Agent-based Modeling of Language Change in a Small-world Network

Dalmo Buzato, Evandro L. T. P. Cunha
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brazil
{dalmobuzato, cunhae}@ufmg.br

Abstract

Language change has been the subject of numerous studies in linguistics. However, due to the dynamic and complex nature of this phenomenon, and to the difficulty of obtaining extensive real data of language in use, some of its aspects remain obscure. In recent years, nonetheless, research has used computational modeling to simulate features related to variation, change, propagation, and evolution of languages in speech communities, finding compelling results. In this article, agent-based modeling and simulation is used to study language change. Drawing on previous studies, a speech community was modeled using Zachary's karate club network, a well-established small-world network model in the field of complex systems. Idiolects were assigned through numerical values for each agent. The results demonstrate that the centrality of each agent in the network, interpreted as social prestige, appears to be a factor influencing change. Additionally, the nature of idiolects also seems to impact the spread of linguistic variants in the language change process. These findings complement the theoretical understanding of the language change phenomenon with new simulation data and provide new avenues for research.

Keywords: agent-based modeling, language change, small-world network, Zachary's karate club

1. Introduction

Since the seminal work of Weinreich et al. (1968), linguists have been increasingly interested in how external aspects of the linguistic system, such as time, space, social structure, and the contact between different cultures and societies worldwide, impact the grammatical system of a language. Building upon this study, language changes that can be observed in different locations or over time have come to be conceived as integral aspects of the linguistic system. Languages are now analyzed as heterogeneous and dynamic objects, in contrast to previous structuralist analyses that were mostly interested in languages as static and homogeneous entities.

From this perspective, one of the questions that emerge is how we accept language change, that is, how we as speakers adopt and propagate new linguistic forms that are in competition during the use of language. As we constantly switch between being speakers and listeners, and because language use itself is a two-way route, studying how a linguistic variant is accepted also implies studying how it propagates.

Some linguists (Osgood et al., 1954; Fagyal et al., 2010; Blythe and Croft, 2012) have proposed that the trajectory of a competing linguistic variant might follow an S-curve trajectory. A variant would initially be in the idiolect (the individual linguistic component of each agent in a speech community) of a few individuals, with the potential to grow based on its propagation and acceptance by other community members. If this form succeeds in the competition process and is continuously propagated, it will reach the top of the curve, being present in the idiolect of the majority of individuals in a given community. However, some linguistic forms never reach this status, either disappearing over time or remaining restricted to the idiolect of only a portion of the speech community.

The discussion about language propagation and change, and its relation to external factors, was considered, for example, by the neolinguistic school, particularly through the work of Italian linguist Matteo Bartoli. Bartoli (1945) posits five principles regarding the relationship between language change and socio-geographic space. Specifically, the second principle states that "if one of two linguistic forms is found in peripheral areas, and the other in a central area, then normally the linguistic form found in the peripheral zone is earlier". Bartoli employs this principle to explain changes in the Latin language considering the expansion of the Roman Empire throughout Afro-Eurasia[1]. This principle can also be applied to studying the propagation and transmission of linguistic items through topology and social structure, taking into account the concepts of centrality and periphery within a network.

To study linguistic phenomena of this nature, which reveal the complex nature and interdepen-

[1]To see a computer simulation of the Bartoli norms applied to the context of Romance languages, please refer to Buzato and Cunha (2024).

LREC-COLING 2024, pages 594–599
20-25 May, 2024.

dence of numerous components in the emergence of a community's language, some linguists, especially those in the field of language dynamics (Abrams and Strogatz, 2003; Wichmann, 2008; Loreto et al., 2011), have relied on contributions from computational modeling and complex systems (Beckner et al., 2009; de Oliveira, 2018). This provides an interdisciplinary research program involving physics, computer science, network science, and sociology to understand the dynamics of language use[2].

Among the methodologies used, agent-based modeling (ABM) stands out. According to Wilensky and Rand (2015), "agent-based modeling is a computational modeling paradigm that allows us to describe how any agent will behave. The methodology of ABM encodes the behavior of individual agents in simple rules so that we can observe the results of these agents' interactions". Linguistic studies have used ABM to investigate the emergence, evolution, and change of grammatical systems, as well as the choice, change, coexistence, or maintenance of languages in the context of language contact or competition, e.g. Harrison et al. (2002); Castelló et al. (2008); Troutman et al. (2008); Fagyal et al. (2010); Castelló et al. (2013); Civico (2019); Dekker and De Boer (2020); Louf et al. (2021); Charalambous et al. (2023); Rosillo-Rodes et al. (2023); Buzato and Cunha (2024).

