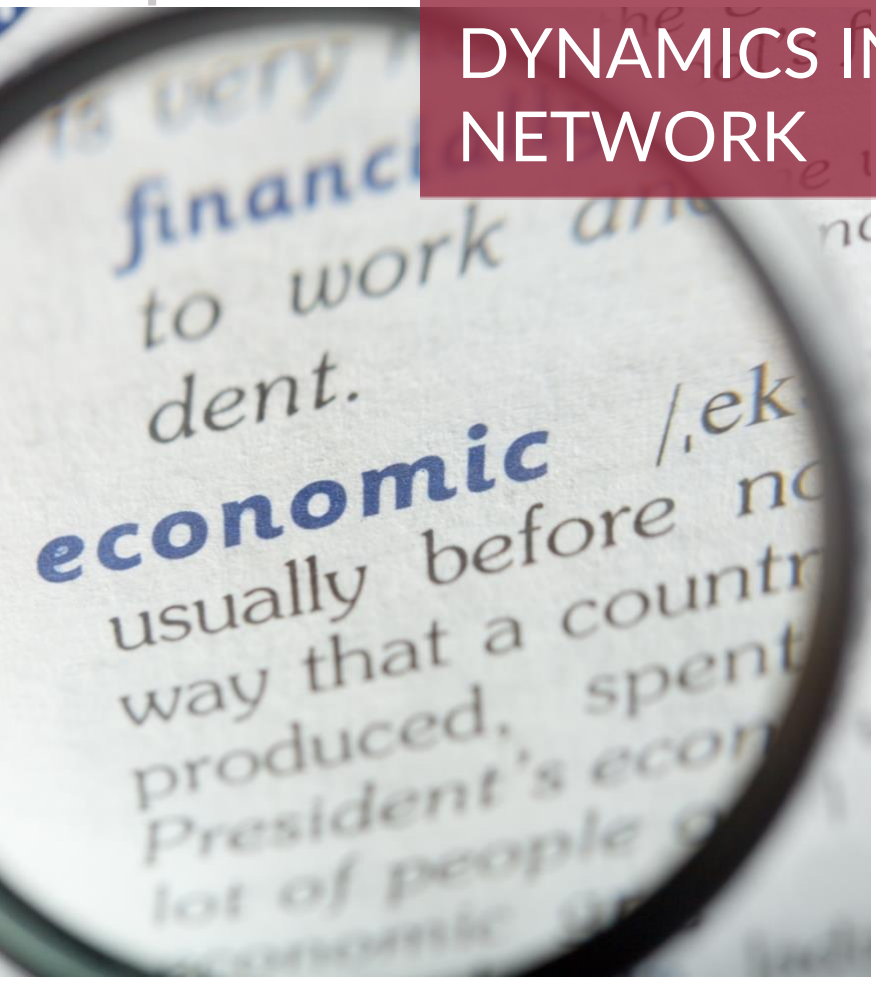
CIRANO
Knowledge into action

A LANGUAGE ECONOMICS PERSPECTIVE ON LANGUAGE SPREAD: SIMULATING LANGUAGE DYNAMICS IN A SOCIAL NETWORK



MARCO CIVICO
FRANÇOIS GRIN
FRANÇOIS VAILLANCOURT

CS
2025s-23
WORKING PAPER

CIRANO
Knowledge into action

The purpose of the **Working Papers** is to disseminate the results of research conducted by CIRANO research members in order to solicit exchanges and comments. These reports are written in the style of scientific publications. The ideas and opinions expressed in these documents are solely those of the authors.

Les cahiers de la série scientifique visent à rendre accessibles les résultats des recherches effectuées par des chercheurs membres du CIRANO afin de susciter échanges et commentaires. Ces cahiers sont rédigés dans le style des publications scientifiques et n'engagent que leurs auteurs.

CIRANO is a private non-profit organization incorporated under the Quebec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the government of Quebec, and grants and research mandates obtained by its research teams.

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du gouvernement du Québec, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO Partners – *Les partenaires du CIRANO*

Corporate Partners – Partenaires Corporatifs	**Governmental partners - Partenaires gouvernementaux**	**University Partners – Partenaires universitaires**
Autorité des marchés financiers	Ministère des Finances du Québec	École de technologie supérieure
Banque de développement du Canada	Ministère de l'Économie, de l'Innovation et de l'Énergie	École nationale d'administration publique
Banque du Canada	Innovation, Sciences et Développement Économique Canada	de Montréal
Banque Nationale du Canada	Ville de Montréal	HEC Montreal
Bell Canada		Institut national de la recherche scientifique
BMO Groupe financier		Polytechnique Montréal
Caisse de dépôt et placement du Québec		Université Concordia
Énergir		Université de Montréal
Hydro-Québec		Université de Sherbrooke
Intact Corporation Financière		Université du Québec
Investissements PSP		Université du Québec à Montréal
Manuvie		Université Laval
Mouvement Desjardins		Université McGill
Power Corporation du Canada		
Pratt & Whitney Canada		
VIA Rail Canada		

CIRANO collaborates with many centers and university research chairs; list available on its website. *Le CIRANO collabore avec de nombreux centres et chaires de recherche universitaires dont on peut consulter la liste sur son site web*.

The observations and viewpoints expressed in this publication are the sole responsibility of the authors; they do not represent the positions of CIRANO or its partners. *Les idées et les opinions émises dans cette publication sont sous l'unique responsabilité des auteurs et ne représentent pas les positions du CIRANO ou de ses partenaires.*

ISSN 2292-0838 (online version)

A language economics perspective on language spread: Simulating language dynamics in a social network

Marco Civico[*], *François Grin*[†], *François Vaillancourt*[‡]

Abstract/Résumé

This paper addresses language dynamics with simulations using an agent-based model (ABM). This model explores language dynamics within a social network. Simulation techniques aim to provide a formalized representation of how factors like language adoption, social influence, economic incentives, and language policies interact, impacting language preferences and fluency over time and, through them, the spread of a language. The ABM developed for this study focuses on complex interactions between agents within a dynamic system. Agents, representing entities that vary according to their level of aggregation (individuals, groups, countries), are endowed with specific linguistic attributes and engage in interactions (communication) guided by predefined rules. A pivotal aspect of our modelling framework is the incorporation of network analysis, where relationships among agents are structured as a network, allowing us to leverage network metrics and measures. The network's dynamic evolution reflects changing inter-agent connections. By combining ABM with network analysis, we gain a nuanced understanding of emergent behaviours and system dynamics, offering insights that extend beyond traditional modelling approaches. This integrative approach proves instrumental in capturing intricate relationships and shedding light on the underlying mechanisms governing complex systems and provides an analytical framework that can be combined with data from sociolinguistic observation.

Ce texte aborde la dynamique linguistique au moyen de simulations basées sur un modèle à base d'agents (MBA). Ce modèle étudie la dynamique linguistique au sein d'un réseau social. Les techniques de simulation visent à fournir une représentation formalisée de l'interaction de facteurs tels que l'adoption d'une langue, l'influence sociale, les incitations économiques et les politiques linguistiques, impactant ainsi l'évolution au fil du temps des préférences linguistiques et de la compétence linguistique et, par conséquent, de la diffusion d'une langue. Le MBA développé pour cette étude se concentre sur les interactions complexes entre agents au sein d'un système dynamique. Les agents, représentant des entités associés à divers niveaux d'agrégation (individus, groupes, pays), sont dotés d'attributs linguistiques spécifiques et interagissent (communication) selon des règles prédéfinies. Un aspect essentiel de notre démarche est l'intégration de relations entre agents structurées en réseau, ce qui nous permet d'exploiter les métriques et les indicateurs de celui-ci. L'évolution dynamique du réseau reflète l'évolution des connexions inter-agents. En combinant MMA et analyse de réseau, nous acquérons une compréhension fine des comportements émergents et de la dynamique du

[*] Researcher, Université de Genève, Faculté de traduction et d'interprétation.
[†] Professor, Université de Genève, Faculté de traduction et d'interprétation and associate researcher, CIRANO
[‡] Emeritus professor Université de Montréal, Faculté des arts et des sciences and Fellow, CIRANO

système, offrant des perspectives qui dépassent les approches traditionnelles de modélisation. Cette démarche intégrative s'avère essentielle pour saisir les relations complexes et mettre en lumière les mécanismes sous-jacents qui régissent les systèmes complexes, et elle fournit un cadre analytique qui peut être combiné avec des données issues de l'observation sociolinguistique.

Keywords/Mots-clés: Sociolinguistics, Networks, Language dynamics, Agent-based Modelling / Sociolinguistique, Réseaux, Modélisation à base d'agents, Dynamique linguistique

JEL Codes/Codes JEL: C63, F15, Z13

Pour citer ce document / To quote this document
Civico, M., Grin, F., & Vaillancourt, F. (2025). A language economics perspective on language spread: Simulating language dynamics in a social network (2025s-23, Cahiers scientifiques, CIRANO.) https://doi.org/10.54932/QXAH2054

1 Introduction

The expansion or spread of languages, as well as their decline or attrition, are all manifestations of language dynamics. In this paper, we use the term "dynamics" to refer to a quantifiable change in the value of variables that index the position of a language relative to other languages. The variables most commonly considered when describing these dynamics are demolinguistic, such as the absolute number or proportion of individuals in a given territory (which may be a region, country, grouping of countries, or the world as a whole) who are able to use the language with some degree of fluency. Sometimes, reference will be made to patterns of language use rather than language skills, whether in general or in specific domains (e.g. economic activity, literature, audio-visual media), or to indicators of a language's visibility in the public space or the Internet.

