Teaching an old dog new tricks? Learning rates, aging, and language change

Ellis Cain

Abstract

Language change can be understood as representing the dynamic relationship between group-level and individual usage patterns. According to statistical learning, language learners are sensitive to the underlying distributional properties of language usage. Therefore, when usage patterns vary across time, it would be expected that the language acquisition of individuals will be impacted by the usage patterns that they experience during their lifetime. While previous research has shown differences in the language representation and cognitive capabilities over the lifespan (i.e., early vocab spurt, late cognitive decline), the impact of varying learning rates on language change is not well understood. Using an agent-based model of (syntactic) language change, this project aims to explore the interaction between learning rates and aging may impact diachronic changes in usage patterns of language. **Main finding: impact on time to equilibrium**

Introduction

Language change

Language can be viewed as a complex adaptive system, where it is a system of speakers in a speech community, whose behavior is based on past interactions, which then influence future behaviors ("Five Graces Group" et al., 2009). Namely, individual usage patterns form the collective language usage patterns, then future individuals learn from these collective usage patterns. Figure 1 below shows an example of this feedback loop between collective usage patterns and individual language usage.

It has been demonstrated before that language changes over historical time in terms of their collective usage statistics across users (Bybee, 2015; Michel et al., 2011). Yet, little is known about the relationship between changes in collective usage patterns and the meaning representations of individual language users.

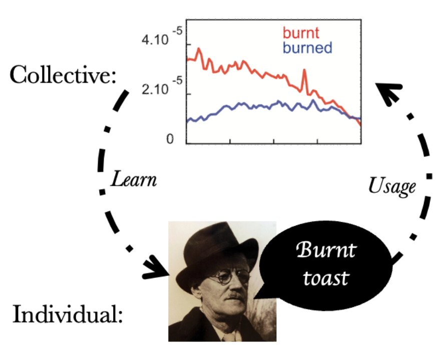


Figure 1: Example of the feedback loop in the complex adaptive system.

Language learning

- Ryskin verb bias
- Brysbaert
- Castro paper

Aging and language

The organization of lexico-semantic representations across the lifespan has been studied using word association data (De Deyne et al., 2019). Dubossarsky et al. (2017) analyzed lexical association networks and found that their organization changes over time. Namely, they observed a U-shaped developmental trajectory: networks are small and sparse (e.g., reduction of average shortest path length and entropy) during language acquisition (10–18 y.o.), they are dense and well-connected in mid-life (e.g., increased in-degree and out-degree), and then they become sparse again, with a larger proportion of isolated, peripheral nodes, in late-life (e.g., reduced in/out degree, increased entropy).

In addition to age-related differences in lexico-semantic organization, previous studies have used electroencephalography (EEG) to explore the relationship between lexico-semantic processing and aging (Federmeier et al., 2010; Wlotko et al., 2012). Some signatures of predictive processing (e.g., increased anterior positivity in response to a word that is less closely related to a preceding cue) are reduced in older adults, suggesting that aging may be related to changes in neural mechanisms of prediction and/or in the predictability relationships among words and meanings.

Model

Overview

The original language change model from Troutman et al. (2008) models the usage and spread of a grammatical variant throughout a population, such as a past tense ending ("-t" or "ed", as in "burnt" or "burned"). I modified the model such that each language user in the community has a learning rate that can be impacted by their age, which allows for an exploration of how changes in learning rates across the lifespan impact the overall collective-level patterns of language change.

The model's assumptions are as follows, modified from Beeksma et al. (2017) and Troutman et al. (2008).

1. Language learning is based on imitating others.
2. Learning rates may change over the lifespan.
3. There are variations between individuals.
4. Language can be influenced by external factors.
5. Language change has multiple stable equilibrium (*no set outcome*).

Here, equilibrium refers to when the community ceases to experience change in individuals' preference for one grammar or the other. For example, one equilibrium could be all language users only using Grammar 1, while another equilibrium could consist of half of the users using Grammar 1 and the other half using Grammar 0, which would correspond to a community with multiple dialect sub-groups.

With this model, the main outcome measures will be:

1. Final state of the system: *mean state of nodes in the population.*
2. Time to equilibrium: *the first time step that the group reaches that mean state.*

Model description

The model initializes a preferential attachment network, where each node in the network represents a language user. There are two grammar variants, which are coded as 0 or 1, and are distributed throughout the population based on a specified percentage. Figure 2 shows an example of the preferential attachment network.

