


## Capturing the diversity of multilingual societies

Thomas Louf <sup>\*</sup>, David Sánchez , and José J. Ramasco 

*Institute for Cross-Disciplinary Physics and Complex Systems IFISC (UIB-CSIC), 07122 Palma de Mallorca, Spain*



(Received 20 July 2021; accepted 8 November 2021; published 30 November 2021)

Cultural diversity encoded within languages of the world is at risk, as many languages have become endangered in the last decades in a context of growing globalization. To preserve this diversity, it is first necessary to understand what drives language extinction, and which mechanisms might enable coexistence. Here, we study language shift mechanisms using theoretical and data-driven perspectives. A large-scale empirical analysis of multilingual societies using Twitter and census data yields a wide diversity of spatial patterns of language coexistence. It ranges from a mixing of language speakers to segregation with multilinguals on the boundaries of disjoint linguistic domains. To understand how these different states can emerge and, especially, become stable, we propose a model in which language coexistence is reached when learning the other language is facilitated and when bilinguals favor the use of the endangered language. Simulations carried out in a metapopulation framework highlight the importance of spatial interactions arising from people mobility to explain the stability of a mixed state or the presence of a boundary between two linguistic regions. Further, we find that the history of languages is critical to understand their present state.

DOI: [10.1103/PhysRevResearch.3.043146](https://doi.org/10.1103/PhysRevResearch.3.043146)

### I. INTRODUCTION

Language, as the basis for communication, is at the heart of the functioning of human societies. It has thus long been an important subject of research, as scientists sought to understand its interactions with society, the internal evolution of a language's aspects with time or how multiple languages interact with one another. The research presented here is concerned with the latter, which emerged a few decades ago as a hot topic when linguists realized that the world may be facing a mass extinction of languages [1–3]. It has been pointed out that the estimated 6000 languages of the world convey a cultural wealth, the loss of which would be irreversible. Hence the need to understand what drives individuals to shift from one language to another.

Modeling language shift has been the subject of much research in the last decades [4,5], which employed various approaches such as the formulation of evolution equations based on ecological models [6–10], of reaction-diffusion equations [11–14], or approaches within the framework of agent-based modeling [14–17]. While global evolution equations determine how the proportions of each language group will evolve in a system, agent-based models (ABMs) describe the shifting mechanisms on an individual level, as they provide probabilities to switch to another language group. These transition probabilities depend on the linguistic environment of the individual, environment which may be defined in many ways. Different networks of interactions can be introduced, ranging

from the simplest (fully connected networks) to more realistic but less tractable ones (like a real-world social network). The former lend themselves easily to mathematical analysis as they can be equivalently written in terms of global evolution equations for large population sizes. As a result, models based on global evolution equations are a subset of the more general, agent-based ones. Moreover, ABMs allow to assess the impact of the social structure on the dynamics. This social structure is closely related to space, but in a nontrivial way, and as there is no model that can claim to be the universal solution to build spatial interaction networks [18], being able to plug in any kind of interaction network is an interesting feature of ABMs. It is for all these reasons that the focus of this article will be on ABMs. The first notable model to mention is the Abrams-Strogatz model [19]. It was the first to attract considerable attention as the authors were able to fit their model to the historical data of multiple languages threatened by extinction, and subsequently predicted their death. The model is very simple as it considers only the monolingual states A and B. The basic principle behind this model is that the more speakers of A, and the more prestigious A is in society, the more B speakers will want to switch to A, and inversely.

However, the existence of around 6000 spoken languages in 200 nations implies that multilingualism is a pervasive phenomenon worldwide. In almost every country, the presence of more than one language naturally leads to speech communities of different sizes. A common situation is that many individuals belonging to these communities use two or more languages independently of the official status and the educational prevalence of those languages. The extent and role of bilingualism is hence a difficult subject. Multiple modeling attempts have been made in that direction [12,15,20,21]. In these models, agents can be in a third state AB through which they have to pass to switch from being monolingual in a

<sup>\*</sup>thomaslouf@ifisc.uib-csic.es

language to another. Apart from Ref. [14], which relied on census data, none of the aforementioned models have been iterated over real-world spatial distributions of speakers, as they were rather implemented in fully connected populations or in toy models, like lattices or random networks. This is a shortcoming we will address here.

Indeed, speech communities are distributed in regions which are heterogeneous and even discontinuous when their boundaries cannot be arranged into a single closed curve. This spatial component cannot be neglected in the study of language dynamics, as the sociolinguistic environment in which individuals interact is of paramount importance for the dynamics. That is why this work also seeks to obtain and analyze the spatial distribution of languages in order to evaluate the models. But despite the ubiquity of language, data on language use have historically been hard to come by. Linguists have mainly relied on data from censuses or surveys which have a limited scope, especially in terms of spatial resolution and sample size. Thus Ref. [22] argued for large-scale data-driven approaches to complement existing sociolinguists' works, in a complementary framework of "computational sociolinguistics." In addition to new tools for speech and text analysis, technological advancements have brought with them the ability to collect unprecedented amounts of data from online communications.

In this work, we combine a large-scale empirical study of the spatial distribution of languages with agent-based modeling. In Sec. II, we show empirically that multilingual societies are characterized by different spatial patterns in the populations of monolinguals and bilinguals, encompassing fully mixed states and segregated distributions with a clear linguistic boundary. As the existing ABMs are not able to explain the range of spatial mixing observed, we introduce in Sec. III a model able to capture the diversity seen in the data. The model also shows how the behavior of bilinguals and the ease of learning a language have their importance for the coexistence of languages. Finally, Sec. IV contains our conclusions.

## II. A DIVERSITY OF MULTILINGUAL SOCIETIES?

As said above, multilingual societies are numerous and thus susceptible to display distinct features. These differences, however, need to be observed and, ideally, quantified, to truly describe the diversity of these societies. Given the very few regions and countries where censuses gather data on language use at a fine enough spatial scale, we choose here to turn to Twitter as an alternative data source. Nonetheless, our analysis can equally be applied to data from surveys and census where available, as shown in Sec. II and Figs. S13 and S14 for Quebec [23].

### A. Twitter data analysis

Twitter is a social networking and microblogging service used worldwide by hundreds of millions of users, who post short messages, called tweets, which can be geolocated. It has thus good potential as a data source to extract spatial distributions of language use, as shown in Refs. [24–29]. Here, we are not so much interested in language distributions fitting perfectly what exists in the offline world, but rather in the

kind of distributions we may encounter. Despite all the biases introduced by the differences of usage of Twitter across the population, it could hence still provide valuable insights for regions in which close to no other data are available. Then to obtain spatial distributions of languages, we selected 16 countries and regions in which there was potential to gather sufficient statistics for multilingual communities (see the list in the Table S1 of Ref. [23]), and analyzed geolocated tweets sent from them from early 2015 to the end of 2019. A regular grid was laid over each area of interest, dividing them in square cells (see for instance the grids laid over Belgium and Catalonia in Fig. 1). The cell size has to be adapted for each studied region, as explained in Sec. I C of Ref. [23]. We have checked the effect of modifying the cell size and made sure that our results are robust (see Figs. S10–S12 [23]). The language of the messages is detected using Chromium's Compact Language Detector (CLD) [30] that provides the most likely language of a text from the messages along with a confidence (see Sec. I C of Ref. [23] for details). After thoroughly cleaning and analyzing the collected tweets, we obtained a sample of local Twitter users to which a cell of residence and a set of languages were attributed. Information about data access and code availability can be found in the Appendix.