2. Methodology

In this study, we modeled a speech community using a small-world network known as the karate club network, which was originally described by Zachary (1977). We chose to use a small-world graph instead of assuming that populations are fully connected because this is not a realistic assumption, as pointed out by Castelló et al. (2013). The choice of Zachary's karate club network is justified by its widespread use in complexity studies, particularly in complex networks and social communities studies. The network consists of 34 members of a karate club as nodes and 78 edges representing friendships between them, as observed over two years.

The assumption that social structure can be modeled as a small-world network is supported by several important studies, such as Granovetter (1973) and Watts and Strogatz (1998). A key feature of this type of graph is the ability of "strangers" to be indirectly connected by a short chain of agents (Barabási, 2002). This feature allows the spread and transmission of innovations and items across the group, as well as the formation of conventions and clusters.

From Figure 1, which displays graphically the karate club network, we can conceive each graph node as representing an individual in a speech community, and each edge of the graph as representing social interaction between two individuals. Each individual in the network has an idiolect, which consists of numerically valued items generated randomly following a normal Gaussian distribution. To computationally model the phenomenon, we utilized the Python programming language, specifically the NetworkX (Hagberg et al., 2008), pandas, and NumPy packages.

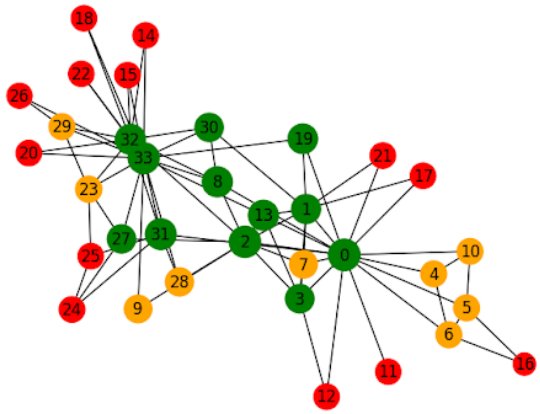

Figure 1: Zachary's karate club network. Nodes represent individuals in the speech community, while edges represent social interaction between them

Since the idiolects are generated from the same function with a single mean and standard deviation parameter, they contain numerically close items. Furthermore, some items are common across the idiolects of multiple agents, while others are present only in the idiolects of one or a few individuals, ensuring diversity of content in individuals' idiolects. This provides greater ecological validity to our model, as linguistic items can exhibit similarity and proximity within agents' idiolects (e.g., lexical proximity of words like *happy*, *unhappy*, *happiness*, *unhappiness*). Additionally, in a society, not all individuals have the same items in their idiolects; certain items are restricted to the idiolects of a few agents or specific groups, while other items are widely shared among individuals within a speech community (e.g., consider the number of individuals in a speech community who have words like *happy* and *strait* in their idiolects).

Each individual in the network also possesses a value representing their social prestige, which is correlated with their centrality in the graph. This

[2]Other approaches, including the analysis of the propagation of linguistic items in online social media (e.g. Cunha et al., 2011), have also been proposed.

value ranges from 0 to 1, with 0 indicating an individual with minimal prestige and 1 representing the maximum possible social prestige. In NetworkX, there are several mathematically distinct ways to calculate node centrality in a network. In this study, we opted to use the *closeness centrality* to generate the prestige values for each individual. For an overview of the different types of centrality measures in complex networks, including quantitative differences within Zachary's karate club network, please refer to: Petrov et al. (2015); Golbeck (2013); Kudělka et al. (2015); Batool and Niazi (2014).

In each round of simulation, when two individuals are connected, they have the opportunity to exchange linguistic items. The initial version of our algorithm can be described as follows. First, a random item is selected from the idiolects of the connected individuals. Then, the prestige of the agents is compared. The speaker with higher prestige transfers the selected item from their idiolect to the idiolect of the speaker with lower prestige. However, if the item already exists in the idiolect of the lower-prestige speaker, no action is taken. The above explanation is exemplified through Figure 2.

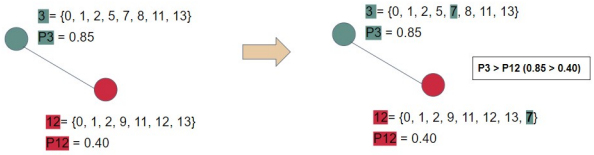

Figure 2: Algorithm version 1. Items in brackets belong to the idiolect of the corresponding agent. P3 and P12 indicate, respectively, the social prestige of agents 3 and 12.