The meaning of the adjective "dynamic" and the noun "dynamics", when used in the sociolinguistics or applied linguistics literature, is often imprecise, variously evoking movement, change, or agency by social actors. In this paper, "language dynamics" refers to the interplay of causal processes whereby the value of certain linguistic variables, such as, for example, the demolinguistic weight of a language relative to others, changes from time t to time $t + 1$. In essence, dynamics therefore refers to a change over time, not just as a once-and-for-all variation, but as a process that goes on over many successive periods.

Interestingly, there is substantially more literature, whether in sociolinguistics or economics, about language decline, often under the name of language shift, than about language spread. The reason for this imbalance may be that the decline of a language is generally considered as problematic, because it may imperil its intergenerational transmission and result in what is often referred to, metaphorically, as "language death". Language *spread*, by contrast, is often seen as an essentially trouble-free mechanism. After all, if more individuals learn and use another language, where is the harm? Only bigoted xenophobes would worry about the spread of a language.

The hitch, of course, is that language spread and language decline, though not strictly correlated, are linked in complex ways. Even if a language spreads because more individuals learn it without relinquishing other parts of their linguistic repertoire (thus exemplifying what is known as additive bilingualism), it does not mean that there is no risk of language displacement, or eviction, particularly in the long term. As we shall see when taking a closer look at the dynamics of language spread, much depends on its nature and manner.

Although the language sciences are the disciplines that are chiefly interested in the spread and decline of languages, the sociolinguistic research into language shift, decline, and attrition tends not to foreground general explanations of these processes. In the main, and with notable exceptions such as Fishman (1991) or Ó Curnáin and Ó Giollagáin, (2023), the sociolinguistic literature has tended to descriptive or interpretive, rather than analytical and predictive, accounts of language decline or shift. These accounts are often case studies, which may offer very cogent explanations of the changes that characterize a given language (e.g. Haugen and McClure, 1982), but they do not provide a general theory of why some languages are spreading and others receding. Alternatively, when they focus on the linkage between events, thus suggesting a focus on causal dynamics, the emphasis is usually placed on a specific chain of events in a given context, which may insightfully highlight parts of the overall process, but does not amount to an integrated, let alone predictive theory.

What is true of accounts of language decline applies all the more to language spread, about which contributions are fewer. This paper therefore proposes an exploration of the dynamics of language spread. We do so using the technique of agent-based modelling. In Section 2, we discuss what may be understood by, and expected from an economic perspective on language dynamics, and explain why novel conceptual and theoretical work is needed. In Section 3, following a short literature review, we present the general rationale for modelling, before explaining our reasons for opting for an agent-based approach. In Section 4, we develop the ABM, whose main strength lies in the conditional predictions that it enables us to make: under given conditions, a language is likely to spread more or less quickly; under another set of conditions, it is not; the challenge then is to identify these conditions. Section 5 links up our results with the language policy, in order to show how a better understanding

of language dynamics can be useful in the selection, design and evaluation of language policies. The systematic examination of the process of language spread indicates that for the latter not to result in the eviction of other languages, it must be carefully curated and accompanied by well-designed language policies.

2 On economic activity and its connection with language dynamics

From outside the specialty, formulating an economic perspective on language spread may seem like a self-evident proposition: aren't patterns of language spread, after all, *obviously* related to economic factors? Indeed, it is highly plausible that there are numerous connections between the spread of some languages and historical developments such as colonial ventures which, in turn, have manifestly economic dimensions. However, this assumption can be misleading. To pre-empt possible confusion, it is important to distinguish between two things: on the one hand, the study of economic activity as it is manifested in production, trade, exchange, and consumption; on the other hand, fundamental economic analysis, which ultimately focuses on causal explanations driven by economic processes.

Surprising as it may seem, the economic literature, including the specialist literature in language economics (Vaillancourt, 1985; Grin, Sfreddo and Vaillancourt, 2010; Gazzola and Wickström, 2016; Ginsburgh and Weber, 2016) is relatively sparse when it comes to relating language spread with economic processes in either of these two perspectives, although several papers offer important cornerstones. Some contributions (e.g. Holden, 2016; Egger & Toubal, 2016) comment on the role of language or languages in connection with historical accounts of the development of international trade, but they do not provide a general explanation of how economic activity affects the fortunes of different languages, nor do they offer an economic theory of language spread itself. In fact, economists are very careful about the claims they make (or abstain from). For example, John (2016) devotes some 20 lines to "how economic forces can influence language dynamics"; he rightly observes, in a section devoted to "feedback mechanisms" that

> *"linguistic factors affect economic decisions in many ways. Many of those decisions can be expected to have positive feedback influences on language dynamics: relatively dominant languages are likely to influence economic choices in ways that further increase the dominance of those languages. Languages that are 'important' or of high value will tend to attract more speakers and thus become yet more important and more valuable" (John, 2016: 103).*

In the same way, Mélitz (2016: 602 ff.) ventures five propositions about language spread, of which only one refers to economic determinants, and his claim remains extremely general and conjectural ("larger trade with speaker of [this] language should do the same" [i.e. "make the language more attractive to learn"]).

These perfectly valid points, however, rather than providing a full-fledged theory of how such-and-such a component of economic activity affects language, or an economic theory of language spread, essentially amount to a call for such questions to be investigated. The contribution that offers the analytical broadest scope, sitting astride the dynamics of language spread and the dynamics of language change, is probably the one by John and Özgür (2020), who propose to view language as an engine of growth, enabling actors who can communicate easily through a shared language to achieve higher prosperity; this can help to explain a certain distribution of the world population between language communities of different sizes. However, their approach focuses on the effects of large-scale forces in the very long term rather than on the processes whereby economic considerations influence actors' language choices (and thus *result* in patterns of language spread).

Observing, in a given case, a plausible connection between economic activity or economic incentives or constraints on the one hand, and language dynamics on the other hand would be an encouraging start. But even if this plausible connection were illustrated by a statistically robust correlation, it would not be sufficient to conclude that we have an economic theory of language dynamics. It is important to beware of anecdotes and one-off examples, particularly in the absence of multivariate quantitative treatment. They induce the classical risk of *affirming the consequent*, that is, of interpreting a given observation pertaining to language dynamics as the result of a particular event or force pertaining to economic activity, when in

fact the observed language dynamics could just as well be explained by another cause altogether. In order to clinch this point with an example, let us simple note that little in the way of actual insights about the role of economics in language spread would be gained by vague claims in the vein of "the economic weight of the USA has resulted in the English language spreading around the world", or by descriptive (and nonetheless speculative) statements such as "the spread of Portuguese has benefitted from the fact that Portuguese crown was awarded the eastern half of the South American continent by the Treaty of Tordesillas in 1494."

In light of the above, there is no quick and easy path to the formulation of a general, consistent economic investigation of language spread; addressing this topic calls for fundamental analytical work, and such an enterprise would require much more than a reasonably sized paper. We have therefore opted for an alternative strategy, in the form of the development of an ABM, which is presented in the following section.

3 Language dynamics and modelling

The formal modelling of language spread, as noted by Pool (1991), offers unique advantages compared to non-formal modelling, essentially by compelling us to derive, at a high level of generality (as opposed to idiosyncratic situations), the implications of the claims made, and hence, to pay particular attention to logical consistency. Pool emphasizes the idea that a model's primary virtue is not to reflect reality but to single out some essential features of reality, in order to build an analytical instrument that helps us understand, reflect, and ultimately act upon reality. This matters when the issue at hand is one as complex as language spread.

Let us begin by defining what we mean by "language spread". A vast array of indicators may be suggested (Gazzola and Iannàccaro, 2023; Grin and Gazzola, 2013). As pointed out above, language spread is usually approached through the increase in the value of some demolinguistic figure over a given period. It may be specified, for a given territory, as:

- the number or percentage of native speakers of *X*;
- the number or % of more or less fluent (level of proficiency) users of *X* (as a first, second or foreign language).
- the frequency of use of language *X* across all or specific (education, work, etc.) domains;
- the visibility of language *X* in various public spaces (e.g. advertising, signage);
- various indicators of the status or prestige of language *X* relative to other languages.

This list is obviously not closed. The absence of a general theory and the lack of robust data probably explain (along with the fact that the spiral of decline is a greater concern than the process of spread) why the study of language spread has tended to focus on basic demolinguistic indicators as the dependent variable (Templin and Wickström, 2023).

Though often primarily concerned with possible applications to minority languages facing attrition, several contributions are formulated in terms of more general dynamics (Abrams and Strogatz, 2003; Castelló, Loureiro-Porto, and San Miguel, 2013; Grin, 1992; John, 2016; Selten and Pool, 1991; Wichmann, 2008; Wickström, 2005). As explained by Templin (2020), they mainly draw their inspiration from physics, biology or economics. In essence, they all model the choices made by actors at time *t* to use language *X* or language *Y*, given certain behavioural assumptions and under certain constraints.

Some rest on a classical economic approach (e.g. Grin, 1992), in which actors are assumed to be driven by the goal to maximize their "utility" (or satisfaction), which they do by performing a certain range of activities; the latter may take place in language *X* or language *Y*; accordingly, a certain amount of time must be spent using either *X* or *Y* to perform these activities, and surrounding circumstances, which are a form of constraint, can make it more or less difficult to do one or the other. Therefore, the amount of time that actors devote to *X*-language activities create a more or less *X*-ish linguistic environment that influences individuals' linguistic choices at time *t+1*.

Other models (e.g. Wickström, 2005) focus on the linguistic composition of adult couples, which may be purely random and reflect an initial demolinguistic distribution between *X*-speakers, *Y*-speakers and

bilinguals, or incorporate a degree of preference for linguistically compatible partners. The dominant language in a couple becomes the primary language of their offspring. This provides a causal framework for intergenerational transmission, which may be enriched by including several other variables such as the role of language acquisition in schools, adult language learning, and migration patterns. Ultimately, these objectives, behavioural rules and constraints operating at time *t* determine intergenerational transmission and, hence, the number or percentage of speakers at time *t+1*.

