Political Complexity's Role in Shaping Language Diversity: An Agent-Based Modelling Approach

Author: Gavin Gormley

Student Number: 1004875

Supervisors: Prof. dr. Derek Karssenberg and Dr. Judith Verstegen

MSc Applied Data Science

Abstract

The uneven global distribution of human languages remains a significant question in linguistics. Prior research suggests that more politically complex societies tend to reduce language diversity by spreading their languages over larger areas through cultural group selection. To explore this, this study refines an existing agent-based model to simulate the emergence and spread of languages, incorporating political complexity and cultural group selection mechanisms. Languages are distinguished in the model using divisive clustering with an optimal Levenshtein distance normalised (LDN) threshold determined by silhouette scores. The model simulates societal interactions over 7,000 years, comparing outcomes with real-world data from West Africa. The results indicated an optimal LDN threshold of 0.624 for distinguishing languages, though a threshold of 0.500 was used to ensure sufficient languages emerged. Further, the results suggested that political complexity might reduce language diversity, though this could not be validated when compared to the real-world data. Despite the studies limitations, it provides insights into the use of agent-based models for simulation language emergence and evolution, along with providing insights into how political complexity and cultural group selection mechanisms shape language diversity.

Table of Contents

1. Introduction

Human languages are distributed unevenly across the globe, with significant concentrations of language diversity located in two primary belts: one spanning West and Central Africa, and another extending through Southeast Asia and the Pacific (Nettle, 1998). Our understanding of the mechanisms contributing to the uneven distribution of languages remains an unresolved issue in linguistics (Gavin et al., 2013). On a micro-level, linguistic features are said to change through diffusion or innovation, where features may be created, adopted, or altered (Nerbonne, 2010). As such, new dialects form, which over time can develop into languages, though distinguishing between both is subject to much scrutiny (Haugen, 1966). Similarly, new languages can emerge through processes of divergence from a common ancestor (Renfrew, 1991). Like the evolution of species, languages produce new languages, some of which become extinct and others of which survive (Pagel, 2000)

Nonetheless, the spread of language cannot solely be explained on the micro-level or as a natural occurrence, it must also be investigated at the macro-level as it is subject to many interacting environmental, social, and cultural factors (Gavin et al., 2013). One socio-cultural factor which has been found to significantly predict language diversity is that of political complexity (Currie & Mace, 2009). It has previously been posited that the emergence of politically complex agricultural societies is a major factor in reducing language diversity (Renfrew, 1994). This extends from the theory that agriculturists fanned out many indigenous languages in their path (Renfrew, 1987), which may explain why Europe has a relatively homogenous and small number of languages (Pagel, 2000). Currie and Mace (2009) found that political complexity is a key predictor of the distribution of ethnolinguistic groups, with more politically complex societies spreading their languages over larger areas. Currie and Mace (2009) posit that a process of cultural group selection favouring more politically complex societies played a significant role in shaping the global distribution of language diversity. They suggest that more politically complex societies tend to replace or absorb less complex groups, thereby spreading their languages over larger areas, which leads to an increase in the proportion of politically complex societies over time and their languages (Currie & Mace, 2009).

Cultural group selection is a hypothesis that suggests that (1) human groups exhibit significant differences in their cultural traits (2) these cultural differences can influence the success and competitiveness of groups and (3) as a result, cultural traits that enhance group survival and success are likely to be passed on and spread, similar to how advantageous genetic traits are favoured in biological evolution (Richerson et al., 2016). As such, it is logical that more politically complex societies would outcompete less politically complex societies, thereby spreading the language of the more politically complex societies. However, this must be investigated. To take a different approach than Currie and Mace (2009), but to also test their hypothesis, this study investigates if mechanisms of cultural group selection can lead to the spread of more politically complex societies and their languages, thereby reducing language diversity.

Language diversity is subject to feedforward and feedback from the speakers in question, rendering the spread of language a complex adaptive system (Beckner et al., 2009). Complex systems are typically referred to as networks of relatively simple components, with respect to their role in the network, where complex behaviours emerge that are not easily predictable from the behaviours of the network's components (Mitchell, 2006). This behaviour can be observed in language diversity globally, which complicates efforts to model language diversity. As such, traditional statistical modelling methods may prove insufficient for understanding the spread of language. Consequently, simulation methodologies are frequently employed in modelling such complex systems. In this context, Steels (1997) advocates for agent-based modelling as an effective approach to exploring both the spread and emergence of language.

Since then, limited research has utilised agent-based modelling in studying language diversity, except for a few works such as de Bie and de Boer (2007), which examines how linguistics patterns and borders evolve from individual actions and social impact theory. However, there have also been studies using agent-based modelling to simulate micro-level language change (see Beeksma et al., 2017; Civico, 2019).

Furthermore, there are few studies investigating the influence of political complexity on language diversity. Specifically, while Currie and Mace (2009) highlighted the potential impact of political complexity on the distribution of ethnolinguistic groups, there is a lack of studies that validate and expand upon their findings. This gap is crucial because understanding how socio-cultural factors affect language diversity can provide deeper insights into the mechanisms of cultural evolution and the emergence of language diversity. Language diversity is declining globally (De Oliveira et al., 2006). As such, expanding this research could inform efforts to preserve endangered languages in more politically complex societies. Therefore, this research aims to fill this gap by integrating political complexity and mechanisms of cultural group selection into an agent-based model to simulate and understand the dynamics of language spread and diversity.

1.1. Problem statement

Given the acknowledged complexity of language diversity as a complex system, traditional statistical models remain inadequate for fully understanding the dynamics of language spread and diversity. Communication through language is a uniquely human factor, and understanding it is essential not only for linguistics, but also for fields such as sociology, anthropology, and cognitive science. Language diversity is influenced by both micro- and macro-level factors, making it essential to develop models that can accurately reflect its adaptive and emergent properties.

1.2. Objective

The primary goal of this study is to explore the impact of political complexity on language diversity patterns. To achieve this, mechanisms of cultural group selection and political complexity will be integrated into an agent-based model that simulates the emergence and dynamics of languages. This approach aims to provide insights into how political complexity influences language evolution and diversity, and whether cultural group selection mechanisms drive this process. A prerequisite for this is to be able to identify languages within the model. Therefore, the first task of this study is to determine a measure for identifying when a different speech forms form languages within the agent-based model. Finally, the model's results will be compared with real-world language diversity in West Africa to validate the findings.

1.3. Research questions

1. How can languages be distinguished in an agent-based model simulating language diversity?
2. How does political complexity influence patterns of language diversity?
3. How closely do the results mirror real-world distributions of language diversity for different political complexity levels?

2. Literature review

2.1. Language emergence

Numerous hypotheses have been suggested to account for the emergence of language during the last million years of human evolution, though a large majority of linguists argue that nothing can be said about languages more than 8,000 years in the past (Coupé & Hombert, 2005). There are two main theories regarding the origins of language: monogenesis; where language was invented at only one prehistoric site, and polygenesis; where language was invented at several prehistoric sites (Freedman & Wang, 1996). Monogenesis is the most assumed theory on probabilistic grounds, though it has been found that polygenesis is also plausible (Freedman & Wang, 1996).

2.2. Dialects and languages

The challenge of differentiating between a language and a dialect is a much-debated topic, with there being no agreed-upon criteria for how to resolve it (Evans & Levinson, 2009). The classification of a speech form as either a dialect or a language is often driven by social or political factors (Nordhoff & Hammarström, 2011). Mutual intelligibility, the ability of speakers of different but related language varieties to understand each other in ordinary conversation (Matthews, 2014), is the key factor linguists have used to differentiate languages from dialects (Van Rooy, 2020). It is generally accepted that if at least 70% of the vocabulary is mutually understandable between people from two regions, they speak different dialects; otherwise, they speak different languages (Kosheleva & Kreinovich, 2013)

Nonetheless, quantitative methods can help provide a more objective approach to this issue. Wichmann (2020) and (Boga, 2020) both provide solutions to this through the use of Levenshtein distance normalised (LDN) measures, with Boga (2020) also employing the Needleman-Wunsch algorithm. Wichmann (2020) uses a large lexical database and identifies a threshold LDN value of 0.51 to distinguish dialects from languages. Boga's (2020) study on Romance languages identifies three clusters: dialect-dialect, language-dialect, and language-language pairs, and found threshold values for their distinction using the Needleman-Wunsch normalised and divided and Levenshtein distances normalised and divided (LDND) (Boga, 2020).

2.3. Linguistic diversity

Linguists identify three types of linguistic diversity: the number of languages *(language diversity*), the number of language families (*phylogenetic diversity*), and the degree of structural differences between languages *(typological diversity* or *disparity*) (Pacheco Coelho et al., 2019). Linguistic diversity is declining globally, with it being estimated that 50% of existing languages may be extinct in the next century (De Oliveira et al., 2006). Furthermore, only one hundred languages are spoken by 90% of the world's population, highlighting the concentrated use of a few dominant languages (De Oliveira et al., 2006). The decline in language diversity in recent years has been attributed to people abandoning their native languages and switching to more dominant languages with larger numbers of speakers, usually due to socio-economic pressures (Harmon & Loh, 2010). However, while the factors shaping language diversity in recent years are largely understood, many details about how language diversity evolved over thousands of years remain unknown.

In early human history, languages spread through initial migration from Africa, approximately 100,000 years ago (Renfrew, 1994). Since then, Renfrew (1994) has emphasised three potential catalysts for the spread of language: the invention of farming, climate/environmental factors, and elite dominance. The transition from hunter-gatherers/foragers to agriculturalists is said to have facilitated the spread of languages through population growth pushing these farmers into wider regions and thereby displacing the languages of pre-existing forager populations (Ross, 2006). Nonetheless, it has been suggested that this transition increased the role of the environment in how human populations are separated into groups, thus shaping the boundaries between languages, and, in turn, the social and economic networks which food-producing populations operate (Derungs et al., 2018).

