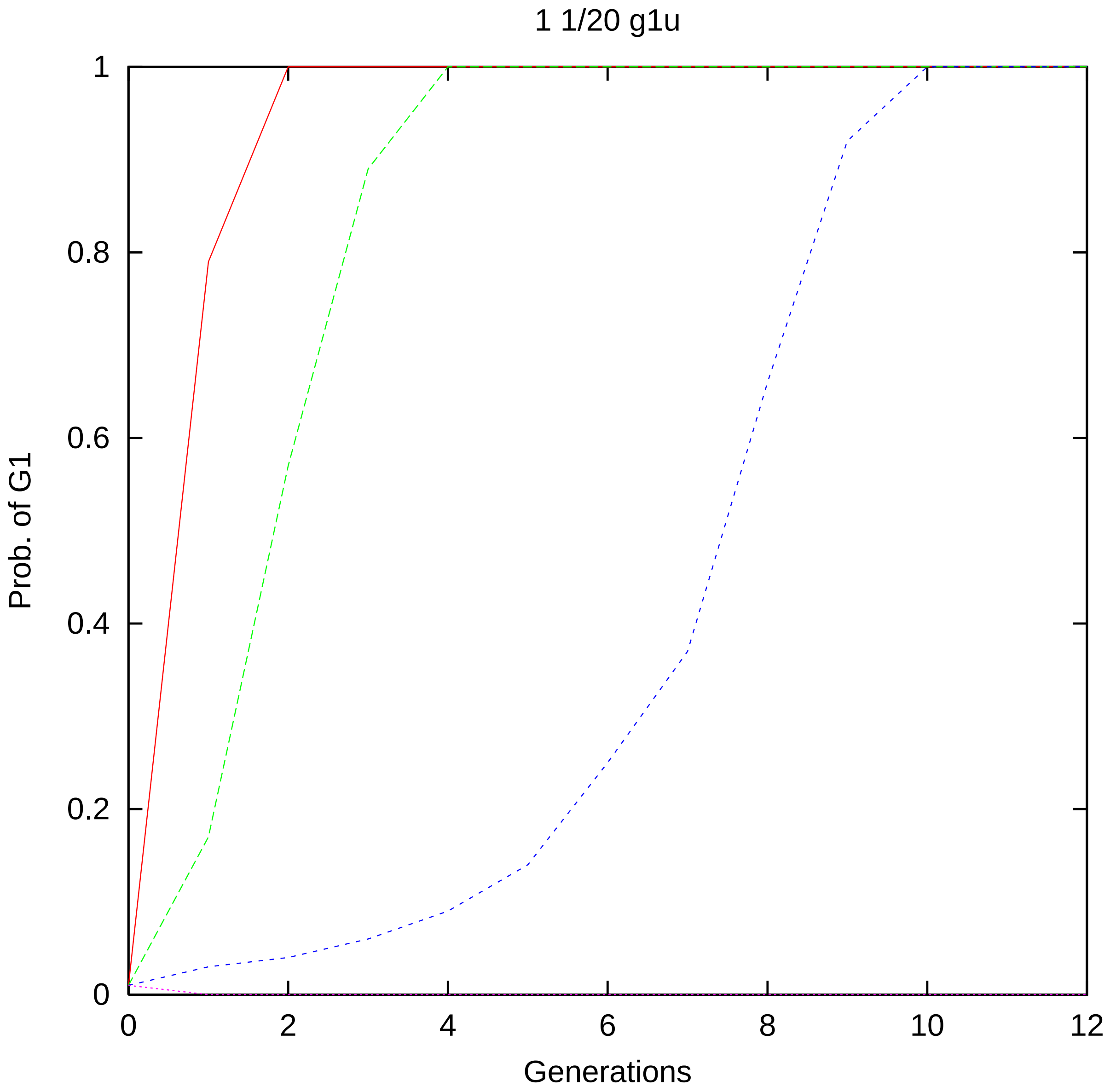
Bayesian Incremental Parameter Setting

BIPS (simplified from Briscoe, 2000c,d)