Clearly, the choice of variables depends on the core concern of the analysis, such as clarifying the conditions for successful policy intervention aimed at protecting and promoting threatened languages (e.g. Minett and Wang, 2008; Templin et al., 2016). Clingingsmith's model stands out in that its focus is on large rather than small languages. He starts out by asking "Why do all individuals not speak the same language?" (Clingingsmith, 2017, p. 143). Indeed, there are substantial advantages to linguistic uniformity, mainly that in the absence of language diversity, there is no need to bridge an interlinguistic gap. The validity of this notion, however, crucially depends on the nature of linguistic uniformity and how it is achieved. Let us leave aside for now the fact that focusing on the savings that result from linguistic uniformity requires us to ignore the material and symbolic adjustment costs befalling, in a transitory period, the entire world population, minus those whose native language is retained in that monolingual world. This entails a massive, uncompensated transfer of resources to the benefit of the latter, an imbalance that can only be avoided if a constructed language like Esperanto is used, or possibly an ancient language which nobody speaks as an L1. These advantages largely rest on what is known as "network externalities": suppose that speakers of languages *Y* and *Z*, having carefully weighed all their pros and cons (encompassing material and symbolic aspects), decide that switching to language *X* delivers a net benefit. Once they make the switch, they reap this benefit; but at the same time, they create a benefit for people who already spoke *X* and who can now, without having incurred any additional expenditure, communicate directly with members of the (formerly) *Y*- and *Z*-speaking communities (Church and King, 1993; Katz and Shapiro, 1985).[1] This phenomenon works to the advantage of large languages, generating the prediction that they should spread, eliminating smaller languages along the way, until only a few are left, and ultimately only one.

Clingingsmith (2017) shows, however, that because of the importance of direct human interaction in patterns of language choice and language use, this quasi-mechanical advantage stops operating beyond a relatively low threshold of about 35,000 speakers. The intuition behind this result is that the communicational gains that flow from being part of a larger speech community tend to fade past that threshold; it follows that the dynamics of growth beyond 35,000 speakers are primarily encouraged by factors other than sheer demography. This suggests that patterns of language spread are largely influenced by other factors, some technological (particularly in connection with the use of ICTs), some economic (such as the growth of the share of international trade in world GDP) and some political. The latter may take various forms, ranging from hard geopolitical domination to the (often deliberate) exercise of soft power (Phillipson, 1992, 2003, 2010). There is broad consensus around the fact that such factors are present, but the complexity of the interplay between them as well as numerous additional factors, such as the provision of language instruction in the education system, or patterns of language transmission in bilingual families, is such that there is still no general theory of language dynamics.

This complexity poses a classical dilemma. On the one hand, researchers may bravely try to do justice to the intricate interplay of a factors influencing language dynamics, but they are soon confronted with the near impossibility of drawing analytically robust, but at the same time general conclusions.[2] On the other hand, precisely in order to circumvent this difficulty, one may opt to sacrifice some of the complexity and pare down the analysis to a few essential forces; this strategy is scientifically sound, but it requires making all kinds of assumptions which often end up weakening the analysis.

However, agent-based modelling offers a third strategy for overcoming some of these difficulties. In a nutshell, its principle is the following: just as in classical approaches, we model reality in terms of variables

[1] Note, however, that there are occasional exceptions to this general principle. For example, the value of a little-known language that can be used as a code to transmit information, as has been the case for Navajo during WWII, goes down if more people learn it.

[2] In classic modelling, where the interplay of factors is frequently formalized with equations, this usually means that is difficult, or impossible to "sign the effects" (positively or negatively), that is, to conclude whether a given change in a variable of interest (for example, the median income of members of the *X*-speaking community) will ultimately result in an increase or a decrease in the demographic weight of the *X*-language community.

and relationships between them, thus producing a system of equations which is nothing but a stylized representation of reality; but then, we do not attempt to solve the system at a general level and try to predict in what direction a change in some explanatory variable will, *in general*, affect a variable we seek to explain. Instead, we run a (very) large number of computer simulations with a range of plausible values for the explanatory variables and observe what values the "explained variables" end up taking, as a result of the dynamic interplay among the many variables featured in the model.

The agent-based approach can therefore be seen as offering an alternative route. It is a deliberately abstract approach and by using it, we do not presume to settle the many unsolved questions that surround processes of language spread. Rather, our aim is to identify some crucial facets of the patterns of spread that can emerge from the combination of numerous explanatory factors of language dynamics.

It is intuitively clear and generally accepted by most scholars that in language dynamics, several factors play a part and interact in complex ways. The combination of these factors materializes through the interaction between agents who meet in networks of language users. Agent-based modelling helps to shed light on the interactions through which some languages end up spreading and gaining ground relative to other languages.

4 The model

In this section, we present the details of the ABM, in which the notion of a *network* of speakers plays a prominent role in the exploration of the dynamics of spread. Let us note that the structuring processes highlighted in this model focus on what happens in the short term or mid-term, as a result of individual actors' language-related choices, as distinct from the emphasis on long-term or even very long-term processes proposed, respectively, by Civico (2019) and by John and Özgür (2020). This strategy enables us to highlight and study the effects of key factors in these dynamics in a way that would be difficult or impossible with other analytical approaches.[3] First, we describe the network of speakers of a language (i.e. our simulation environment), highlighting how this network is created in the model setup, and how certain properties (in particular those that pertain to linguistic behaviour) are initially assigned to the agents.[4] Next, we detail the simulation loop. This includes the rules that govern interactions among agents and how they update their behavior to adapt to changes in their linguistic environment. Essentially, we outline how agents communicate and influence each other linguistically within the simulation. Finally, we describe the main trends resulting from the simulations.

4.1 The network

The first step is to create a network of agents.[5] In plain terms, a network is made up of a certain number of individuals, called "nodes" in network theory. Each agent has specific characteristics. Then, we postulate connections, called "edges" in network theory, between them. Three languages, a number allowing for some diversity yet not too large for simulation purposes, are spoken in the simulated world where our agents live: Alphish, Betish, and Gammish. Because we assign native languages randomly, we expect that each language will be spoken by about one-third of the population. This could be a country with three language communities or three separate countries, each with a different language.

Each node represents an agent who has attributes related to their language skills (including their fluency), their preferences regarding languages, and their financial status. In the model setup, before the simulation begins, each agent is assumed to have full proficiency (fluency = 1) in one of these three languages, selected at random. This language is considered the agent's *native* language. Additionally, each agent is assigned a random level of fluency in the other two languages, with values ranging from 0 (no knowledge) to 1 (full proficiency).[6] These attributes determine how agents connect with other agents.

[3] The model is written in the Python programming language. The complete code can be found in the following repository: https://www.comses.net/codebases/f8590435-ed56-4364-83f0-2e5ffee7c558/releases/1.0.0/

[4] The word "agent", incidentally, could be replaced by the word "actor", which is more commonly used in the social sciences; however, "agent" is the standard term in the discipline, and we shall keep using it in this paper.

[5] The network is developed using the NetworkX library (Hagberg, Swart, and Chult, 2008).

[6] The level of fluency is selected at random from a uniform distribution, but other distributions can be considered. For example, one may want to try to select fluency according to a normal distribution centered around a specific value of fluency.

Agents also have a *preferred* language for communication. This is the language they prefer to use in conversations, though they can switch to another language if necessary and if their fluency allows it. Initially, the preferred language for each agent is their native language, but it may change over time. Additionally, each agent starts with the same amount of capital, representing their socioeconomic status.

Connections between agents are based on their linguistic profile. Agents who share the same native language are connected with probability p, while agents with different native languages are connected with probability q. Both probabilities (p and q) can be adjusted when creating the network. By structuring the model in this way, we can observe how agents with different language skills interact and form communities, which helps us understand the dynamics of language competition and coexistence in a society.

4.2 Agent interactions

When the simulation loop is launched, agents start interacting with each other in pairs. At each step of the simulation, all agents have the opportunity to interact with every other agent. During each step, an agent is assumed to perform a number of actions.

1. *Building new connections*: an agent may create new connections ("edges") with other agents. This happens with a certain probability and only if both agents are fluent in at least one shared language. The probability of creating a new connection and the fluency level required for this is set before the simulation starts.
2. *Updating the preferred language for communication*: an agent may update their preferred language for communication. The new preferred language will be the one in which the agent's neighbours (the agents he or she is directly connected to) have the highest average fluency, as long as the agent is fluent enough in that language. If the agent's current preferred language already matches this, no change occurs.
3. *Starting conversations*: Each agent tries to start a pairwise conversation with his neighbours. By default, this happens according to the following conditions:
    a. if two agents share the same native language, they use that language for their interaction;
    b. if two agents do not share the same native language but have the same preferred language for communication, they use that language;
    c. if neither of these conditions is met, they look at the other languages in which both are fluent and choose the language in which they have the highest average fluency;
    d. if they have no language in common in which they both have an adequate level of fluency, no conversation happens.[7]
4. *Updating language fluency*: after each conversation, agents update their language fluency in the language used for communication, increasing their fluency in that language. Conversely, their fluency in unused languages (including their native language) will decrease.
5. *Removing connections*: if two agents no longer share any languages in which both are sufficiently fluent, the connection between them (if it exists) may be removed with a fixed probability.

Based on these rules, we can highlight several features of this artificial society:

1. *Changing language preferences*: while they continue to use their native language with members of their own language community, they might prefer a different language when interacting with speakers of other languages. This change depends on their fluency in the other languages and the average fluency within their sub-network (the part of the network made up of an agent's connections).
2. *Feedback loops*: as the simulation progresses, feedback loops quickly develop. Using a particular language increases an agent's fluency in that language. This, in turn, raises the average fluency within their network, influencing their preferred language for communication.[8]

[7] Since new connections are only formed between agents who share at least one language in which they are fluent, this occurrence is relatively rare. However, it can happen between agents whose connection was established during the initial setup phase, before the simulation began.

