Emergence of a language dialect shaped by the geographical terrain

Agent Technology Practical

Arsenijs Golicins , Mansur Nurmukhambetov ,
Teodora Stereciu , Franciszek Szewczyk

1 Introduction

In this section, we introduce the goal of our project and the associated topic that we investigated. Previous literature is linked to the problem, and the relevance of the agent-based modelling method is highlighted.

1.1 Topic

The topic that our project aimed to model is the evolution of language, in particular the convergence of two languages into one. It is remarkable that while belonging to the same biological species, humankind, over tens of centuries of its evolution, has developed means of communication so different from each other. It is no less intriguing how the inner processes of language evolution give rise to the formation of dialects, which, in the long run, can be split into distinct languages. One peculiar example is the Creole languages—languages that are formed on the basis of languages very different from each other. Our project goal was to investigate the emergence of this type of new language using agent-based modelling. We aimed to understand the conditions that allow for the language's emergence out of the convergence of the "parent" languages.

Language phenomena have been investigated in the realm of agent modelling before. Some earlier works include the modelling of an artificial Creole language evolution [1], and the influence of geographic terrain variations on language diversity [2]. In particular, authors were studying which factors allow for creating language imbalance, either by creating new languages or the domination of one language over others, in the case where there exists a language contact [3]. Such factors were identified as, among others, population size and lexical similarity. Some studies were dedicated to the phenomenon of a dialect continuum. As authors note, *dialect continuum* occurs when "languages gradually merge into each other over an extended geographical range" [4].

Agent-based modelling is the appropriate method for researching the stated problems, as languages that are necessary have speakers, which can be seen as the agents. The whole process of communication is the agent's interaction with others. However, what is interesting about the language diversity of speakers is that it can be hardly explained atomistically. Rather, the language becomes an entity itself. This is a revealing manifestation of an emergent phenomenon. Language diversity, dialectal continuum, and language borders are in fact the result of the behaviour of the system with the individual entities as its actors. Since language development is the emergent behaviour of such systems, agent-based modelling offers a framework to model such complex behaviour.

1.2 Innovation

The goal of this model is, just like the previous research, to get insights about the conditions that allowed for the development of language diversity in the real world. However, while the previous

research focused on sustaining language multiplicity and diversity, our project investigates the opposite phenomenon. We are curious about the conditions that allow two languages to eventually merge into one, and precisely because of this, we think that our model is interesting. This, naturally, can be seen as the case for the formation of a dialect in a *dialect continuum* [4]. However, we are not modelling a specific dialect continuum, but, rather, our new idea is investigating the particular, isolated case of the convergence that leads to a specific form of a dialect. Since this new language, which appears after the mixing of its parents, is a dialect in between these languages, in the report we will use the terms "dialect" and "language" interchangeably. The model is not trying to replicate some particular existing phenomena; rather, the validation is concerned with relational alignment, namely the influence of geographical and intergroup language complexity factors on language diversity.

Our research question can then be formulated as follows:

To what extent do geographic barriers and initial intratribal language complexity affect the convergence of languages spoken by agents from two neighbouring tribes?

To answer this question, we decided to look at the behaviour happening within a specific geographic terrain that is inhabited by two different tribes. We would introduce a barrier between these two populations and check which values of the barrier bandwidth and the language complexity within both of the tribes would result in faster convergence between the two languages. Our hypothesis is that a barrier that is easier to cross and a low language complexity of both languages would lead to faster convergence, as there would be fewer obstacles between intertribal communication.

The model's simplicity makes it a base-case scenario for the interaction of two languages. However, we acknowledge that assumptions were made that decreased the ecological validity of the model, and we will discuss this aspect further in the paper. Nevertheless, this model could provide insights into the broad dynamics that affect language interactions in the long run. The relevance of our model includes filling the gap in the emergence of Creole languages. Previous studies that focused on the evolution of Creole languages, such as Furman & Nitschke (2020), studied the effect of population size and lexical similarity [1]. Our study, however, expanding on the methods of Magee & Hendery (2015), explores this emergence in the context of geographical constraints and intra-tribal parameters. This constitutes the innovation of our approach.

2 Methods

This section describes the conceptual model, the implementation details, and the experimental setup of our project. We argue that this methodology reflects the proposed goal and the model is well-suited for it.