### B. Metrics

Before introducing any metric, we specify our definition of language groups. First, we focus only on the languages considered local. For instance, the use of English is widespread on Twitter, but we do not register those tweets unless English is one of the local languages (e.g., in Canada or Malaysia). A user is classified as a speaker of a language if at least 10% or 5 of their tweets are detected in that language. One individual can thus be naturally in a monolingual or in a multilingual group if they fulfill the condition in more than one language. The groups defined here are mutually exclusive: each user must be in one of the monolingual and multilingual groups that are possible to form with the given set of local languages. For the purposes of our work, we consider language as a social phenomenon. Thus, we do not take into account the individual proficiency, which is indeed interesting in other fields of study [31], but instead observe the language production of a speech community defined inside every cell, based on their use of one or more languages. Thereafter, we will talk of  $L$  speakers instead of "individuals who belong to the  $L$  group" for simplicity.

Starting from the counts  $N_{L,i}$  of  $L$  speakers residing in cell  $i$  obtained from the data, we wish to gain insights on the spatial distributions of language use. To do so we need to define a few basic metrics:

- (1) concentration in cell  $i$  of  $L$  speakers:

$$c_{L,i} = \frac{N_{L,i}}{N_L}, \quad (1)$$

- (2) proportion of  $L$  speakers in  $i$ 's population:

$$p_{L,i} = \frac{N_{L,i}}{N_i}, \quad (2)$$

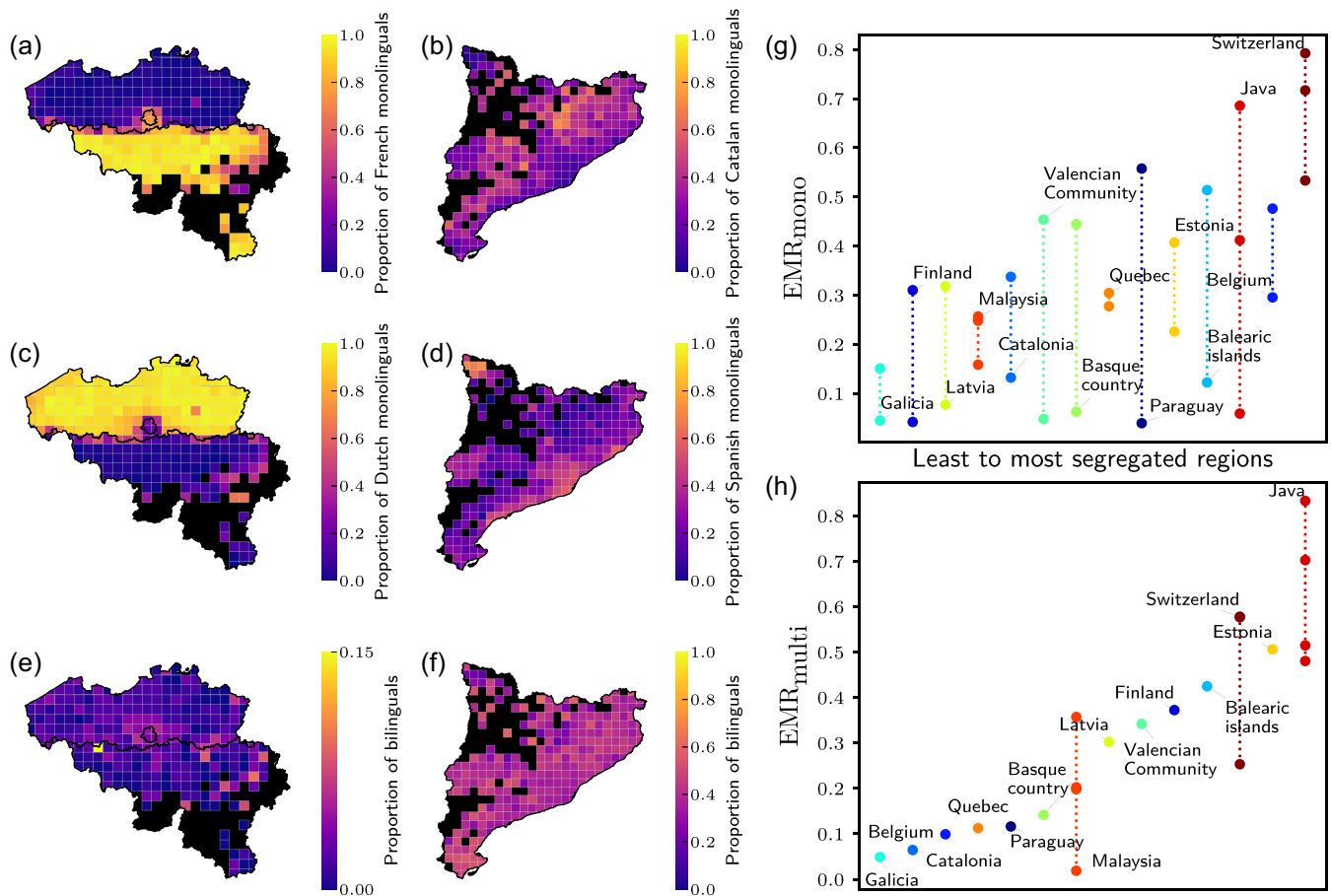


FIG. 1. Visualization of the diversity of multilingual societies. For each cell of  $10 \times 10 \text{ km}^2$ , the proportions  $p_{L,i}$  of monolinguals in (a) French, (b) Catalan, (c) Dutch, and (d) Spanish in Belgium (left) and Catalonia (right) are shown. The maps (e) and (f) show the proportion of bilinguals [note the different scale needed in (e)]. In the case of Belgium, the border between Flanders (North) and Wallonia (South) is drawn, and the Brussels region too. In black are cells in which fewer than ten Twitter users speaking a local language were found to reside, consequently discarded for the insufficient statistics. A clear separation of language groups is visible in Belgium following the linguistic regions, displaying mixing mainly around the border and in Brussels, while mixing in Catalonia is much more widespread, with a slight difference between the countryside and the large cities of the coast (East). [(g) and (h)] Earth mover's ratios of respectively the monolingual and multilingual groups of multilingual regions of interest, ranked left to right by increasing average of the y-axis values. In (h), the point for trilinguals in Switzerland is not displayed because its value was deemed unreliable (for more details see Sec. I F [23]). A rich diversity of mixing patterns is shown, beyond the two paradigmatic cases of Catalonia and Belgium.

where  $N_L$  are all the users classified as  $L$  speakers in the country or region considered, and  $N_i$  is the population of Twitter users residing in cell  $i$  speaking any of the local languages. As in Ref. [24], we can define the polarization of a language  $A$  for every cell  $i$  in a bilingual system with languages  $A$  and  $B$  as

$$\theta_{A,i} = \frac{1}{2}(1 + p_{A,i} - p_{B,i}). \quad (3)$$

The polarization vanishes when there are only  $B$  monolinguals, takes the neutral value of 0.5 when there are as many  $A$  speakers as  $B$  speakers, and goes to 1 when there are only  $A$  monolinguals. We will use this metric in bilingual regions as an indication of the mixing at the cell level.

Building further upon proportions and concentrations, we want to be able to measure the spatial mixing of language groups, or inversely, their spatial segregation. We define segregation as the difference in how individuals of a given group are spatially distributed compared to the whole population. Segregation is thus conceptualized as the departure from a

baseline, the unsegregated scenario, in which regardless of the group an individual belongs to, they would be distributed according to the whole population's distribution. Explicitly, the concentrations corresponding to this baseline, or null model, are  $c_i = N_i/N$ . To quantify language mixing, we would then like to measure a distance between the spatial distribution of a given language group and that of the whole population.