[8] We shall note in passing that feedback loops are a characteristic feature of complex systems, where the output of a process influences the operation of the process itself, either amplifying it (positive feedback) or dampening it (negative feedback). In social systems, this can lead to self-reinforcing behaviors or equilibrating dynamics, which are essential to understanding how such systems evolve over

3. *Integration and isolation*: an agent who becomes more integrated into a group with a specific preferred language is more likely to lose fluency in other languages. Over time, this can lead to losing connections with agents in other language groups.

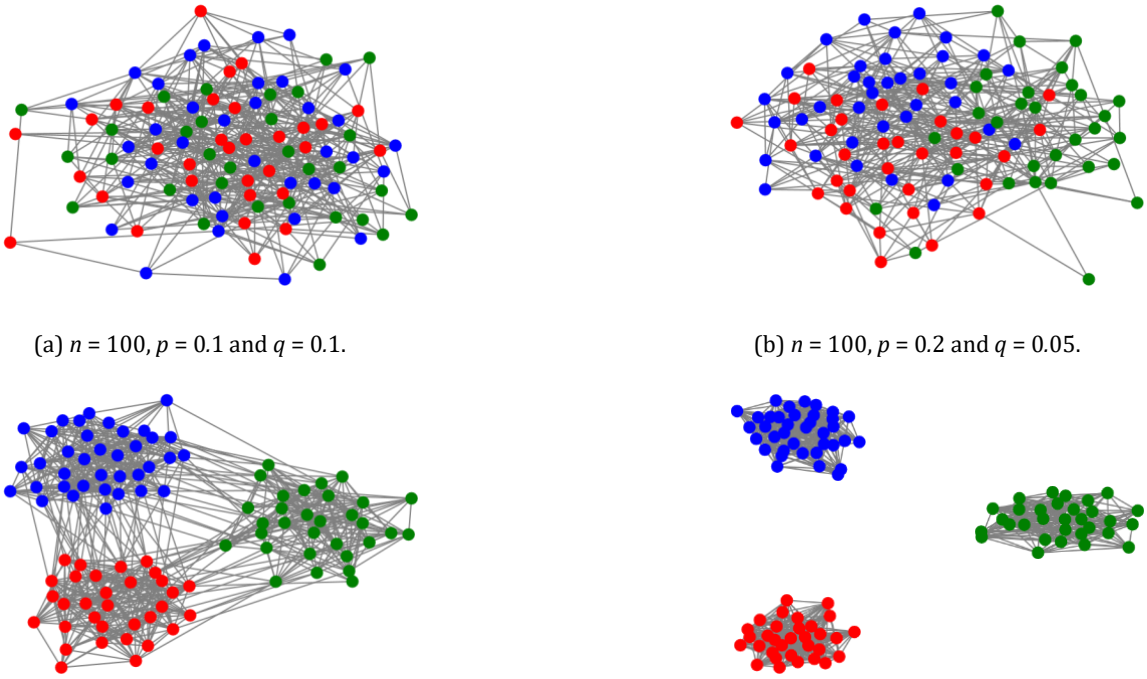
By observing these interactions, we can gain insights into how language preferences and fluency evolve in a simulated society, providing a better understanding of the dynamics of language spread. For instance, we can think of an environment where different language communities coexist. Over time, some agents might develop a preference for a language spoken by a larger or more influential group, much as an agent in our simulation might shift their preferred language based on the fluency of their neighbours. This can lead to linguistic *convergence*, where different groups begin to use a common language more frequently, leading to the emergence of a domestic or international *lingua franca*. Conversely, in more isolated communities within the same environment agents might retain their native language as the primary means of communication. This would lead to linguistic *divergence* and less exchanges of goods, services or information.

4.3 Socio-cultural factors

During the simulation, several factors come into play to account for the socio-cultural impact of languages. In the network setup, we use two key parameters, p and q, which represent the probabilities of creating a connection with agents from one's own language group (p) or from a different language group (q). These parameters reflect the interaction potential of sharing a common language. We set q to be less than p, indicating that agents who share the same native language are more likely to connect. This reflects a common real-world scenario where individuals tend to form economic and social bonds with others who share similar linguistic and cultural backgrounds – a concept known as homophily (McPherson, Smith-Lovin, and Cook, 2001). To better understand this, one can imagine a city where individuals who speak the same language often live in the same neighbourhoods, attend the same social events, and form close-knit communities. For instance, in many multicultural cities, one may find neighbourhoods where the residents share a common language and cultural heritage. These communities often have strong internal connections and fewer links with individuals outside their linguistic group. The larger the difference between p and q, the stronger this tendency, meaning that individuals are more likely to make connections within their own language group, similar to how individuals in these neighbourhoods often maintain stronger ties within their heritage community.

Figure 1 illustrates this concept with four networks created under the same conditions (n = 100 agents) but with varying values of p and q. In these networks, the nodes are coloured according to their native language: blue for Alphish, green for Betish, and red for Gammish. These colours help visualize how different values of p and q create societies where communities have varying levels of internal cohesion (connections within the same language group) and external cohesion (connections between different language groups). As the gap between p and q widens from Figure 1a to Figure 1d, language clusters become more distinct. This can be compared to a scenario where, for example, social groups in a city become increasingly insular, interacting primarily within their own cultural and linguistic community and less with others. For example, when p equals q (Figure 1a), there is no distinct language clustering because connections are equally likely across all language groups. Conversely, when q is set to zero (Figure 1d), the three language clusters are entirely isolated from each other, as in a society with strict language-based segregation, where individuals from different language communities have little to no interaction with one another.

time. The presence of feedback loops often leads to emergent properties, i.e. behaviors or outcomes that are not easily predictable from the individual components alone but arise from the interactions within the system (Civico, 2021).



(a) $n = 100$, $p = 0.1$ and $q = 0.1$.

(b) $n = 100$, $p = 0.2$ and $q = 0.05$.

(c) $n = 100$, $p = 0.4$ and $q = 0.025$.

(d) $n = 100$, $p = 0.5$ and $q = 0$.

FIGURE 1: AGENT NETWORKS GENERATED USING DIFFERENT VALUES FOR THE *P* AND *Q* PARAMETERS.

These figures represent the starting conditions of the simulation, showing four potential scenarios before any dynamic process begins. The impact of these initial settings extends beyond the setup phase, influencing the entire simulation. As the simulation progresses, the probability of creating new connections can be adjusted based on whether the agents share the same native language. This adjustment is controlled by a parameter we call *s*, which increases the likelihood of forming new bonds when the two agents share a common native language. This is similar to how, in the real world, a shared language and culture often enhance economic and social bonding, strengthening the ties within linguistic communities while potentially reducing the frequency and strength of cross-community interactions.

4.4 Specific language dynamics

The model simulates the impact of four language promotion mechanisms on the linguistic behaviour of a population. These mechanisms provide incentives for agents to prioritize a promoted language over their preferred one (but not their native language), provided they are sufficiently fluent in it. This mirrors real-world scenarios where government campaigns or societal trends promote a particular language, such as English for international business or a regional language in a specific area with a high population of native speakers. As the simulation progresses, agents gradually improve their fluency in the language promoted, reflecting how market forces and government policies (including educational programs and language courses) can lead to increased proficiency in a targeted language over time. By incorporating these dynamics, the model allows us to explore how different language policies may influence linguistic behaviour and fluency, providing insights into the mechanisms through which languages gain or lose prominence.

The first two mechanisms for language adoption are based on an economic rationale. They assume that language behaviour is linked to the economic power associated with a language, either through the total economic power (GDP, assets, income) of its speakers (Economic 1, or E1) or its average per fluent agent (Economic 2, or E2). In E1, the language with the highest total becomes the economically attractive language, reflecting real-world scenarios where languages like English dominate global business due to the economic power of their speakers. This approach simulates situations where the economic attractiveness of a language group/country induces others to learn and use that language. This is similar to the outcome of

gravity models in international trade, where larger economies trade more among themselves, being attracted to each other. In contrast, E2 focuses on the average economic power of fluent speakers, where the language with the highest average value becomes the preferred one. This reflects scenarios where individuals choose to learn and use a language spoken by wealthier segments of the population to enhance their economic opportunities, a situation of perhaps greater relevance in multilingual countries.

In this model, agents interact if they share at least one language in which they are both fluent; they will interact in the language where their combined fluency is the highest. This mirrors real-world situations where agents choose to conduct business in the language they are most comfortable with, ensuring clear communication and reducing misunderstandings. The probability of an interaction occurring depends on the fluency levels of the agents in the shared language. The simulation examines a set of potential languages, calculating the average fluency for each. If at least one language exceeds a specified fluency threshold, the simulation evaluates whether the interaction will take place. The likelihood of a successful interaction increases with the agents' fluency in the chosen language – the more fluently they can communicate, the higher the chances that an interaction will occur. This approach emulates a simplified, yet realistic representation of economic interactions influenced by language dynamics within the evolving social network, demonstrating how language proficiency not only affects communication but also plays a crucial role in economic opportunities and outcomes within a society.

