Math 243 Final Project

Trace Baxley, Caitlin Hauser, Bright Liu

May 2023

1 Abstract

How multilingualism is developed and maintained is an important research subject in linguistics and social studies (Lieberman, Erez, et al 2007). In this paper, we explore the notion of language competition, the process through which multiple languages compete for dominance within a given population. Understanding the factors that influence language competition may help provide insight on language evolution and extinction, and yields important results for mathematicians and linguists alike.

We are going to examine the language evolution models proposed in Steven Strogatz's 2002 paper *Modelling the dynamics of language death* and Zhijun Wu's 2020 paper *Why multilingual, and how to keep it — An evolutionary dynamics perspective*. Specifically, we are going to offer an alternative solution to the problem of how bilingualism could coexist stably long-term by introducing a more realistic evolution model that takes language enclaves into account (Auer, Peter; Schmidt, Jurgen Erich. 2009).

We discover that the model proposed results in equilibria similar to that of language enclaves in real life. Moreover, we find that increasing population and increasing the enclave size, up to a point, helps in keeping the language alive.

Later on, we lay important theoretical grounding for how we could implement this evolutionary dynamics for trilingualism and other future expansions to our project.

2 Background and Literature

Steven Strogatz's minimal model for language change (Nature, 2003), uses a simple rate of change expression and minimal assumptions to reach the conclusion that "two languages cannot coexist stably - one will eventually drive the other to extinction." At the same time, he conjectures that by incorporating control on the relative status of X, "active feedback does indeed show stabilization of a bilingual feedback point." This raises an interesting question: how could multilingualism persist when our models predict it should have gone extinct?

Moreover, most language evolution models rely on a crucial assumption: well-mixed populations. This assumption is imperfect and doesn't necessarily apply to the real world. We seek to explore the case of **language enclaves** to understand the evolution of languages in non-well-mixed populations. We will do this by incorporating enclaves into our model and simulations. An enclave is defined as a portion or territory within or surrounded by a larger territory whose inhabitants are linguistically distinct. Thus, by having linguistically distinct subpopulations clustered into a certain region of the greater population, we eliminate the assumption that the population is well-mixed and explore how languages evolve when modifying the number and sizes of enclaves.

A research paper titled "Why multilingual, and how to keep it—An evolutionary dynamics perspective" explores that by combining societal interventions and concluding that the stable co-existence of languages in the multilingual form is possible (Zhijun Wu, 2020). We will examine a variation of the problem posed in this research paper where we relax the assumption that the model is well-mixed populations. Our adjusted model may help provide insight on the evolution of language when contact between individuals is limited (e.g. during the pandemic) or not evenly distributed.

3 The Basic Model of Language Competition (Strogatz, Wu)

3.1 Strogatz's Model for Language Death

$$\frac{dx}{dt} = yP_{yx}(x,s) - P_{xy}(x,s)$$

We have two competing languages X and Y and the equations of change depend on two factors: number of speakers and its perceived status. $P_{yx}(x,s)$, x represents the fraction of the population speaking X and $0 < s < 1$ measures X's relative status, with $y = 1-x$. Then $P_{xy}(x,s) = P_{yx}(1-x, 1-s)$, $P_{yx}(0,s) = 0$, $P_{yx}(x,0) = 0$ and P_{xy} are smooth and monotonically increasing in both arguments. Strogatz's model has two stable fixed points: $x = 0$ and $x = 1$. Thus, according to the model, two languages cannot coexist stably long term and one will eventually drive the other to extinction.

3.2 Wu's Evolutionary Dynamics of Language Competition

(1): Wu introduced two sets of change equations, one for language competition and the other for purely societal influences. This leads to the net payoff functions:

$$\pi((x_A, x_B), (y_A, y_B)) = x_A P_A(y_A) + x_B P_B(y_B)$$
$$\bar{\pi}((x_A, x_B), (y_A, y_B)) = x_A \bar{P}_A(y_A) + x_B \bar{P}_B(y_B)$$

(2): Wu assumed P is smooth and increasing, P^* is smooth and decreasing, so the actual payoff functions are:

$$P_A(y_A) = cy_A^{\alpha-1} s_A, P_B(y_B) = cy_B^{\alpha-1} s_B$$

$$\tilde{P}_A(y_A) = \tilde{c}y_A^{\tilde{\alpha-1}} \tilde{s_A}, \tilde{P}_B(y_B) = \tilde{c}y_B^{\tilde{\alpha-1}} \tilde{s_B}$$

4 Relaxing the Well-Mixed Assumption

4.1 Using Wu's model Without Societal Intervention and Applying it to Language Enclaves

An important assumption that the Wu study makes note of is the "well-mixed" assumption. That is, every individual in the population has an equal chance of interacting with one another. However, in practice, we observe that societies are often separated based on language preference, such as the language and cultural enclaves in large cities like New York or Boston.