To this end, at a full country or region scale, we define the so-called Earth mover's distance (EMD). This metric allows us to quantify the discrepancy between two distributions embedded in a metric space of any number of dimensions. It has mainly been used within the field of computer vision [32], and it was shown to be a proper distance (in the metric sense) between probability distributions [33]. Here, we consider the distributions defined by the signatures  $P = \{(i, c_i)\}$  and  $Q_L = \{(i, c_{L,i})\}$ . We then define  $\text{EMD}_L$  as

$$\text{EMD}_L \equiv \text{EMD}(P, Q_L) = \sum_{i,j} \hat{f}_{ij} d_{ij}, \quad (4)$$

with  $d_{ij}$  the distances between cells  $i$  and  $j$ , and  $\hat{f}_{ij}$  the optimal flows to reshape  $P$  into  $Q_L$ , obtained by minimizing  $\sum_{i,j} \hat{f}_{ij} d_{ij}$  under the following constraints:

$$\begin{cases} f_{ij} & \geq 0, \forall i, j \\ \sum_j f_{ij} & = c_{L,i}, \forall i \\ \sum_i f_{ij} & = c_j, \forall j \\ \sum_i \sum_j f_{ij} & = \sum_i c_{L,i} = \sum_j c_j = 1, \end{cases} \quad (5)$$

where  $c_i$  and  $c_{L,i}$  are the concentrations of the population and  $L$  speakers in every cell  $i$ , as defined above.  $\text{EMD}_L$  quantifies thus the distance between the concentration distributions of  $L$  speakers and of the whole population, as needed. The computation of the EMD was implemented with Ref. [34], which uses the method of Ref. [35]. However, in its raw form, it is dependent on the spatial scale of the system considered. Hence the need for a normalization factor  $k_{\text{EMD}}$  in order to enable comparisons between regions of different sizes. The first, obvious choice for  $k_{\text{EMD}}$  would be the maximum distance between two cells of the region. However, such a choice would neglect the disparities of population density existing between different regions. The factor would be very high in Quebec, for instance, since the geographical scales are large even though its northern part is scarcely populated. This is why we choose instead the average distance between individuals:

$$k_{\text{EMD}} = \frac{\sum_i \sum_j N_i N_j d_{ij}}{(\sum_k N_k)^2}. \quad (6)$$

Our final metric is then the normalized version of the EMD, the EMR (Earth mover's ratio), defined as

$$\text{EMR}_L = \frac{\text{EMD}_L}{k_{\text{EMD}}}. \quad (7)$$

The EMR is a global parameter. The higher it is, the more segregated a linguistic community. On the contrary, if the EMR is close to zero this community is distributed according to the total population and the mixing is complete. As shown in Fig. S13 and Sec. S14 [23], the EMR is cell size invariant and, quite generally, a reliable metric when a careful statistical analysis is made.

### C. Empirical results

We propose a first visualization of the collected data in Figs. 1(a)–1(d), where the proportions of monolinguals in Dutch and French, Catalan and Spanish, are displayed for Belgium and Catalonia, respectively. The cell size is here of  $10 \times 10 \text{ km}^2$  (see Figs. S10 and S11 [23] for equivalent maps with cells of  $5 \times 5$  and  $15 \times 15 \text{ km}^2$ ). The maps already show two configurations that frequently appear across the world in multilingual societies: either a marked boundary between mostly monolingual domains (Belgium) or high mixing in every cell with local coexistence (Catalonia). The population of bilingual users concentrate in the border in the first case (especially in the region around Brussels and in the southern border with Luxembourg), and it is widespread in the second [Figs. 1(e) and 1(f)]. Results for the other multilingual regions listed in Table S1 are shown in Figs. S1–S14 [23]. These findings are summarized in Figs. 1(g) and 1(h), which presents the ranges of values reached by the EMR of respectively the monolingual and multilingual groups in 14 of our 16 regions

of interest. We filtered out regions where we deemed not sufficient the statistics gathered from Twitter (see Table S2 for all measured metrics and cell sizes used [23]). A wide diversity of situations can be observed. Multilingual societies may have rather balanced monolingual groups separated by a clear-cut border, which have thus high but quite similar EMR values, like in Belgium and Switzerland. One can also see unbalanced situations where one language is majoritarian, and has thus a much lower EMR than the monolinguals and multilinguals of other smaller, isolated languages. This is for example the case on the island of Java, where Indonesian is widespread, and Javanese and Sundanese are more localized. Multilinguals may also be mixing well in the whole population, like the bilinguals in Galicia and Catalonia. These groups can thus be of completely different natures from one region to another, from sustaining a minority language while being spatially mixed or isolated, to standing at the border between monolingual communities.

The metrics introduced to evaluate the spatial mixing of languages can be calculated using similar data taken from other sources. Although data on language use on a fine enough spatial scale are difficult to find, it can, for instance, be obtained for Quebec from the Canadian census of 2016. Maps equivalent to the ones of Fig. 1 are shown using both data from the census and from Twitter for Quebec in Figs. S13 and S14 [23]. Similar mixing patterns can be observed from both data sources.

## III. MODELS CAPTURING DIVERSITY

As language use in a society only sees significant changes on a time scale of generations [36], the maps obtained from Twitter are only snapshots of the situation around the years 2015 to 2019 (synchronic viewpoint). We do not have access to data providing the longitudinal evolution (diachronic framework), but the models at hand do describe the dynamics of the system. Since some of the multilingual societies we study have had the same kind of spatial pattern of language coexistence for generations (Belgium with a separation and Catalonia with mixing), it is natural to ask whether these states are stable solutions of a model describing language competition. We will check, in the first place, if the existing models meet the basic requirement of reaching the observed stable states. Crucially, if they do not fulfill it, the underlying mechanisms of language shift are not therein fully captured, missing a significant element that could be key to language preservation.

### A. Previous models

The individuals in a population can be in states representing their use of one or several languages. Under this framework, the dynamics are governed by the permitted transitions between states and their corresponding probabilities of occurring. Figure 2 displays the states: monolingual in A and B, and bilingual AB, with the associated transition probabilities in two previous models and in our proposal. We denote  $p_A$  and  $p_B$  the proportions of monolinguals in A and B, respectively, and  $p_{AB}$  the proportion of bilinguals. Within a mean-field approximation, and all the population being mixed,

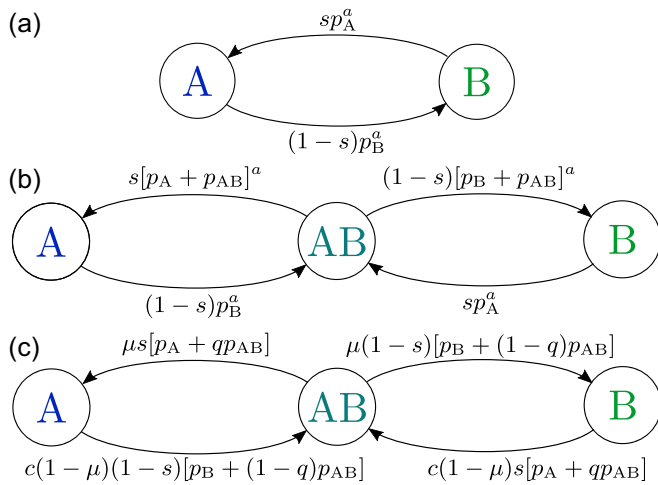


FIG. 2. Diagrams of the models presented in the text, showing the transition probabilities from one state to another. (a) Abrams-Strogatz model from [19]. (b) Bilinguals model from [15]. (c) Our model of bilinguals including both their preference and the ease to learn the other language [see Eq. (10)].

all equations can be written in terms of the proportions, which satisfy the equality  $p_A + p_B + p_{AB} = 1$ . Within this notation, a state of coexistence is a state in which the two languages remain spoken, which corresponds to either  $p_{AB} > 0$ , or  $p_A > 0$  and  $p_B > 0$ . Extinction of A (B), for instance, corresponds to  $p_A = p_{AB} = 0$  ( $p_B = p_{AB} = 0$ ).