The other two mechanisms consider demographic factors that influence language use. They identify languages based on two different criteria: either the language with the highest number of fluent speakers, including both native (L1) and second-language (L2) speakers, within the population (Demographic 1, or D1) or the language with the lowest number of speakers who prefer to use it for communication (Demographic 2, or D2).[9] These criteria lead to radically different outcomes, even though they share a similar underlying rationale. The D1 criterion focuses on promoting a language that is already widely spoken fluently by a significant portion of agents. This approach assumes that prioritizing a language with a relative majority of fluent speakers is attractive because it minimizes the time needed for all agents to achieve high fluency in a common language. However, this comes at the potential cost of sidelining other languages. This scenario reflects real-world situations where majority languages gain further prominence due to their practicality and widespread use, either in international institutions or countries. Conversely, the D2 criterion supports the language with the fewest speakers who not only know the language but also prefer to use it for communication. This scenario suggests a context where the focus is placed on supporting or revitalizing a language that is less widespread, as may occur when a deliberate effort is made to promote an endangered language or sustain linguistic diversity.

These demographic scenarios reflect not just language preferences among agents but also the language that would be prioritized in education. Each scenario can be interpreted either as a natural societal trend, where these dynamics arise organically within the population, or as the result of deliberate policy interventions, such as government-backed promotion campaigns or educational initiatives aimed at influencing language use. By distinguishing between these dynamics and potential policy impacts, the model helps us better understand how different factors shape the linguistic landscape of a society, revealing the complex interplay between demographic forces and language policy.

A critical aspect of language dynamics is the role of exposure in language acquisition and fluency. The model accounts for this by allowing agents to improve their fluency in a language through two primary methods: active use in conversations or deliberate learning. Additionally, each agent adapts their preferred language based on the most fluently spoken language within their sub-network, aligning with the dominant linguistic trend in their social circle. Even if an agent is not fully fluent in the most prevalent language within their network, their fluency in that language will still gradually increase through passive exposure. Simply being surrounded by and hearing the language contributes to incremental improvements in their fluency. This mirrors real-world scenarios where individuals begin to understand and use a language more confidently simply by being in an environment where it is frequently spoken, even if they are not actively studying it, or using it regularly. For instance, someone living in a multilingual community might pick up a second language over time due to consistent exposure, even if they are not initially fluent. This aspect of the

[9] Note that in the latter case, the model looks at language preference rather than native language. This choice is made to avoid considering a language with few L1 speakers but many L2 speakers a minority language.

model highlights the powerful influence of social environments on language learning and preference. It shows how language exposure, both active and passive, can drive linguistic shifts within a population, leading to changes in language preferences and fluency levels. These dynamics are particularly relevant in multicultural and multilingual settings, where the language spoken in social networks can significantly impact an individual's language skills and choices.

Overall, the model provides a comprehensive framework for understanding how economic incentives, demographic forces, and language exposure interact to shape linguistic behaviour and fluency within a society. By simulating these dynamics, the model offers valuable insights into the factors that contribute to language adoption and retention, helping us understand the implications for cultural cohesion, social integration, and economic opportunity. Whether through the lens of economic power, demographic trends, or the everyday experience of language exposure, this model helps us explore the complex and often subtle forces that influence the linguistic landscape of a population, including through the phenomenon of language spread.

5 Simulations and results

The model's simulations and results provide a range of insights into the complex dynamics of language spread, fluency, and social connectivity within a population. By simulating various scenarios, the model helps us understand interactions between multiple variables that are otherwise difficult to analyze due to their complexity and interdependence. Additionally, these simulations generate artificial data allowing us to perform regression analyses that enhance our understanding of the factors influencing language behavior. Figure 2 presents, for illustrative purposes, how two simulations with 100 agents, 500 iterations and identical parameters except for the propensity to connect (parameter *s*) may be graphed in the case of a trilingual society. The left-hand side panel shows a clear dominance of Alphish as a language of communication over time while the right-side panel does not display the same evolution.

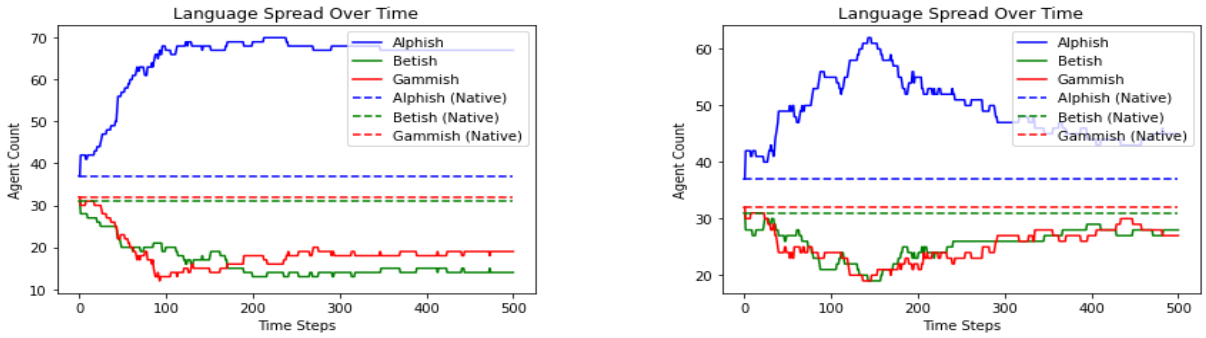


FIGURE 2 EXAMPLES OF SIMULATION PATH AND OUTCOME – LOW (LEFT) VS HIGH (RIGHT) PREFERENCE FOR IN-GROUP CONNECTIONS

The simulations incorporate both fixed and variable parameters such as population size, fluency thresholds, and the effects of economic and demographic factors. By changing these parameters, we can observe how they influence the evolution of language preferences and fluency among agents over time, with ensuing effects on the spread or decline of the languages present. The simulations' primary advantage is that they allow us to explore how complex interactions between numerous variables unfold in a controlled environment. This helps to reveal patterns and trends that would be difficult to identify otherwise.

One significant finding from the simulations is the impact of population size on the speed of language spread, particularly under different assumptions about the role of linguistic commonality in forming social bonds. When the population is small, even a slight demographic advantage can quickly lead to the dominance of one language, especially when agents show only a weak preference for forming connections within their own language community. However, in larger populations, language communities tend to be more self-reliant and resistant to the spread of a single dominant language, suggesting that larger populations provide a buffering effect that helps maintain linguistic diversity. This difference in language

spread is closely related to the persistence of minority language groups and its impact on network connectivity, measured by network density.[10] As minority language groups persist, network density decreases, indicating a more fragmented network structure. This fragmentation occurs because minority groups tend to form tighter, more insular connections, leading to a sparser overall network in their case.

When education strategies are introduced, the impact on language fluency and network structure becomes more pronounced. Under the Demographic 1 strategy, the dominant language rapidly becomes more prevalent, especially when the education policy has a strong influence. This scenario mirrors real-world situations where an education system prioritizes a dominant language, potentially at the expense of minority languages. The resulting increase in network density suggests that promoting a common language facilitates broader communication but also reduces linguistic diversity. Conversely, when the education strategy supports the minority language (Demographic 2), the impact on linguistic diversity is less straightforward. Although the overall fluency in all languages decreases, the decline is less pronounced for smaller languages, ensuring that they maintain a presence within the population. This dynamic is similar to real-world efforts to preserve minority languages through education, where continuous promotion helps sustain a strong proficiency base even if the language is underutilized in daily interactions.

Another critical aspect of the simulations is the relationship between the spread of a language as the preferred means of communication and the resulting fluency in that language. In scenarios without external economic incentives or demographic pressures, fluency levels stabilize around values that reflect the distribution of language preferences in the population. Agents initially start fully proficient in their native language, with fluency in other languages distributed randomly. Over time, as agents specialize in one language based on their social network, their fluency in other languages declines, reflecting real-world scenarios where individuals may lose proficiency in less frequently used languages.

When education strategies are implemented, the dynamics of fluency shift significantly. Promoting the dominant language leads to increased proficiency in that language, while fluency in others diminishes. This scenario is common in global contexts where languages like English are promoted, leading to increased fluency at the expense of other languages. On the other hand, supporting a minority language ensures that its fluency remains relatively high among all agents, even if it is not their preferred language for communication.

Let us note that the familiarity and plausibility of the simulation results provide a form of validation for the model, indicating that it behaves as one would expect based on real-world observations. This consistency gives us confidence that the model accurately captures essential dynamics, allowing us to consider the results of the regression analyses conducted in the subsequent section as more reliable.

Let us now turn to the regression analysis. The variables impacting language choices, along with the three dependent variables, are presented in Table 1.[11] The second column indicates the range of values assigned to these model parameters. Certain parameters are fixed across all simulations. These are the

[10] See Appendix A for the technical definition of "network density." In short, network density is a measure of how many connections exist in a network relative to the total possible number of connections. It quantifies the extent to which agents in the network are interconnected. A higher density indicates a more interconnected network, where a greater proportion of possible relationships between agents actually exist. Conversely, a lower density suggests sparser connections, meaning that fewer of the possible relationships are realized. For example, imagine a community of 10 people, where each person is connected to every other person by some form of social or linguistic interaction. In this case, the network density would be high because nearly all possible connections are present. However, in a larger community of 100 people, where only small groups are interconnected (perhaps due to language barriers or social divisions), the network density would be much lower. This concept is particularly useful in sociolinguistics when analyzing how tightly knit different linguistic communities are within a larger society. For instance, a high-density network in a bilingual community might suggest strong interaction and language exchange between speakers of different languages, while a low-density network could indicate that the groups are more isolated from each other, with fewer opportunities for cross-linguistic interaction.

[11] For variable parameters, the ranges in which they vary are indicated as "first value-final value, step value." For example, a range described as "a-b, c" means that the parameter varies from a to b in increments of c. These variable parameters change across different simulations. For instance, the total number of agents in the population (n) might vary, with simulations being run where n is equal to 50, 100, 150, and so on. Conversely, parameters that are not given a range of values remain constant and only take on the specified value(s). For example, the initial probability of creating a bond with an allophone (q) is set to 0.05 for all simulations, meaning that this parameter does not change across different runs of the simulation. Similarly, r_{ed} only takes on values of 0.001, 0.005 and 0.01 across the simulations.

number of time steps (t), the probability of creating connections between agents with the same native language (p), the probability of creating connections between agents with different native languages (q), the increase in fluency when exposed to a language (r_{ex}), and the increase or decrease in fluency when using or not using a language to communicate (r_c).