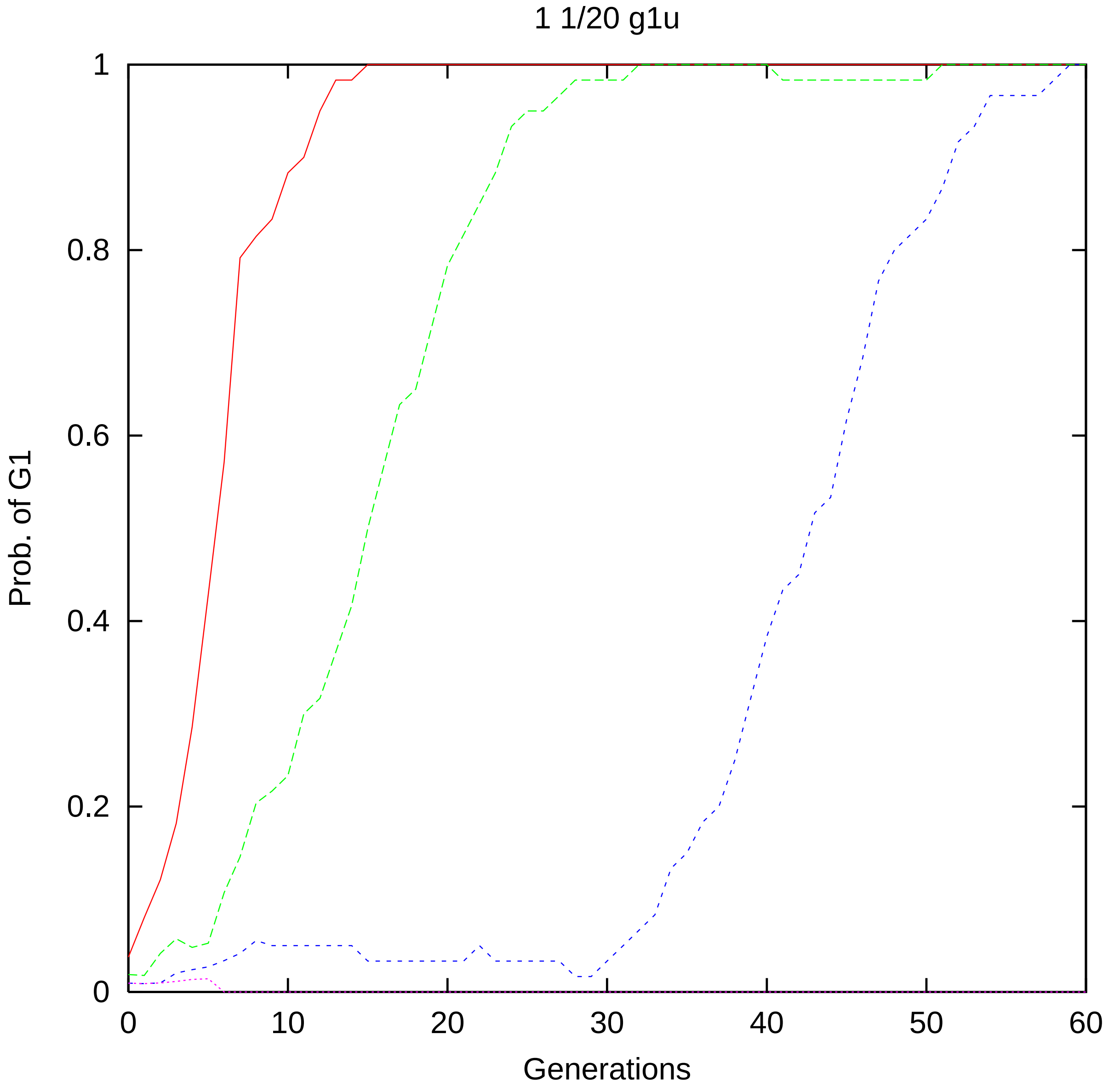
1) Start with all parameters of UG unset ($P = 0.5$) or default ($P = 0.8$);
2) Randomly sample a trigger (degree-0/1, SF-LF pair), t_i from $P(T(L(g^t))$;
3) If the current grammar parses/generates t_i, increment the probabilities of the parameter settings expressed by t_i, goto-2, else:
4) Select one random parameter, flip its setting, increment the probability of this new setting and of the other parameter settings expressed by t_i if that change allows t_i to be parsed/generated, select the most probable $g \in UG$, goto-2.
5) Stop if for trigger, $t_i, i = n$ (i.e. at end of critical period).