The first model to mention is the one introduced in Ref. [19] by Abrams and Strogatz [Fig. 2(a)]. The model only contains monolinguals, who can change their languages with a probability that depends on the proportion of speakers of the other language to an exponent  $a$  (called volatility), which controls if the dependence on the proportion of the other language group is linear ( $a = 1$ ), sublinear ( $a < 1$ ) or superlinear ( $a > 1$ ). Besides, they also include a parameter  $s$  between zero and one, which stands for the prestige of the language A. If  $s$  is close to one, all the individuals will forget B and start to speak A alone. Set in a single population and in mean field, this model was shown to fit historical data of the decline of minority languages in Ref. [19]. It was thoroughly analyzed in Ref. [21], where it was first shown that its stable state is extinction of one language for  $a \geq 1$ , and coexistence for  $a < 1$ , independently of the prestige. In complex contact networks, the coexistence region in the  $(s, a)$  space shrinks, as not all values of prestige enable coexistence for  $a < 1$ . It is important to note that the linear version of the model does not predict coexistence.

Later, an extended model with bilinguals was proposed by Castelló *et al.* [15] [see Fig. 2(b)]. The transitions to lose a language are there related to the proportion of bilinguals besides the monolinguals of the other side. The idea is that since A can be spoken to both A and AB individuals, the utility to retain B decreases with an increasing proportion of these two types of individuals. An analysis of the stable states of this bilinguals model performed in Ref. [21] shows that the coexistence only occurs if  $a < 1$  and that the area of parameters allowing it is reduced compared to the Abrams-Strogatz model. Again,

the linear ( $a = 1$ ) version of the model does not allow for language coexistence.

Several concerns may be raised about these models. The first one is that for languages with equal prestige ( $s = 1/2$ ) and with equal social pressure (same proportion terms), learning and forgetting a language is equiprobable, while they result from two completely different processes. People may inherit a language from their parents, use it for endogenous communication, and they could be driven to learn a new one for work or education purposes, which corresponds to exogenous communication. This is a typical diglossic situation [37] with a linguistic functional specialization. A difference in prestige favors this process, but losing a language, especially in the presence of cultural attachment, can be more difficult. In the case of bilingualism, once someone masters a new language to a bilingual level they will not forget their first. Besides, it seems reasonable to assume that most of the time, a language is lost when it is not passed from one generation to the next [3,38]. A second concern we raise here is that both models only find stable coexistence in a nonlinear configuration, when  $a < 1$ . These values of  $a$  imply easier transitions overall, and thus that coexistence is favored when speakers are more loosely attached to their spoken languages. This nonlinearity is hence hard to explain from a practical point of view and it has the effect of making the transitions less dependent on the actual proportions of speakers. Thirdly, it is important to note that the bilingual model of Fig. 2(b) is not able to produce a stable solution in which the bilinguals coexist with monolinguals of a single language.

## B. Our model

Our proposal stems from the realization of this last point: there are several bilingual societies where the monolinguals of one language, e.g., B, are virtually extinct (e.g., Catalonia, Quebec, or the Basque Country). However, the bilinguals continue to use B and keep it alive for decades if not centuries due to cultural attachment. This “reservoir effect” must be incorporated in models of language shift. The other ingredient that we will include concerns demographics, in relation with the first concern raised above: language loss mostly occurs between generations. For this, we get inspiration from the work of Ref. [16] that sets a rather generic framework for models differentiating horizontal and vertical transmission.

We thus first distinguish generational, or vertical, transmission, which corresponds to the death of a speaker replaced by their offspring. If the speaker was monolingual, their single language is transmitted. If they were bilingual, one of their two languages might get lost in the process of transmission. This loss occurs according to the following transition probability:

$$P(\text{AB} \rightarrow X) = \mu s_X [p_X + q_X p_{AB}], \quad (8)$$

where, as in the other models,  $s_X$  refers to the prestige of language X, which can be either A or B. The other parameters are  $\mu \in [0, 1]$ , that is the fixed probability for an agent to die at each step; and,  $q_X \in [0, 1]$  that reflects the preference of bilinguals to speak X. So bilingual speakers may be more inclined to transmit only language X when it is more prestigious, preferred by other bilinguals, and more spoken around them.

The second kind of transition is horizontal, it is related to the learning of a new language by a monolingual in the course of their lives. This transition occurs according to the following transition probability:

$$P(X \rightarrow AB) = c(1 - \mu) s_Y [p_Y + q_Y p_{AB}], \quad (9)$$

where  $Y$  is the language other than  $X$ , and, critically,  $c \in [0, 1]$  is a factor adjusting the learning rate. The timescales of the learning process and of a generational change are completely different, hence the need to adjust  $(1 - \mu)$  by this factor  $c$  here. It depends on the similarity between the two languages and on the implemented teaching policies. For the sake of simplicity and to avoid the inclusion of more parameters, we assume that the process is symmetric between learning  $A$  when  $B$  is spoken and vice versa. This is not necessarily true in all cases, but it can easily be solved by splitting  $c$  in more parameters for each transition. To translate this expression of the transition probability into words, a monolingual in  $X$  will be more willing to learn  $Y$  as it is easier to learn, more prestigious, preferred by bilinguals, and more spoken around them.

We define  $s$  and  $q$  as symmetric around  $1/2$ , and thus define  $s = s_A = 1 - s_B$  and  $q = q_A = 1 - q_B$ . The transitions in our model are illustrated in Fig. 2(c) and we explicit here below the transition probabilities that define it:

$$\begin{cases} P(A \rightarrow AB) &= c(1 - \mu)(1 - s)[p_B + (1 - q)p_{AB}] \\ P(B \rightarrow AB) &= c(1 - \mu)s[p_A + q p_{AB}] \\ P(AB \rightarrow A) &= \mu s[p_A + q p_{AB}] \\ P(AB \rightarrow B) &= \mu(1 - s)[p_B + (1 - q)p_{AB}] \end{cases} \quad (10)$$

An important aspect of the model is that the use of a language by bilinguals contributes potentially unequally to the sizes of each language community. The neutral case occurs when  $q = 1/2$  and bilinguals on average contribute equally to both groups. It is however natural that even if bilinguals are fluent in both languages, individually they may have a certain preference for one of them and their language use is not necessarily balanced [39]. Even if one of the two languages is in a minority or suffers from a lack of prestige, appropriate values of  $q$  may maintain it alive. The most extreme example occurs when the monolinguals of  $B$ , for example, are extinct ( $p_B = 0$ ). Still, the use of  $B$  by the bilinguals keeps attracting monolinguals of the group  $A$  proportionally to  $(1 - q)p_{AB}$ .

Finally, we chose not to include nonlinearities in the model ( $a = 1$ ), as it turned out not to be necessary to capture the diversity we observed, and it would only add unnecessary complexity.

### C. A single population

We first analyze the model in the simplest setting of a single well-mixed population to determine the typology of possible solutions. Given the normalization condition  $p_A + p_B + p_{AB} = 1$ , the system dynamics can be described by a set of two coupled equations, let us say, for  $p_A$  and  $p_B$  (see Sec. III [23]). Fixed points are the solutions for which  $\partial p_A / \partial t = \partial p_B / \partial t = 0$ . The stability of these points is studied by performing a linear perturbation analysis around them, which requires the calculation of the Jacobian of the linearized equations and of its eigenvalues. Points for which all the

eigenvalues have strictly negative real parts are stable, while if any eigenvalue's real part is zero or positive the fixed point is unstable. Stream plots in Fig. 3 show where the model converges to in three characteristic examples, depending on the model parameters. In the first one [Fig. 3(a)], the stable (blue) points lie over the axis at values 1 and the system has as only solution the extinction of one of the two languages. In Fig. 3(b), the stable fixed point falls in the middle of the diagram and, therefore, the solution is symmetric coexistence with a majority ( $\sim 1/2$ ) of bilinguals. Finally, in Fig. 3(c), we find a stable fixed point over the  $x$  axis that represents the extinction of monolinguals  $B$  but coexistence between  $A$  monolinguals and bilinguals. Surprisingly enough, this represents the survival of a less prestigious language within a relatively small bilingual community. These results show already the flexibility of the model even in a single population.