Other parameters vary across simulations and are the focus of our analyses. These include the number of agents in the network (n), the minimum level of fluency required for a user to be considered fluent (f), the increase in the probability of creating new connections during the simulation phase when two agents share the same native language (s), and the increase in fluency due to learning (r_{ed}). Additionally, the type of strategy (D and E) adopted to promote a particular language also varies across simulations, with different scenarios exploring the effects of no strategy, economic strategies, and demographic strategies. By analyzing the impact of these variable parameters, we can better understand how changes in population size, fluency thresholds, social preferences, and language policies influence the linguistic dynamics within the society that the ABM simulates.

Parameter	Value range, value step	Description
	Variables unchanged across simulations	
t	500	The number of time steps.
r_{ex}	0.0001	The increase in fluency when exposed to a language.
r_c	0.0005	The increase (decrease) in fluency when using (not using) a language to communicate.
p	20%	The probability of creating a connection between agents who speak the same native language during the setup phase.
q	5%	The probability of creating a connection between agents who do not speak the same native language during the setup phase.
	Variables changing between simulations	
f	0.4-0.7, 0.1	The minimum level of fluency at which a user is considered fluent.
n	50-300, 50	The number of agents in the network.
s	10%-50%, 10%	The *increase* in probability of creating new connections during the simulation phase when two agents speak the same native language.
r_{ed}	0.001 (low), 0.005 (medium), 0.01 (high)	The increase in fluency when learning a language.
D1	Low, medium, high	Demographic 1 promotion mechanism
D2	Low, medium, high	Demographic 2 promotion mechanism
E1	Low, medium, high	Economic 1 promotion mechanism
E2	Low, medium, high	Economic 2 promotion mechanism

TABLE 1: SUMMARY OF THE PARAMETER VALUES USED FOR THE SIMULATIONS.

In essence, regression analysis offers a quantitative measurement of how much the level of a dependent variable (for example, linguistic diversity) responds to changes in the value of each independent variable (for example, the level of language skills at which a person is considered to be fluent in a language), given that all other independent variables remain unchanged (which we may view as "the context").

The total number of simulations conducted in this study results from the combinations of various parameter values and yields a total of 2880 observations. In this section, we examine the impact of three key independent variables, total population size (n), the fluency threshold (f), and the language education and promotion strategy (D and E) along with their effectiveness (r_{ed}), on three dependent variables, linguistic diversity, network density, and clustering. These dependent variables represent different aspects

of the social and linguistic dynamics within the simulated population.

For simplicity, we combine the effect of the r_{ed} parameter (which represents the effectiveness of the education strategy) with the variables *D1 and D2* (the type of education strategy). We then categorize each strategy based on its effectiveness, treating them as distinct factors in the analysis. For example, the strategy "Demographic 1" (*D1*) will be referred to as "Demographic 1 (low)" when the effectiveness (r_{ed}) is set to 0.001, "Demographic 1 (medium)" when r_{ed} is 0.005, and "Demographic 1 (high)" when r_{ed} is 0.01. This categorization shows more precisely how different levels of effectiveness in the education policy influence the outcomes of the model. The results of all three regression analyses are reported in Table 2.[12]

	Estimated change in…		
	linguistic diversity index G_r	**network density**	**clustering coefficient**
(Intercept)	0.611***	0.3386***	0.4156***
Number of agents (n)	0.0011***	-0.0004***	0.0003***
Minimal fluency to communicate (f)	0.2703***	-0.1166***	0.0703***
Probability of connection given same native language (s)	0.0631***	0.0478***	0.1767***
Education strategy favouring the dominant language D1			
D1 (high)	-0.6881***	0.1707***	-0.0782***
D1 (medium)	-0.2975***	0.0826***	-0.0589***
D1 (low)	-0.0411***	0.016***	-0.0152***
Education strategy favouring the minority language			
D2 (high)	0.0641***	0.0549***	-0.0561***
D2 (medium)	0.0561***	0.01634***	-0.0283***
D2 (low)	0.0044	0.0032	-0.0055***
Highest total economic attractivity			
E1 (high)	-0.707***	0.1734***	-0.0784***
E1 (medium)	-0.3677***	0.0932***	-0.0635***
E1 (low)	-0.0488***	0.0171***	-0.014***
Highest average economic attractivity			
E2 (high)	-0.6959***	0.1693***	-0.0807***
E2 (medium)	-0.3332***	0.089***	-0.0609***
E2 (low)	-0.0297**	0.0137***	-0.0136***

TABLE 2: IMPACT ESTIMATES FROM REGRESSION ON DIVERSITY INDEX, NETWORK DENSITY AND CLUSTERING COEFFICIENT.

Let us first discuss the impact of these variables on linguistic diversity (Table 2, column 2). Linguistic

[12] All coefficients are significant at 99.9% (***) or 99% (**), except for one. However, statistical significance should be interpreted with caution. These results are based on data generated by an agent-based simulation model, in which all variation arises from the structure and parameters defined by the modeler. There is no sampling from an external population, and the usual assumptions underlying statistical inference do not apply. In this context, significance levels reflect patterns internal to the model, not inference about real-world relationships. A lack of statistical significance may simply indicate high variance in the model's outcomes for a given parameter setting, rather than the absence of an effect.

diversity is measured by comparing the ratio of the linguistic diversity index at the end of the simulation with the same index at the beginning.[13] The higher linguistic diversity, the smaller the spread of any specific language. We refer to this ratio as G_r. Let us interpret the estimates in Table 2.[14]

- **Intercept**: the value of 0.611 indicates that linguistic diversity tends to decrease moderately over time. Specifically, this value indicates that, whatever the simulation, the level of linguistic diversity is approximately 61% after 500 periods (end point) of what it was at the beginning. This baseline implies that, regardless of other factors, some reduction in linguistic diversity is expected. This context is crucial for understanding how various variables influence the maintenance or decline of linguistic diversity within a population.
- **Population size (*n*):** The positive coefficient of 0.0011 for population size implies that larger populations are better at maintaining linguistic diversity throughout the simulation. If the population of a specific language group goes from 17 to 100 (the initial distribution between language groups is one third of the total population, which goes from 50 to 300), the linguistic diversity associated with the larger value increases by 27.5%[15] compared to the lower value, suggesting that a more substantial population base provides a buffer that helps preserve diverse languages over time. In real-world terms, this could be compared to multilingual cities or regions where large, diverse populations help sustain multiple languages through rich and varied social interactions.
- **Fluency threshold (*f*):** The positive coefficient of 0.2703 for the fluency threshold indicates that setting a higher threshold for fluency before agents adopt a new language supports greater linguistic diversity. This suggests that when individuals need to achieve a higher level of proficiency before they feel confident enough to start using a new language regularly, a broader array of languages continues to be spoken within the community. Thus, going from the lowest (0.4) to the highest (0.7) fluency requirement increases linguistic diversity by 8.1%. This is analogous to societal perceptions of "fluency" in multilingual societies, where individuals may only switch to using a second language in public or professional settings once they reach a sufficient level of proficiency. This cultural expectation helps preserve linguistic diversity, as individuals continue to rely on their native languages or other familiar languages until they meet that perceived fluency threshold. Conversely, less demanding standards in a *lingua franca* abets spread and are detrimental to diversity.
- **Preference for in-group connections (*s*):** The positive coefficient of 0.0631 suggests that a strong preference for in-group connections within the same language community fosters linguistic diversity and slows the spread of a dominant language. When this preference goes from 10% to 50%, this leads to a 2.5% increase in linguistic diversity. While it might seem that encouraging in-group connections would lead to linguistic isolation, it actually helps maintain the presence of multiple languages within the broader social network. This mirrors situations in multicultural societies where tight-knit language communities support the use and preservation of their native languages, helping to sustain a diverse linguistic environment.
- **Language promotion strategies (*D1*, *E1*, *E2*, *D2*):** The negative coefficients for the "Demographic 1" strategy (D1) and the two economic strategies (E1, E2) indicate that these policies, which promote already dominant languages, tend to reduce linguistic diversity to a strikingly, even unexpectedly high extent. The demographic or economic weight of a language leads to a more homogeneous linguistic environment, as illustrated by the global spread of English at the expense of less widely spoken languages, where promotion of a dominant language often leads to the erosion of linguistic diversity. Although this observation is *per se* rather obvious, the model offers a way to gauge the relative importance of these various factors. This observation is borne out by the coefficients associated with the "Demographic 2" strategy, which supports minority languages: it displays positive coefficients (0.0044 to 0.0641), indicating a positive effect on maintaining

[13] See Appendix A for the derivation of this metric.

[14] The intercept indicates the value of the dependent variable in the absence of any change in the independent variables. As regards the influence of the latter, we simply multiply the value of the estimated parameter for each variable by an exogenous change affecting this variable. For example, if population size increases by 300 people, the effect on G_r, as shown in the second row of Table 2 is equal to 300 × 0.0011 = 0,33. Since this figure represents the ratio of final to initial linguistic diversity, the final linguistic diversity would be 33% higher than it would be without this population increase. This implies that larger populations tend to better maintain linguistic diversity over time.