Furthermore, many other studies have examined the impact of environmental and climatic factors on language diversity. Nettle (1998) found a correlation between the number of languages spoken and the climatic variability in major tropical countries, suggesting that regions with stable climates that allow year-round food production tend to have higher language diversity because small, self-sufficient groups can maintain distinct languages. Similarly, Gavin et al. (2013) used a process-based modelling approach which found that regions with higher precipitation can support more individuals, thus enabling self-sufficient groups to preserve distinct languages. Hua et al. (2019) further corroborated these findings, highlighting that year-round productivity plays a significant role in promoting language diversity. The study also found that climatic features have a much more pronounced influence on language diversity than landscape features (Hua et al., 2019). These studies highlight the importance of environmental factors in enabling societies to be self-sufficient, thereby allowing them to maintain distinct languages, and thus promoting language diversity. In contrast, Gavin & Sibanda (2012) examined the degree to which area, isolation, environmental conditions, and time since first settlement explained variation in language richness among Pacific islands. The study found although environmental productivity may influence language diversity patterns at a global scale, it plays a minimal role on Pacific islands, with approximately half of the variance in language richness remaining unexplained, suggesting that language diversity patterns are also influenced by economic, political, and social factors (Gavin & Sibanda, 2012).

Nonetheless, there has been limited research into how social and cultural factors shape language diversity. Coulmas (2013) suggests that migration, language loyalty, and language utility play a role in language shift and maintenance, which ultimately affect language diversity. Further, it has been found that the rate of language evolution is affected by the population size of groups speaking a given language (Bromham et al., 2015). Currie and Mace (2009), perhaps the most notable work on socio-cultural factors and political complexity, found that political complexity is a key predictor for the distribution of ethnolinguistic groups.

2.4. Political complexity

Political complexity is typically measured by the number of hierarchical levels of decision-making within a society. Societies without permanent leadership beyond the local community are acephalous, those with one or two hierarchical levels are termed simple chiefdoms and complex chiefdoms, respectively, while those with more than two levels are classified as states (Currie et al., 2010). These levels are summarised and defined in *Table 1*.

Political complexity	**Jurisdictional hierarchical levels beyond the local community**
Acephalous	0
Simple chiefdoms	1
Complex chiefdoms	2
States	3
Big states	4

Table 1: Political complexity levels (Murdock et al., 1999)

Acephalous societies are distinguished by the lack of hierarchical leadership or centralized power structures, with the term acephalous originating from the Greek word for 'headless' (Townsend, 2018). It is argued that chiefdoms then emerged through a combination of social, economic, and political pressures, often driven by population and resource concentrations within societies (Grinin & Korotayev, 2012). Before the rise of bureaucratic states, chieftaincies were the primary political institutions, developing independentlu of one another around 7,000 years ago across the world, and in some regions, they have maintained power into modern times (Earle, 2011). Chiefdoms are centralised tribes led by a chief, a term reflecting their leadership structure (Turchin & Gavrilets, 2009). They are regionally organised societies with a centralised decision-making hierarchy that coordinates multiple village communities, ranging in size from simple chiefdoms of around a thousand people to complex chiefdoms with populations in the tens of thousands (Earle, 1987). States represent the highest level of political complexity, characterised by many jurisdictional decision-making levels and a bureaucratic form of governance (Covey, 2008). According to Carneiro's (1970) theory, states emerged due to environmental or social circumscription, where population pressure and competition for limited resources led to increased conflict and warfare, resulting in the subjugation of defeated groups and the consolidation of power under a central authority. The complexity of this political structure enables states to govern large and diverse populations over extensive territories, providing a higher degree of organisation compared to the lower levels of political complexity.

As such, political complexity can provide insight into the political organisation and governance within societies. There have been many theoretical studies regarding political complexity (e.g. Grinin & Korotayev, 2011, 2012) and qualitative research on the political complexity of early societies (e.g. Dueppen, 2018; Vasyutin, 2019).

However, few studies have investigated how political complexity affected other factors of these societies or how those factors influence political complexity. Abrahamson (1969) found that political complexity in pre-industrial societies positively correlates with social differentiation, demographic complexity, and socioeconomic development. Further, Hamilton et al. (2020) found that as societies increase in political complexity, both population size and geographic range scale predictability, with a four-fold increase in population size and a two-fold increase in the geographic range observed with each additional level of complexity. Moreover, it is suggested that more politically complex societies face endogenous factors such as population growth and more complex institutions which might contribute to further increases in political complexity (Hamilton et al., 2020). Currie and Mace (2009) examined various cultural and environmental factors to understand what influences the area covered by human languages. They found that political complexity is the most significant factor in predicting the distribution of ethnolinguistic groups. These findings supported their cultural group selection hypothesis, which posits that more politically complex societies tend to occupy larger areas and dominate less complex ones. This political dominance often results in the spread of their languages and cultural traits across broader regions. Further, Currie et al. (2010) found that while the political complexity of societies typically increases incrementally through small steps, decreases in complexity can occur both sequentially and in larger drops, though these are less frequent.

2.5. Cultural group selection

Many social scientists have suggested that cultural inheritance may be controlled by a mechanism very similar to natural selection (Richerson, 1977). This follows from Darwin's (1871) suggestion that group evolution could impact individual survival. As such, the cultural group hypothesis posits that social groups lacking group-beneficial traits that support essential societal functions will become extinct, leaving only those societies with effective cultural attributes to survive (Soltis et al., 1995).

Among the cultural group selection literature, many group-beneficial mechanisms have been posited. Henrich (2004) identifies three processes where group-beneficial behaviour can spread: demographic swamping, prestige-biased group selection, and selective migration between groups. Additionally, Richerson et al. (2016) highlight three similar mechanisms: natural selection, selective imitation of groups, and selective migration between groups.

a. Natural selection / demographic swamping

Groups display differences in behaviours that affect the rate at which they grow, form new groups, overcome resource issues, resolve internal conflict, and other issues, which Richerson et al. (2016) refer to as *natural selection*. As found in Soltis (1995), these differences result in a relatively slow selection process that contributes to group success. Similarly, Henrich (2004) describes a process of *demographic swamping* which changes the frequency of cultural traits in a population because some groups reproduce faster due to some cultural practices, which is found to be the slowest form of cultural group selection, usually unfolding over millennia.

b. Selective imitation of successful groups / prestige-biased selection

If people prefer to imitate successful individuals and have contact with out-group members, those in less successful groups will often copy those in more successful groups, promoting the spread of beneficial norms and institutions, referred to as *selective imitation of successful groups* (Richerson et al., 2016). *Prestige-biased group selection* describes the same concept, where individuals preferentially copy individuals who get higher payoffs, with the higher an individual's payoff, the more likely that individual is to be imitated, with this process being the fastest of the group selection mechanisms, probably occurring on the time scales of decades or centuries (Henrich, 2004).

c. Selective migration

Individuals in dysfunctional groups may move to another group if they see that it maintains more group-beneficial behaviours, resulting in the reduction in size and competitiveness of unsuccessful groups and the possibility of the group imitating more successful groups as described previously (Richerson et al., 2016).

d. Intergroup competition

Intergroup competition explains how different groups compete through warfare and raiding, which leads to the proliferation of cultural practices that enhance competitive success, with the less effective group being defeated, absorbed, or dispersed (Henrich, 2004).

3. Conceptual framework

The conceptual framework of the model will be primarily based on the culture group selection mechanisms discussed in Henrich (2004) and Richerson et al. (2016). These mechanisms will inform a rule-based approach to an agent-based model which simulates the spread and emergence of language. The area will be divided into societies (regions), amongst which agents will be randomly distributed. As such, political complexity will be assigned to different societies, where they will start as either acephalous or simple chiefdoms. The maximum level of political complexity in this model will be complex chiefdoms, to replicate the maximum level in the validation dataset in West Africa. Agents will originally take the political complexity of the society in which they are initially assigned to. This is visualised in *Figure 1*.

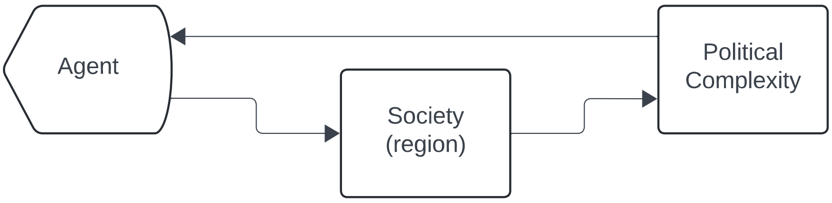


Figure 1: Political complexity mechanism in the model

As such, the finalised rules are as follows:

1. *Agents in less politically complex societies are more likely to adopt linguistic features from more politically complex societies they encounter.* This is due to selective imitation of successful groups (Richerson et al., 2016) and prestige-biased selection (Henrich, 2004).
2. *Societies' political complexity can grow at an incremental rate, where societies will have a fixed probability of increasing in political complexity in each time step.* This aligns with the findings of Currie et al. (2010), where the political complexity of societies was found to increase incrementally through small steps.
3. *More politically complex societies have a higher probability of advancing in political complexity*. This stems from the suggestion that more politically complex societies face endogenous factors which further contribute to advancements in political complexity (Hamilton et al., 2020).
4. *Societies will have a higher growth rate in terms of political complexity if they neighbour a more politically complex society*. This follows Henrich's (2004)

explanation of intergroup competition, which leads to societies adopting practices that enhance their competitiveness.

5. *Agents from less politically complex societies migrate to more successful societies they interact with.* This rule also stems from Henrich's (2004) intergroup competition mechanism, where agents can be "absorbed" by more politically complex societies. Further, this rule also represents the selective migration mechanism described in Richerson et al. (2016)., where agents may migrate to more politically complex societies due to the unsuccessfulness of their original society.
6. *Agents who move more than one region away from their original society will be reassigned to the society in which they currently reside.* This rule ensures that agents are correctly associated with their respective regions only after significant geographical movements. This enables agents to still interact with agents outside of their societies, while also maintaining the functionality of the model. As such, this is primarily a technical mechanism rather than conceptual.

These rules are illustrated in *Figure 2*.