This first version of the algorithm has some limitations, particularly regarding the ecological validity of the data. Communication and social interaction through language is a two-way route, and the initial version of our algorithm does not account for this, as only the speaker with higher prestige has the chance to transmit and propagate their items – thus, only these items are considered in the interactional situation. In the real world, a lower-prestige speaker can also pass items from their idiolect to the idiolect of a higher-prestige speaker, albeit with a lower probability.

Our proposed solution to address these issues is through a probabilistic approach. We introduce, into the simulation process, a probabilistic variable, which is a randomly generated value between 0 and 1 (i.e., within the same range as the centrality degree values). Figure 3 and Figure 4 illustrate the process described above, this time taking into account the newly introduced probabilistic item. In this updated version, the prestige values are not directly compared to each other. Instead, they are compared with the randomly generated probabilistic item for each round: if the prestige is higher than the probabilistic item, then propagation occurs (again, only if the linguistic item did not exist already in the other agent's idiolect). This enables dual-route communication and allows the propagation of items from lower-prestige to higher-prestige speakers as well.

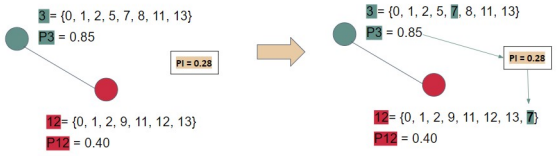

Figure 3: Algorithm version 2, transfer from agent 3 to 12. PI = probabilistic item

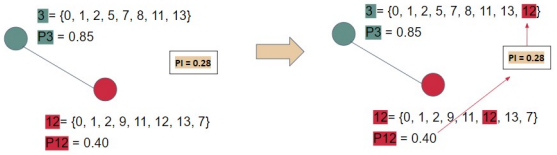

Figure 4: Algorithm version 2, transfer from agent 12 to 3. PI = probabilistic item

3. Results

For each simulation, our second algorithm (including the probabilistic item) was executed over 100 rounds, with each round representing a potential interaction between connected agents. Namely, there were 100 opportunities for mutual exchanges between each pair of agents. To ensure robustness in the quantitative analysis, 1,000 simulations were performed, allowing for reliable observations and measurements. Thus, in total, 100,000 rounds were computed and analyzed. Each agent's performance was evaluated based on three key factors: 1) the number of donated items, indicating their contribution to the propagation of linguistic elements; 2) the number of items received, reflecting their engagement in acquiring linguistic elements from other agents; and 3) their prestige level, serving as a measure of social influence within the network. These parameters provide valuable insights into the dynamics of linguistic propagation and exchange of items within the simulated community.

Figure 5 shows the results for speech communities in which idiolects deviate from each other by a standard deviation of 2.5. Firstly, there is a notable

high correlation (Pearson correlation coefficient) between the quantity of items donated and the centrality of the agent. This suggests that agents with higher centrality tend to contribute a larger number of linguistic elements to the community. Secondly, we find a low correlation, and in some cases, a negative correlation between the number of items received and the centrality of the agent. This indicates that agents with higher centrality may not necessarily receive a higher quantity of linguistic items from others in the community.

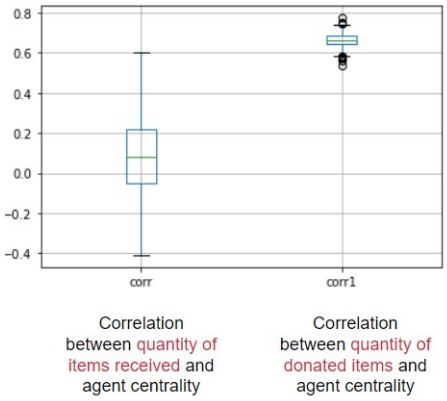

Figure 5: Correlations regarding idiolets with an sd=2.5

In speech communities where idiolects exhibit significant divergence from each other, characterized by a standard deviation of 5 (Figure 6), there is again a notable high correlation between the number of items donated by agents and their centrality within the community. This suggests again that agents with higher centrality tend to contribute a larger number of linguistic items to the overall language dynamics. Also, a medium correlation emerges between the number of items received by agents and their centrality. This implies now that agents with higher centrality are more likely to receive a greater number of linguistic items from others within the community.