[15] This number results from the multiplication of the estimated impact times the change in population: 0.0011 × (300-50) = 0.275.

linguistic diversity; however, the impact of the high version of D2 is substantially smaller in absolute terms (one tenth) than the impact of the "high" version of D1, E1 and E2. This gives quantitative substance to the uphill character of attempts to halt and reverse language shift, but confirms that under very general conditions, consistent support for minority languages helps sustain a diverse linguistic landscape. Understanding these dynamics is crucial, particularly in contexts such as international trade, where maintaining a diverse linguistic environment can foster cultural exchange and inclusive communication (in which a wider range of languages receive appropriate recognition). The analysis reveals how different factors – such as population size, fluency requirements, and targeted language policies – can either support or undermine linguistic diversity, ultimately shaping the sociolinguistic fabric of a population. By examining these relationships, we gain insights into how to balance language promotion strategies to preserve linguistic diversity while also accommodating the practical needs of communication within large and complex societies.

Let us now discuss the impact of the same independent variables on network density (Table 2, column 3). Network density is a measure of how interconnected the agents in the network are. Specifically, it reflects the proportion of actual connections between agents relative to all the possible connections that could appear in the network. This concept is crucial because it gives us insights into the overall level of interaction and connectivity within a population, which can be particularly relevant for understanding social dynamics, communication patterns, and economic exchanges within a community. Here are the key findings:

- **Intercept:** The intercept term in our analysis suggests that when all other variables are held constant at zero, the estimated network density at the end point of simulations is 0.3386. This indicates that, under baseline conditions, about 33.86% of all possible connections between agents are realized. This provides a starting point to evaluate how different factors influence the overall connectivity of the network.
- **Population size (*n*):** For every increase of 50 individuals in the total population, the estimated network density decreases by approximately 2%. This reduction in network density implies that in larger populations, there are relatively (though not absolutely) fewer connections between agents, leading to sparser interaction networks. This is similar to how in larger cities or countries, individuals may have fewer close connections compared to those in smaller, tighter-knit communities, where interactions are more frequent and networks are denser.
- **Fluency threshold (*f*):** An increase in the fluency threshold is associated with a decrease in network density, with each unit increase in the threshold leading to a reduction in density by approximately 0.0117. A higher fluency threshold means that agents must reach a higher level of proficiency before they are comfortable using a language for communication. This results in fewer interactions, as only those who meet this higher standard will engage in conversations, thereby reducing overall connectivity within the network.
- **Preference for in-group connections (*s*):** The variable *s* shows a positive association with network density. Specifically, for each unit increase in the impact of belonging to the same language group on connection creation, the estimated network density increases by approximately 0.0478. This means that when the value of *s* goes up from 10% to 50%, network density is about 2% higher. Although the impact is modest, it is logical that a stronger preference for in-group connections – where agents are more likely to connect with those who speak the same language – increases overall network density. In real-world terms, this reflects how a shared language within a community fosters closer ties and more frequent interactions.[16]
- **Language promotion strategies (*D1*, *D2*, *E1*, *E2*):** As compared with the situation that obtains in the absence of any language promotion strategies, higher levels of impact from such strategies (that is, D1, E1 and E2, which support or promote the dominant language) result in a higher estimated network density of around 0.17. The policy targeting the minority language (D2) is much less

[16] At first glance, this result may appear to contradict the finding that higher fluency thresholds reduce network density. However, the two mechanisms operate at different levels. The fluency threshold limits *who* can participate in interactions by setting a minimum proficiency requirement: raising it excludes some agents from the network altogether. In contrast, the in-group preference parameter *s* affects *how likely* agents are to connect once they are already part of the communicative pool. A stronger in-group preference concentrates links within language communities, which can increase the density of connections among the eligible agents, even if the total pool is smaller. Therefore, while *f* restricts access, *s* enhances cohesion within the accessible part of the network, explaining why both effects can occur simultaneously without contradiction.

effective. This suggests that promoting an already attractive language – whether because of its demolinguistic dominance or economic influence – encourages more interactions within the network, particularly if the language is already widely spoken or linked to influential social groups. This can be seen in real-world situations where language promotion policies, such as those in education or media, lead to greater interconnectedness among speakers of the language promoted, facilitating interaction within the population, but at the same time abetting the spread of the language. This is, *per se*, unsurprising, but grasping these dynamics is essential, particularly for singling out and weighing their respective influence on aggregate interactions, on the exchange of cultural values, and on the maintenance of social cohesion. Network density provides insights into how connected and cohesive a population is, which can have significant impacts on everything from the spread of information to the efficiency of economic transactions. By analyzing how different factors affect network density, we can better understand the underlying social structures that support or hinder these interactions.

Finally, let us examine the impact of the variables on the clustering coefficient (Table 2, column 4). This metric quantifies the degree to which nodes in the network tend to cluster together. Specifically, it measures the likelihood that two randomly chosen neighbors of a node are also connected to each other. This provides insight into the local structure of the network, indicating the extent of connectivity and cohesion within communities or subgroups.[17] . Here are the key findings:

- **Intercept:** The intercept value of 0.4156 suggests that, on average, about 41% of an agent's neighbors are also connected to each other when other variables are not influencing the model. This indicates a moderate to high level of local clustering or community structure similar to the one illustrated by Figures 1b and 1c within the network, reflecting a tendency for agents to form close-knit groups.
- **Population size (*n*):** Though small, the positive coefficient for population size indicates that as the network grows in size, the clustering coefficient slightly increases. This suggests that in larger networks, there is a growing tendency for nodes to form local clusters or communities. This finding reinforces our previous observations that larger, self-reliant communities are more likely to develop tightly connected subgroups.
- **Fluency threshold (*f*):** The positive coefficient of 0.0703 for the fluency threshold implies that when agents only use a language once they are very fluent in it, they are more likely to cluster together with others who share that language proficiency. This leads to higher overall levels of clustering, as agents tend to form close-knit groups where fluency in a particular language is a common bond.
- **Preference for in-group connections (*s*):** The positive coefficient of 0.1767 for the same-language impact variable highlights the significant role of linguistic similarity in shaping network structure. When agents have a strong preference for in-group connections, they are more likely to form cohesive clusters or communities, characterized by dense interconnections. This underscores how shared language facilitates closer social bonds and community formation.
- **Language promotion strategies (*D1*, *D2*, *E1*, *E2*):** All language promotion strategy-related variables, whether focused on demographic majority (*D1*), economic factors (*E1*, *E2*), or minority language support (*D2*), exhibit negative coefficients, suggesting that they tend to reduce the clustering coefficient relative to the absence of any strategy. This means that when such mechanisms are in place, there is a slight decrease in the tendency of agents to form tight-knit clusters. However, the Demographic 2 strategy, which supports minority languages, is associated with higher clustering coefficients compared to other strategies, as indicated by its relatively smaller negative coefficients. This suggests that while all strategies may reduce clustering to some extent, those that support minority languages do so to a lesser degree. Nevertheless, the results suggest that local clustering is more likely to happen in the absence of any language promotion mechanism.

These findings provide insights into how different factors influence the structure and cohesion of social networks, with implications for the distribution of the population in different language communities – in a word, language spread. The clustering coefficient is particularly relevant in understanding how

[17] See Appendix A for more detail on the clustering coefficient.

communities form and sustain themselves, whether through linguistic similarity, shared fluency levels, or policy interventions. The observation that clustering tends to decrease under various language policies suggests that while such policies may encourage broader communication and integration, they might also disrupt existing close-knit communities. This has important implications for the design and implementation of language policies, as it highlights the trade-off between promoting linguistic unity and preserving local community structures.

6 Conclusion

As noted above, language spread is a complex process, making it hard to circumscribe or define, also for analytical purposes. However, within a closed universe (i.e. a geographic area with a fixed population), it is inversely related to language diversity. The greater the diversity, the less dominance any single language will have. Our empirical findings suggest that:

1. in a closed universe, the larger each language group is in absolute numbers, the more difficult it becomes for one specific language to spread. Consequently, increasing the number of languages in such a setting may facilitate the spread of a dominant language. Therefore, preserving a large number of languages spoken by very few individuals is not an efficient policy if one is trying to avoid linguistic hegemony or spread. Pro-natality policies of reasonably large populations can promote diversity and slow down language spread;
2. the higher the level of fluency expected or required to use a language for communication, the less likely it is that a specific language will spread. Linguistic distance between one's mother tongue and another language is likely to affect fluency acquisition in another language: the greater the distance, the more costly is it to become fluent. Thus, one would expect less spread in a Chinese-English-French world than in a French-Italian-Spanish one. While linguistic differences are not easily altered by policy, the codification of various dialects into one language may influence this;
3. a stronger preference for individuals with the same mother tongue results in more diversity and less language spread. This preference can be strengthened by actions of civil society bodies, such as those active in the Baltic states in the 19th century, and by national or subnational (e.g. Catalonia, Flanders, Québec) governments either by specific linguistic policies or by policies promoting the socio-economic status of the speakers of a specific language;
4. language promotion strategies of the same intensity are much more efficient in accelerating the dominance of a language that in preserving or promoting a minority language. Thus, achieving similar results in minority language preservation requires significantly more resources.

Overall, while the 21st century may witness the death of numerous languages, it is unlikely that one language will replace all others, as the size of the remaining languages will provide protection against it.

Acknowledgments

The authors would like to thank Dr Lucía Ormaechea Grijalba of the University of Geneva for her support in the use of university computing resources.

References

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

Barabási, Albert-László and Márton Pósfai (2016). *Network science*. Cambridge: Cambridge University Press.

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

Church, Jeffrey and Ian King (1993). "Bilingualism and network externalities". In: *Canadian Journal of Economics*, pp. 337–345.

Civico, Marco (2019). "The Dynamics of Language Minorities: Evidence from an Agent-Based Model of Language Contact". In: *Journal of Artificial Societies and Social Simulation* 22 (4) 3.

Civico, Marco (2021). "Language policy and planning: a discussion on the complexity of language matters and the role of computational methods". In: *SN Social Sciences*, 1, 197.

