Language Change Model: Final Report

Introduction

Natural language changes over time, and it tends to follow one of several patterns. But what causes each of these patterns to happen is still not entirely understood. In this paper, I attempt to find out what causes these patterns to happen through the use of agent based modeling.

Motivation

One phenomena I attempt to model is the concept of a dialect continuum. A dialect continuum is a situation where speakers of each dialect in the continuum can easily talk to and understand their neighbors but the edges of the continuum have little to no ability to talk to each other. Generally, dialect continua in the real world are caused by a single language changing in different ways in different places, as happened with Arabic evolving into the various Arabic dialects, and with Latin evolving into the Romance languages. One of the goals of this model is to discover why dialect continua form and under what conditions.

Sometimes, it happens that languages diverge very slowly, and other times they diverge quickly. Old English had three descendents: Modern English, Scots, and an extinct Southern Irish language called Yola. Here is a sample of Scots[1]:

We twa hae run about the braes,

and pou'd the gowans fine;

But we've wander'd mony a weary fit,

sin' auld lang syne.

1 This is the 4th verse from the song Auld Lang Syne, written by the Scottish poet Robert Burns in 1788.

And here is a sample of Yola[2]:

A moan vrim a Bearlough an anoor vrim a Baak,

Thaye zhoult upan oother at high Thurns o Cullpaak,

Themost wi egges an heimost wi thick,

Fan a truckle ee zhoulthered too nigh upa ditch.

Clearly, Yola is more distant from Modern English than Scots is, even though all three languages began diverging at about the same time. One of the goals of this model is to figure out what causes languages to diverge more quickly or more slowly than average.

Model Construction and Measurement

In this model, speakers of a language are modeled as patches, on a 16x16 landscape of patches that does not wrap. Each patch has a language variable, which is represented as a list of numbers. The model has several parameters, which will be discussed in detail below. When the model runs, on each tick each patch first attempts to mutate itself. Then it attempts to talk to a number of its neighbors; successfully talking to a neighbor makes the language of both patches more similar.

NUMBER-OF-ATTRIBUTES is a parameter of the model which corresponds to the number of attributes in the patch's language list. Each attribute is a number from 0 to 255. If the number of attributes is exactly three, the patch's color will be set to its language; otherwise the patch's color will be set to the mean of all numbers in the list. Because the special case visualization with exactly three attribute is generally more intuitive, I usually run the model with

2 This sample is from a folk song called the “Song of Two Market Women”. Lyrics found in Shiels, Damian. "Column: Yola and Fingalian – the Forgotten Ancient English Dialects of Ireland." TheJournal.ie. Accessed June 6, 2015. http://www.thejournal.ie/readme/column-yola-and-fingalian-–-the-forgotten-ancient-english-dialects-of-ireland-985649-Jul2013/.

NUMBER-OF-ATTRIBUTES = 3.

SPEECH-RADIUS is the maximum distance in space that patches can talk to each other at. The theoretical justification for this was that sometimes, particularly with modern communication, people talk to people who are far away from them. However, I usually run the model with SPEECH-RADIUS = 1, because higher distances produce very strange and unrealistic patterns of language spread.

NUM-INTERACTIONS is the number of patches each patch attempts to talk to on each tick. It varies between 0 and 8. At zero, patches do not attempt to talk to each other at all, and at eight, they attempt to talk to all of their neighbors.

MUTATION-RATE is the maximum amount a patch will attempt to change its own language per tick. Every tick, each patch will attempt to change its own language a random amount between 0 and the mutation rate. Although language attributes range from 0 to 255, MUTATION-RATE ranges from 0 to 100 simply because values over 100 are not noticeably different from each other. No matter what the other parameters are set to, values of mutation rate above 100 knock out all other effects and end up in a situation where each patch's language is essentially random.

INFLUENCE-FACTOR is a multiplier of the amount each patch will attempt to change its neighbors on each tick. The other factor in this equation is the vector-space distance between the two languages. The full equation is that the amount a language changes when it talks to another languages is equal to the distance between the two languages times INFLUENCE-FACTOR divided by 100. It varies from 0 to 100, but because it is divided by 100 in the equation, it actually varies between 0 and 1.

THRESHOLD-OF-INTELLIGIBILITY is the percent of the maximum possible lexical

distance (linguistic similarity, computed as Euclidean distance) at which two languages can talk to each other. (It is a percentage because the maximum possible lexical distance changes depending on NUMBER-OF-ATTRIBUTES.)