Based on the results presented above, we can observe that besides the prestige variable (node centrality in the network), another factor appears to significantly influence the outcomes: the difference between individuals' idiolects, produced by the variation in standard deviation within the function that generates each individual's idiolect. It is possible to speculate that a greater divergence between the idiolects of high prestige (acrolect) and low prestige (basilect) speakers would impact the propagation of linguistic items. More central agents are more likely to donate and receive a higher quantity of items in their idiolects, while more peripheral agents exhibit lower levels of item exchange, resulting in conservative clusters with minimal changes to their idiolect

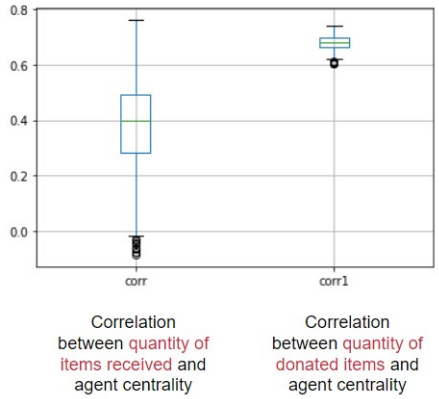

Figure 6: Correlation regarding idiolets with an sd=5

or limited transmission of their items to the rest of the network. Consequently, these clusters maintain linguistic forms in a stable state with minimal alterations to their individual language.

Considering that more central agents donate and receive a higher number of items compared to more peripheral agents, and extrapolating the results of the models to reflect the functioning of language in concrete usage, we can hypothesize that, after a certain period of time, peripheral agents would become isolated and likely carry the oldest, most stable linguistic elements of the language. In other words, they represent the more conservative version of the language, as Bartoli postulates (Bartoli, 1945). Furthermore, when we increase the standard deviation (sd) to 10 in the function that generates idiolects, we observe a strong correlation between prestige and the quantity of both donated and received items.

It is pertinent to add that, since the conception of language extends beyond the grammatical limits of linguistic item similarity and mutual intelligibility, and encompasses a more political concept, we can computationally conceive idiolects with considerable variation among them and still understand these individuals as speakers of the same language. This is precisely the scenario that the latest result seems to demonstrate.

4. Conclusion

In this study, we employ agent-based modeling in a small-world network to simulate language change, specifically the propagation of linguistic items within a speech community. To achieve this, a probabilistic algorithm was created to facilitate item exchanges among individuals. It was observed that prestige, modeled as the degree of centrality of individuals in the graph, appears to influence the dynamics of linguistic item transmission. Addition-

ally, the degree of difference between individuals' idiolects also seems to significantly impact the process of language change. These results complement the theoretical understanding of the language change phenomenon with new simulation data.

The use of Zachary's karate club network on the simulations reported here allows for greater control and interpretability of results due to its small-world nature. Also, this a network built using real social data, thus representing a real small speech community. However, we acknowledge that this is a simple model of social networks, which may not capture the complexity of real-world speech communities. Thus, the findings may not be easily generalizable to larger or more diverse speech communities. For this reason, future studies could explore more complex network structures in order to enhance the realism and applicability of the findings.

Finally, it is important to mention that we consider that our model, in the current format, is "feature agnostic", meaning that any linguistic feature (be it phonological, lexical or syntactical) could be considered. Future work could use the model presented here to specifically investigate the role of different linguistic features in the simulations, using additional linguistic information. Hence, we understand our study as a door-opener for additional research using the proposed model.

5. Bibliographical References

Daniel M Abrams and Steven H Strogatz. 2003. Modelling the dynamics of language death. *Nature*, 424(6951):900.

Albert-László Barabási. 2002. *Linked: the new science of networks*. Perseus, Cambridge.

Matteo Giulio Bartoli. 1945. *Saggi di linguistica spaziale*. Rosenberg & Sellier, Torino.

Komal Batool and Muaz A Niazi. 2014. Towards a methodology for validation of centrality measures in complex networks. *PLOS ONE*, 9(4):e90283.

Clay Beckner, Richard Blythe, Joan Bybee, Morten H Christiansen, William Croft, Nick C Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, and Tom Schoenemann. 2009. Language is a complex adaptive system: position paper. *Language Learning*, 59:1–26.

Richard A Blythe and William Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language*, 88(2):269–304.

Dalmo Buzato and Evandro L T P Cunha. 2024. Bartoli's areal norms revisited: an agent-based modeling approach. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese (PROPOR)*, pages 422–431.

