**KRETZSCHMAR — Computer**

<3 Figures>

**Computer Simulation of Diffusion: New Suggestions about the Process of Language Change**

Kretzschmar, W. A. and Juuso, I.

In two previous papers at DH we have presented work in progress on computer simulation of language diffusion. In this paper, we offer the results from our completed research program, highly suggestive findings about how the process of linguistic change may operate.

Computer simulation is the only practical way to model linguistic diffusion. We have successfully simulated diffusion with a cellular automaton, which uses update rules with respect to the status of its neighboring locations to determine the status (whether a linguistic feature is used or not) at a given location. As shown in Figure 1, each target cell may become live (if dead) if a certain number of neighbors is live, whether from the eight neighbors immediately next to the target, or alternatively from the 24 neighbors in the first and second rows around the target. The same calculation takes place for a target cell to stay live if it is already live.

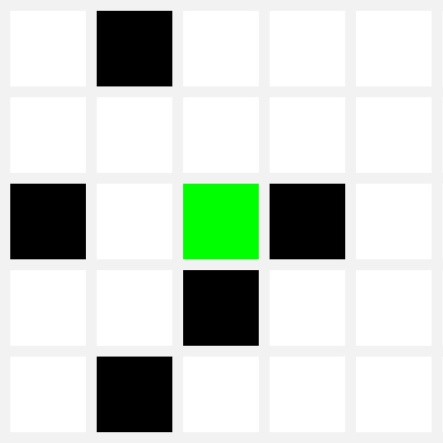


Figure 1. Cellular automaton (green target cell, evaluated by status of eight first-order neighbors, or 24 first- and second-order neighbors).

All locations in a matrix are evaluated, and then the new status for each one is displayed all at once (one generation). Throughout hundreds of generations we can watch regional distributional patterns emerge. In so doing we model human interactions, as speakers talk or write to each other and change their behavior based on that of their neighbors. We validate our results by comparison to actual linguistic data from survey research: we always observe clustered patterns in the survey, and we know that our simulation is successful if similar clusters emerge from the cellular automaton, as shown in Figure 2, the status of our simulation after 1,000 generations with a random factor of .01% (one decision overturned randomly in 10,000). This sort of clustered behavior is characteristic of complex systems (Kretzschmar, 2009), as they are studied in physics, evolutionary biology, economics, and other fields, where nonlinear (or ‘fractal’) distributions of variants regularly emerge at every level of scale in scale-free networks.

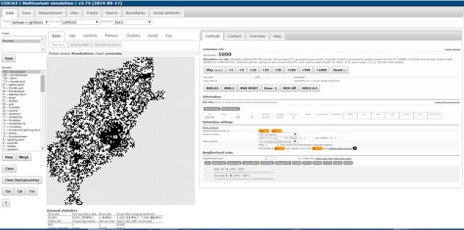


Figure 2. Simulation after 1,000 generations.

The cellular automaton is not the only form of simulation that could be applied to speech, but it is perhaps the simplest. Stanford and Kenny (2013), for example, employed an agent-based model to model chain-shifts, where ‘agents’ could move between locations to spread a linguistic feature, whereas in a cellular automaton the location of each cell is fixed. The Stanford and Kenny model, however, uses numerous unvalidated assumptions about how information is shared and about how agents move. Similarly, Baxter et al. (2009), Blythe and Croft (2009), and Ellis and Freeman-Larsen (2009a) all create more complex simulations that do achieve results but lack validation. Our cellular automaton shows that the complexity of agent movement or other similar parameters in such simulations may not be necessary as well as being unvalidated. After extensive testing of possible rule sets in our two-dimensional model, only one rule set produces stable, clustered results of the kind we always observe from real data (2,3,4 live neighbors to become live, 5,6,7,8 neighbors live to stay live). And this rule set, with suitable adjustment by social weighting and a small random factor, is sufficient to produce results that match the clustered patterns that arise in real survey data.

After substantial experience with the computer simulation, we have observed a number of characteristics that are highly suggestive for how the complex system of speech may operate in actual human populations of speakers:

1. While we have only ever found one rule set that produces clusters, the Bailey set (2,3,4/5,6,7,8, for N=1), other rule sets may be useful, such as proportional rules 90/10, 75/25, and 60/40 that all produce estimates of ‘Where people say X’. The Bailey set, however, eventually produces stable clusters of locations on the grid that match the kind of clustering we observe in Density Estimation statistical processing from the same data.