2.1 Conceptual model

Our model simulates a geographical environment initially populated with two groups (*tribes*) of agents speaking different languages. They are situated at opposite ends of the environment and are separated by the border. The border is supposed to signify a geographical constraint. Historically, the difference in the terrain defined the border areas of the habitat for the people. Hence, the border in our simulation has a certain bandwidth, which serves the purpose of the agent's movement limitation. This bandwidth is one of the operationalized parameters of the model. The environment the agents inhabit is hence a rectangular area, which is split in the middle. The border is not curved but a straight line, which is an assumption that the interaction is happening at a geographic interval where the curvature of the border does not matter.

Agents will occasionally cross this border and interact with agents from another language group. This will, in the long term, result in the washing away of the strict differential border between lan-

guages and the emergence and development of a dialect of the initial two languages.

Any given agent is an entity capable of observing its environment, namely the vicinity of other agents around it and their language of communication. It also manipulates its environment by walking and engaging in social interaction with its neighbours. The outcome of social interaction is dependent on the difference in language of communication. The more different the languages are, the more effect it would have on the communicators; namely, they would have to adapt and develop their own language, deviating from the norm they were speaking at first. Agents change their language by modifying their vocabulary. In particular, they choose one word to communicate about per interaction and change it accordingly for it to be mutually understandable for the next time of the interaction.

The second operationalized parameter is the language complexity within the initial tribe. Language complexity is to be understood as the *intra-tribal language diversity*, *vocabulary size* and the *words' length*. The first variable is measured by the standard deviation from the norm at the very start of the simulation. An increase in diversity theoretically leads to more complex patterns on the dialectical continuum. Vocabulary size is the number of words representing discrete concepts that the population of the world knows. Finally, the word length is the length of each word. Each letter has many possible values, which this letter may be changed to, which will be changed as agents try to communicate with others whose words for a given concept would be different.

What the model will measure is the time of convergence of the population over the long run. That is, the eventual variation of the whole population should start decreasing and reach the convergence value.

2.2 Implementation details

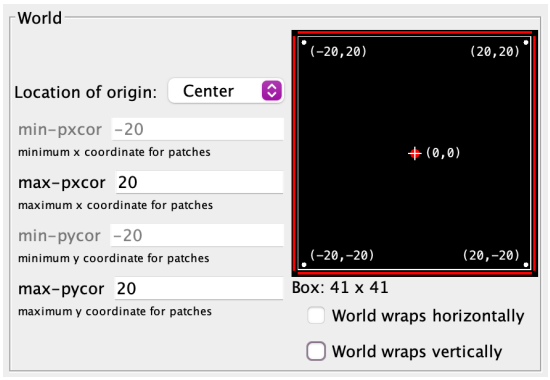
The model was implemented in NetLogo 6.4.0 from scratch. This section will cover how we handled the world and its agents, along with how we adjusted the interface.

2.2.1 Agents

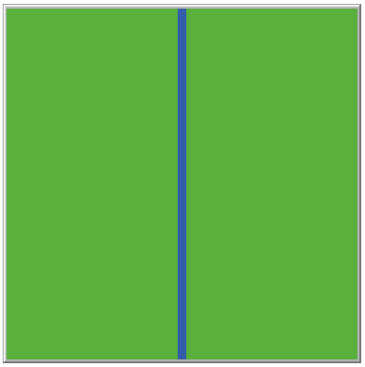
In our model, agents are 'occupied' patches. *Unoccupied* patches make up the world. Each patch keeps track of its status with an `occupied` attribute. Patches also have their own vocabulary `language` of size `number-of-words`, wherein words have length `word-length`. Those size parameters are shared across all patches, meaning all agents know the same number of words with the same number of letters. For agents, the `language` is a list of lists of floats. Patches are also given a `probability-of-entering`, which is 1 for patches with agents on them. The agents of our model are reflexive, acting upon relatively simple rules. Their behaviour will be described in detail later in this section. Overall, they move randomly and talk to a neighbour. They also adjust their vocabulary to be more similar to that of their conversation partner. We do not model any goals or preferences, since we are looking to see how sheer proximity affects language exchanges made more difficult by a border.

2.2.2 Environment

The environment was implemented as a 41×41 square lattice on a box, as shown in Figure 1a. Every square patch has a discrete two-dimensional location. The world does not wrap around vertically in either direction to mimic a limited region. The world is (optionally) split into two vertical halves by a border, made up of a column with the width of one patch (Figure 1b). The border corresponds to a physical barrier, such as a river or mountains. This environmental setup was appropriate for our modelling goals, since precise movement dynamics were not a priority and our aim was to simulate geographic separation.



(a) The grid world dimensions and settings.



(b) The environment without agents, showing the centre border.