Civico, Marco (2025). "A language economics perspective on language spread: Simulating Language Dynamics in a Social Network" (Version 1.0.0). *CoMSES Computational Model Library*. Retrieved from: https://www.comses.net/codebases/f8590435-ed56-4364-83f0-2e5ffee7c558/releases/1.0.0/

Clingingsmith, David (2017). "Are the world's languages consolidating? The dynamics and distribution of language populations". In: *The Economic Journal* 127.599, pp. 143–176.

Egger, Peter H. and Farid Toubal (2016). "Common Spoken Languages and International Trade". In: *The Palgrave Handbook of Economics and Language*. Ed. by Victor Ginsburgh, Shlomo Weber, et al. Basingstoke: Palgrave MacMillan, pp. 263-289.

Fishman, Joshua A. (1991). *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages*. Vol. 76. Multilingual matters.

Gazzola, Michele and Gabriele Iannàccaro (2023). "Indicators in language policy and planning". In: *The Routledge Handbook of Language Policy and Planning*. Ed. by Michele Gazzola, François Grin, Linda Cardinal, and Kathleen Heugh. London/New York: Routledge, pp. 331-347.

Gazzola, Michele and Bengt-Arne Wickström (2016). *The Economics of Language Policy*. Boston: The MIT Press.

Greenberg, Joseph (1956). "The measurement of linguistic diversity". In: *Language* 32.1, pp. 109–115.

Grin, François (1992). "Towards a Threshold Theory of Minority Language Survival". In: *Kyklos* 45, pp. 69–97.

Grin, François and Michele Gazzola (2013). "Assessing Efficiency and Fairness in Multilingual Communication". In: *Exploring the dynamics of multilingualism*. Ed. by Anne-Claude Berthoud, François Grin, and George Lüdi. Amsterdam/Philadelphia: John Benjamins, pp. 365–385.

Grin, François, Claudio Sfreddo, and François Vaillancourt (2010). *The Economics of the Multilingual Workplace*. London: Routledge.

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

Haugen, Einar and J. Derrick McClure, eds. (1982). *Minority Languages Today: A Selection from the Papers Read at the First International Conference on Minority Languages Held at Glasgow University from 8 to 13 September 1980*. Edinburgh: Edinburgh University Press.

Holden, Nigel (2016). "Economic Exchange and Business Language in the Ancient World: An Exploratory Review" in: *The Palgrave Handbook of Economics and Language*. Ed. by Victor Ginsburgh, Shlomo Weber, et al. Basingstoke: Palgrave MacMillan, pp. 290-311.

John, Andrew (2016). "Dynamic models of language evolution: The economic perspective". In: *The Palgrave Handbook of Economics and Language*. Ed. by Victor Ginsburgh, Shlomo Weber, et al. Basingstoke: Palgrave MacMillan, pp. 101–120.

John, Andrew and Onur Özgür (2020). "Linguistic Diversity in the Very Long Run". In: *The Economic Journal* 131, 1186-1214.

Katz, Michael L and Carl Shapiro (1985). "Network externalities, competition, and compatibility". In: *The American Economic Review* 75.3, pp. 424–440.

McPherson, Miller, Lynn Smith-Lovin, and James Matthew Cook (2001). “Birds of a Feather: Homophily in Social Networks”. In: *Annual Review of Sociology* 27, pp. 415–444.

Mélitz, Jacques (2016). “English as a Global Language”. In: *The Palgrave Handbook of Economics and Language*. Ed. by Victor Ginsburgh, Shlomo Weber, et al. Basingstoke: Palgrave MacMillan, pp. 583-615.

Minett, James W. and William S. Wang (2008). “Modelling endangered languages: The effects of bilingualism and social structure”. In: *Lingua* 118, pp. 19–45.

Ó Curnáin, Brian and Conchúr Ó Giollagáin (2023). “Minority language protection and promotion”. In: *The Routledge Handbook of Language Policy and Planning*. Ed. by Michele Gazzola, François Grin, Linda Cardinal and Kathleen Heugh. London/New York: Routledge, pp. 396-415.

Phillipson, Robert (1992). *Linguistic imperialism*. Oxford and New York: Oxford University Press.

Phillipson, Robert (2003). *English-only Europe? Challenging language policy*. London: Routledge.

Phillipson, Robert (2010). *Linguistic imperialism continued*. New York: Routledge.

Pool, Jonathan (1991). “A tale of two tongues”. in: *Manuscript, Political science department, University of Washington (Seattle)*.

Selten, Reinhard and Jonathan Pool (1991). “The distribution of foreign language skills as a game equilibrium”. In: *Game equilibrium models IV: Social and political interaction*. Ed. by Reinhard Selten. Berlin: Springer, pp. 64–87.

Templin, Torsten, Andrea Seidl, Bengt-Arne Wickström, and Gustav Feichtinger (2016). “Optimal language policy for the preservation of a minority language”. In: *Mathematical Social Sciences* 81, pp. 8– 21.

Templin, Torsten and Bengt-Arne Wickström (2023). “Language competition models”. In: *The Routledge handbook of language policy and planning*. Ed. by Michele Gazzola, François Grin, Linda Cardinal, and Kathleen Heugh. London/New York: Routledge, pp. 66–86.

Vaillancourt, François (1985). *Économie et langue*. Québec : Conseil de la langue française.

Wichmann, Søren (2008). “The emerging field of language dynamics”. In: *Language and Linguistics Compass* 2.3, pp. 442–455.

Wickström, Bengt-Arne (2005). “Can bilingualism be dynamically stable? A simple model of language choice”. In: *Rationality and Society* 17.1, pp. 81–115.

A Network metrics

Throughout the simulation, a number of values are recorded, which will allow us to make comparisons across different scenarios. Concerning language use, the model keeps track of:

1. the total number of agents who have each language as preferred means of communication over time;
2. the average fluency of all agents in each language; and
3. the average fluency in each language, but only for agents having that language as their preferred language.

If the transaction system is enabled, the model also keeps track of the distribution of capital over time.

Additionally, the model calculates a number of network-related metrics.[18] The degree k_i of node i is the number of links of that node, while the total number of links L is *half* the sum of the links of each node.[19] That is, in an undirected network of n nodes, we have:

$$L = \frac{1}{2}\sum_{i=1}^{n} k_i$$

The average degree $\langle k \rangle$ is calculated as the average number of links across all nodes in the network:

$$\langle k \rangle = \frac{1}{n}\sum_{i=1}^{n} k_i = \frac{2L}{n}$$

The network density D is the ratio of the existing edges to the total of potential edges:

$$D = \frac{L}{n(n-1)}$$

When calculated at the level of a single node, this metric is called the "local clustering coefficient" and is calculated as follows:

$$C_i = \frac{2L_i}{k_i(k_i - 1)}$$

where L_i is the number of edges between k_i's neighbours. This metric can then be calculated for all nodes. The average local clustering coefficient is the clustering coefficient of the network:

$$\langle C \rangle = \frac{1}{n}\sum_{i=1}^{n} C_i$$

The average degree is also calculated within and between language groups (e.g., the average degree within the group Alphish speakers and the average degree between the groups of Alphish and Betish speakers). This can be done by considering native speakers as well as preferred language speakers. Additionally, the clustering coefficient is also calculated for the three language groups, whether they speak it as a native language or a preferred language.

We also include a measure of diversity G. One of the simplest metrics to compute the degree of linguistic diversity is the linguistic diversity index (Greenberg, 1956). In its basic version, the index takes into account two dimensions of diversity: richness, i.e. the absolute number of languages present in a given territory, and evenness, i.e. how balanced or unbalanced the distribution of speakers is. Given a country where n languages (with $n > 1$) are spoken, each individual has only one native language, and the proportion of native speakers

[18] For an in-depth reviews of networks and their properties, we refer the reader to Barabási and Pósfai (2016).

[19] This is clearly the case in an *undirected* network, such as this one, where edges are symmetrical. If there exists an edge between A to B, then it goes from A to B and from B to A, and the total number of edges is one. However, in the case of a *directed* network, the existence of an edge pointing from A to B does not imply the existence of an edge pointing from B to A. If they both exist, then the total number of edges in this two-node network is two.

of language i is d_i (with $i = 1,2,...,n$ and $0 < d_i < 1$), the total probability of randomly picking two individuals who have the same native language is given by the sum of this event happening for each single language (i.e., picking two individuals speaking language 1, two individuals speaking language 2, and so on), that is $\sum_{i=1}^{n} d_i^2$. The Greenberg index, being a metric of diversity rather than uniformity, is equal to:

$$G = 1 - \sum_{i=1}^{n} d_i^2$$

The linguistic diversity index ranges from 0 to 1, with higher values indicating greater linguistic diversity. A value of 0 indicates complete linguistic homogeneity (everyone speaks the same language or has the same native language), while a value of 1 indicates complete linguistic heterogeneity (every individual speaks a different language). Hence this index will increase as n increases and as the weight of one language relative to the total (p_i) increases. The index can be calculated considering either native languages (which we shall call G_n) or preferred languages for communication (G_p). As G is a *static* metric (i.e. it refers to a moment in time), we turn it into a *dynamic* one by looking at the ratio of G_p at the end of the simulation to its value at the beginning.[20] We call this value G_r. Values below (above) 1 indicate that diversity has decreased (increased).[21]

[20] Note that at the beginning of the simulation $G_p = G_n$.

[21] Note that we could also simply look at the value of G_p at the end of the simulation, as the initial conditions are roughly the same for all simulations, except for the total number of agents. However, for small populations, even small absolute differences between groups imply big changes in the value of G. Therefore, in order to make all values comparable and isolate the change in diversity that does not depend on the initial population, we prefer to look at the value of G_r.