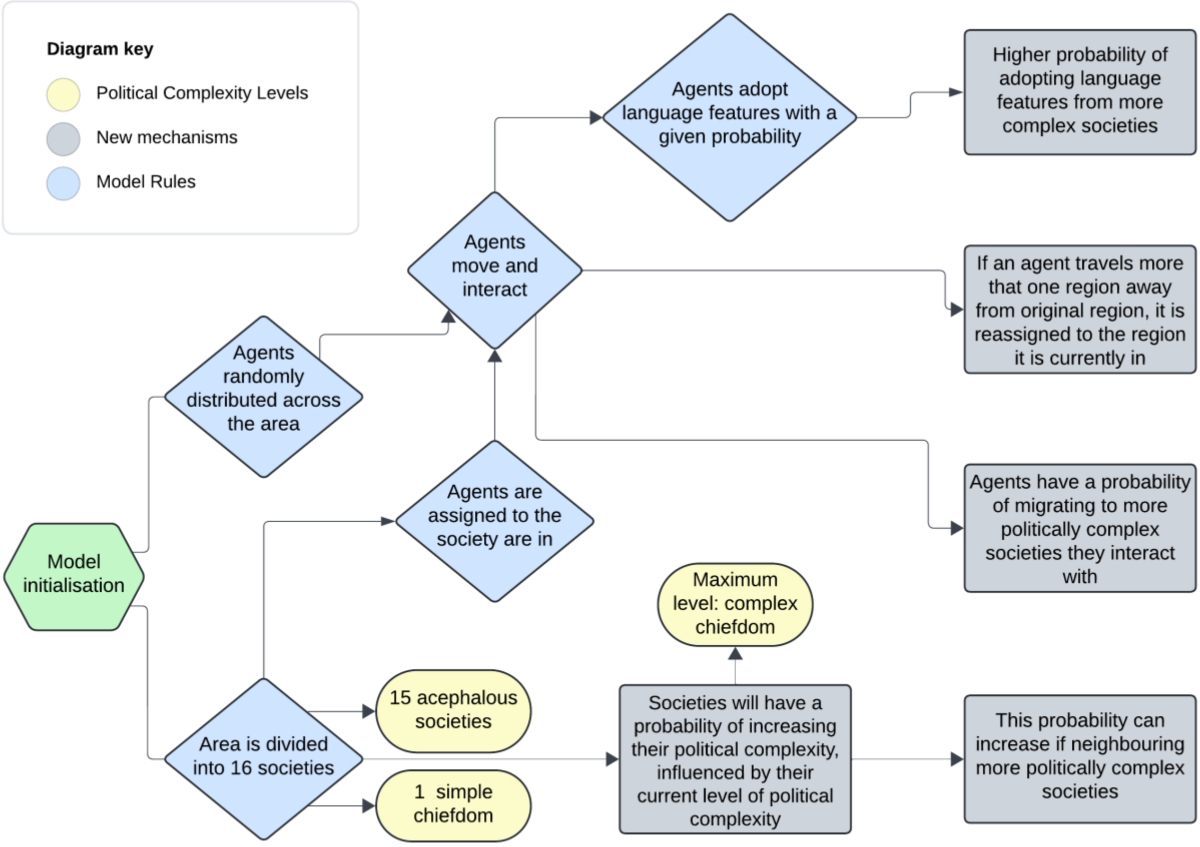


Diagram key
Political Complexity Levels
New mechanisms
Model Rules
Model initialisation
Agents randomly distributed across the area
Area is divided into 16 societies
15 acephalous societies
1 simple chiefdom
Agents are assigned to the society are in
Agents move and interact
Agents adopt language features with a given probability
Higher probability of adopting language features from more complex societies
If an agent travels more that one region away from original region, it is reassigned to the region it is currently in
Agents have a probability of migrating to more politically complex societies they interact with
Maximum level: complex chiefdom
Societies will have a probability of increasing their political complexity, influenced by their current level of political complexity
This probability can increase if neighbouring more politically complex societies

Figure 2: Conceptual diagram

4. Methodology

The methodology section of this thesis is structured to address the three primary research questions through an agent-based modelling approach. This section begins with an overview of the original model from which was built upon, followed by detailed explanations of its key components and the modifications made to incorporate mechanisms of cultural group selection and political complexity. As such the approaches to answering *RQ1*, *RQ2*, and *R3* will be explained in *"Distinguishing languages"*, *"Political complexity and cultural group selection"*, and *"Model validation"*, respectively.

4.1. Model overview

Karssenberg (2024) developed an agent-based model to simulate the dynamics of language evolution and diversity within a defined population over time. The model simulates agents' interactions, language mutation, and spatial movements within a fixed area. The simulation initialises with a defined number of agents distributed randomly across a fixed area. Each agent is assigned a random language configuration, which is represented as a list of integers. These integers represent different linguistic properties. Each agent in the model represents a society of 100 individuals. A more technical overview is available in the *Appendix*. A more conceptual overview is provided as follows.

4.1.1. Language representation

In this model, language is conceptualised as a collection of discrete features. Each agent in the population possesses a unique linguistic profile represented by a sequence of classes. Each property in this sequence can take on one of several possible classes. These classes collectively encode the linguistic attributes of an agent. To facilitate the analysis and clustering of languages, each linguistic profile is converted into a unique identifier. This identifier is derived by concatenating the values of the individual properties of the sequence and interpreting the resulting string as a number in a specified base.

4.1.2. Model mechanisms

a. Movement

Agents move within the simulated area according to a specified standard deviation in x- and y-direction. The movement of agents is constrained within the geographic boundaries of the model.

b. Neighbourhood

The agents choose interaction partners within a predefined neighbourhood radius.

c. Mutation

Each agent has a predefined probability of undergoing a mutation in any given year, where one or more elements of their language configuration might randomly change. This simulates natural or spontaneous changes in language over time.

d. Interaction (chatter)

Agents have a fixed probability of engaging in a language exchange (chatter) with nearby agents within a specified radius. During such interactions, agents may adopt language elements from each other, simulating social learning.

e. Chatter effect

After an interaction (chatter), agents have a probability of taking one linguistic class and property combination from their interacting agent, leading to the recalculation of their language representation.

4.2. Model modification

4.2.1. Distinguishing languages

To address *RQ1 – "How can languages be distinguished in an agent-based model simulating language diversity"* – the original model is enhanced to differentiate languages. In the original model, each agent has linguistic features which make up its speech form. As such, if two agents have very similar speech forms, they are classed as two different languages, however, this is not the case in practice. Therefore, for this model to be able to answer our first and subsequent research questions, it must be able to distinguish when a speech form becomes a new language. To resolve this, the model will make use of LDN as used in Wichmann (2020) to provide a quantitative measure for distinguishing different languages via clustering, and then utilise silhouette scores to obtain the optimal LDN threshold for the clustering.

a. Levenshtein distance

Levenshtein distance is a similarity metric which allows for the quantification of similarity between strings (Yujian & Bo, 2007). It is the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into another. As such, this metric is useful in the agent-based model for measuring the differences between linguistic properties represented as a combination of integer values. Initially, each agent's speech form configuration is represented as a list of integers, and this is converted into string format for the computation of the Levenshtein distances. The Levenshtein distance between the language strings of all agents is then calculated and normalised, resulting in a distance matrix.

b. Divisive hierarchical clustering

To be able to split agents into different language groupings, divisive hierarchical clustering is employed. This method is appropriate for identifying linguistic groups as it divides the dataset into increasingly smaller clusters. The process begins with all agents grouped into a single cluster. From there, the clusters will be recursively split based on an LDN threshold metric. Wichmann (2020) found that an LDN value of 0.51 could be used to distinguish dialects from languages. However, this measure will likely not prove useful for the agent-based model, as it will not reflect language in the same manner given as Wichmann (2020) uses an actual lexical database to derive the obtained threshold of 0.51, yet the model uses sequences of integers to represent language. As such, the use of silhouette scores was explored to find optimal thresholds for the clustering process.

c. Silhouette scores

Silhouette scores are a method to evaluate the quality of clusters. The silhouette score measures how similar an object is to its own cluster compared to other clusters and therefore measures the cohesion and separation of clusters (Rousseeuw, 1987). The silhouette score is determined by the average intra-cluster distance (a) and the average nearest-cluster distance (b) for each data point, calculated as $(b - a)/\max(a, b)$ (Shahapure & Nicholas, 2020). A score close to +1 indicates perfect clustering, a score near 0 suggests the data point could belong to another cluster, and a score near -1 implies incorrect clustering (Shahapure & Nicholas, 2020). As such, silhouette scores will be utilised to determine the optimal LDN threshold for the divisive clustering algorithm. The model will first run to calculate the optimal range of thresholds as per the silhouette scores, then, the model will run again using the optimal silhouette score.

d. Cluster continuity

In this model, the clusters are created independently at fixed interval time steps. However, this introduces a lack of continuity in the visualisation of clusters, with languages likely being assigned different colours at each interval. Consequently, a continuity mechanism using Jaccard similarity is employed. Jaccard similarity is a metric used to compare the similarity between sets, which in this case are the sets of agents in each cluster. It is calculated as $J(A, B) = |A \cap B|/|A \cup B|$, where A and B are both sets, and in this case clusters. By comparing the sets of agents, the number of agents shared between clusters in different time steps can be determined. As such, clusters from the previous interval are matched with clusters from the current interval based on the highest Jaccard similarity scores, and then take that cluster’s previously assigned colour. If there is an increase in the number of clusters from the previous interval, the clusters which were not matched to a previous cluster are assigned a new colour.

4.2.2. Political complexity and cultural group selection

The model will be modified to incorporate political complexity, whereby societies will be assigned into four distinct types: acephalous, simple chiefdoms, complex chiefdoms, and states. This will be implemented at the spatial level, where all regions (societies) will initially be assigned as acephalous or simple chiefdoms. As such, an agent's behaviour will be affected by the political complexity of the society to which it belongs. The rules below follow from the rules outlined in the conceptual framework, where cultural group selection is hypothesised to be a driver of political complexity. The incorporation of political complexity into the model enables *RQ2 – "How does the integration of political complexity into the agent-based model influence the simulation outcomes in terms of language diversity?* – to be answered. The model will be simulated including and excluding political complexity to explore the impacts of the integration of political complexity into the model.

a. Initial assignment of Political Complexity

Each region within the simulation area will be assigned an initial political complexity level based on the proportions that reflect historical make-ups of political complexity. The model will be simulated over 7,000 years, to reflect when simple chiefdoms first emerged. As such, one region will be assigned as a simple chiefdom, and the rest will be assigned as acephalous societies.

b. Influence of political complexity on adopting linguistic features

Agents will have a probability of adopting linguistic features from other agents they interact with. This process is influenced by two factors. First, there is a base probability that governs the likelihood of an agent adopting a linguistic feature from another agent. Second, this probability is based on the difference in political complexity between the interacting agents, where a greater difference in complexity increases the likelihood of adoption. This aligns with the first rule outlined in the conceptual framework where agents in less politically complex societies are more likely to adopt linguistic features from more politically complex societies due to selective imitation of successful groups (Richerson et al., 2016) and prestige-biased selection (Henrich, 2004). As such, the effective probability of an agent adopting a linguistic feature during an interaction (P_{adopt}) is governed by the following equation:

$$P_{adopt} = P_{chatter\ effect} \times \alpha^{\Delta C}$$

In this equation, the term $P_{chatter\ effect}$ is the base probability of adopting a feature during an interaction. The parameter α is a scaling factor that adjusts the probability based on the complexity difference. The variable ΔC denotes the difference in political complexity between the interacting agents.