We change now the viewpoint from the phase space to the parameter space. In Fig. 4, we plot the region of parameters where the model converges to stable coexistence. Since  $c$  and  $\mu$  act over the stability only in a combined form, their contributions can be merged into a new variable  $r$  defined as  $r = \mu / (c(1 - \mu))$ , which stands for the ratio between the mortality and learning rates. The other two parameters,  $s$  and  $q$ , are considered independently. We observe that the coexistence region expands when  $r$  decreases. This means that increasing the ease to learn one language when knowing the other (with a fixed mortality rate) makes coexistence more likely. Additionally, coexistence occurs more frequently when both prestige and bilingual preference are neutral,  $s = q = 1/2$ , which is expected. When the prestige of language  $A$  is lower than that of  $B$ , we find that there exists an optimal value of  $q$  making possible the coexistence,  $q^{opt} > 1 - s$ . For  $q < q^{opt}$ ,  $A$  is more at risk of extinction whereas for  $q > q^{opt}$ , the endangered language is  $B$ . There is thus a balance between prestige and bilingual preference that enables coexistence.

This model opens up unique classes of stable solutions: from the extinction of a language to coexistence when prestige is neutral, but also when it favors one of the two languages, and even only through a community of bilinguals. However, these analytic results in a fully-connected population do not suffice, as they do not show if the model is able to reproduce a case such as Belgium, where in the majority of cells there remains almost exclusively one language, except on the boundary between the two large communities. Consequently, we now analyze the model in a metapopulation framework to uncover the effect of including space and check whether this pattern can arise.

### D. The model in space

The idea of introducing a metapopulation framework in order to study interaction dynamics in space has been extensively exploited in ecology [40] and epidemiology [41,42]. In our context, we would need some information to build the extended model. The basic ingredients are a spatial division, the population in each division, the mobility between them and the characteristics of the populations in terms of language groups. Since we are interested in the phase space of the model, it is possible to use a completely abstract setting. However, this would require the generation of reasonable data

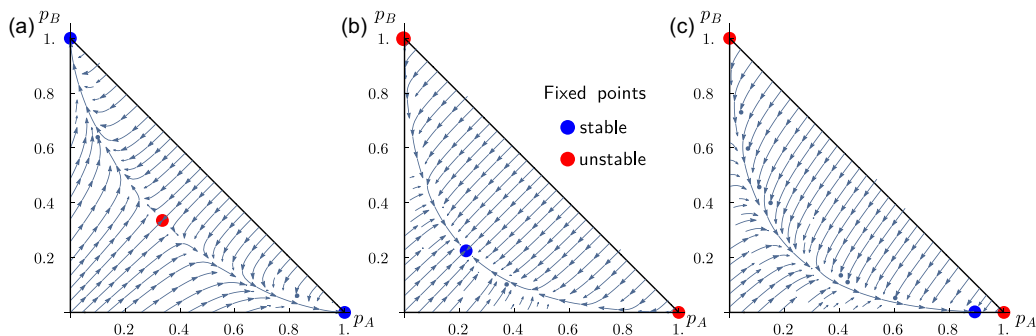


FIG. 3. Flow diagrams for the dynamics of two languages according to our model described in Eq. (10) set in a well-mixed population.  $p_A$  and  $p_B$  denote the proportions of monolinguals in A and B, respectively, and the proportion of bilinguals  $p_{AB}$  is such that  $p_A + p_B + p_{AB} = 1$ . The mortality rate is fixed at  $\mu = 0.02$ . (a) For  $s = q = 1/2$  and  $c = 0.02$ , the stable outcome is extinction of one of the two languages. (b) For  $s = q = 1/2$  and  $c = 0.05$ , the higher learning rate leads to a solution featuring stable coexistence. (c) For  $s = 0.57$ ,  $q = 0.45$  and  $c = 0.05$ , despite the lower prestige, B survives in a small community of bilinguals as it is the preferred language among them.

in terms of population and mobility, while this information is easily accessible from census data in many countries. Since we wish here to study the stability of the present, observed state, to make metapopulations interact with one another we use readily available commuting data from the census, as commuting is the backbone of everyday mobility. Some further work could include other kinds of mobility, like migrations, in order to investigate long-term time evolutions. We have thus chosen to use census data in Catalonia and Belgium as benchmarks, although it is important to stress that the intention is not to produce accurate predictions. Alternatively, the spatial interactions could be estimated from the population data us-

ing a model of human mobility, such as gravity, radiation or distance-kernel-based models [18,43,44].

The populations and commuting are thus obtained from the national census at municipality scale (see Methods for how to access them). We implement a mapping process from municipalities to our cells based on area overlap (details in Sec. IV A [23]). Regarding the language groups, Twitter data may suffer from different sociodemographic biases [45,46], and besides tweets reflect language use online, not necessarily the offline practices in the full population. Since in the census we found information on the total number of persons per language group and of residents per municipality, we have scaled the  $L$  speakers that we find on Twitter to match these two sets of marginal sums via iterative proportional fitting (IPF) [47,48].

Once the metapopulation has been initialized, the model can be simulated. As in Ref. [49], the day is divided in two parts: the individuals first start in their residence cells and interact with the local agents following the rates of Eq. (10), and then move to their work cells where again they interact with the local population. The agents encounter thus different environments characterized by diverse proportions  $p_{L,i}$  in the two parts of the day. Even if they live and work in the same cell, the local population changes from one part of the day to the next.

In order to analyze the stability of the steady states reached by the extended model, we derive an approximate master equation for the full metapopulation setting. To this end, we adapt the methodology described in Refs. [41,42] for epidemiological models (see Sec. IV B for details [23]). The equations obtained are only approximated but since they are analytic we can integrate them and calculate the Jacobian at their fixed points. To check the consistency of both approaches and that the fixed points of the dynamics are the same, we also introduce the initial conditions in the master equation, to then integrate it numerically using a standard Runge-Kutta algorithm. The fixed points reached by the simulations turn out to be fixed points as well for the equations. Not only that, all the eigenvalues of the Jacobian at these states have negative real parts and they are thus stable fixed points.

To explore the parameter space systematically, we perform a number of simulations until convergence to a stable

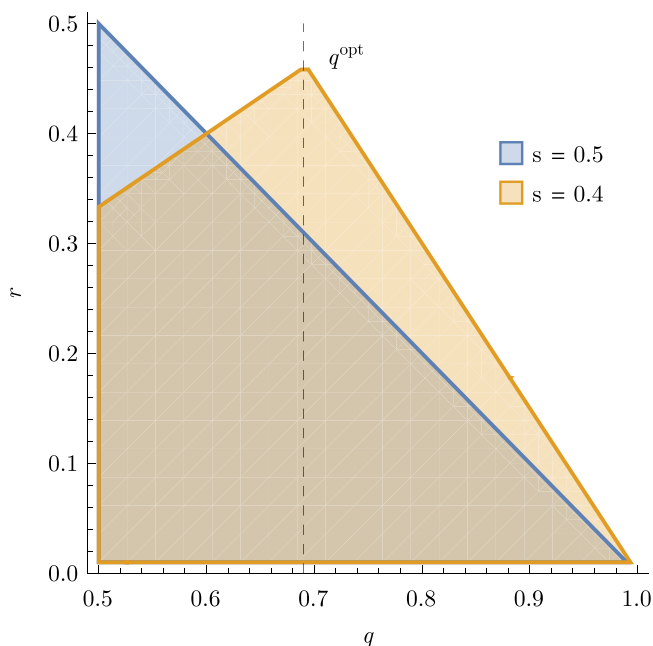


FIG. 4. Region of the parameter space where the dynamics of our model in a single population converge to stable coexistence of languages. We show two 2D cuts of the coexistence region in the  $(q, r)$  space for fixed values of  $s = 0.5, 0.4$ , with  $r = \mu/(c(1 - \mu))$ . Lower values of  $r$  favor coexistence, as well as a neutral prestige and bilingual preference  $q$ . When  $s < 0.5$ , coexistence is favored for an optimal value  $q^{opt} > 1 - s$ .