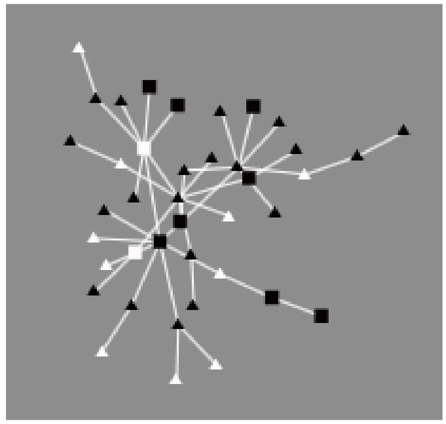


Figure 2: Example of the type of network generated in the model.

Nodes have the following properties; state, spoken state, age, cohort, and gamma. *State* represents the node's current grammar state, ranging from [0,1], which influences their likelihood to use one of the two grammar variants. It can be understood as the bias towards a specific grammar (e.g., node with state of 0.25 would be biased towards Grammar 0 when speaking). *Spoken state* represents the output of each node's speech, as (grammar) 0 or 1.

The node's *age* should be self-explanatory, and it increases by 1 each tick. During initialization, the age of the nodes can be distributed according to a uniform or normal distribution. Likewise, there are two possible *cohorts*, which are initialized at the start. The cohort ages can be specified, such that Cohort 1 starts at *age = 0* and Cohort 2 starts at *age = 50*. Moreover, the grammars can be distributed at different percentages for each cohort, such that Cohort 1 starts with 25% using Grammar 1 and Cohort 2 starts with 50% using Grammar 1. Lastly, *anti-preference* can be specified, such that for each "communication", a given node has a chance to ignore out-group members.

Gamma is the learning rate of a given node. Higher gamma values (0.08) mean that the node will adjust their state more when listening, as opposed to a lower gamma (0.01). It can be initialized according to a uniform or normal distribution, or it can be set to decrease with age.

Model dynamics

There are three main steps in the dynamics of the model: speaking, listening, and aging.

Regarding *speaking*, each node asynchronously generates an 'utterance', which consists of a discrete output (either 0 or 1, think of 'burnt' or 'burned'). This is based on a logistic curve that relates a node's internal state (weight of using Grammar 1) to the probability of using Grammar 1.

Regarding *listening*, all of the connected neighboring nodes will listen and use a linear reward-penalty algorithm to update their state (Bush & Mosteller, 1951, 1955; Yang et al., 2002). For this, they select a grammar that will be used to interpret the utterance (0 or 1). If the listener's selected grammar matches the heard utterance, they successfully interpret the utterance and update their internal state towards the heard state, otherwise, it will be updated away from the heard state. Here, the gamma parameter modifies the step size that the listener node takes when they update their internal state. As previously mentioned, if *anti-preference* is specified, there may be a chance that a given node does not update their internal state if the speaker is a member of the other cohort.

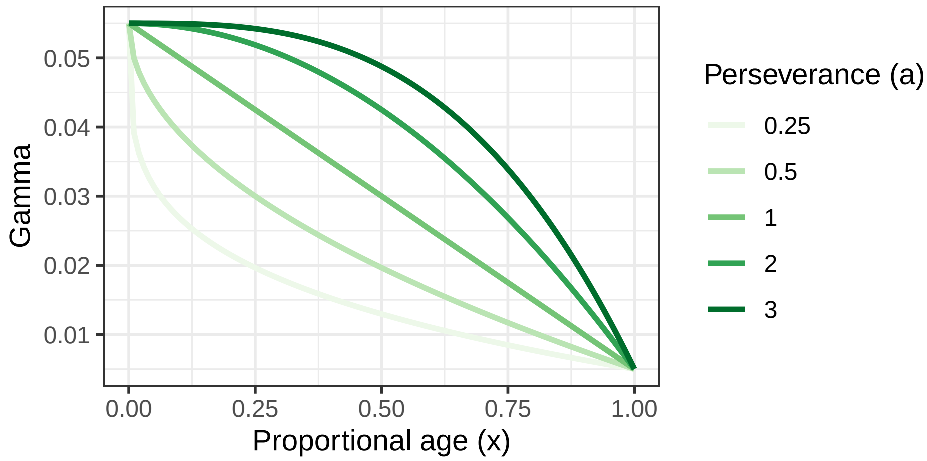


Figure 3: Graph of learning curve, which describes how learning rate may change with age. Colors represent different perseverance values.

Regarding *aging*, every node ages with each tick. If gamma is set to decrease with age, it decreases based on a "learning curve" function. This "learning curve" is defined using a power law relationship between *age* and *gamma* (learning rate), and is as follows:

$y = -0.05(x^a) + 0.005$

The variable x is the normalized age of a given node, such that it is within [0,1]. The variable a represents the "perseverance" of learning, such that higher perseverance values ($a = 3$) indicate slow initial change that speeds up closer to x = 1, and lower perseverance values ($a = 0.25$) indicate a quick initial change that slows down later on. Graphs for the (decreasing) learning curves can be seen in Fig. 3.