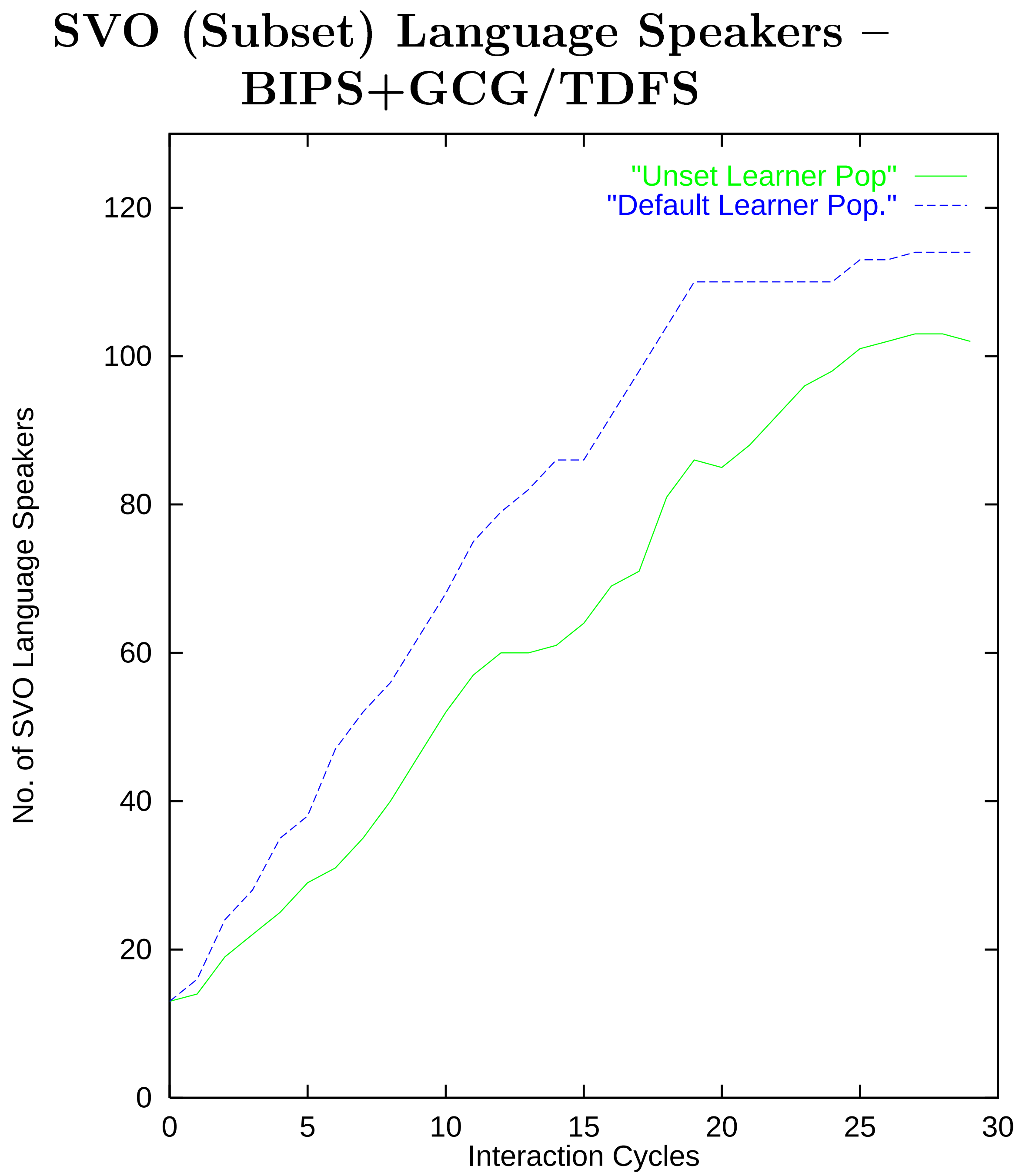
Local, incremental, conservative, memory-limited but frequency-sensitive search for most probable grammar from 'agnostic' / 'biased' starting point(s)

Stochastic NB Model – BIPS



Stochastic OG Model – BIPS





SVO (Subset) Language Speakers –
BIPS+GCG/TDFS
"Unset Learner Pop"
"Default Learner Pop."
No. of SVO Language Speakers
Interaction Cycles
0
20
40
60
80
100
120
0
5
10
15
20
25
30

Conclusions

1. Macro / deterministic model derives logistic spread from unrealistic assumptions of infinite non-overlapping populations and TLA $n = 2$
2. Micro / stochastic non-overlapping finite population model with TLA doesn't predict logistic spread, but does with BIPS
3. Micro / stochastic overlapping finite population model also predicts slower logistic spread with BIPS
4. Micro stochastic models are harder to analyse but not impossible – statistical sampling and population genetics
5. More analytically tractable macro deterministic models will emerge when we have good stochastic models and want to approximate them to prove qualitative behaviour is not accidental 'sampling' of such models and derive laws for behaviour of linguistic dynamical systems (E-languages)
6. Better modelling and better understanding of I-language acquisition is (still) the critical key to understanding (major) linguistic change

Draft paper and other related papers at:
http://www.cl.cam.ac.uk/users/ejb/papers.html