In this model, there are three setup methods. The first one is “setup-uniform”. When the model is setup in this way, each patch's language is set to the same list composed of NUMBER-OF-ATTRIBUTES random numbers. This setup method is very useful for modeling language divergence, which makes this setup method the one that I use the most.

The next setup method is “setup-random”. This setup method, like setup-uniform, sets each patch's language variable to a randomly generated list. However, unlike setup-uniform, with this method each patch is assigned a different randomly generated list, which generates a very chaotic starting situation where starting intelligibility is low. This setup method would be useful for modeling dialect convergence; however, since that happens relatively rarely in real life without outside forces, in practice I mostly use this setup method for testing purposes.

The last setup method is setup-contact. It is very similar to setup-uniform, except that instead of all patches being given the same list, patches left of the origin are given one list and patches right of the origin are given a different list. How far these are from each other is set by the CONTACT-MAX-DISTANCE parameter. I use this setup method partially for testing the model and partially for exploring situations of language contact.

In addition to these setup methods, there is a setup-like method called spread-obstacle-with-mouse. When this button is on, whatever patch the user clicks will become an “obstacle” with all language attributes set to -1, which other patches cannot talk to. This allows the user to model situations of language isolation.

When discussing my results I will refer to several metrics I defined to measure the

behavior of the model. The first is called the “average language”. Theoretically, I would like to make this the language that the highest number of other languages can understand. However, this would require comparing every patch's language to every other patch's language, making it computationally intractable to compute every tick. So instead I approximate this by finding the language where the mean of the attributes of that language is closest to the mean of the attributes of all languages.

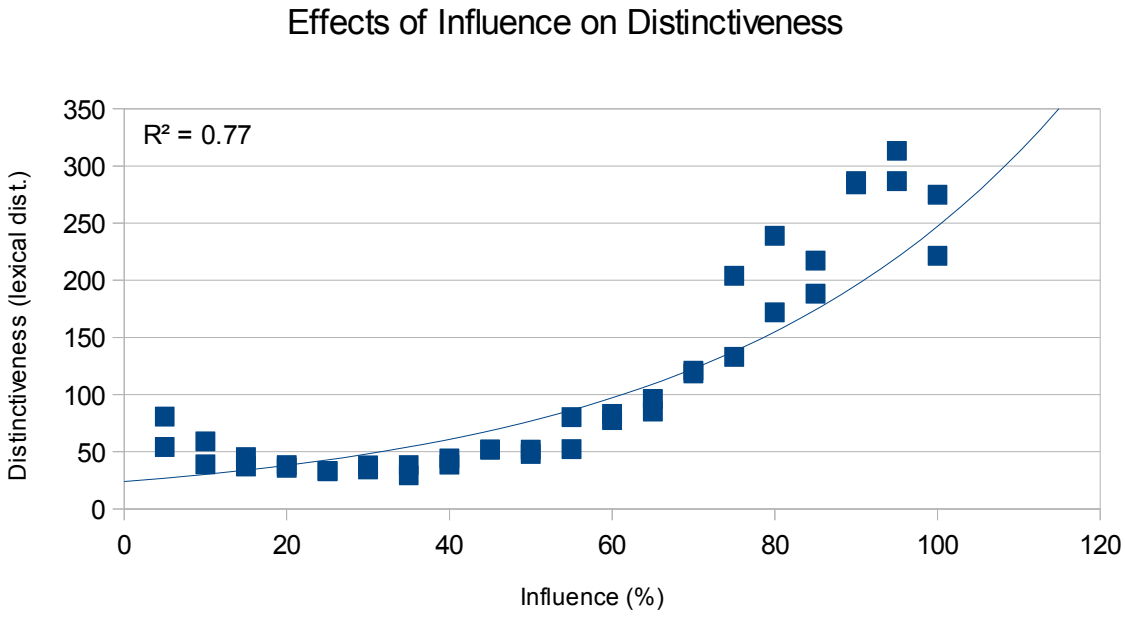
From that measure naturally follows the “intelligibility to the average language”. This is the percentage of patches whose language can be understood by (i.e. is under the intelligibility threshold from) the average language. I use this as one measure to determine how fragmented the data is; the nearer this metric is to 0%, the more fragmented the data is.

Another new measure is called “distinctiveness”. If the most distinctive language is the language furthest from the average language, the distinctiveness of the system is the distance between the most distinctive language and the average language. This measure can be used to see how much language change there has been in the system. Related to distinctiveness is the distance between the edges: this is the distance between the max language (the language with the greatest mean of attributes) and the min language (the language with the least mean of attributes). Distinctiveness and distance between edges are very strongly correlated with each other.