Results

Each model configuration was run 50 times, with 1000 steps total per run. The networks were generated with 40 nodes, based on the original Troutman et al. paper. Using this agent-based model, I looked at the impact of learning rate and learning perseverance, aging and cohort age difference, and cohort anti-preference. Again, 'equilibrium' will refer to the stable state that the system reaches.

Impact of learning rate

Figure 4 shows the impact of learning rate on the final state of the networks. When the initial percentage of Grammar 1 is low ($<$ 35% of nodes), it rarely spreads through the population, as almost all of the runs end with grammar 0 as the majority. On the flip side, when Grammar 1 is present in the majority of agents ($>$ 60% of nodes), it will spread through the network the majority of the time. However, when it is present in half of the nodes, dialect subgroups will emerge (Fig. 5), where different areas of the network will settle on different grammars.

For the variable that we manipulated between runs, gamma, there is no systematic impact on the final state that the system reaches. Regardless of how quickly nodes learn, the final result is mainly determined by the initial percentage of Grammar 1. However, when we look at the individual runs, we see a marked difference as gamma is manipulated. Figure 6 plots the mean state of the system over time for individual runs, based on various configurations of gamma and initial percent of Grammar 1. Namely, we see that while

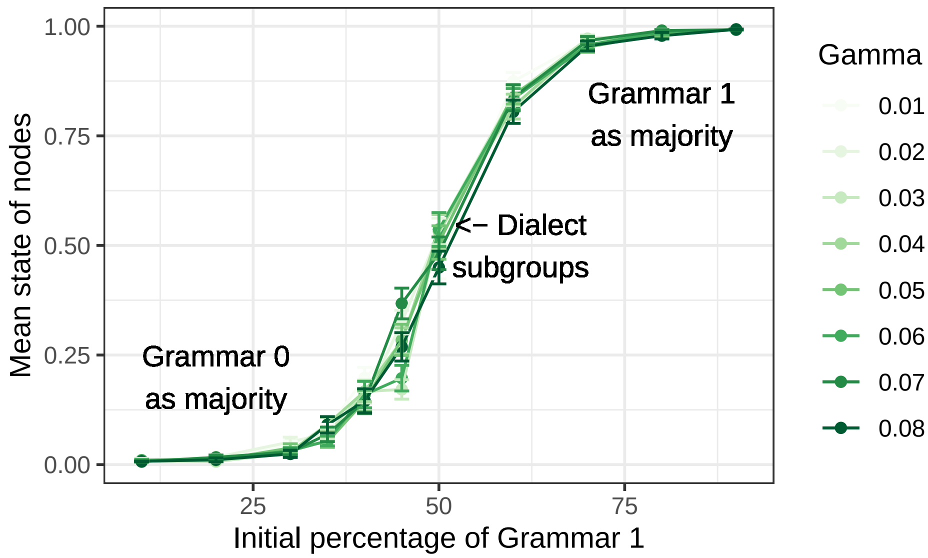


Figure 4: Final state of the langauge community, as measured by the mean state of nodes. Color of lines represents the gamma, or learning rate, of the nodes. Gamma is constant throughout the lifespan. Bars represent standard error.

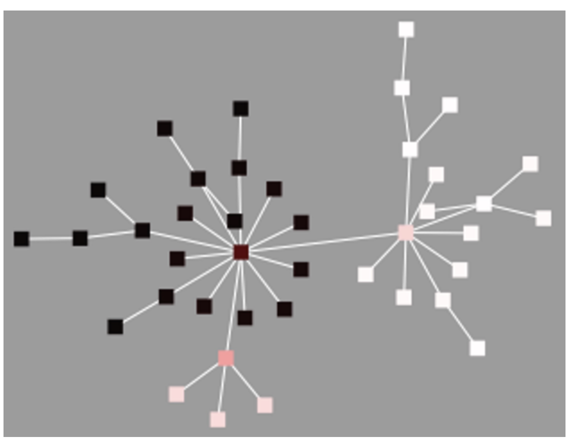


Figure 5: Dialect sub-groups.

the equilibrium reached by the system does not differ between different gamma values, the speed at which it reaches that equilibrium decreases as gamma increases.