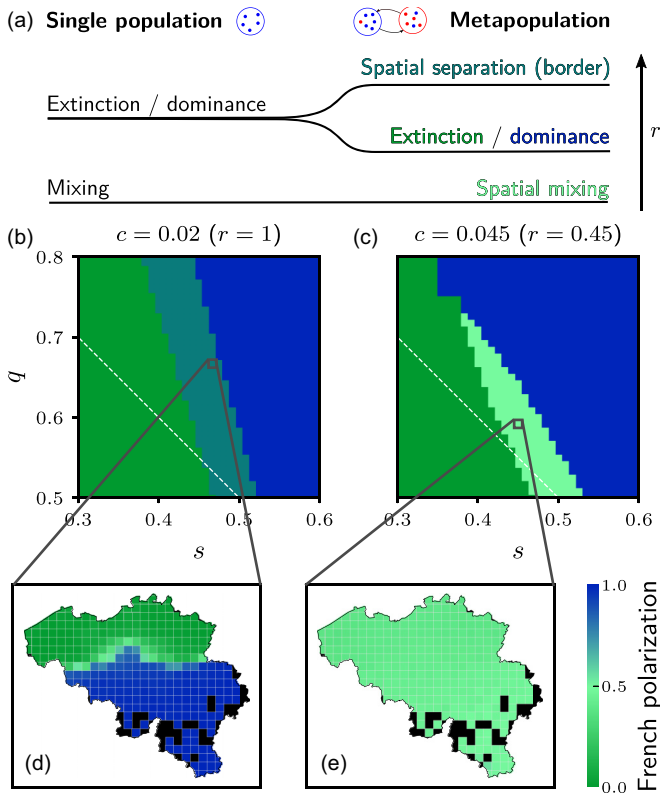


FIG. 5. Types of stable states of convergence of our model in a metapopulation set for Belgium. (a) Diagram illustrating the effect of adding metapopulations in the stable states of a single population: the former extinction state bifurcates in full extinction and in a boundary-like state with monolinguals separated in space. Larger values of  $r$  favor homogeneity, either by full extinction or by separation states (see Fig. S16 for more  $r$  values [23]). Below are the regions of the parameter space ( $s, q$ ) where these stable solutions emerge, (b) with  $r = 1$  and (c) with  $r = 0.45$ . Finally, two polarization maps show examples of states the model converges to, (d) a boundary-like state for  $r = 1, s = 0.467, q = 0.667$ , and (e) complete mixing for  $r = 0.45, s = 0.45, q = 0.592$ .

state. We show the results for the metapopulation setting of Belgium in Fig. 5. Remarkably, a new kind of stable state emerges. While in a single population we had only two stable configurations: extinction or mixing, here we can find full mixing [Fig. 5(e)], global extinction and local extinction of a language in part of the territory leading to a boundarylike state [Fig. 5(d)]. This state of convergence is similar to the initial conditions, corresponding to the language border we observe today. We have thus checked that our model, in these conditions, is able to obtain the present state as a stable solution. A surprising aspect of the results is that increasing  $r$ , or in other words making it easier or more common to learn the other language, does not necessarily favor coexistence. Indeed, as  $r$  increases, at one point boundary states become unstable and this may not necessarily lead to fully mixed states. When  $r$  grows bilinguals become more numerous on the boundary, until they expand beyond the boundary and spread bilingualism across the region. Still, if this happens when  $r$  is not high enough, the two languages cannot coexist and one ends up extinct, as the coexistence region of the parameter space in a single population shown in Fig. 4 may not have been reached.

We also wished to explore the possibility of having a hybrid state, consisting in an area where a minority language survives through bilinguals within an otherwise monolingual region. This is the case of Sundanese and Javanese in Java for instance (see Fig. S7 [23]). We initialized a hypothetical population in Belgium, with only monolinguals in Dutch, except in a pocket of cells in the South of the country, where there are only bilinguals. The latter were attached a  $q = 0.62$ , while  $q = 0.5$  for the rest. Iterating the model yields a stable solution similar to this initial state, with a mix of bilinguals and Dutch monolinguals in the pocket, and only Dutch monolinguals elsewhere (see Fig. S15 [23]).

### E. Dynamics in the parameters

The effect of multilingual education or, in general, policies favoring the use of one or several languages can alter the values of our model parameters. For example,  $c$  represents how monolinguals learn the other language. This process can be facilitated by the similarity between the languages or by teaching in both languages at school, for instance. Next, we investigate whether a parameter changing in time can perturb the system out of a stable state, and how the transition to a completely different configuration occurs. To this end, we run a simulation for 23 000 steps and present the results in Fig. 6. To explore the effects of the  $c$  parameter evolution alone, we fix the other parameters  $s = q = 1/2$  and  $\mu = 0.02$ . We start from our initial conditions with  $c = 0.005$ , which converges to a stable state with a boundary [see the first map of Fig. 6(c)]. After 2200 steps, we then increase  $c$  by 0.005 every 400 steps until we reach  $c = 0.055$ . The system converges quickly to a state of mixed coexistence, with a majority of bilinguals and equal proportions of monolinguals, like in Fig. 5(e).  $c$  is then decreased at the same rate as before to reach its initial value of 0.005. The system eventually converges to a state displaying a boundary, but displaced compared to its initial position. A visualization of this evolution is proposed in movie S1. Also, the resulting trajectory in the EMR space in Fig. 6(b) shows that the final stable state exhibits more segregation for both monolinguals and bilinguals, since the boundary between communities lies in the countryside, and not around Brussels as in the original scenario. The importance of the history of languages is hence clearly shown by this experiment.

The seemingly random placement of the boundary may be owed to the absence of constraints on the system, which is completely closed. In reality a country is an open system with exterior influences, notably from its direct neighbors. Thus we ran the same simulation with trans-border proportions  $p_{TB}$  equal to 0.5% and 0.2% of the population of the border municipalities of France and the Netherlands commuting to Belgium. These commuters act as a fixed population of monolinguals interacting only during the workday with the local population (for more details, see Sec. IV D [23]). These boundary conditions stabilize the final state of convergence, as the linguistic boundary resulting from the process of varying  $c$  is similar for the two values of  $p_{TB}$ , following the orientation of the two opposite borders (see Figs. 6(d)–6(e)). This positioning is a clear improvement over the closed-system simulation, albeit still not quite the one we observed in Fig. 1. In Fig. 6(b), the positions of these two states in the EMR space