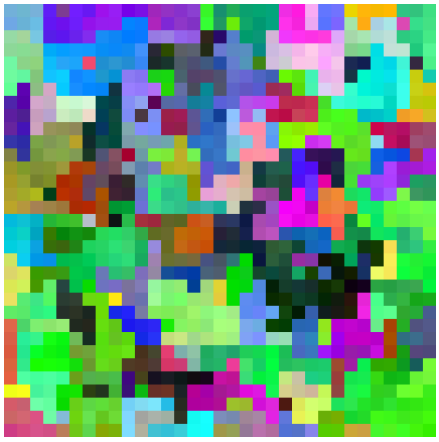
Finally, I approximate the number of distinct languages by gathering a list of all patches' languages, removing those that are intelligible to the average language, and then repeatedly removing those that are intelligible to the first item in the list until no more remain. This metric is much more computationally intensive than the others to compute, so I only collect it at the end of a run instead of at every tick.

Results

The most obviously counter-intuitive result I found is that, as long as it is greater than zero, INFLUENCE-FACTOR is strongly positively correlated with both distinctiveness and distance between edges, and somewhat negatively correlated with intelligibility to the average language and distance between averages. (If it is zero, all these values mimic the values they would take if influence rate was 100, as mutation rate dominates.) Or in other words, higher influence actually increases the rate at which languages diverge.

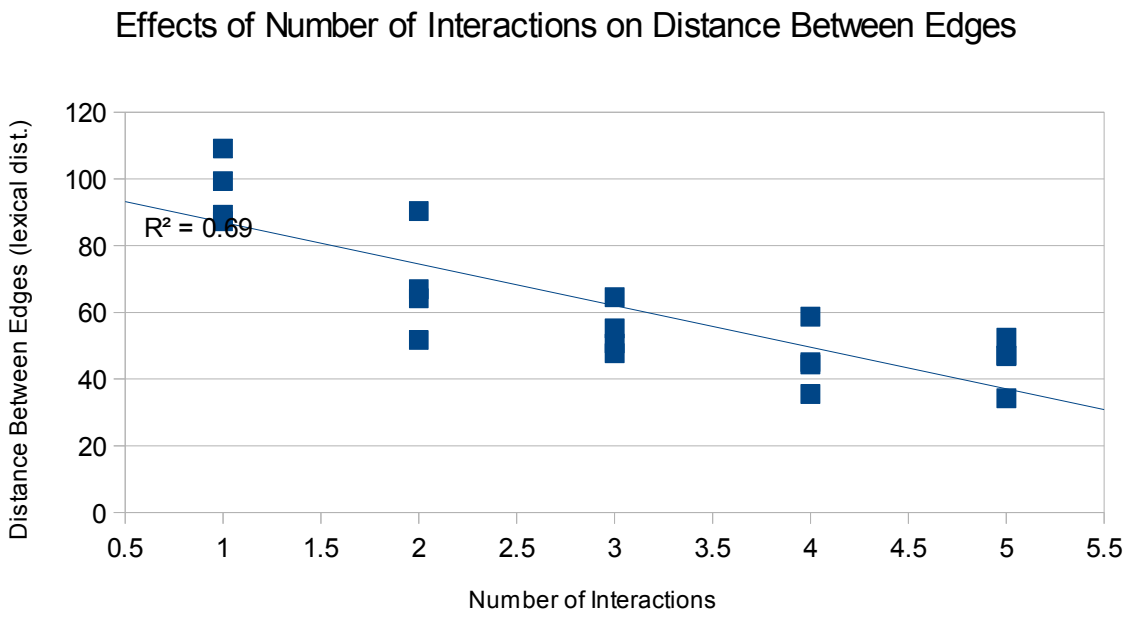


Below, the screenshot on the left is from tick 100 on a run of the model at INFLUENCE-FACTOR = 25, while the screenshot on the right is from the same tick with the same other parameters but INFLUENCE-FACTOR = 100. It can be seen that distinctiveness is not just high at high influence, it also produces patterns where patches near each other tend to share a language much more than patches further away.