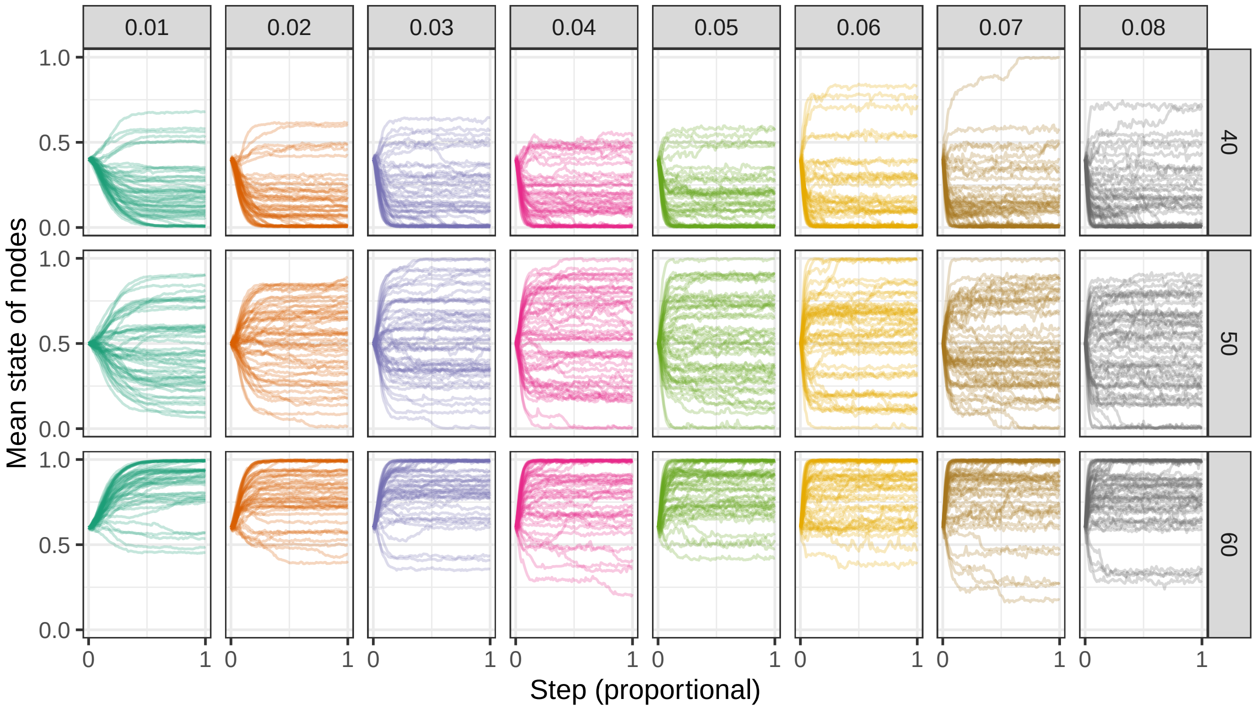


Figure 6: Each line represents the average state of the system for individual model runs. Here, the equilibria can be seen when the mean state flattens out. As gamma increases, the time to the equilibrium decreases. Rows faceted by initial percentage of Grammar 1, columns by the gamma value.

Figure 7 shows the proportional time to equilibrium for different values of gamma. The initial percentage of Grammar 1 does not impact the time to equilibrium. When the learning rate is the slowest (gamma = 0.01), the system will reach equilibrium at a little less than 400 steps (0.4), while the fastest learning rate will cause the system to reach equilibrium in about 100 steps. As gamma increases, the decrease in time to equilibrium also decreases as well, seemingly approaching a limit of information transfer through the system.

In summary, gamma (learning rate) impacts the time to equilibrium of the system, and this impact decreases as gamma increases.

Learning perseverance

For these next simulations, gamma will be determined by age, and the perseverance of learning (rate of decrease) will be manipulated. Refer to Fig. 3 for the relationship between age and gamma, where decreasing perseverance increases the speed at which learning rate decreases with age. Figure 8 shows the impact of learning perseverance on the mean state of the nodes at the end of the simulation. Similar to before, there does not seem to be a systematic relationship between perseverance and the final state.

However, when we plot the relationship between perseverance and time to equilibrium (Fig. 9), we see that most of the learning rate curves ($0.5 \leq$ perseverance ≤ 3), the time to equilibrium is very similar. Only when the learning rate quickly decreases (perseverance = 0.25) does the time to equilibrium increase. The range of time to equilibrium for these persevernace variations are similar in range to when gamma is varied.

In summary, we see little difference of perseverance's impact on time to equilibrium when perseverance is greater than 0.5.