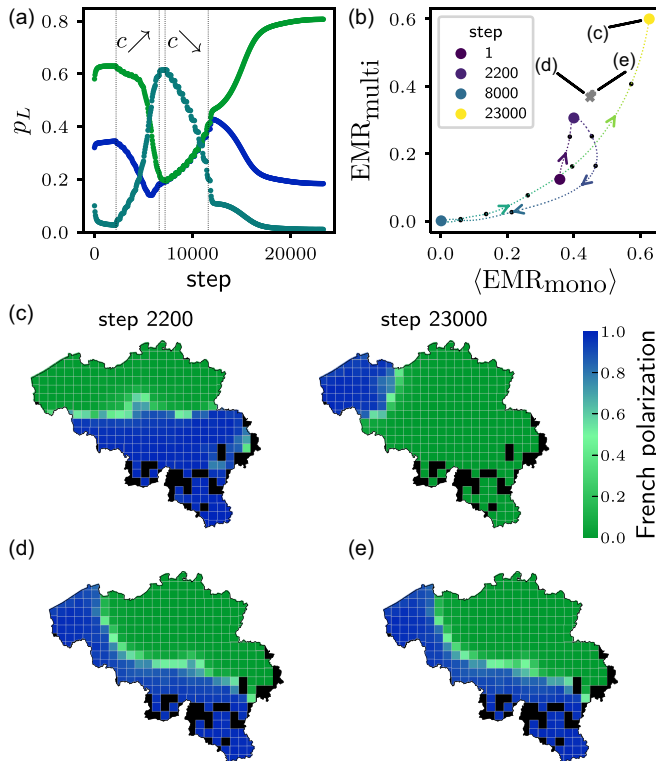


FIG. 6. Evolution of the state of the metapopulation model in Belgium when  $c$  varies, first slowly increased and then decreased to recover the original value. We fixed  $s = q = 1/2$  and  $\mu = 0.02$ . (a) Evolution of the global proportions  $p_L$  of individuals belonging to each  $L$  group. The blue curve corresponds to French monolinguals, light green to Dutch monolinguals and dark green to bilinguals. (b) Trajectory of the system in the EMR space: on the  $x$  axis the average of the EMR between each monolingual community and the whole population, and on the  $y$  axis the one between bilinguals and the whole population. The initial state and the stable states the system went through are marked by colored circles, while black ones mark additional points where the EMR was calculated, and the dashed line the interpolation between them. (c) Polarization maps of French in the initial and final states, both featuring a boundary but located in different areas, thus showing the irreversibility of the dynamics. [(d) and (e)] Polarization maps of French in the final states of simulations including trans-border commuters from France and the Netherlands, respectively with proportions  $p_{TB}$  equal to 0.5% and 0.2% of the population of the border municipalities of these two countries. The points in the EMR space corresponding to these final states are also represented in (b).

are also shown to be much closer to the original state than the final state of the first trajectory.

More complex settings could be envisaged to get closer to a realistic solution. A space-dependent prestige could be introduced, taking different values in Flanders, Wallonia, and Brussels, for instance. Also, we here considered only the commuting part of human mobility, but other kinds of mobility like migrations may have their importance. This is especially true for attractive metropolises like Brussels, which are typically places of intense language contact [50]. However, in this simulation the aim was to check the irreversibility of a change when increasing the ease to learn the other language

and subsequently decreasing it to its original value, which was indeed confirmed.

#### IV. DISCUSSION

In summary, we have explored the spatial distribution patterns of language competition and coexistence in multilingual societies. We first did so by introducing the Earth mover's ratio, a metric capable of measuring the spatial segregation of a group in a given society, starting from a distance between its distribution and that of the whole population. Two main configurations have thus been observed: either spatial mixing with multilinguals widespread or separate linguistic groups with a clear boundary between them and multilinguals concentrating around it.

Despite the ubiquity of these two configurations and their apparent temporal stability, the models introduced in the literature were not able to offer clear solutions capturing them. As we show, the main difficulty comes from the role of bilinguals in keeping languages alive. In many occasions, the monolingual community of one of the languages may become virtually extinct and its use relies only on the bilingual group. We have introduced a model taking this into account and have shown that it is able to produce naturally both configurations as stable solutions without the need for artificial nonlinearities. The model features a parameter considering the preference of bilinguals for one of the two languages. This preference actually acts as a kind of defense mechanism since the use by bilinguals of the endangered language may be enough to save it, countering a possibly lower prestige of the language within society as a whole. The ease to learn the other language also has a role in the model. It may be influenced by both the similarity between languages, which can hardly be controlled, but also by the policies put into place to facilitate its learning. We have shown that this parameter is critical to determine whether languages can coexist. The parameters of the model could be estimated using longitudinal data. The scope of this work was not predictive, but rather to study stable solutions of the model, so we leave it here for future work.

When spatial interactions are taken into account via the commuting patterns of individuals, the model is able to reach a stable state where two language communities are separated by a boundary around which they coexist. In this case, however, we have shown that, quite counter-intuitively, increasing this ease to learn the other language may break the existing boundary and lead to extinction, and not to the desired coexistence with mixing of the languages. This calls for caution when designing policies since the final state is strongly history-dependent.

Overall, our findings shed light on the role of heterogeneous speech communities in multilingual societies, and they may help shape the objectives and nature of language planning [51] in many countries where accelerated changes are threatening cultural diversity.

#### ACKNOWLEDGMENTS

The authors acknowledge funding from the project PACSS (RTI2018-093732-B-C22) of the

MCIN/AEI/10.13039/501100011033/ and of the EU through FEDER funds (A way to make Europe) and also from MCIN/AEI/10.13039/501100011033/ under the Maria de Maeztu program for Units of Excellence in R&D (MDM-2017-0711). This work has been carried out within the COST Action Nexus Linguarum CA 18209.

## APPENDIX

### 1. Data access

The geolocated tweets used to map language use were collected through the streaming API of Twitter, and more specifically using the “statuses/filter” endpoint Ref. [52]. This endpoint provides a sample of tweets in real time matching some provided filters. For the purpose of this work, bounding box filters were set to collect tweets from a set of countries of interest. Before reproducing this method of data collection, one should bear in mind that the current form and even the availability of this endpoint is subject to future changes intro-

duced by the Twitter Developer’s team. The aggregated data giving the counts of local users by language group by cell have been deposited on figshare [53]. The data on commuting patterns at the municipality level in Belgium were obtained from the 2011 census [54]. The population per municipality in France and in the Netherlands were obtained respectively, see Refs. [55,56]. The data about the knowledge of official languages (English or French or both) by census subdivisions in Quebec were obtained from the 2016 Canadian census, and can be downloaded directly from [57].

### 2. Code availability

The data processing, the plotting of results and the simulations were carried out in Python with the help of open-source libraries. All of the Python code used for this work is hosted on GitHub [58]. Mathematica was used to carry out part of the analytic work on the models and to generate the associated Figs. 3 and 4. The corresponding code is also hosted on GitHub, available at [59].