We hypothesize that the counter-intuitive result of the dynamics in the Strogatz paper (that one language will eventually dominate another) is largely a result of this well-mixed assumption, and that adding another set of equations to the payoff function to represent societal influence (as seen in Wu's paper) is not necessary to form a non-(0,1) equilibrium (meaning that the population eventually becomes monolingual). Additionally, we will see what kind of language enclaves (changes in size, population, etc.) lends itself to more attractive equilibria.

4.1.1 Set-Up

Parameters:

- Population size: P
- Enclave population size: E, $(E < P)$
- Number of enclaves: k
- Total Area: a_T
- Enclave Area: a_E, $(a_E < a_T)$

- The typical parameters associated with payoff function as stated in Wu's standard model (α, s_A, s_B)

In our model, we treat every individual in the population as some $v \in V(G)$. Then, we randomly assign each v to an ordered pair in

$$\{(x, y) | 0 \leq x \leq \sqrt{a_T}, 0 \leq y \leq \sqrt{a_T}\}$$

via a uniform distribution. We assign $P - E$ nodes ordered pairs not contained in the enclave subset, and E nodes ordered pairs contained in the enclave subset (the formalization behind this depends heavily on the type of model discussed below).

Additionally, we assign normally distributed x_A and x_B, restricted on $[0, 1]$, and such that $1 - x_A = x_B$, where (x_A, x_B) is the strategy of an individual in a population. x_A is the probability that an individual will speak language A in an interaction, and x_B is the probability that an individual will speak language B in an interaction. We assign those in the enclave an x_A value via a normal distribution where the mean is 0.25 , and assign those not in the enclave an x_A value via a normal distribution where the mean is 0.85. We create a complete graph and initialize edge weights using the subsequent construction.

4.1.2 Weighted adjacency matrix

Typically, gravitational models are meant to model population flow between cities, dependent on both the population size of the cities and the geographic distance between them (Poot, Jacques, et al, 2016). For our model, we want the edge weights to represent a probability that two individuals will interact in a population in any given generation. A typical gravitational model looks like the following:

$$\frac{m_1 m_2}{\text{dist}(x_1, x_2)^2}$$

where m_1 and m_2 are the populations of the cities and **dist** refers to the standard Euclidean distance function. For the sake of our model, our adjustment will reward speakers who tend towards being pure A speakers or pure B speakers, and offer additional reward for similar strategies. We have

$$w_{ij} = k \frac{\max\{\sqrt{(x_A^j)^2 + (x_A^i)^2}, \sqrt{(x_B^i)^2 + (x_B^j)^2}\}}{\text{dist}(x^i, x^j)^2}$$

where

$$k = \frac{1}{\max\{w_{ij}\}}$$

.

because we want $w_{ij} \in [0,1]$. We then create a weighted Y_A^i value for a given node i. In the standard model, (Y_A, Y_B) is considered to be the average strategy of all of the nodes in the population:

$$Y_A^i = \frac{\sum_{k=1}^{P-1} w_{ik} x_A^k}{P-1}$$

Our weighted Y_B^i is similarly defined. Note that these values should be considered the weighted average strategy for each individual i.

4.1.3 Abrams-Strogratz Method for Simulation

We will be using the Abrams-Strogratz method for simulation to carry out our dynamic simulation. A graph of P nodes is first constructed. A random individual node is then selected repeatedly from the graph, where a game is played for the node against the rest of the nodes of the graph. We let (x_A, x_B) be the current strategy for the node, and (y_A, y_B) be the strategy for the population (the rest of the nodes). Then $p_A = P_A(y_A)$ and $p_A = P_A(y_A)$ are the payoffs for A and B speakers respectively. Then we compute the payoff for the individual node $\pi = x_A P_A(y_A) + x_B P_B(y_B)$. If $p_A > \pi$, we increase x_A by setting $x_A = y_A$ if $x_A < y_A$. However if $p_A < \pi$, we decrease x_A by setting $x_A = y_A$ if $x_A > y_A$.