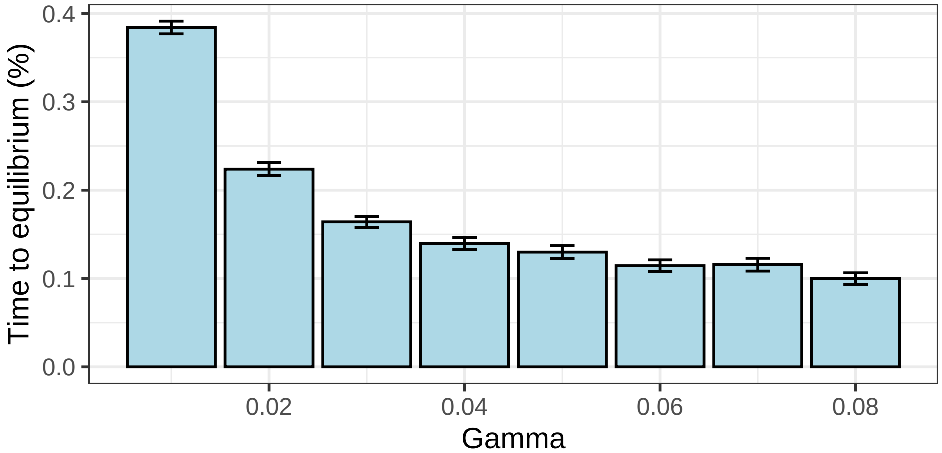


Figure 7: Time to equilibrium across the different values of gamma. Bars represent standard error.

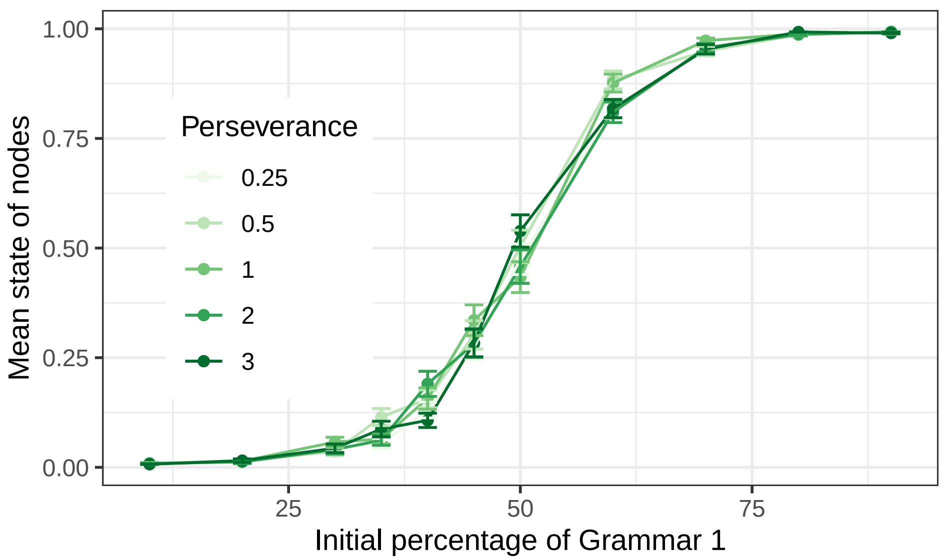


Figure 8: Impact of learning perseverance on the final outcome of the system.

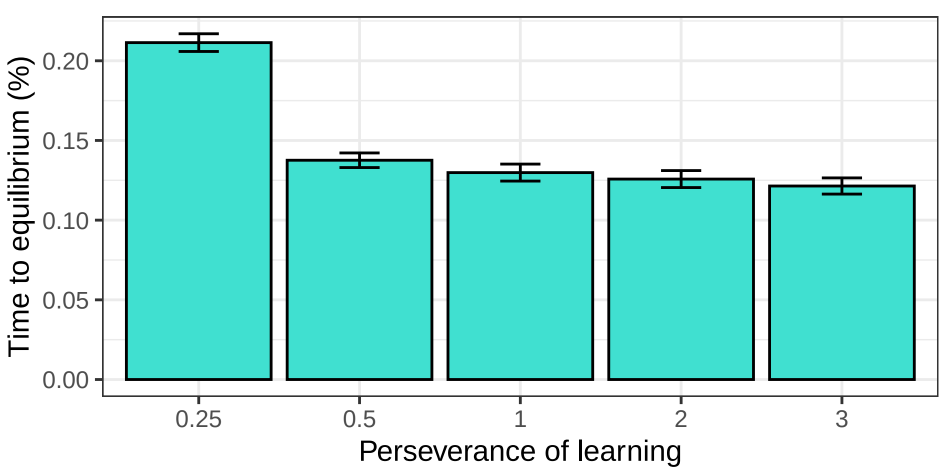


Figure 9: Time to equilibrium across the different values of perseverance, with standard error bars.

Aging and cohort equilibria

For this next set of comparisons, gamma will be set to decrease with age, with perseverance set at 1. In the simulations, there are two age cohorts, which the age gap between is varied. Similar to before, the initial starting percentage of Grammar 1 will be varied as well.

Figure 10 shows the time to equilibrium for different combinations of the initial percentage of Grammar 1 and population composition (percent of nodes in cohort 2). Cohort 1 started at age 0, so the different facets represent the age gap between the two cohorts. First, as the age difference between cohorts increases, the simulations where Grammar 1 was used initially by half the population increasingly takes longer on average to reach equilibrium. This effect generally seems to be more pronounced for the simulations where the majority of nodes were cohort 2. Then, when the initial percentage is at an extreme (10% or 90%), the time to equilibrium increases as well and has a larger impact on simulations with majority cohort 2.