- 
- [1] M. Krauss, The world’s languages in crisis, *Language* **68**, 4 (1992).
  - [2] L. A. Grenoble and L. J. Whaley, *Endangered Languages: Language Loss and Community Response* (Cambridge University Press, 1998), p. 384.
  - [3] D. Crystal, *Language Death* (Cambridge University Press, 2000).
  - [4] C. Castellano, S. Fortunato, and V. Loreto, Statistical physics of social dynamics, *Rev. Mod. Phys.* **81**, 591 (2009).
  - [5] M. Boissonneault and P. Vogt, A systematic and interdisciplinary review of mathematical models of language competition, *Humanit. Soc. Sci. Commun.* **8**, 21 (2021).
  - [6] J. Mira and Á. Paredes, Interlinguistic similarity and language death dynamics, *Europhys. Lett.* **69**, 1031 (2005).
  - [7] J. Pinasco and L. Romanelli, Coexistence of Languages is possible, *Physica A* **361**, 355 (2006).
  - [8] A. Kandler and J. Steele, Ecological Models of Language Competition, *Biological Theory* **3**, 164 (2008).
  - [9] R. V. Solé, B. Corominas-Murtra, and J. Fortuny, Diversity, competition, extinction: the ecophysics of language change, *J. R. Soc., Interface* **7**, 1647 (2010).
  - [10] E. Heinsalu, M. Patriarca, and J. L. Léonard, The role of bilinguals in language competition, *Advances in Complex Systems* **17**, 1450003 (2014).
  - [11] A. Kandler, Demography and Language Competition, *Human Biology* **81**, 181 (2009).
  - [12] M. Patriarca and E. Heinsalu, Influence of geography on language competition, *Physica A* **388**, 174 (2009).
  - [13] N. Isern and J. Fort, Language extinction and linguistic fronts, *J. R. Soc., Interface* **11**, 20140028 (2014).
  - [14] K. Prochazka and G. Vogl, Quantifying the driving factors for language shift in a bilingual region, *Proc. Natl. Acad. Sci. USA* **114**, 4365 (2017).
  - [15] X. Castelló, V. M. Eguíluz, and M. S. Miguel, Ordering dynamics with two non-excluding options: bilingualism in language competition, *New J. Phys.* **8**, 308 (2006).
  - [16] J. W. Minett and W. S. Wang, Modelling endangered languages: The effects of bilingualism and social structure, *Lingua* **118**, 19 (2008).
  - [17] X. Castelló, L. Loureiro-Porto, and M. San Miguel, Agent-based models of language competition, *Int. J. Sociol. Language* **2013**, 21 (2013).
  - [18] H. Barbosa, M. Barthélemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, Human mobility: Models and applications, *Phys. Rep.* **734**, 1 (2018).
  - [19] D. M. Abrams and S. H. Strogatz, Modelling the dynamics of language death, *Nature (London)* **424**, 900 (2003).
  - [20] M. Patriarca, X. Castelló, J. R. Uriarte, V. M. Eguíluz, and M. S. Miguel, Modeling two-language competition dynamics, *Advances in Complex Systems* **15**, 1250048 (2012).
  - [21] F. Vazquez, X. Castelló, and M. San Miguel, Agent based models of language competition: Macroscopic descriptions and order-disorder transitions, *J. Stat. Mech.: Theory Exp.* (2010) P04007.
  - [22] D. Nguyen, A. S. Doã ruöz, C. P. Rosé, and F. de Jong, Computational sociolinguistics: A survey, *Computational Linguistics* **42**, 537 (2016).
  - [23] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevResearch.3.043146> for detailing the methods for the data analysis and simulations, providing derivations of analytic results as well as more maps of languages distributions. A Supplemental Movie S1 presents the simulation corresponding to Figs. 6(a)–6(c).
  - [24] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani, The twitter of Babel: Mapping world languages through microblogging platforms, *PLoS One* **8**, e61981 (2013).
  - [25] U. Pavalanathan and J. Eisenstein, Confounds and Consequences in Geotagged Twitter Data, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics (ACL), 2015), pp. 2138–2148.

- [26] B. Gonçalves and D. Sánchez, Crowdsourcing Dialect Characterization through Twitter, *PLoS One* **9**, e112074 (2014).
- [27] Y. Huang, D. Guo, A. Kasakoff, and J. Grieve, Understanding U.S. regional linguistic variation with Twitter data analysis, *Computers, Environment and Urban Systems* **59**, 244 (2016).
- [28] B. Gonçalves, L. Loureiro-Porto, J. J. Ramasco, and D. Sánchez, Mapping the Americanization of English in space and time, *PLoS One* **13**, e0197741 (2018).
- [29] J. Dunn, Mapping languages: the Corpus of Global Language Use, *Language Resources and Evaluation* **54**, 999 (2020).
- [30] R. Al-Rfou and B. Solomon, Python bindings for the Compact Language Detector 2 (2014).
- [31] C. Baker, *Foundations of Bilingual Education and Bilingualism*, 5th ed., Bilingual education and Bilingualism (Multilingual Matters, Bristol, UK, Tonawanda, NY, 2011), Vol. 79.
- [32] Y. Rubner, C. Tomasi, and L. J. Guibas, A metric for distributions with applications to image databases, in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE, 1998), pp. 59–66.
- [33] E. Levina and P. Bickel, The Earth Mover’s distance is the Mallows distance: Some insights from statistics, in *Proceedings of the IEEE International Conference on Computer Vision* (IEEE, 2001), Vol. 2, pp. 251–256.
- [34] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. H. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong *et al.*, POT: Python Optimal Transport, *J. Machine Learning Res.* **22**, 1 (2021).
- [35] N. Bonneel, M. van de Panne, S. Paris, and W. Heidrich, Displacement Interpolation Using Lagrangian Mass Transport, in *Proceedings of the 2011 SIGGRAPH Asia Conference on - SA '11* (ACM Press, New York, 2011), Vol. 30, p. 11.
- [36] W. Labov, *Sociolinguistic Patterns* (University of Pennsylvania Press, 1973).
- [37] C. A. Ferguson, Diglossia, *Word* **15**, 325 (1959).
- [38] A. Portes and L. Hao, E pluribus unum: Bilingualism and loss of language in the second generation, *Sociology of Education* **71**, 269 (1998).
- [39] S. Romaine, The Bilingual and Multilingual Community, in *The Handbook of Bilingualism and Multilingualism* (John Wiley & Sons, Ltd, Chichester, UK, 2012), pp. 443–465.
- [40] I. Hanski, Metapopulation dynamics, *Nature (London)* **396**, 41 (1998).
- [41] L. Sattenspiel and K. Dietz, A structured epidemic model incorporating geographic mobility among regions, *Math. Biosci.* **128**, 71 (1995).
- [42] D. Balcan, B. Gonçalves, H. Hu, J. J. Ramasco, V. Colizza, and A. Vespignani, Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model, *Journal of Computational Science* **1**, 132 (2010).
- [43] J. Burridge, Spatial Evolution of Human Dialects, *Phys. Rev. X* **7**, 031008 (2017).
- [44] J. Burridge and T. Blaxter, Inferring the drivers of language change using spatial models, *Journal of Physics: Complexity* **2**, 035018 (2021).
- [45] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, Understanding the Demographics of Twitter Users, in *Proceedings of the Fifth International AAAI Conference on Web and Social Media, ICWSM 2011* (AAAI Press, 2011), pp. 554–557.
- [46] D.-P. Nguyen, R. B. Trieschnigg, and L. Cornips, Audience and the Use of Minority Languages on Twitter, in *Proceedings of the Ninth International AAAI Conference on Web and Social Media, ICWSM 2015* (AAAI Press, 2015), pp. 666–669.
- [47] W. E. Deming and F. F. Stephan, On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *Ann. Math. Stat.* **11**, 427 (1940).
- [48] S. E. Fienberg, An iterative procedure for estimation in contingency tables, *Ann. Math. Stat.* **41**, 907 (1970).
- [49] J. Fernández-Gracia, K. Suchecki, J. J. Ramasco, M. San Miguel, and V. M. Eguíluz, Is the Voter Model a Model for Voters? *Phys. Rev. Lett.* **112**, 158701 (2014).
- [50] S. Simon, *Cities in Translation: Intersections of Language and Memory* (Routledge, London, 2011) p. 224.
- [51] R. B. Kaplan and R. B. Baldauf, *Language Planning from Practice to Theory* (Multilingual Matters, Bristol, UK, Tonawanda, NY, 1997).
- [52] <https://developer.twitter.com/en/docs/twitterapi/v1/tweets/filter-realtime/overview>.
- [53] [https://figshare.com/articles/dataset/Spatial\\_distributions\\_of\\_languages\\_on\\_Twitter/14339321](https://figshare.com/articles/dataset/Spatial_distributions_of_languages_on_Twitter/14339321).
- [54] <https://statbel.fgov.be/en/open-data/census-2011-matrix-commutes-sex>.
- [55] <https://www.insee.fr/fr/statistiques/4989724>.
- [56] <https://opendata.cbs.nl/statline/#/CBS/nl/dataset/70072ned/table?dl=3B993>.
- [57] <https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dt-td/Index-eng.cfm>.
- [58] <https://github.com/TLouf/multiling-twitter>.
- [59] <https://github.com/TLouf/multiling-analytical>.