There is one major difference between our model and the Abrams-Strogratz original model. Firstly, our adjusted Y_A^i, Y_B^i values are chosen given depending on which i value is chosen. We then replace values of x_A and x_B depending on comparison to adjusted Y_A^i, Y_B^i and update the other aspects of our simulation accordingly.

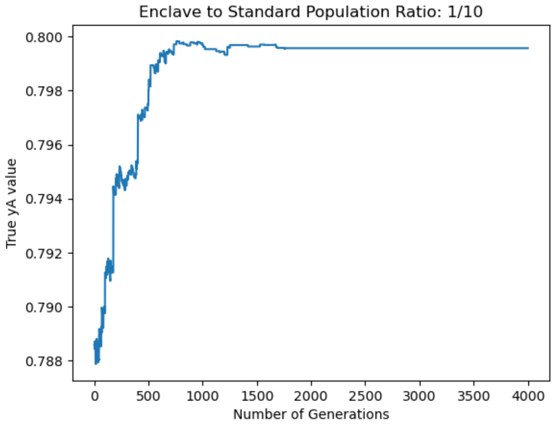
Note that in all of our simulations, we imposed the following restrictions on some of the aforementioned parameters: $s_A = s_B = 0.5$, $\alpha = 1.5$, $P = 300$, $a_T = 20$.

Our simulation runs for 4000 generations, with each generation having a total of P^2 "interactions" (possibilities of updating). Given some set of parameters, we run the simulation 10 times and take the average value of y_A (and thus y_B) for each generation. Ultimately, we are concerned with the equilibrium strategy (Y_A^*, Y_B^*), or the average value from the 10 iterations that occurs at the end of the 4000 generations. In many of our sets of simulations, our initial average strategy changes depending on our variable parameters. Thus, for many of our plots we will be concerned with $Y_A^* - Y_{A0}$, which is meant to show the difference between the equilibrium point and the initial strategy. In our discussion about bilingual populations, it will be helpful to uncover the types of populations that allow for stable equilibria that are similar to the initial distribution of the population.

5 Results

5.1 Adjusting Population Ratio

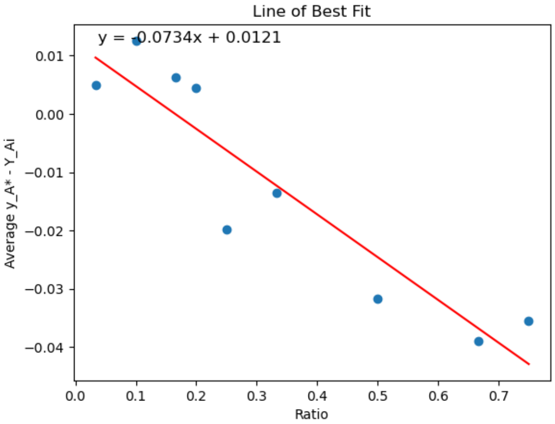
Initially, we intended on adjusting the ratio of the population size inside of the enclave ($\frac{E}{P}$) to observe a possible change in equilibrium. We add the additional constant condition that $a_E = 1$. For example, after 10 runs of the simulation, here is the our mean strategy for language A after 4000 generations for $\frac{E}{P} = \frac{1}{10}$:



We ran 10 simulations with these conditions in place and calculated the average values, for

$\frac{E}{P} = \frac{1}{30}, \frac{1}{10}, \frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}$

Plotting the difference in language distribution, against the population ratio, we are given the following line of best fit, we have the correlation coefficient $r = -0.999$:

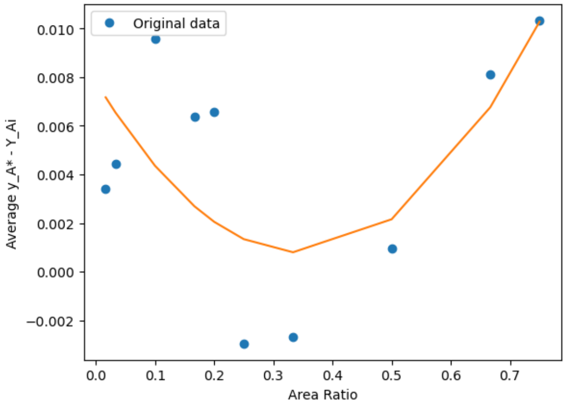


5.2 Adjusting Area Ratio

Now, we keep $\frac{E}{P}$ at a constant value $(\frac{1}{4})$ and adjust the enclave area. We ran 10 simulations with these conditions in place and calculated the average values, for