c. Probabilistic Increase in Political Complexity

Regions will have a probability of increasing their political complexity over time. There will be two mechanisms by which this will occur. First, regions have a base probability of increasing in political complexity. Second, this growth will also be influenced by the regions, where if a region borders a more politically complex region, it's probability of increasing in political complexity increases, as it seeks to maintain competitive against its neighbours. These mechanisms reflect that societies tend to increase incrementally in terms of political complexity (Currie & Mace, 2011) and that societies must adapt due to intergroup competition (Henrich, 2004). As such, the increase in political complexity (C) for a region i from time t to $t+1$ is governed by the following equation:

$$P\left(C_{i,t+1} > C_{i,t}\right) = (P_{base\ growth} \times (1 + C_{i,t}) + (\beta\ \times N_{more\ complex\ neighbours})$$

In this equation, $P\left(C_{i,t+1} > C_{i,t}\right)$ represents the probability that the political complexity of region i at time $t+1$ is greater than at time t. The term P_{base} is the base probability of the region increasing its complexity. This base probability is then scaled directly by the current complexity level $C_{i,t}$. This means that when the political complexity of region i is higher, the base probability increases proportionally, making the probability of increasing complexity higher. In the current model, this only affects simple chiefdoms, as the maximum political complexity level in the model is complex chiefdoms. This approach addresses the issue of complex chiefdoms rarely emerging due to the random nature of increments in the absence of neighbouring more politically complex societies. Moreover, the parameter β is the probability increase factor due to the influence of neighbouring regions. The variable $N_{more\ complex\ neighbours}$ is the number of neighbouring regions that have a higher complexity than region i.

d. Migration of agents to new societies

The migration of new agents to new societies reflects the processes of intergroup competition (Henrich, 2004) and selective migration (Richerson et al., 2016). This mechanism ensures that agents can move between societies, influenced by the political complexity and interactions within the simulation. When an interaction occurs, the migration decision is influenced by the difference in political complexity between the interacting agents' societies. The probability of migrating to a new society is determined by the same scaling factor used to determine the probability of adopting linguistic features. As such, the migration probability $(P_{migration})$ is governed by the following equation:

$$P_{migration} = P_{base\ migration} \times \alpha^{\Delta C}$$

Here, $P_{base\ migration}$ is the base probability of migration, α is a scaling factor, and ΔC is the difference in political complexity between the interacting agents. Migration is only considered if there is a complexity difference between the interacting agents' societies. Only agents belonging to a less politically complex society can migrate to a more politically complex society.

Finally, the migration of agents to different societies will only occur every 25 years in the model. This is due to the computational intensity of agents migrating based on interactions. As such, this will be taken into consideration when configuring the parameters for determining the probability of migration, $P_{base\ migration}$ and α.

e. Reassignment of agents to new societies

Agents in the model are reassigned to other regions based on the distance they travel from their assigned region. If the agent travels a distance equivalent to at least one region away from the centre of its assigned region, the agent will be reassigned to the region it currently resides in. As such, the agent then behaves according to the political complexity of its new region.

4.3. Parameter configuration

The parameters selected for the agent-based model are crucial for ensuring that the simulation represents the dynamics of language evolution and political complexity as closely as possible. These are summarised in *Table* 1 for an overview of the parameters.

Parameter	Definition	Value
General parameters		
T	Total number of years	7,000
$N_{population}$	Population size	1,500
Linguistic parameters		
N_p	Number of linguistic properties	6
N_c	Number of linguistic classes	6
$P_{mutation}$	Probability of linguistic mutations	0.0002
Interaction parameters		
$P_{chatter}$	Probability of agents engaging in chatter	1.0
$P_{chatter\ effect}$	Probability of adopting linguistic feature from chatter	0.8
Geographical and movement parameters		
$N_{regions}$	Number of regions	16
r	Neighbourhood size (radius)	10km
σ	Standard deviation of agent movement	5km
Political complexity parameters		
α	Scaling factor for complexity differences	1.05
β	Probability of political complexity increase per complex neighbour	0.000075
$P_{base\ growth}$	Base probability of political complexity growth	0.000025
$P_{base\ migration}$	Base probability of migration	0.0005

Table 2: Agent-based model parameters

The model runs over a total of 7,000 years (T), a period chosen to reflect the time since chiefdoms began to emerge. The population size ($N_{population}$) is set at 1,500, representing a population of 1,5000,000 individuals. This value was selected as higher values resulted in a much increased computational time.

The model utilises six properties (N_p) to represent linguistic features, where each property can take on one of six possible classes (N_c). This combination allows for 46,656 unique speech forms. As such, this number is large enough to ensure that each agent initially will have its own unique speech form. The probability of

linguistic mutations ($P_{mutations}$) is set at 0.0002, reflecting the rarity of such mutations in groups of individuals.

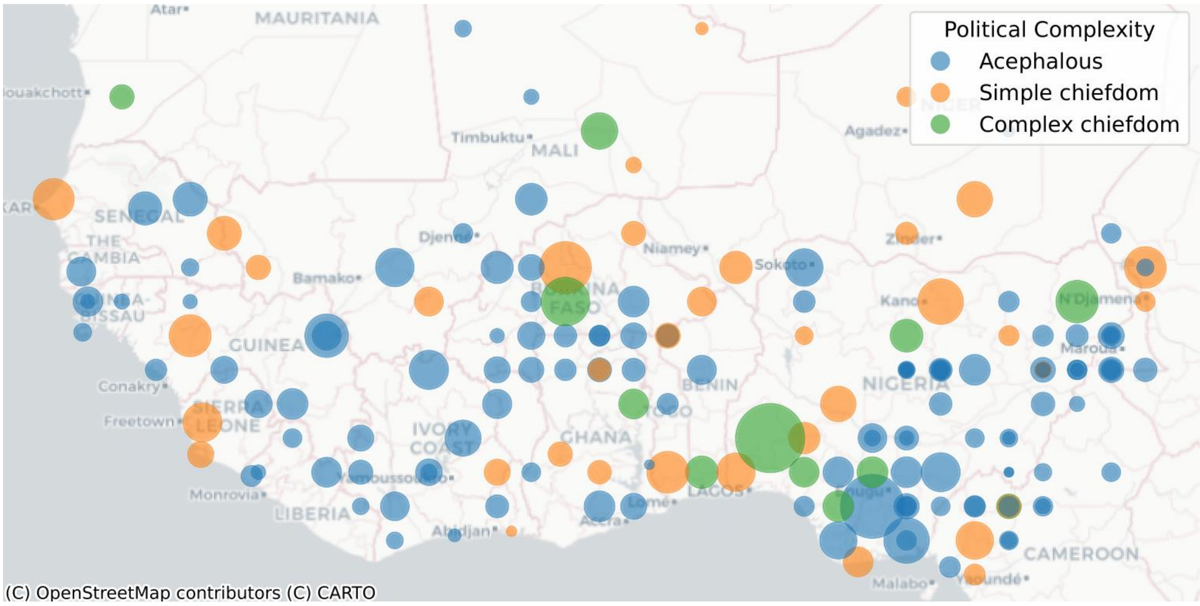
To ensure constant interaction among agents, the probability of agents engaging in chatter ($P_{chatter}$) is set at 1.0. This means that every agent interacts at each time step. The probability of adopting linguistic features from these interactions ($P_{chatter\ effect}$) is set at 0.8, simulating the influence of social interactions on language adoption and evolution.

The geographical area in which agents interact is defined as 1000km by 1000km, providing a large enough space to mimic real-world regions. Agents' movements are modelled with a standard deviation (σ) of 5km per year, representing realistic pre-modern societal movements. A neighbourhood radius (r) of 10km is used to determine which agents are considered neighbours for interactions. The area is divided into 16 regions $(N_{regions})$ to provide a manageable number of regions for analysis and to simulate regional differences in interactions and linguistic evolution.

The model also incorporates political complexity parameters to simulate the evolution of the political structures in these societies. The base probability of political complexity growth $\left(P_{base\ growth}\right)$ is set at 0.000025, reflecting the slow and gradual nature of political complexity evolution. The scaling factor for complexity differences (α) is 1.05, and the probability of political complexity increase per complex neighbour (β) is set at 0.000075. Additionally, the base probability of migration $\left(P_{base\ migration}\right)$ is set at 0.0005, allowing for migration, but with a low probability to reflect the rarity of migration in pre-modern societies.

4.4. Data

The data regarding political complexity originates from the *Ethnographic Atlas* (Murdock et al., 1999). This dataset contains a variable, *"Jurisdictional hierarchy beyond local community"*. This variable indicates the number of jurisdictional levels beyond the local community, where 1 corresponds to stateless societies (0 levels), 2 for chiefdoms (1 level), 3 for complex chiefdoms (2 levels), 4 for states (3 levels), and 5 for larger states (4 levels). As such, it acts as a measure of political complexity. This data has been aggregated on *D-PLACE* (Kirby et al., 2021) along with the Glottolog code for the language spoken by the societies in question, which will provide us with data to validate the results of the model. *D-PLACE* (Database of Places, Language, Culture, and Environment) is an open-access database that compiles information on the geography, language, culture, and environment of over 1,400 human societies, enabling researchers to explore drivers in cultural change and global patterns of cultural diversity (Kirby et al., 2016).



(C) OpenStreetMap contributors (C) CARTO

Figure 3: Geographical Distribution of Societies in West Africa (1870-1960) (Kirby et al., 2021)

For this study, the dataset was filtered to include only societies within the region of West Africa, resulting in 173 societies observed. Within this, there are 125 acephalous societies, 36 simple chiefdoms, and 12 complex chiefdoms. The location of these societies is illustrated in *Figure 4*, where the size of the points depicts the population size. The political complexity of societies was measured from 1870 to 1960, with an average year of approximately 1922 and a median year of 1920. Furthermore, the population of the societies in this data varies, spanning from 1,000 to 5,500,000 individuals. The mean population across all societies is 202,542, while the median population is 55,000, indicating a substantial variability in the population sizes of these societies. The population of these societies was recorded between 1902 and 1960, with an average year of approximately 1944 and a median year of 1949. The distribution for the years of measurement for population and political complexity is visualised in *Figure 5*.