Xavier Castelló, Víctor M Eguíluz, Maxi San Miguel, Lucía Loureiro-Porto, Riitta Toivonen, Jari Saramäki, and Kimmo Kaski. 2008. Modelling language competition: bilingualism and complex social networks. In Andrew D M Smith, Kenny Smith, and Ramon Ferrer i Cancho, editors, *The evolution of language*, pages 59–66. World Scientific.

Xavier Castelló, Lucía Loureiro-Porto, and Maxi San Miguel. 2013. Agent-based models of language competition. *International Journal of the Sociology of Language*, 2013(221):21–51.

Christos Charalambous, David Sanchez, and Raul Toral. 2023. Language dynamics within adaptive networks: an agent-based approach of nodes and links coevolution. *Frontiers in Complex Systems*, 1:1304448.

Marco Civico. 2019. The dynamics of language minorities: evidence from an agent-based model of language contact. *Journal of Artificial Societies and Social Simulation*, 22(4).

Evandro Cunha, Gabriel Magno, Giovanni Comarela, Virgilio Almeida, Marcos André Gonçalves, and Fabricio Benevenuto. 2011. Analyzing the dynamic evolution of hashtags on Twitter: a language-based approach. In *Proceedings of the Workshop on Language in Social Media (LSM)*, pages 58–65.

Marco Antônio de Oliveira. 2018. Origem, propagação e resolução da variação linguística na perspectiva da linguagem como um sistema adaptativo complexo. *Caletroscópio*, 6:11–36.

Peter Dekker and Bart De Boer. 2020. Neural agent-based models to study language contact using linguistic data. In *4th NeurIPS Workshop on Emergent Communication: Talking to Strangers: Zero-Shot Emergent Communication*.

Zsuzsanna Fagyal, Samarth Swarup, Anna María Escobar, Les Gasser, and Kiran Lakkaraju. 2010. Centers and peripheries: network roles in language change. *Lingua*, 120(8):2061–2079.

Jennifer Golbeck. 2013. Network structure and measures. In *Analyzing the Social Web*, pages 25–44. Morgan Kaufmann Boston.

Mark S Granovetter. 1973. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380.

Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. Exploring network structure, dynamics, and function using NetworkX. Technical report, Los Alamos National Lab (LANL), Los Alamos, NM (United States).

K David Harrison, Mark Dras, and Berk Kapicioglu. 2002. Agent-based modeling of the evolution of vowel harmony. In *North East Linguistics Society*, volume 32, pages 217–236.

Miloš Kudělka, Šárka Zehnalová, Zdeněk Horák, Pavel Krömer, and Václav Snášel. 2015. Local dependency in networks. *International Journal of Applied Mathematics and Computer Science*, 25(2):281–293.

Vittorio Loreto, Andrea Baronchelli, Animesh Mukherjee, Andrea Puglisi, and Francesca Tria. 2011. Statistical physics of language dynamics. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(04):P04006.

Thomas Louf, David Sánchez, and José J Ramasco. 2021. Capturing the diversity of multilingual societies. *Physical Review Research*, 3(4):043146.

Charles E Osgood, Thomas A Sebeok, John W Gardner, John B Carroll, Leonard D Newmark, Susan M Ervin, Sol Saporta, Joseph H Greenberg, Donald E Walker, James J Jenkins, Kellogg Wilson, and Floyd G Lounsbury. 1954. Psycholinguistics: a survey of theory and research problems. *The Journal of Abnormal and Social Psychology*, 49(4, Pt.2):i–203.

Dmitry Petrov, Yulia Dodonova, and Andrey Shestakov. 2015. Magolego SNA – Lab 4. Online. https://rpubs.com/shestakoff/sna_lab4.

Pablo Rosillo-Rodes, Maxi San Miguel, and David Sánchez. 2023. Modelling language ideologies for the dynamics of languages in contact. *Chaos*, 33(11):113117.

Celina Troutman, Brady Clark, and Matthew Goldrick. 2008. Social networks and intraspeaker variation during periods of language change. *University of Pennsylvania Working Papers in Linguistics*, 14(1):325–338.

Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440–442.

Uriel Weinreich, William Labov, and Marvin Herzog. 1968. *Empirical foundations for a theory of language change*. University of Texas Press, Austin.

Søren Wichmann. 2008. The emerging field of language dynamics. *Language and Linguistics Compass*, 2(3):442–455.

Uri Wilensky and William Rand. 2015. *An introduction to agent-based modeling: modeling natural, social, and engineered complex systems with NetLogo*. MIT Press, Cambridge.

Wayne W Zachary. 1977. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473.