Universitat de les Illes Balears

DOCTORAL THESIS
2023

COMPLEXITY IN LANGUAGE VARIATION:
EXPLORING THE INTERPLAY BETWEEN GEOGRAPHY, CULTURE AND THE SOCIAL FABRIC

THOMAS LOUF

DOCTORAL THESIS
2023

Doctoral programme in Physics

COMPLEXITY IN LANGUAGE VARIATION:

EXPLORING THE INTERPLAY BETWEEN GEOGRAPHY, CULTURE AND THE SOCIAL FABRIC

THOMAS LOUF

Director: José Javier Ramasco
Director: David Sánchez
Tutor: Cristóbal López

Doctor by the Universitat de les Illes Balears

Thomas Louf: *Complexity in language variation:* Exploring the interplay between geography, culture and the social fabric, © 2023

SUPERVISORS:
José Javier Ramasco
David Sánchez

LOCATION:
Palma, Spain

Ohana means family.
Family means nobody gets left behind, or forgotten.

— Lilo & Stitch

Dedicated to the loving memory of Rudolf Miede.

1939 – 2005

ABSTRACT

Short summary of the contents in English. . . a great guide by Kent Beck how to write good abstracts can be found here:

https://plg.uwaterloo.ca/~migod/research/beckOOPSLA.html

RÉSUMÉ

Mon résumé

PUBLICATIONS

Most of the ideas, results and figures presented in this thesis have appeared previously in the following publications:

[1] Emir Ganić, Nico van Oosten, Luis Meliveo, Sonja Jeram, Thomas Louf and Jose J. Ramasco. 'Dynamic Noise Maps for Ljubljana Airport'. In: *10th SESAR Innovation Days*. Dec. 2020.

[2] Thomas Louf, Bruno Gonçalves, Jose J. Ramasco, David Sanchez and Jack Grieve. *American Cultural Regions Mapped through the Lexical Analysis of Social Media*. Aug. 2022. DOI: 10.48550/arXiv.2208.07649.

We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.

— Donald E. Knuth [70]

ACKNOWLEDGMENTS

Put your acknowledgments here.

Many thanks to everybody who already sent me a postcard!

Regarding the typography and other help, many thanks go to Marco Kuhlmann, Philipp Lehman, Lothar Schlesier, Jim Young, Lorenzo Pantieri and Enrico Gregorio[1], Jörg Sommer, Joachim Köstler, Daniel Gottschlag, Denis Aydin, Paride Legovini, Steffen Prochnow, Nicolas Repp, Hinrich Harms, Roland Winkler, Jörg Weber, Henri Menke, Claus Lahiri, Clemens Niederberger, Stefano Bragaglia, Jörn Hees, Scott Lowe, Dave Howcroft, José M. Alcaide, David Carlisle, Ulrike Fischer, Hugues de Lassus, Csaba Hajdu, Dave Howcroft, and the whole LaTeX-community for support, ideas and some great software.

Regarding LyX: The LyX port was intially done by *Nicholas Mariette* in March 2009 and continued by *Ivo Pletikosić* in 2011. Thank you very much for your work and for the contributions to the original style.

[1] Members of GuIT (Gruppo Italiano Utilizzatori di TeX e LaTeX)

CONTENTS

ACRONYMS

ABM	agent-based model	19
API	application programming interface	12
CLD	Compact Language Detector	26
EMD	earth mover's distance	27, 29
EMR	earth mover's ratio	29
NLP	natural language processing	17
POI	point of interest	15
QRT	quote retweet	12
RT	retweet	12
SES	socio-economic status	6

Part I

INTRODUCTION

1 WHERE LANGUAGE DIVERSITY COMES FROM

Language can generally be defined as a structured system that human beings use to communicate. More specifically here, we use the term language to refer to natural languages, meaning languages which have evolved naturally, or, said differently, that have not been designed intentionally — as opposed to programming languages, for instance. Our objects of study are thus languages in the common sense of the word, that is coherent systems that define words and how their combinations convey meaning — like English, Mandarin Chinese or Hindi, to cite the three most spoken nowadays. As a primary means of communication, language is ubiquitous in any individual's life and in the workings of any human society. It is so much so that it is considered a "cultural universal", meaning all known human societies have some form of language [20, 62]. And it is so much so that researchers are unable to trace back to the origin of such a structured system of communication [57, 70, 109, 141]. Those who have ventured into this kind of inquiry have estimated that language dates back tens or even hundreds of thousands of years [18, 28, 33, 115]. One fact is for certain though: for what could be colloquially called *a very long time*, human beings have come up with, innovated upon, used, and more generally interacted with languages. It is then safe to say that human history must have seen a huge diversity of languages emerge. What is more ambiguous, though, is how this diversity is shaped through individuals' interactions, as they form societies. This is the central question that defines the whole field of sociolinguistics [27, 90, 91, 147, 153], which is also the broad subject of this thesis. In the following sections, we will touch on the different roles of language in society that may bring about variation, or, on the contrary, reduce existing linguistic diversity.