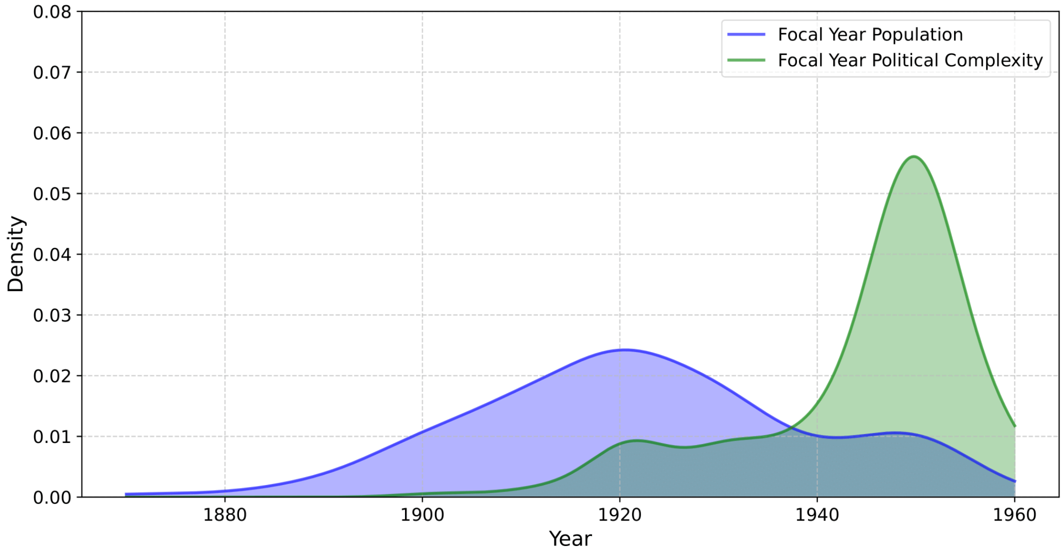


Figure 4: Density Plot of Recorded Years for Population and Political Complexity

As such, the data to validate the model has several limitations. The population and complexity data were recorded in different years, with there being a 29-year difference in medians between the recording of both variables. Societal characteristics can change significantly over time, and as such, this temporal difference could affect the results. Furthermore, the data obtained from *D-PLACE* is not complete, as not every society that existed during this period had both language and political complexity recorded. Nettle (1996) found 708 distinct languages in the region of West Africa, obtaining this information from Moseley & Asher's (1994) linguistic atlas. By this measure, the data at hand only covers 24.4% of the languages in the region. Nonetheless, it provides a reasonable sample size that can offer insights into the dynamics of language diversity and political complexity.

4.5. Model validation

Validating the model ensures that the agent-based model represents real-world outcomes of language emergence and diversity. The results of the simulation will be validated against the data previously described. To validate the model, data will be obtained from the last time step and compared with the empirical data. Language diversity will be calculated using the Shannon index, an ecological diversity measure which has been commonly adopted in linguistic diversity studies (e.g. Grin & Fürst, 2022; Väisänen et al., 2022). It is defined as:

$$S = -\sum_{i=1}^{R} p_i \times \ln(p_i)$$

where p_i represents the proportion of the population that speaks language i and R refers to the number of unique languages (Grin & Fürst, 2022). Higher scores indicate greater linguistic diversity within a given population. As such, language diversity will be calculated for each political complexity level for both the validation and the simulation data. If the model is accurate, the Shannon scores from the simulation should closely mirror the patterns in the empirical data. Consequently, the validation process will address *RQ3 – "How closely do the results mirror real-world distributions of language diversity for different political complexity levels?"*

5. Results and discussion

5.1. Optimal LDN threshold

The model was first run across 30 evenly spaced LDN thresholds between 0.4 and 0.65 to determine the optimal value based on silhouette scores. *Figure 5* depicts the relationship between the LDN threshold values and the corresponding silhouette scores, indicating how changes in the threshold affect the clustering performance. The silhouette scores varied across different thresholds, displaying the most fluctuation between thresholds of 0.400 and 0.503, starting at 0.14 and ending at 0.17. From a threshold of 0.512, the scores declined to approximately 0.16 and remained stable within this range until 0.590. Beginning at 0.598, the scores increased from around 0.16 to 0.18, before experiencing another rise and peaking at 0.624 with a silhouette score of approximately 0.205. As such, the results demonstrate that the optimal threshold for distinguishing languages in the agent-based model is an LDN of 0.624. However, using the optimal LDN threshold resulted in too few languages emerging, which did not align well with real-world observations. This reflects the difficulty in distinguishing languages. To address this issue, a threshold of 0.500 was selected, with this value obtaining a silhouette score of approximately 0.17. As such, using this value balances the need for an adequate number of languages while maintain distinctions between languages.

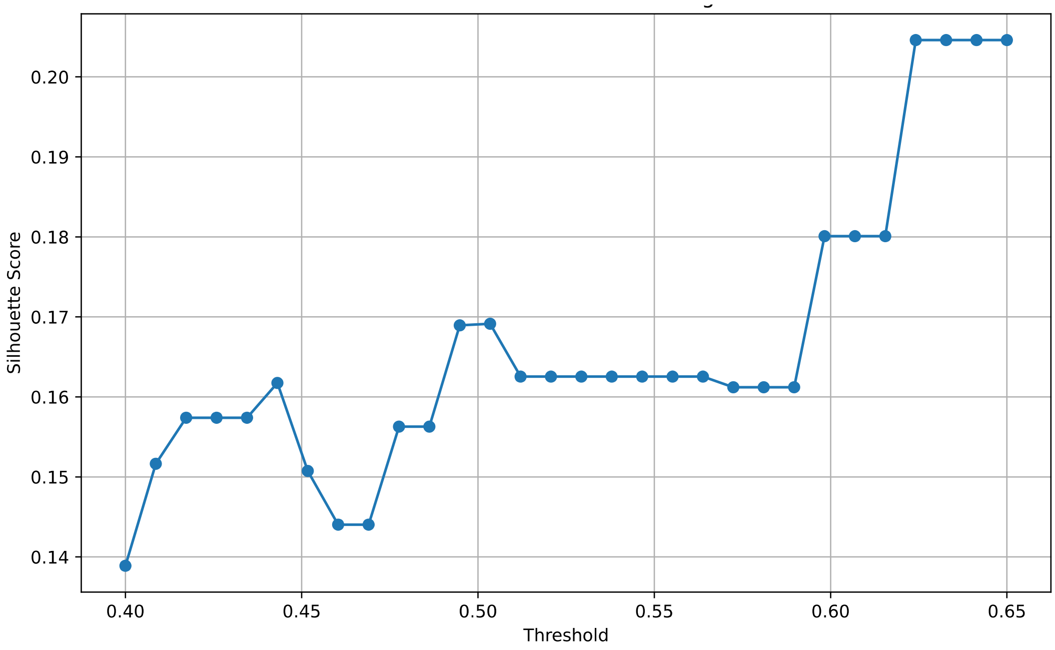


Figure 5: Silhouette Scores for Different LDN Clustering Thresholds

5.2. Language diversity and model validation

To explore the impact of political complexity on language diversity, an agent-based model was run for a period of 7,000 years, simulating the evolution and spread of languages among societies with varying levels of political complexity. As discussed above, an LDN threshold value of 0.500 was used. Throughout the simulation, agents interacted, migrated, and adopted linguistic features based on the political complexity of their respective societies, leading to significant changes in the number and distribution of languages.

The plots in *Figure 6* illustrates the first language cluster distribution, where 74 unique language clusters formed. These spatial distribution maps show the simulated spread and diversity of languages, with each colour corresponding to a distinct language cluster. The right plot displays the location of societies classified by their political complexity level, with one simple chiefdom and fifteen acephalous societies present.

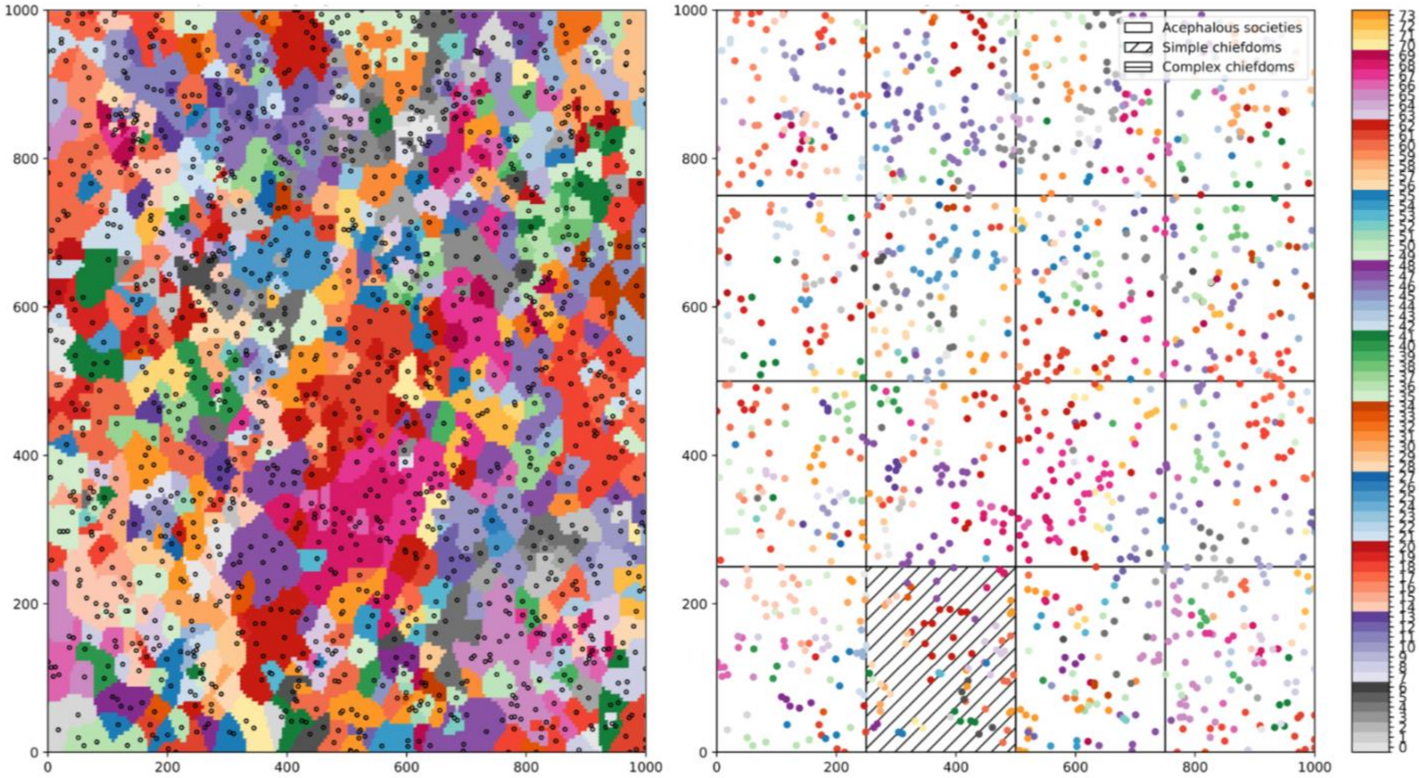


Figure 6: Spatial distribution of language clusters and complexity levels in year 250

Figure 7 illustrates the changes in language diversity and political complexity over a period of 7,000 years. The graph depicts the number of languages and the number of agents within different political complexity levels throughout the simulation.