2. The relative ages of locations (how many consecutive generations a location has been live) always occur in a nonlinear distribution, with the most one generation old, then many two generations, then small numbers of older locations. This suggests that the persistence of features, not just use of features, is important in language diffusion. Persistence is what accounts for the creation of long-term stable clusters of locations.

3. Inclusion of a random factor overturning decisions from the rules up to .06% (six decisions in 10,000) slows down the process of cluster formation, but more than .06% randomness throws the simulation into a chaotic (everchanging) condition where no stable clusters form. This suggests that proximity, not random decisions by speakers, controls language diffusion. However, inclusion of a small random factor preserves nearly all of the long tail of infrequent responses in the nonlinear distribution after 1,000 generations, and so it is necessary to include random decisions by speakers in order to achieve the nonlinear distributions we know to exist in survey data.

4. Inclusion of a social factor also creates clustered behavior (N=2, 25% social weighting). Clusters appear in different places for the social groups defined by characteristics such as age or level of education. Clusters also appear in different places for the same social groups depending on different social information in different seeds, where social information proportional to original survey speakers is added randomly to empty matrix cells. Social ‘proximity’ is thus important to the creation of nonlinear clustering in scale-free networks.

5. When variants fill the grid, they rapidly increase in number of locations up to about 4,000 locations (c. 50%), then hit a plateau where the number of locations only rises very slowly. Persistence in the plateau stage produces stable clusters. The simulation thus has a life cycle for all surviving variants: constant motion across the grid, smaller temporary clusters for up to 250 iterations as a variant builds density across the grid, and (in addition to smaller temporary clusters) larger stable clusters after a variant reaches 50% density, a process that make take 1,000 iterations. This suggests that features in actual speech may also show a life cycle, e.g., common use across wide areas, temporary small areas in which particular features become very common for a time, and stable, potentially large areas in which features are persistent for long periods.

6. Running the simulation across all the variants does not produce the A-curve of values that we see in the survey data, so the A-curve in the survey data does not arise merely from the effects of proximity. However, we can create a separate array that represents what speakers remember in their *active* speech, while allowing the rules to run across all the variants represents what is available to speakers in their *passive* speech. As shown in Figure 3, this ‘active memory’ array does show an A-curve across all surviving variants when a small random factor is included, which acts on all variants but preserves many low-frequency variants that would otherwise die out. This suggests that the operation of the complex system preserves variants across wide areas in passive understanding of language, but that active use of language involves common use of a smaller number of variants per speakers with a nonlinear preservation of variants across the whole population; this active use of language is what a survey normally elicits. The simulation thus addresses individual human cognitive capacity.

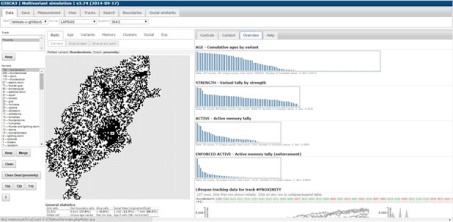


Figure 3. Chart of ‘active memory’ for simulation with random factor of .01% after 1,000 generations (see ‘active’ and ‘enforced active’ tallies at right).

Our use of a simple cellular automaton in a successful simulation suggests how we might better understand the survey and other data we have already collected, and also suggests how we might do a better job of collecting additional empirical data about language in future. The simulation indicates that we should use care in creation of overly complicated simulations when a simpler one will do.

**References**

**Baxter, G., Blythe, R., Croft, W. and McKane, A.** (2009). Modeling Language Change: An Evaluation of Trudgill’s Theory of the Emergence of New Zealand English. *Language Variation and Change,* **21**: 257–96.

**Blythe, R. and Croft, W.** (2009). The Speech Community in Evolutionary Language Dynamics. In Ellis and Larsen-Freeman 2009b, pp. 47–63.

**Ellis, N. and Larsen-Freeman, D.** (2009a). Constructing a Second Language: Analyses and Computational Simulations of the Emergence of Linguistic Constructions from Usage. In Ellis and Larsen-Freeman 2009b, pp. 90–125.

**Ellis, N. and Larsen-Freeman, D.** (2009b). *Language as a Complex Adaptive System*. Wiley-Blackwell, Oxford.

**Kretzschmar, W.** (2009). *The Linguistics of Speech*. Cambridge University Press, Cambridge.

**Kretzschmar, W.** (In press). *Language and Complex Systems*. Cambridge University Press, Cambridge.

**Stanford, J. and Kenny, L.** (2013). Revisiting Transmission and Diffusion: An Agent-Based Model of Vowel Chain Shifts across Large Communities. *Language Variation and Change,* **25**: 119–53.