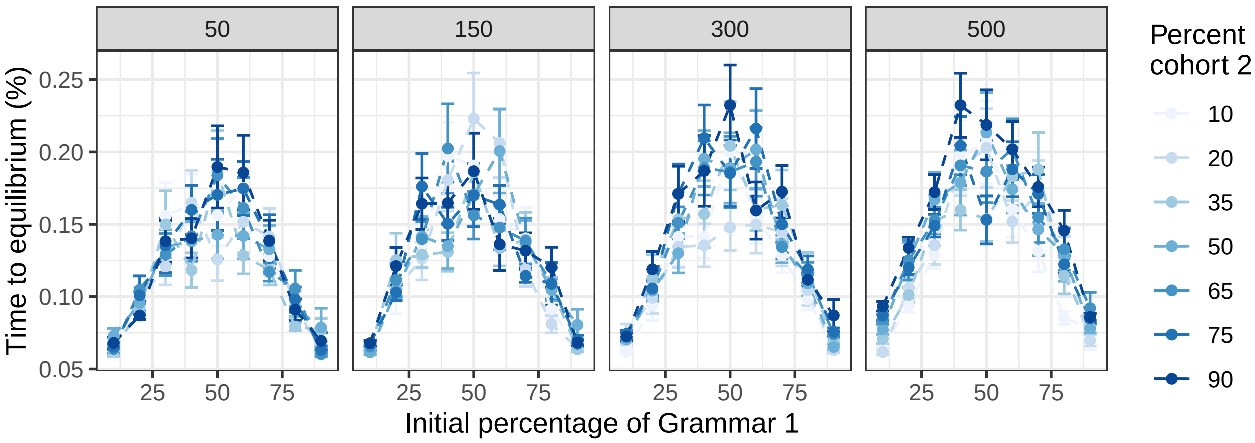


Figure 10: Time to equilibrium as a function of initial percentage of Grammar 1, across different population compositions. Faceted by cohort 2 age, which is the age gap between the two cohorts.

Figure 11 shows the separate equilibrium for each cohort, based on the initial percentage of cohort 2. The time to equilibrium seems to be generally consistent across the different variations, regardless of the age gap. Cohort 2, however, takes longer to reach equilibrium as they become more numerous in the population. Then, as the age gap increases, this impact on time to equilibrium increases as well.

In summary, the older cohort slows down the time to equilibrium, and as the age gap increases, the impact on time to equilibrium increases as well.

Cohort anti-preference

Brief explanation:

What is kept constant, what varies:

Description of plot one:

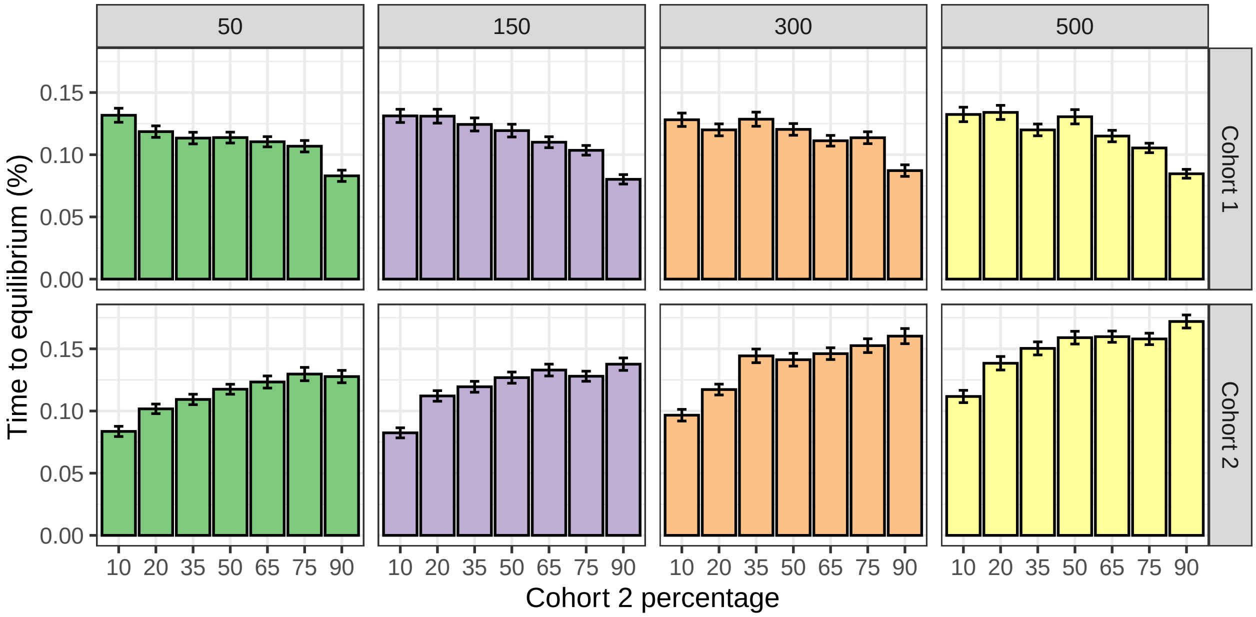
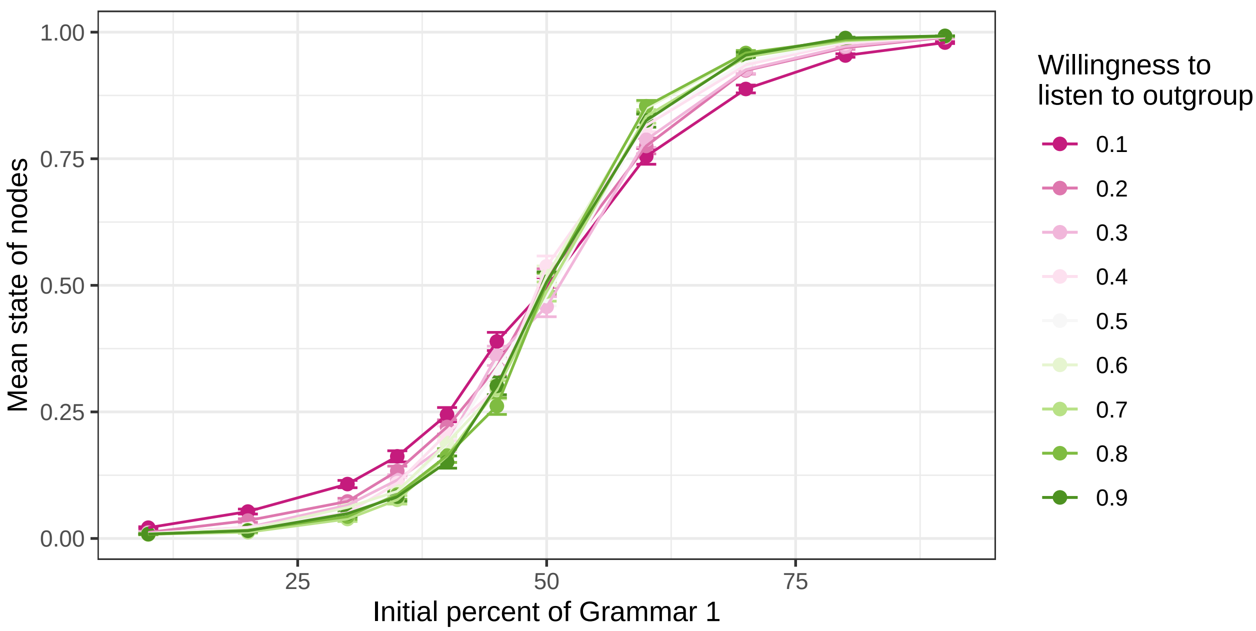
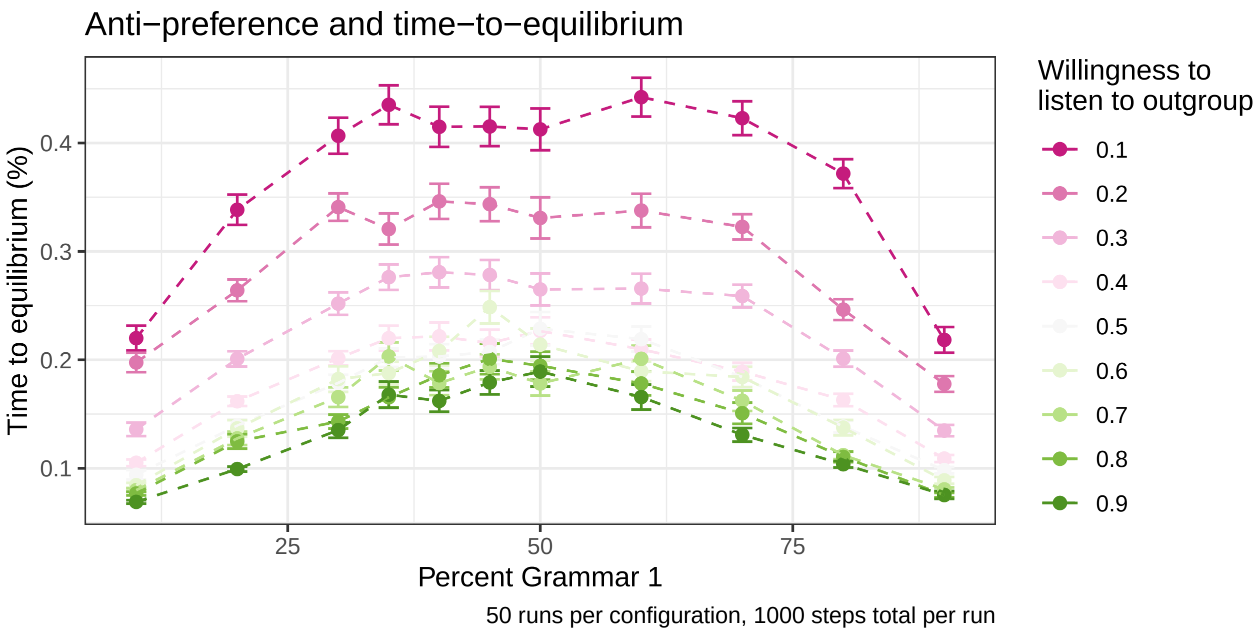