$\frac{a_E}{a_T} = \frac{1}{30}, \frac{1}{10}, \frac{1}{6}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}$

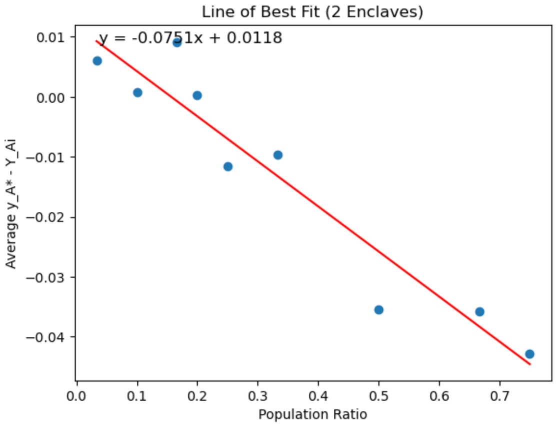
We plotted a quadratic line of best fit to account for the decreasing then increasing quality of our data. (Note that this is a quadratic approximation, but for values that are close together the plot function does not give a sufficient visualization of the parabola), we have the correlation coefficient $r = 0.2$, thus the model is not very accurate compared to the real data.



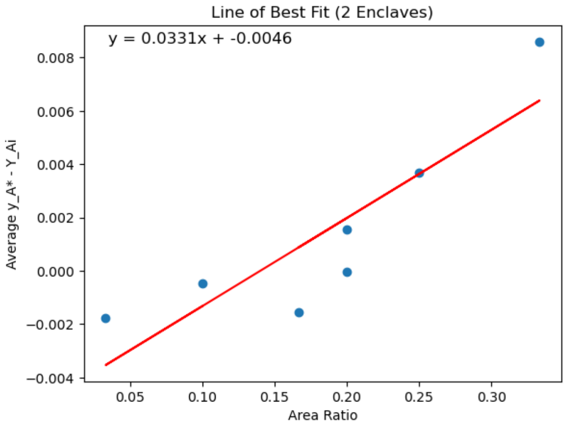
5.3 The Case of 2 Enclaves

Now, we simulated the language evolution in the same way, but did so for 2 enclaves whose total area added up to a_E. These enclaves were non-intersecting and square in shape (as before) in the set with area a_T, we have the correlation coefficient $r = -0.92$:

Plotting the difference in language distribution against the population ratio, we are given the following line of best fit:

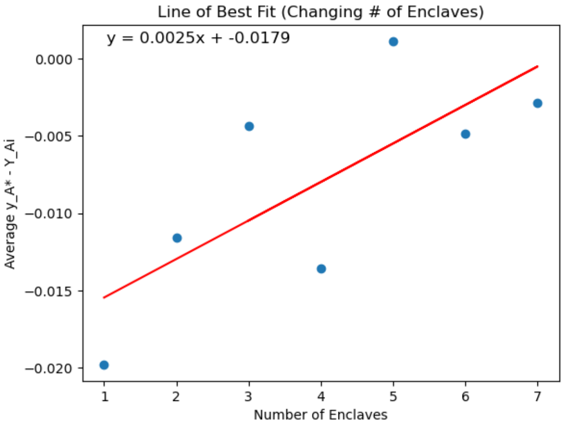


Note that due to the randomly generated nature of the enclaves and the non-intersection restriction, it was not possible to test values near or above $\frac{1}{2}$, we have the correlation coefficient $r = 0.8769$:



5.4 The Case of k Enclaves

Finally, we kept $\frac{P}{E} = \frac{1}{4}$ and kept $\frac{a_E}{a_T} = \frac{1}{10}$, and adjusted the amount of enclaves generated (again, whose area add up to a_E), we have the correlation coefficient $r = 0.741$.



6 Discussion/Conclusion

We will start by discussing the intuition behind the behavior in our regression analyses.

We found a medium negative correlation between the ratio of the population size and our dynamic difference (we will refer to $Y_A^* - Y_A 0$ as dynamic difference from here on out). By the way we set up our simulation, the greater the population in the enclave, the less likely that those outside the enclave will be in spots that are near the enclave (because there are less values that are randomly generated). As a result, the adjusted Y values for the vertices that are in the enclave should have a larger influence from the non-enclave nodes, given that there are more of them in the surrounding areas in the enclave (note that x_A is the surrounding language and x_B is the enclave language).