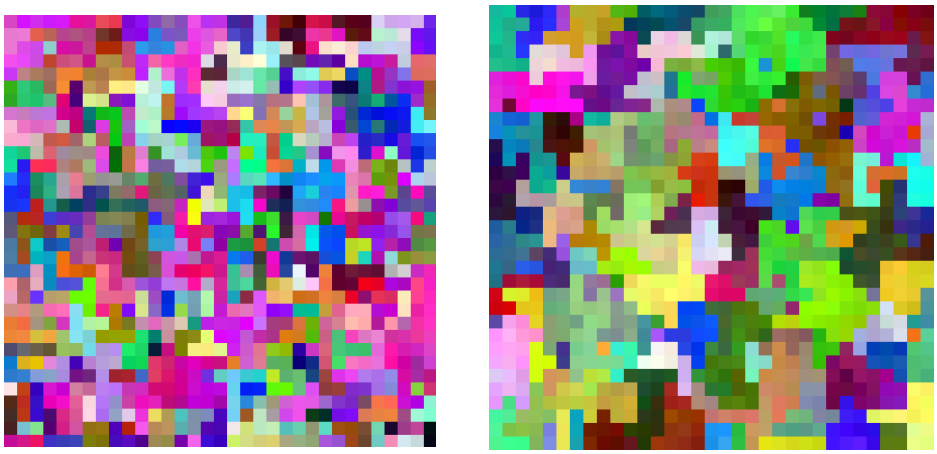
What seems to be happening here is that at very high influence, languages that mutate a little bit are better able to drag their neighbors with them than at lower influence. So as long as influence is not zero (in which case mutation dominates), two areas in the same run of the model get dragged further away from each other at high influence than low influence.

NUM-INTERACTIONS is negatively correlated with both distinctiveness and distance between edges, although not as strongly so:

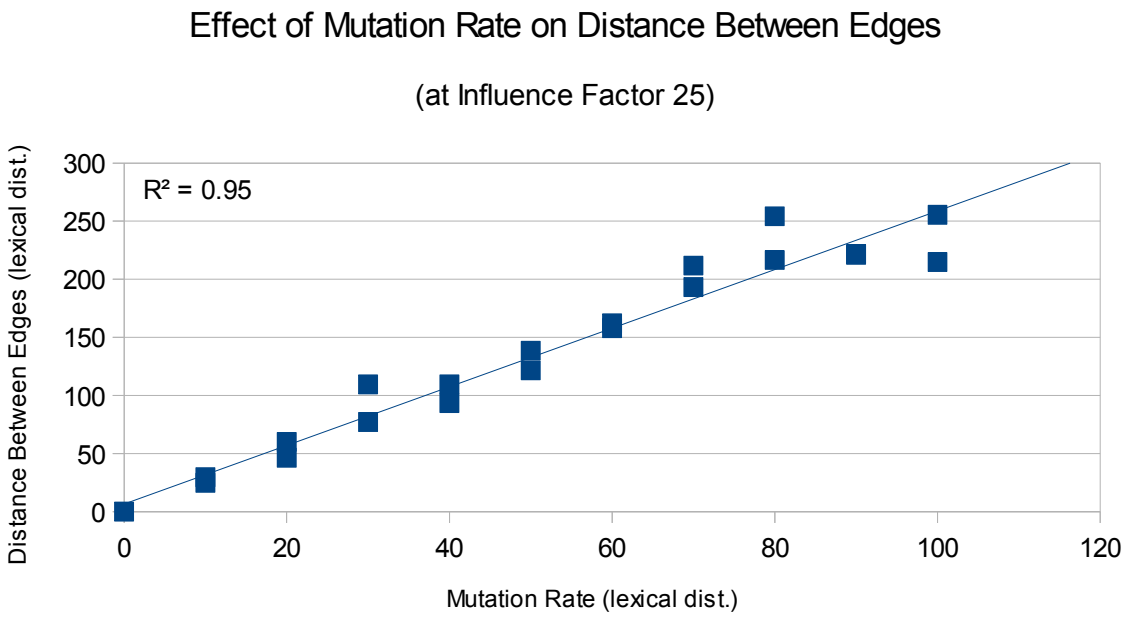


The following two views are both from tick 200. The one on the left is NUM-INTERACTIONS = 1, and the one on the right is NUM-INTERACTIONS = 2. As can be seen,