Figure 1: Setup of the model world.

2.2.3 View

The view from NetLogo is illustrated in Figure 2. There are buttons to set up and run simulations, sliders to adjust parameter values, the graphic representation of the world, a tick counter, and results are being recorded with a reporter and monitor.

The world parameters are:

- `population-density`: probability of an agent spawning at a green field;
- `intra-language-variation`: standard deviation of the zero-mean distribution from which individual language variation is sampled. The variation is added to the default language vector of each population initially;
- `number-of-words`: size of vocabulary;
- `word-length`: how long words in the vocabulary are;
- `border-difficulty`: probability of agent not moving to a border patch once it decides to move there;
- `learning-rate`: how much vocabulary changes after a conversation to be more similar to the partner's.

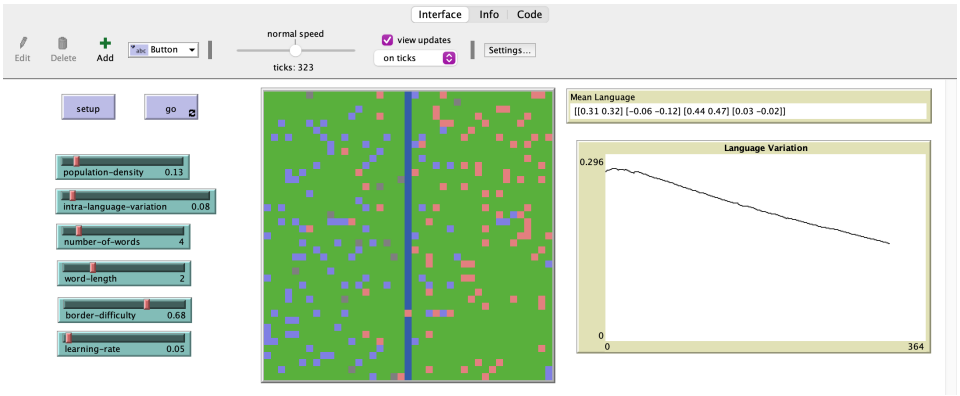


Figure 2: The model view. On the left, there is a setup button and a go button to control the flow of the simulation. Model parameters were set via the sliders below the buttons. In the centre, the world is shown. On the right, a reporter shows the mean language and a monitor follows the changes in language variation. At the top, a tick counter keeps track of the time steps.

2.2.4 Model Setup

The model can be set up via the setup button, which calls the `setup` procedure. At setup, the world is cleared and the tick counter reset. The procedure also initializes two languages, i.e., two vocabulary vectors with random values in $[-1, 1]$ of the specified shape (`number-of-words` $\times$ `word-length`). Then, it initializes the patches to be occupied with agents, empty, or serve as the border. It first initializes the border, then the patches on either side of the border separately.

The border is always at the same coordinates, along $x = 0$, in the middle, as shown in Figure 1b. Border patches are blue, unoccupied initially, have `probability-of-entering` set to $1-$`border-difficulty`, and have an empty vocabulary that is not used. Similarly, grass patches are empty of an agent, rendered as green, have `probability-of-entering` equal to 1, and have an empty vocabulary.

On the other hand, agent patches are coloured based on the mean value of their first three word vectors, have `probability-of-entering` equal to 1 (just to differentiate between border patches and non-border patches, in reality a patch can only be occupied by one agent), and have non-empty vocabularies. Agents have probability `population-density` to spawn at a patch instead of plain grass. The vocabulary of an agent is initialized in the following way. Starting from the base language of the side the agent is on, a small number is added to each value in a word vector. The small number is sampled from a zero-mean distribution with standard deviation `intra-language-variation`. This initialization process aims to replicate interspeaker variation, which occurs in real languages.

The last step of the initialization process is setting the colours corresponding to each patch type: blue, green, or vocabulary-dependant.

2.2.5 Agent Behaviour

The overall agent behaviour is to move randomly, have a conversation, and adapt its language based on the conversation. At each tick, an agent talks, moves, and the world is recoloured.

To talk, an agent randomly chooses one of its occupied Von Neumann neighbours to talk to. Talking in our model means the agent who initiated the conversation will adapt its `language` to be more similar to its partner. The updated `language` is the weighted average of the one prior to the conversation and the mean `language` between the two agents:

$$\alpha * \text{my-language} + (1 - \alpha)\mathbf{mean}(\text{my-language}, \text{partner-language}),$$

where α is the `learning-rate`, and the mean is computed element-wise.