The behavior of the data in the single enclave case decreases and then increases with the increasing area ratio. As for the case before where the area of the enclave was relatively small ($\frac{1}{10}$), the high dynamic difference value is attributed to the fact that for nodes generated around the edge of the enclave, they will be relatively close to every single node in the enclave, and thus have a large influence on each of the Y_A^i values in the enclave. As the area gets larger, there is more ”room” in the enclave so that every non-enclave node does not have as much influence on, let's say, the other side of the enclave. When the enclave grows to a size such that it takes up almost the entire region, the randomly generated non-enclave nodes have no other option but to be close to the edge of the region (and thus, many of the enclave nodes). Further research is necessary to determine if a quadratic approximation is a proper one.

For the behavior of the change in the third graph, because there are two enclaves, there is more of a chance for randomly generated nodes in the non-enclaves area to interact with those inside of the enclaves, a bigger parameter and we indeed see there is more noise, causing a lower correlation in this graph.

The data from the two enclaves that varied the area of the regions had a linear regression with a medium-high positive correlation (.92 r-score). In our assumption, our population in our enclaves was 1/4 of the total population. Intuitively, we can use similar rationale from the increasing part of the function with one variable. However, because two enclaves were randomly generated in the space, the smaller area of the graphs do not affect the enclave speakers as much – one non-enclave node being around another enclave node will not affect every single Y_A^i value there – only approximately half of them. Additionally, due to the small size of the enclaves, the probability that a non-enclave node gets randomly generated near the enclave is very low in comparison to higher values.

For the behavior of the graph of the case of k enclaves, we see that in the one-enclave situation, it is a central block in the population, however for k enclaves we are still keeping the total number of people in enclaves constant, which may cause two or more small enclaves to border each other just by accident. Following the same logic as the third graph, we see our correlation is indeed lower.

When deciding on whether this result can show real world applicability, we must consider the language enclaves that exist in the real world. Language enclaves like French in Brussels and Pennsylvania Dutch in parts of Pennsylvania seem to exist for a very long time, even though their surrounding state/country speak completely different languages. Our simulation shows that there is a very small dynamic difference, affirming the model's accuracy. For small increases in enclave language prominence, we can determine that increasing the population of the region, increasing the area of the region (up until around $\frac{1}{2}$ of the region) will make the language in the enclave best suited for not going extinct. However, further research can be done. For example, experimenting with parameters of the payoff function (α, s_A, s_B), can result in more drastic changes in the true Y_A value in the simulation that may be more applicable to real-life language models. Moreover, a higher α will result in a more drastic decrease and increase in the dynamics, and increasing s_A will place a higher payoff for language A. This is also where social factors may come into play, and where Wu's model may be applicable.

7 Future Extensions

We are laying the theoretical groundwork of three languages evolution with societal influences instead of two, with a particular emphasis on the dynamic simulation component. In order to understand the model that the simulations are based off of, we need to define the payoff function for a general (x_A, x_B, x_C)-speaker in a (y_A, y_B, y_C)-population, which represents the average use of languages A and B and C by the speaker:

$$\pi((x_A, x_B, x_C), (y_A, y_B, y_C)) = x_A P_A(y_A) + x_B P_B(y_B) + x_C P_C(y_C)$$

In our project, we focused on the theoretical framework of trilingual evolutionary dynamics with societal interventions, aiming to describe the dynamic behaviors of multilingual populations using a combination of the following two models:

$$\begin{cases} \dot{y}_A = y_A y_B(\tilde{P}_A(y_A) - \tilde{P}_B(y_B) + y_A y_C \tilde{P}_A(y_A)) - \tilde{P}_C(y_C)) \\ \dot{y}_B = y_B y_A(\tilde{P}_B(y_B) - \tilde{P}_A(y_A) + y_B y_C \tilde{P}_B(y_B)) - \tilde{P}_C(y_C)) \\ \dot{y}_C = y_C y_A(\tilde{P}_C(y_C) - \tilde{P}_A(y_A) + y_C y_B \tilde{P}_C(y_C)) - \tilde{P}_B(y_B)) \end{cases} . \tag{1}$$

$$\tilde{\pi}((x_A^*, x_B^*, x_C^*), (x_A^*, x_B^*, x_C^*)) \geq \tilde{\pi}((x_A, x_B, x_C), (x_A^*, x_B^*, x_C^*)) \tag{2}$$

Here, we consider an evolutionary game, where every individual tries to maximize their payoff. (x_A, x_B, x_C) represents the strategy of an individual and (y_A, y_B, y_C) represents the strategy of the overall population. There is an optimal strategy for every individual, defined as (x_A^*, x_B^*, x_C^*), which becomes the strategy for the whole population when the Nash equilibrium is reached. $\tilde{P}$ represents the payoff functions when taking into account societal influences and interventions.