Figure 11: Time to equilibrium for the different age cohorts, faceted by cohort 2 age.



Description of plot two:



In summary, cohort anti-preference will delay the equilibrium, and slightly impact final equilibrium.

Discussion

Summarize the results, in the context of previous literature:

- Learning rate impacts time to equilibrium, but not the final system outcome
- Except for extreme decay (0.25), majority of learning perseverance curves have similar time to equilibrium
- Older cohort delays group equilibrium
- Cohort anti-preference increases the threshold for full spread

Language change

- Language change is gradual (Cain Ryskin 2023)
- Language change in this model changes very quickly

Language learning over the lifespan

- Verb bias stuff?

Previous models

- Consisted of a uniform population, built in a bias towards a specific grammar
- Variation among the population, no bias towards a specific grammar
- Opinion dynamics models?

Extensions

Network structure. Implement a small world network structure, where the parameters can be used to vary between fully random and a small world network. I suspect that the emergence of dialect sub-groups (fig. n) is an artifact of the preferential attachment structure, such that you have one well connected node with some other off-shoots. Based on this, there may be a section of the network that is only connected by one node, which will by structure block the grammar from spreading through to the distant group. However, if dialect sub-groups still appear when using a small world network, this would be interesting, though unlikely.

Cohort-based network structure. In addition to a small world network, potentially implement a connection preference, such that nodes will be more connected to other nodes in the same cohort. Something similar to a stochastic block model?

Generation modifications and node death, potentially integrating different types of learning (vertical, oblique, horizontal) since the current model just has one type. The anti-preference parameter was aimed at differentiating horizontal and vertical, such that nodes would prefer horizontal to vertical learning.

Currently, the learning algorithm is a linear reward-penalty algorithm (citation), which was one of the defaults from the original model and was tied to the speaking algorithm. One of the implications is that when a listener node fails to comprehend an utterance (the grammar it chooses to interpret does not match the utterance), it is punished and the internal state is updated away from the heard grammar. Future iterations could change this learning algorithm, such that failure to comprehend is not punished.

Conclusion.

Multiple possible equilibria: full spread of one grammar variant or a population with dialect subgroups.

Learning impacts time to equilibrium.

If learning rate is based on age, the age gap between groups will modulate the impact of having more older members of the speech community when it comes to language change.

Cohort anti-preference something.

References

Beeksma, M., Vos, H. de, Claassen, T., Dijkstra, A., & Kemenade, A. van. (2017). *A probabilistic agent-based simulation for community level language change in different scenarios.*

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, *58*(5), 313.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning.*

Bybee, J. (2015). *Language change.* Cambridge University Press.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006. https://doi.org/10.3758/s13428-018-1115-7

Dubossarsky, H., De Deyne, S., & Hills, T. T. (2017). Quantifying the structure of free association networks across the life span. *Developmental Psychology*, *53*(8), 1560–1570. https://doi.org/10.1037/dev0000347

Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and Language*, *115*(3), 149–161.

"Five Graces Group", Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., et al. (2009). Language is a complex adaptive system: Position paper. *Language Learning*, *59*, 1–26.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, *331*(6014), 176–182. https://doi.org/10.1126/science.1199644

Troutman, C., Clark, B., & Goldrick, M. (2008). Social networks and intraspeaker variation during periods of language change. *University of Pennsylvania Working Papers in Linguistics*, *14*(1), 25.
Wlotko, E. W., Federmeier, K. D., & Kutas, M. (2012). To predict or not to predict: Age-related differences in the use of sentential context. *Psychology and Aging*, *27*(4), 975–988. https://doi.org/10.1037/a0029206
Yang, C. D. et al. (2002). *Knowledge and learning in natural language*. Oxford University Press on Demand.