To move, an agent randomly decides on one of four directions up, down, left, right, which correspond to its Von Neumann neighbourhood. If the patch it is headed to is occupied, the agent does not move. If the neighbour patch where it intends to go is empty and not a border patch, it occupies it and frees up the one where it previously was. If the destination is unoccupied but a border patch, then it has `probability-of-entering` to move there, and otherwise it stays in place.

The world (including the agent patches) is recoloured such that the changes can be visualized.

2.3 Experimental setup

The experiments have been set up and ran through BehaviorSpace in NetLogo. To answer the research question, we will vary the border difficulty and initial intra language diversity.

The experiments run an exhaustive combination search through the following parameters that can be found in the table 1.

Each experiment has been run 5 times and has been limited to 1000 time steps. This is due to the fact that, in an infinite amount of time, all agents would interact with each other so much that the language will converge to the mean, and there won't be any differences.

During the experiment, we will observe how the initial conditions have effect on the diversity of the languages at the last step. Another metric we observe is how long does it take the language to converge given the initial setup. We define a language to be converged by having its rate of change of variance below 10^{-4}.

Parameter	Start	End	Increment
`population-density`	0.13	-	-
`intra-language-variation`	0.01	1	0.25
`number-of-words`	4	-	-
`word-length`	2	-	-
`border-difficulty`	0.01	1	0.2
`learning-rate`	0.05	-	-

Table 1: Parameter space of the model explored.

3 Results

In this section, we analyse the effect of varying border difficulty and language diversity on the convergence time of language.

3.1 Experimental results

Our results provide insight into the dynamics of language convergence under various conditions of border difficulty and intra-language variation. The outcomes of the experiment are displayed through figures 3, 4, and 5, with each elucidating different aspects of the language evolution process.

In figure 3 there is a heatmap of the variance of words at the final step, which is set at the 1000 time step. The heatmap illustrates the end state of language diversity, dependent on the initial intra-language variation and border difficulty. As expected, higher intra-language variation leads to greater diversity at the end of the simulation, whereas increased border difficulty seems to limit this diversity. The darkest red areas, which indicate the highest variance, are predominantly seen at higher levels of intra-language variation and lower levels of border difficulty.

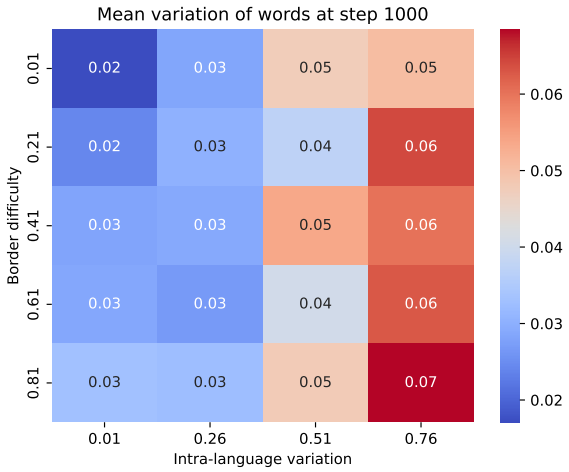


Figure 3: Heatmap of the variance of the words at the final step, where the final step is at 1000th time step.

In figure 4 plots the mean variation of words over time, coloured by border difficulty. The graph provides a visual trajectory of how language variation diminishes over time under different conditions. Notably, the curves are steeper at lower border difficulties, signifying a faster rate of convergence. This suggests that when borders present less of a challenge to communication, languages tend to homogenize more rapidly.

The figure 5 is a heatmap of the time of convergence, which further supports this observation. The convergence time is measured by the timestep at which the rate of change of variance drops below 10^{-4}. The cooler blue tones represent faster convergence times, while the warmer red tones indicate a slower convergence. It is evident from the pattern that lower intra-language variation and border difficulty contribute to a quicker convergence.

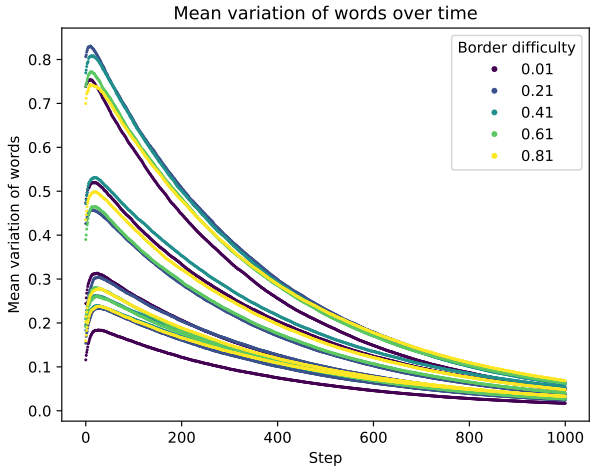


Figure 4: Language variances over time, coloured by border difficulty. Here you can observe that they have three groups, they correspond to the initial language variances.