A real-world example of a societal intervention is that in Switzerland, the official language taught in school is German whereas the language spoken by local people is Swiss German, a dialect of German.

Equation (2) represents the dynamical behaviors of the trilingual population when giving equal weight to competition and intervention. Equation (3) represents the payoff function of the game. Using a combination of these two models, we will create dynamic solutions of multilingual populations, and investigate how the dynamical behavior changes as we modify certain parameters, such as the number of languages spoken in the population, the degree to which languages are "well-mixed", and the respective weights of competition and intervention in the model.

Some other extensions we have considered include:

- Combining Wu's societal influences factor into our language enclaves model and observe the evolutionary dynamics
- Incorporate carrying capacities into the system
- Code more graphics to visualize the change over time and see the weights and the nodes
- Explore other changes in parameters to our simulation and further simulations
- Adjust the model for determining probability of interaction
- Analyzing real-world language groups to see how their evolutionary dynamics compare to our more flexible theoretical model

8 References

Z;, Wu. “Why Multilingual, and How to Keep It-an Evolutionary Dynamics Perspective.” PloS One, U.S. National Library of Medicine,

https://pubmed.ncbi.nlm.nih.gov/33171482/.

Abrams, Daniel M., and Steven H. Strogatz. “Modelling the Dynamics of Language Death.” Nature News, Nature Publishing Group,

https://www.nature.com/articles/424900a.

Castello, Xavier. “Ordering Dynamics with Two Non-Excluding Options: Bilingualism in Language Competition.” Institute of Physics,

https://iopscience.iop.org/article/10.1088/1367- 2630/8/12/308.

Verma, Shresth. “Emergence of Multilingualism in Population Based Referential Games.” Semantic Scholar, 1 Jan. 1970,

https://www.semanticscholar.org/paper/Emergence- of-Multilingualism-in-Population-based-Verma/5af5c07fc348d5b7ca4f6652e18fbf5dc3c9485b.

Hauser, Marc D., et al. “The Mystery of Language Evolution.” Frontiers, Frontiers, 16 Apr. 2014,

https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00401/full.

Lieberman, Erez, et al. “Quantifying the Evolutionary Dynamics of Language.” Nature News, Nature Publishing Group, 11 Oct. 2007,

https://www.nature.com/articles/nature06137.

Nowak, Martin A., et al. “Computational and Evolutionary Aspects of Language.” Nature, U.S. National Library of Medicine, 6 June 2002,

https://pubmed.ncbi.nlm.nih.gov/12050656/.

Allen, Benjamin, et al. “Evolutionary Dynamics on Any Population Structure.” Nature News, Nature Publishing Group, 29 Mar. 2017,

https://www.nature.com/articles/nature21723.

Poot, Jacques, et al. “The Gravity Model of Migration: The Successful Comeback of an Ageing Superstar in Regional Science” IZA World of Labor, Oct. 2016, https://ftp.iza.org/dp10329.pdf.

Nowak, Martin A. “Evolutionary Biology of Language.” JSTOR, The Royal Society, 29 Nov. 2000,

https://www.jstor.org/stable/3066890.

Valverde, Sergi, and Ricard V. Sole. “Punctuated Equilibrium in the Large-Scale Evolution of Programming Languages.” The Royale Society Publishing, 6 June 2015,

https://royalsocietypublishing.org/doi/10.1098/rsif.2015.0249.

Nowak, Martin A. “Evolutionary Dynamics: Exploring the Equations of Life.” JSTOR, Harvard University Press, 2006,

https://www.jstor.org/stable/j.ctvjghw98.

Auer, Peter; Schmidt, Jürgen Erich. ”An International Handbook of Linguistic Variation” Section ”The history of language island research (Sprachinselforschung)”, Volume 1, 2009.