Furthermore, the final spatial distribution of language clusters and complexity levels for the year 7000 is illustrated in *Figure 8*.

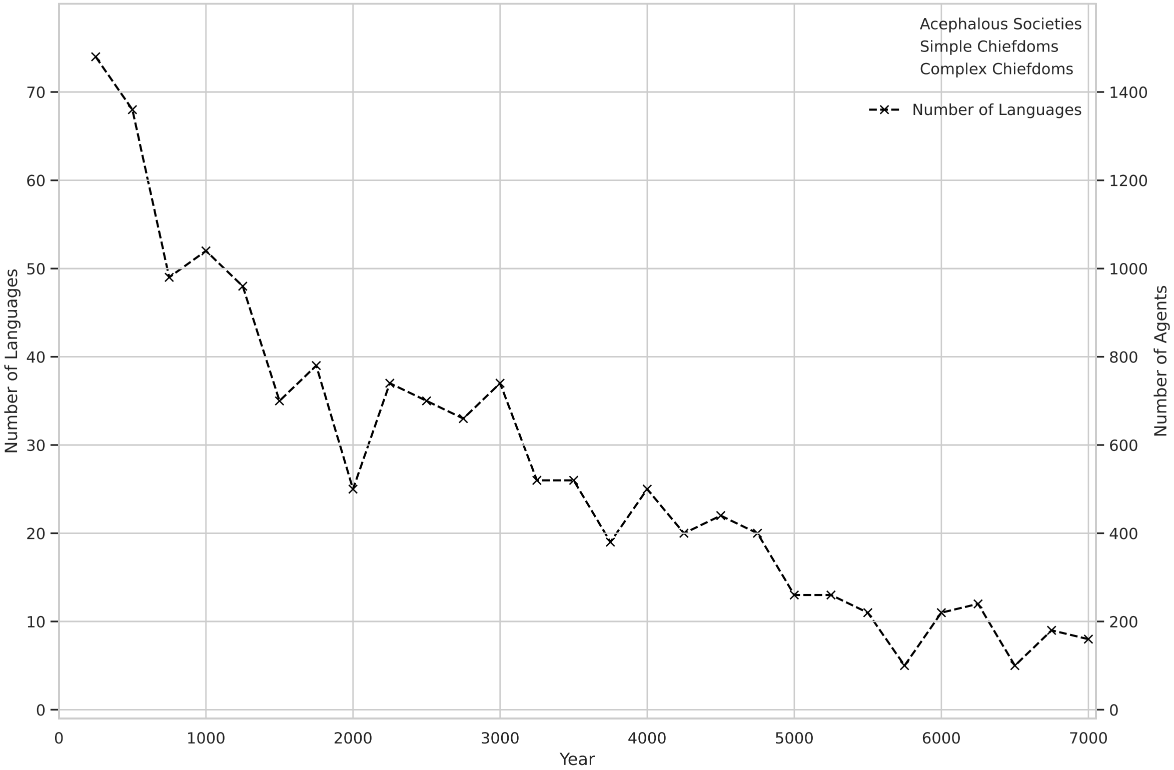


Figure 7: Number of agents per political complexity level and languages every 250 years

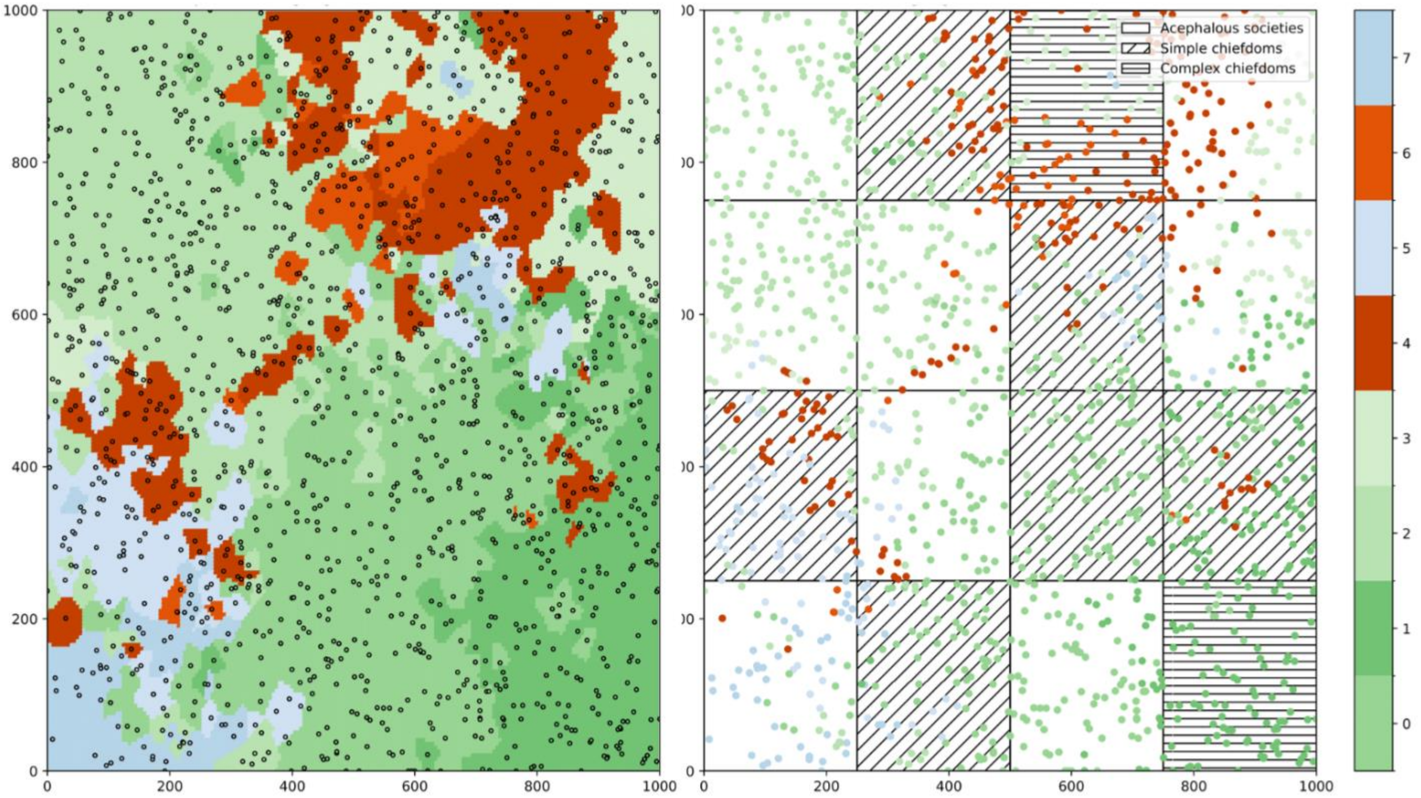


Figure 8: Spatial distribution of language clusters and complexity levels in year 7000

Initially, the simulation shows a high number of distinct languages, corresponding to a large proportion of acephalous societies. As time progresses, there is a large decline in the number of languages which also coincides with the emergence and growth of more politically complex societies. This trend supports the hypothesis by Currie and Mace (2009) that political complexity leads to a reduction in language diversity through processes of cultural group selection. Nonetheless, the initial decline in language diversity may also be attributed to the model stabilising given that agents were initially randomly assigned languages.

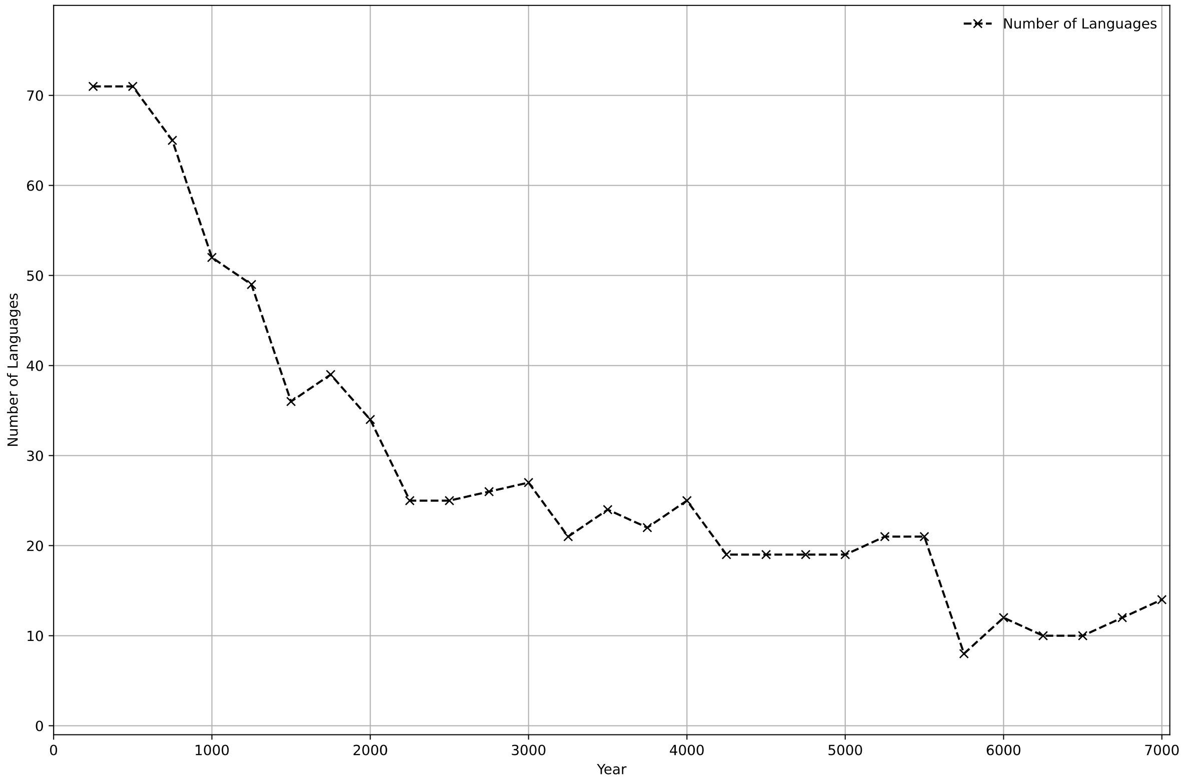


Figure 9: Number of languages every 250 years (without political complexity)

To isolate the impact of political complexity, the model was run again without any of the original mechanisms in place and the absence of political complexity levels, using the same LDN threshold. The number of languages over time for this scenario is illustrated in *Figure 9*. In comparison to *Figure 8, Figure 9* shows the number of languages over time without considering the political complexity levels. Initially, the number of languages is quite high, similar to *Figure 8*. However, without the influence of political complexity, the decline in the number of languages is more gradual and does not reach as low of a number as in the original simulation. This suggests that political complexity and the cultural group selection mechanisms play a role in reducing language diversity. Nonetheless, this cannot be confirmed without running multiple iterations of both models, which was not possible due to time constraints and a lack of computational resources.