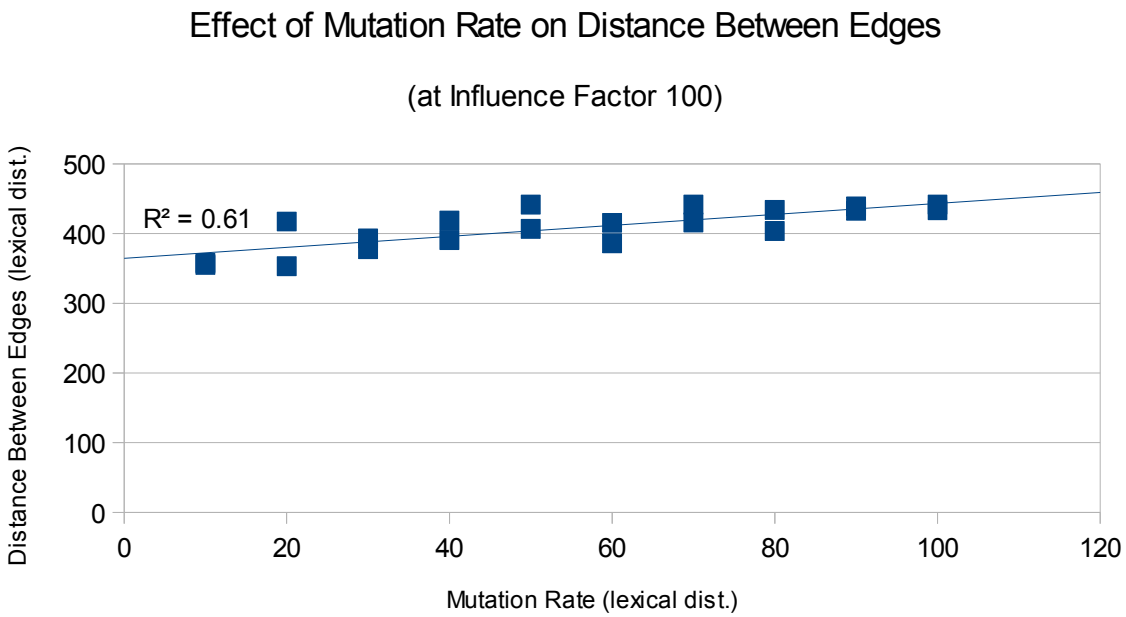
high NUM-INTERACTIONS generally makes the landscape less choppy and more composed of contiguous blobs of the same language.



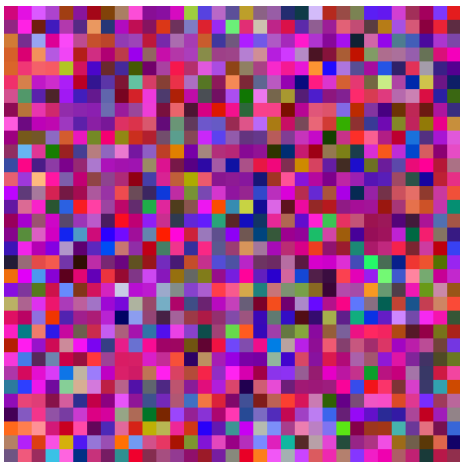
At low levels of influence, distinctiveness and distance between edges are almost entirely due to mutation rate:

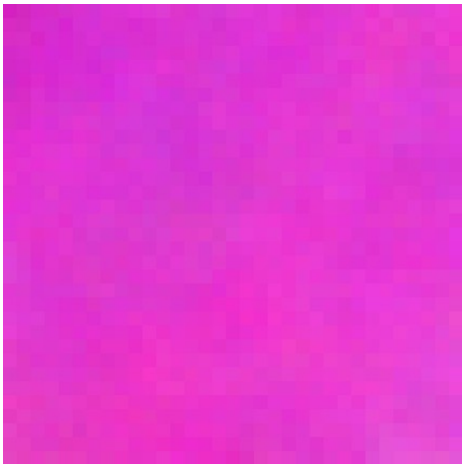


However, at higher levels of influence, influence rate quickly drowns that effect out:



THRESHOLD-OF-INTELLIGIBILITY is almost entirely uncorrelated with distinctiveness and distance between edges. It obviously has effects on average intelligible neighbors and intelligibility to the average, but no non-trivial ones. Anecdotally, a high THRESHOLD-OF-INTELLIGIBILITY reduces choppiness in a similar manner to how NUM-INTERACTIONS does. The image on the left is at THRESHOLD-OF-INTELLIGIBILITY = 5, the image on the right is at THRESHOLD-OF-INTELLIGIBILITY = 95, both at 150 ticks.





Finally, the set of parameters that best causes a dialect continuum-like situation is a high num-interactions (4 is good, 8 is best), a low-ish mutation rate (10 - 15) and influence factor (15-25%), and a low-ish threshold of intelligibility (25-35%). These conditions combine to form a situation like the bottom, where (for example) the purple at the top and the cyan in the center-left can just barely not understand each other.

Conclusions

From my results, it seems that a dialect continuum is caused when speakers of a language change their language relatively slowly and do not exert special pressure on others to speak their language. When speakers do exert special pressure on others to speak their language, the language fractures into new languages fairly quickly.