Principled Assessment of Population Structure in Models of Language Change

Jordan Kodner & Christopher Cerezo Falco (University of Pennsylvania)

We discuss unintended effects that common assumptions have in computational models of language change. In particular, the assumptions about population size and single-grammar speakers in a recent paper by Kauhanen (*Journal of Linguistics* 2016) interact to produce unpredictable transient behavior which undermines the paper's point.

Kauhanen seeks to address the impact that social networks have on neutral selection or neutral interactor selection (Baxter et al. *LVC* 2009, Blythe & Croft *Language* 2012). These are patterns of change in which learners internalize their grammars in proportion to their distribution in the linguistic input. This contrasts with all other types of change in which there is some kind of valuation or quantifiable fitness among competing variants. Neutral change has been largely dismissed as a likely kind of change because of its inability to produce S-curves (Blythe & Croft 2012 addressing Trudgill *LiS* 2000). Kahaunen proposes that network rewiring, that is changes in the network structure over time, improve the smoothness and monotonicity of neutral change.

To model the behavior of neutral change, Kauhanen sets up a series of agent-based simulations on centralized network clusters and lays out metrics to quantify their degree of well-behavedness. He finds evidence that rewiring does quantitatively improve neutral change, but does not show that it produces an S-curve. He studies populations of size $n = 200$ and assumes that learners settle on single grammars categorically rather than entertaining multiple competing grammars. We contend that the baseline dynamics which these simulations uncover are primarily a function of these assumptions, and not of neutral change in general. Additionally, we show that under his simulation assumptions, classic S-curves fail to materialize even under grammar competition.

Mathematical models of language change often assume infinitely large populations (Niyogi 1996, Yang *LVC* 2000, etc.). Under such a model, it does not matter whether individuals are categorical or internalize a probability distribution of grammars because an infinite population of categorical individuals can perfectly approximate any distribution. So, as finite n approaches infinity, a categorical population begins to approximate the "true" distribution of grammars calculated through these mathematical models. This makes sense intuitively. A network of $n = 100$ can only capture probabilities in increments of 0.01, while $n = 10000$ has a resolution of 0.0001, and so on.

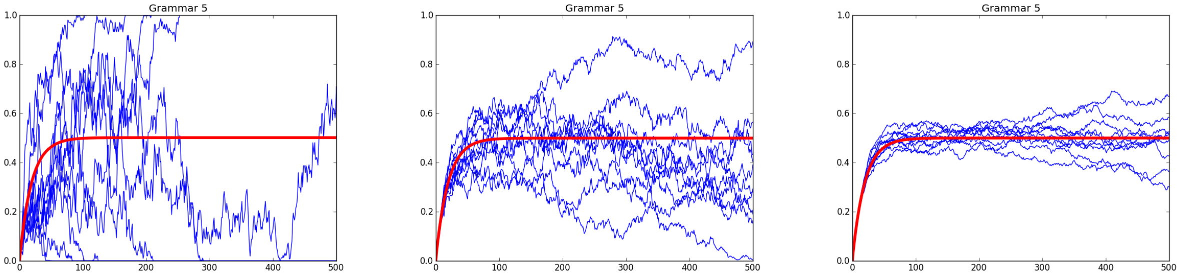

Figure 1: Improving approximations of neutral change with $n = 200$, $n = 2000$, $n = 20000$.

We simulate neutral change in a two-community two-grammar network using the dynamical system model from Niyogi & Berwick (1996, 1997, 2009, etc) augmented with an adjacency matrix to describe network structure. This model is general enough to recreate Kauhanen's network assumptions and efficient enough to compute with large n. Each community begins at 100% monolingual. If $n = \infty$ or competing grammars are allowed, each community homogenizes at a 50/50 grammar distribution (red curve). We then simulate the change for 10 trials with categorical speakers at $n = 200$ (cf. Kauhanen), $n = 2000$, and $n = 20000$ (blue curves). At $n = 200$, the results are chaotic and unpredictable, and most simulations fix at 0 or 100% within 500 iterations. But in line with our intuitions, the path of change begins to approximate the mathematically predicted curve as n increases. Thus the categorical speaker assumption and population size dominate in determining the behavior of this neutral change. Kauhanen's $n = 200$ is just too small to produce well-behaved neutral change among categorical individuals.

Kauhanen draws conclusions about neutral change and network rewiring based on his small simulation, but this experiment shows that small populations under his assumptions do not behave like large populations, either transiently or in the limit. This is troublesome if one seeks to connect the simulation to empirical data from sociolinguistics or historical corpora. Furthermore, neutral change *without* rewiring is itself well-behaved under his metrics when the populations is large. It would be important to test whether rewiring has the same effect in these large populations.

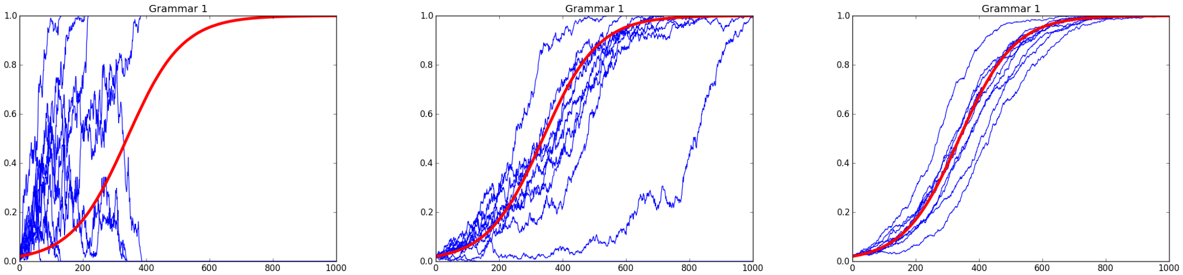

Figure 2: Improving approximations of S-curve change with $n = 200$, $n = 2000$, $n = 20000$.

But the population size effect does not end with neutral change. To demonstrate the problem, we repeat the above experiment with a change including differential fitness in order to produce an S-curve. The logarithmic S-curve is emblematic of language change, and has been empirically observed dozens of times (Kroch *LVC* 1989, Labov 1994, Poplack & Malvar 2006, etc). In a two-grammar system, if one grammar has an advantage over the other, a logarithmic S-curve is all but guaranteed. The dynamics of this curve are well understood (Niyogi & Berwick 1997, Yang 2000) (red curves; $\alpha = 0.31$, $\beta = 0.30$, innovative grammar initialized at 5%). Any computational model of change should be able to produce an S-curve in situations where one variant has an advantage.

We simulate S-curve change in populations of $n = 200$, $n = 2000$, $n = 20000$. The blue lines in the plots represent the first 10 trials which did not fix at 0% within 20 iterations. A familiar pattern emerges. At large n, change is well behaved and approximates a logarithmic curve. For small n, however, change is chaotic. No trials have well-behaved dynamics and most fix at 0% despite the innovative grammar's advantage. So small population and single-grammar speakers conspire to prevent S-curve change.

The question of how, if at all, neutral processes are involved in language change is an interesting one. Simulation under reasonable assumptions can complement empirical work on the histories of attested changes. However, Kauhanen's assumptions concerning population size and categorical learning prevent him from modeling realistic language change. As a result, it is unclear how conclusions about evolving network structures apply beyond his simulations.