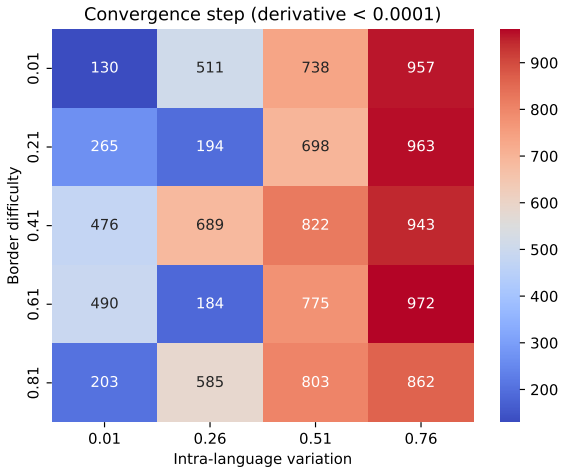


Figure 5: Heatmap of the time of convergence of languages, where convergence is when the rate of change of variance is less than 10^{-4}.

From these results, we can infer that the initial setup plays a significant role in the speed and extent of language convergence. When border difficulty is low and intra-language variation is minimal, languages converge quickly, reducing diversity. Conversely, when there is high intra-language variation and border difficulty is increased, convergence is slowed, preserving language diversity over the 1000 timestep duration of our simulation.

4 Discussion

In this section, we discuss the implications of the model, acknowledge limitations of our approach, and outline the ground for the future research.

The results of this model run provide insight into which factors prevent language convergence in the short run. In particular, those are the difficulties of crossing the geographical border between two groups and the high language variation within the initial tribe. Then, our research question can be answered. The easier it is to cross the border, the faster is the development of intermediate dialects, with convergence into one language ultimately. Similarly, the less complex the initial languages are — that is, the less vocabulary there is and the less standard deviation from the norm within the language there is — the faster this language will converge with the others. In the Figure 4, one can notice a small bump in variation at the start, before it starts to decrease. This behaviour can be

explained by the instant of the emergence of the third language. This instance of a first in-between language introduces some variation previously not present at the setup. However, after this, the variation demonstrates stable decrease.

There are certain limitations to our model that concern the conditional nature of the implementation. Firstly, simulation is based on the assumption of first contact. That is, the two tribes have never interacted before, and hence there was no influence of one language over the other. While this might be the perfect case for the modelling of the Creole language [1], it reduces the model's ecological validity to other circumstances. More serious assumptions are that the diversity of languages within a tribe at the start of the simulation is homogenous. There are no clusters of intra-tribal dialects, but rather they are distributed equally. Another limitation of the model is that all words are assumed to be of equal size. Finally, both intra-language diversity and vocabulary parameters are shared between the populations. It is not obvious that this is always the case, especially in the real-life formations of creole language, where colonizers, with a much larger conceptual vocabulary, were interacting with the indigenous people. Incorporating these extensions in our model would contribute to more precise answer to the stated research question.

Future directions of research could include extending the model by introducing more intratribal parameters, such as openness to contact versus unfriendliness or the nomad versus settler factor. These goals or preferences can play a detrimental role in the process of the interaction with one's neighbours. The research questions could then be generalised as studying the intratribal parameters effect on the convergence of the languages. Also, while our model largely explores the phenomenon of *integration*, where two languages are being merged in some form of a dialect, future research could explore the phenomenon of *assimilation*, where one language fully overtakes the other.

Finally, the implications of our results might suggest that, given no conservative forces that are aimed at language preservation, over the long run, all interacting languages in the world will merge into a single language. While our model works with the condition of a necessary physical interaction, our epoch provides an alternative in the form of online interactions. We suggest that this factor will increase the convergence rate of languages even more significantly.

References

[1] Gregory Furman and Geoff Nitschke. Evolving an artificial creole. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion*, pages 85–86, 2020.

[2] Magee L. Hendery, R. Geo-language games: An agent-based model of the role of terrain in language diversity. In *Digital humanities*, 2015.

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

[4] Pieter de Bie and Bart de Boer. An agent-based model of linguistic diversity. *Language, Games, and Evolution*, page 1, 2007.