To compare the observed language diversity quantitatively, the Shannon index was used, where higher scores indicate greater diversity. These results were recorded in the last time step against real-world data in West Africa between the years 1870 and 1960. However, the agent-based model simulation did not reveal significant patterns in language diversity across different levels of political complexity. *Table 3* presents the Shannon index scores, indicating the level of language diversity for acephalous societies, simple chiefdoms, and complex chiefdoms, both in the simulation and validation data. As previously mentioned, the Shannon index provides a quantitative measure of language diversity, where higher scores indicate greater language diversity.

	Acephalous societies	**Simple chiefdoms**	**Complex chiefdoms**	**Overall**
Simulation	1.89	1.87	1.90	1.88
Validation	3.60	2.93	1.51	3.92

Table 3: Shannon index scores for simulation and validation data

The simulation results show a relatively consistent level of language diversity across different levels of political complexity, with the Shannon index ranging from 1.87 to 1.90. This consistency highlights that the simulated societies maintain a similar degree of language diversity regardless of their political complexity level. In contrast, the validation data, derived from historical records in West Africa, demonstrates more significant variation in language diversity. Acephalous societies exhibit the highest diversity with a Shannon index of 3.60, followed by simple chiefdoms at 2.93, and complex chiefdoms at 1.51. Further, the overall language diversity in the validation data is noticeably higher at 3.92 than in the simulation at 1.88. Though the agent-based model mechanisms in the model did not manage to produce more language diversity within less politically complex societies, the validation results show support for Currie and Mace's (2009) findings which emphasise that more politically complex societies tend to reduce language diversity.

5.3. Limitations

While the results provide insights into language diversity and how political complexity influences it, there are several limitations to consider. Firstly, the model itself has several constraints. Agent-based models generally require greater levels of abstraction to be practical, due to the computation restraints of simulating systems which in reality have large amounts of components (Rhodes et al., 2016). As such, this model offers a simplistic simulation of language diversity, where many interacting components contributing to language diversity have not been included. These include environmental and other socio-cultural factors.

Other rules could have been incorporated in this model to better represent mechanisms which were not initially considered. For instance, politically complex regions are more likely to standardise language for social and governance purposes, with the conscious promotion of language convergence playing a factor in the development of the nation state (Wright, 2016). This standardisation reduces linguistic diversity within their boundaries as dominant languages or dialects are promoted at the expense of minority languages. The absence of a mechanism imitating this is likely the reason behind the lack of language diversity between different political complexity levels as per the Shannon index, as more complex regions do not have a single standardised language for less complex regions to adopt linguistic features. Instead, as observed in the model results, each political complexity type has a multitude of languages which they use.

Additionally, the model fails to simulate geographical changes in regional borders over time, though this was attempted to be mimicked through the migration mechanism. More politically complex regions have the capacity and resources to exert power and influence over less complex neighbouring regions, leading to territorial expansion (Henrich, 2004). Moreover, regions with a higher concentration of agents are more likely to transition to a higher level of political complexity. This reflects the idea that densely populated areas require more sophisticated jurisdictional structures to manage resources and social interactions effectively (Hamilton et al., 2020). As such, future studies could seek to include these mechanisms. Including these additional mechanisms could enhance the accuracy of the model relative to the validation data.

As previously mentioned, the computational requirements of agent-based models lead to constraints in running repeated simulations. Running this agent-based model required access to third-party GPU solutions. Reduced computational constraints would enable the model to be run multiple times, where the average number of languages could be calculated across multiple simulations which would enhance the validity of the model's results.

6. Conclusion

This study set out to explore the impact of political complexity on language diversity using an agent-based model, aiming to test and expand upon the findings of Currie and Mace (2009). Integrating mechanisms of cultural group selection and political complexity into an agent-based model simulating language emergence and evolution has demonstrated their role in shaping language diversity.

The first research question addressed how languages can be distinguished in an agent-based model simulating language diversity. The study combined the use of LDN, divisive hierarchical clustering, and silhouette scores to differentiate languages within the model. This provided a quantitative basis for distinguishing languages within the model. An optimal LDN threshold of 0.624 was provided using silhouette scores, though this resulted in too few languages generated in the model. As such, an LDN threshold of 0.500 was used to strike a balance between creating distinct clusters and simulating enough languages.

Regarding the second research question, which explored how political complexity influences patterns of language diversity, the findings suggest that political complexity through processes of cultural group selection might contribute to a reduction in language diversity. However, this could not be confirmed definitively due to the need for repeated simulations to verify the model's results. Therefore, while the results support the hypothesis by Currie and Mace (2009) to some extent, further research with more iterations and the inclusion of more mechanisms is necessary to confirm these findings.

The third research question examined how closely the results mirror real-world distributions of language diversity for different political complexity levels. The simulation results revealed consistent Shannon index values of language diversity across different levels of political complexity, indicating that there are no differences in language diversity within the model intrinsic to the society type. In contrast, the validation data showed a clear pattern, where more politically complex societies have lower values of language diversity. As previously discussed, this would likely be solved by integrating a language standardisation mechanism within the model.

The limitations identified in this study highlight the need for future research. Future work could explore optimising the model to reduce the computational load or utilising more advanced GPUs to enable more simulation runs. Additionally, incorporating other socio-cultural and environmental factors and other mechanisms discussed in the limitations section could enhance the model which might emulate mechanisms of cultural group selection and political complexity more accurately. Moreover, a sensitivity analysis could be conducted to assess the robustness of the model and how variations in the model parameters influence the model results.

In conclusion, this study has illustrated the potential of using agent-based models in simulating language evolution and diversity. It is also the first study investigating how processes of cultural group selection might facilitate the spread of more politically complex societies' languages over larger areas. Future research should build upon these findings by expanding the model to incorporate additional factors, such as environmental influences and language standardisation. Otherwise, more studies investigating cultural group selection mechanisms and political complexity are required. This will enable linguists and anthropologists to achieve a better understanding of the mechanisms driving language evolution and diversity, and what factors have led to the uneven distribution of language globally.

Bibliography

Abrahamson, M. (1969). Correlates of Political Complexity. *American Sociological Review*, *34*(5), 690–701.

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language Is a Complex Adaptive System: Position Paper. *Language Learning*, *59*(s1), 1–26. https://doi.org/10.1111/j.1467-9922.2009.00533.x

Beeksma, M., Vos, H. de, Claassen, T., Dijkstra, T., & Kemenade, A. van. (2017). A Probabilistic Agent-Based Simulation for Community Level Language Change in Different Scenarios. *Computational Linguistics in the Netherlands Journal*, *7*, 17–38.

Boga, H. I. (2020). What is a Language? What is a Dialect?*. *Studies in the Linguistic Sciences: Illinois Working Papera*, 1–31.

Bromham, L., Hua, X., Fitzpatrick, T. G., & Greenhill, S. J. (2015). Rate of language evolution is affected by population size. *Proceedings of the National Academy of Sciences*, *112*(7), 2097–2102. https://doi.org/10.1073/pnas.1419704112

Carneiro, R. L. (1970). A Theory of the Origin of the State. *Science*, *169*(3947), 733–738.

Civico, M. (2019). The Dynamics of Language Minorities: Evidence from an Agent-Based Model of Language Contact. *Journal of Artificial Societies and Social Simulation*, *22*(4), 3. https://doi.org/10.18564/jasss.4097

Coulmas, F. (Ed.). (2013). Language spread, shift and maintenance: How groups choose their language. In *Sociolinguistics: The Study of Speakers' Choices* (2nd ed., pp. 163–188). Cambridge University Press. https://doi.org/10.1017/CBO9781139794732.012

Coupé, C., & Hombert, J.-M. (2005). Polygenesis of Linguistic Strategies: A Scenario for the Emergence of Languages. In J. W. Minett & W. S.-Y. Wang (Eds.), *Language Acquisition, Change and Emergence: Essays in Evolutionary Linguistics*. City University of HK Press.

Covey, R. A. (2008). Political complexity, rise of. In *Encyclopedia of Archaeology* (pp. 1842–1853). Elsevier. https://doi.org/10.1016/B978-012373962-9.00249-1

Currie, T. E., Greenhill, S. J., Gray, R. D., Hasegawa, T., & Mace, R. (2010). Rise and fall of political complexity in island South-East Asia and the Pacific. *Nature*, *467*(7317), 801–804. https://doi.org/10.1038/nature09461

Currie, T. E., & Mace, R. (2009). Political complexity predicts the spread of ethnolinguistic groups. *Proceedings of the National Academy of Sciences*, *106*(18), 7339–7344. https://doi.org/10.1073/pnas.0804698106

Darwin, C. (1871). *The descent of man, and Selection in relation to sex, Vol 1.* John Murray. https://doi.org/10.1037/12293-000

de Bie, P., & de Boer, B. (2007). An agent-based model of linguistic diversity. *Language, Games, and Evolution*, 1–8.

De Oliveira, V. M., Gomes, M. A. F., & Tsang, I. R. (2006). Theoretical model for the evolution of the linguistic diversity. *Physica A: Statistical Mechanics and Its Applications*, *361*(1), 361–370. https://doi.org/10.1016/j.physa.2005.06.069

Derungs, C., Köhl, M., Weibel, R., & Bickel, B. (2018). Environmental factors drive language density more in food-producing than in hunter–gatherer populations. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1885), 20172851. https://doi.org/10.1098/rspb.2017.2851

Dueppen, S. (2018). The Archaeology of Political Complexity in West Africa Through 1450 CE. In S. Dueppen, *Oxford Research Encyclopedia of African History*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190277734.013.140

Earle, T. (1987). Chiefdoms in Archaeological and Ethnohistorical Perspective. *Annual Review of Anthropology*, *16*, 279–308.

Earle, T. (2011). Chiefs, Chieftaincies, Chiefdoms, and Chiefly Confederacies: Power in the Evolution of Political Systems. *Social Evolution & History*, *10*(1), 27–54.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 429–448. https://doi.org/10.1017/S0140525X0999094X

Freedman, D. A., & Wang, W. S.-Y. (1996). Language Polygenesis: A Probabilistic Model. *Anthropological Science*, *104*(2), 131–137. https://doi.org/10.1537/ase.104.131

Gavin, M. C., Botero, C. A., Bowern, C., Colwell, R. K., Dunn, M., Dunn, R. R., Gray, R. D., Kirby, K. R., McCarter, J., Powell, A., Rangel, T. F., Stepp, J. R., Trautwein, M., Verdolin, J. L., & Yanega, G. (2013). Toward a Mechanistic Understanding of Linguistic Diversity. *BioScience*, *63*(7), 524–535. https://doi.org/10.1525/bio.2013.63.7.6

Gavin, M. C., & Sibanda, N. (2012). The island biogeography of languages. *Global Ecology and Biogeography*, *21*(10), 958–967. https://doi.org/10.1111/j.1466-8238.2011.00744.x

Grin, F., & Fürst, G. (2022). Measuring Linguistic Diversity: A Multi-level Metric. *Social Indicators Research*, *164*(2), 601–621. https://doi.org/10.1007/s11205-022-02934-5

Grinin, L. E., & Korotayev, A. V. (2011). Chiefdoms and their Analogues: Alternatives of Social Evolution at the Societal Level of Medium Cultural Complexity. *Social Evolution & History*, *10*(1), 276–335.

Grinin, L. E., & Korotayev, A. V. (2012). Emergence of Chiefdoms and States: A Spectrum of Opinions. *Social Evolution & History*, *12*(2), 191–204.

Hamilton, M. J., Walker, R. S., Buchanan, B., & Sandeford, D. S. (2020). Scaling human sociopolitical complexity. *PLOS ONE*, *15*(7), e0234615. https://doi.org/10.1371/journal.pone.0234615

Harmon, D., & Loh, J. (2010). *The index of linguistic diversity: A new quantitative measure of trends in the status of the world's languages*. http://hdl.handle.net/10125/4474

Haugen, E. (1966). Dialect, Language, Nation1. *American Anthropologist*, *68*(4), 922–935. https://doi.org/10.1525/aa.1966.68.4.02a00040

Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior & Organization*, *53*(1), 3–35. https://doi.org/10.1016/S0167-2681(03)00094-5

Hua, X., Greenhill, S. J., Cardillo, M., Schneemann, H., & Bromham, L. (2019). The ecological drivers of variation in global language diversity. *Nature Communications*, *10*(1), 2047. https://doi.org/10.1038/s41467-019-09842-2

Karssenberg, D. (2024). *Agent-based model for simulating language evolution [Unpublished model]*.

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Bibiko, H.-J., Blasi, D. E., Botero, C. A., Bowern, C., Ember, C. R., Leehr, D., Low, B. S., McCarter, J., Divale, W., & Gavin, M. C. (2016). D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity. *PLOS ONE*, *11*(7), e0158391. https://doi.org/10.1371/journal.pone.0158391

Kirby, K. R., Gray, R. D., Greenhill, S. J., Jordan, F. M., Gomes-Ng, S., Hans-Jörg Bibiko, Blasi, D. E., Botero, C. A., Bowern, C., Ember, C. R., Leehr, D., Low, B. S., McCarter, J., Divale, W., & Gavin, M. C. (2021). *D-PLACE/dplace-data: D-PLACE – the Database of Places, Language, Culture and Environment* (v2.2.1) [dataset]. [object Object]. https://doi.org/10.5281/ZENODO.5554395

Kosheleva, O., & Kreinovich, V. (2013). Dialect or a New Language: A Possible Explanation of the 70% Mutual Intelligibility Threshold. *Departmental Technical Reports (CS)*. https://scholarworks.utep.edu/cs_techrep/802

Matthews, P. H. (2014). Mutually intelligible. In *The Concise Oxford Dictionary of Linguistics*. Oxford University Press. https://www.oxfordreference.com/display/10.1093/acref/9780199675128.001.0001/acref-9780199675128-e-2153

Mitchell, M. (2006). Complex systems: Network thinking. *Artificial Intelligence, 170*(18), 1194–1212. https://doi.org/10.1016/j.artint.2006.10.002

Moseley, C., & Asher, R. E. (1994). *Atlas of the World's Languages*. Routledge.

Murdock, G. P., Textor, R., Barry, H. I., White, D. R., & Divale, W. T. (1999). Ethnographic Atlas. *World Cultures*, *10*, 24–136.

Nerbonne, J. (2010). Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1559), 3821–3828. https://doi.org/10.1098/rstb.2010.0048

Nettle, D. (1996). Language Diversity in West Africa: An Ecological Approach. *Journal of Anthropological Archaeology*, *15*(4), 403–438. https://doi.org/10.1006/jaar.1996.0015

Nettle, D. (1998). Explaining Global Patterns of Language Diversity. *Journal of Anthropological Archaeology*, *17*(4), 354–374. https://doi.org/10.1006/jaar.1998.0328

Nordhoff, S., & Hammarström, H. (2011). *Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources*. First International Workshop on Linked Science 2011 - In conjunction with the International Semantic Web Conference (ISWC 2011). https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_1752673

Pacheco Coelho, M. T., Pereira, E. B., Haynie, H. J., Rangel, T. F., Kavanagh, P., Kirby, K. R., Greenhill, S. J., Bowern, C., Gray, R. D., Colwell, R. K., Evans, N., &

Gavin, M. C. (2019). Drivers of geographical patterns of North American language diversity. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1899), 20190242. https://doi.org/10.1098/rspb.2019.0242

Pagel, M. (2000). The History, Rate and Pattern of World Linguistic Evolution. In C. Knight, J. Hurford, & M. Studdert-Kennedy (Eds.), *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form* (pp. 391–416). Cambridge University Press. https://doi.org/10.1017/CBO9780511606441.023

Renfrew, C. (1987). *Archaeology and Language: The Puzzle of Indo-European Origins*. Cambridge University Press.

Renfrew, C. (1991). Before Babel: Speculations on the Origins of Linguistic Diversity. *Cambridge Archaeological Journal*, *1*(1), 3–23. https://doi.org/10.1017/S0959774300000238

Renfrew, C. (1994). World Linguistic Diversity. *Scientific American*, *270*(1), 116–123. https://doi.org/10.1038/scientificamerican0194-116

Rhodes, D. M., Holcombe, M., & Qwarnstrom, E. E. (2016). Reducing complexity in an agent based reaction model—Benefits and limitations of simplifications in relation to run time and system level output. *Biosystems*, *147*, 21–27. https://doi.org/10.1016/j.biosystems.2016.06.002

Richerson, P. (1977). ecology and human ecology: A comparison of theories in the biological and social sciences [1]. *American Ethnologist*, *4*(1), 1–26. https://doi.org/10.1525/ae.1977.4.1.02a00010

Richerson, P., Baldini, R., Bell, A. V., Demps, K., Frost, K., Hillis, V., Mathew, S., Newton, E. K., Naar, N., Newson, L., Ross, C., Smaldino, P. E., Waring, T. M., & Zefferman, M. (2016). Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. *Behavioral and Brain Sciences*, *39*, e30. https://doi.org/10.1017/S0140525X1400106X

Ross, M. (2006). Examining the Farming/Language Dispersal Hypothesis (review). *Language*, *82*(3), 628–648. https://doi.org/10.1353/lan.2006.0163

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Shahapure, K. R., & Nicholas, C. (2020). Cluster Quality Analysis Using Silhouette Score. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 747–748. https://doi.org/10.1109/DSAA49011.2020.00096

Soltis, J., Boyd, R., & Richerson, P. J. (1995). Can Group-Functional Behaviors Evolve by Cultural Group Selection?: An Empirical Test. *Current Anthropology*, *36*(3), 473–494. https://doi.org/10.1086/204381

Steels, L. (1997). The Synthetic Modeling of Language Origins. *Evolution of Communication Journal*, *1*. https://doi.org/10.1075/eoc.1.1.02ste

Townsend, C. (2018). Egalitarianism, Evolution of. In H. Callan (Ed.), *The International Encyclopedia of Anthropology* (1st ed., pp. 1–7). Wiley. https://doi.org/10.1002/9781118924396.wbiea1826

Turchin, P., & Gavrilets, S. (2009). Evolution of Complex Hierarchical Societies. *Social Evolution & History*, *8*(2).

Väisänen, T., Järv, O., Toivonen, T., & Hiippala, T. (2022). Mapping urban linguistic diversity with social media and population register data. *Computers, Environment and Urban Systems*, *97*, 101857. https://doi.org/10.1016/j.compenvurbsys.2022.101857

Van Rooy, R. (2020). Mutual intelligibility: The number one criterion? In R. Van Rooy, *Language or Dialect?* (1st ed., pp. 256–262). Oxford University PressOxford. https://doi.org/10.1093/oso/9780198845713.003.0020

Vasyutin, S. A. (2019). Political Complexity in Nomadic Empires of Inner Asia. *Social Evolution & History*, *18*(2). https://doi.org/10.30884/seh/2019.02.05

Wichmann, S. (2020). How to Distinguish Languages and Dialects. *Computational Linguistics*, *45*(4), 823–831. https://doi.org/10.1162/coli_a_00366

Wright, S. (2016). Language Planning in State Nations and Nation States. In S. Wright, *Language Policy and Language Planning* (pp. 47–77). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-137-57647-7_3

Yujian, L., & Bo, L. (2007). A Normalized Levenshtein Distance Metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), 1091–1095. https://doi.org/10.1109/TPAMI.2007.1078