1.1 LANGUAGE AS A VECTOR FOR COMMUNICATION

The first obvious function that language serves is to facilitate communication between individuals, more specifically the kind of communication called *verbal communication*. To optimize language with regard to that function, there should only be one single language, shared homogeneously among all individuals. This has not been the case historically though, for many reasons, including historical and political ones, but also a very down-to-earth one. It is the very simple fact that, for most of its history, humanity has been spread around the Earth and unable to communicate at long distances. There is one very well known example that illustrates this. Humans have been in America for thousands of years: according to recently-found evidence, they have for more than 21 000 years [10]. Yet, the first lasting contact between Europeans and indigenous Americans came less than 600 years ago. During all this time, people on the two continents have had ample time to come up with new languages, innovate upon existing ones, and mix within their own continent, at least partially. Thus, on the scale of all these languages' histories, it is only very recently that the two groups came into contact. Since then, things have accelerated extremely fast though. First, transport has allowed long distance communication on the scale of months with boats for roughly the past 500 years, and then on the scale of hours with planes since the start of the last century. In the last two centuries, telecommunication has enabled long distance and near-real-time communication, and it has truly been widely democratized with the Internet in the last two decades. On the technical front, the communication barriers between individuals across the globe seem to have come down. But does this imply a push towards a reorganization of the world in what the philosopher Marshall McLuhan called a *global village* [100]? Does this imply a more interconnected world, and in turn, that we will naturally tend towards the communication-optimal state of homogenous language?

A physicist's intuition would say that the more individuals interact with one another, the more language should *thermalize*, or reach an equilibrium state of spatially uniform and temporally constant language. Another view would be that, because it costs energy for humanity to maintain language diversity, homogenization of language would be both desirable and inevitable. Ferdinand de Saussure, a prominent linguist of the late XIX$^{\text{th}}$ - early XX$^{\text{th}}$ century, seems to echo this view:

> Among all the individuals that are linked together by speech, some sort of average will be set up: all will reproduce — not exactly of course, but approximately — the same signs united with the same concepts. [35]

But this relies on the hypothesis that the global society would tend toward complete interconnectedness. This idea was for instance challenged by the anthropologist Robin I. M. Dunbar, when he suggested the existence of a maximum number of people one can maintain stable social relationships with, which is known as Dunbar's number. Its existence was first hypothesized [39, 40], and later demonstrated, not only on real-world social networks [75, 98], but also for a massive, online one [58]. On the other hand, while all individuals may not be completely interconnected, any two individuals may be closer on the social network than one would expect. This is the claim behind the famous idea of the six degrees of separation, or of the small world property of social networks [36, 102, 146, 154]. This idea and the experiments behind it have received some criticism though [85]. Also, more recent works focusing on online communication networks have consistently found that distance still plays a major role in defining both strong and weak ties [52, 93, 144].

The existence of a process of globalization is undisputed, but the idea that more interconnectedness would create a global village has been challenged [13, 116]. In the words of the sociolinguist Jan Blommaert:

> The world has not become a village, but rather a tremendously complex web of villages, towns, neighbourhoods, settlements connected by material and symbolic ties in often unpredictable ways. That complexity needs to be examined and understood. [13]

Nonetheless, there is some evidence of a homogenizing trend. English is on a steady path to become a global language [31]. Most of the estimated 6000 languages that exist in the world today are endangered [30, 63, 87] and getting replaced by a few dominant languages [67, 152].

One can then start to see the limitations of considering language as a neutral means of communication. As important as the act of communicating itself are the reasons and contexts of this communication.

1.2 LANGUAGE AND POLITICS

As Pierre Bourdieu argued, language is not only a means of communication but also a medium of power [19]. Some aspect of this idea has been popularized in the world of fiction with the concept of the *Newspeak* language in George Orwell's *1984* [119]. It illustrates how a control over the language spoken in society implies a better control over society itself. This has some echo in the real world [50], and not necessarily with dystopian, *Big Brother*, intentions of total control over individuals. When nation-states were built, a common language was seen as a means to unite a nation [157]. This was particularly the case in post-revolution France and imperial Great Britain [67, 74], where respectively French and English were heavily pushed as the languages of higher status. Still today, most constitutions specify one or a maybe a few languages as the languages of the state. In theory, this would be beneficial as it enables the state to build a common ground to guarantee equal opportunity, for instance with public education and the rule of law. It also would not necessarily mean dropping local languages, but only learning a common one, resulting then in a large population of multilinguals. In practice though, these political pushes for a shared language have lead to the near extinction of many regional languages and dialects. It is however not uncommon that a later reaction tried to reverse this process, with policies switching roles completely. Politics can then oppose homogenization and protect language diversity.

Current examples include the policies introduced to protect national languages against global English [140], and also the ones to preserve regional languages and dialects within nations [84].

1.3 LANGUAGE AS A COMMODITY

Language can also be seen as part of the set of skills that an individual possesses and may need to perform their job. Knowing a language has thus an economic value, and this is particularly true in a globalized economy [73]. Indeed, the world has not only become more interconnected in terms of communication, but also in terms of trade. It is the aspect of globalization that has had the most impact on our contemporary societies: in fact, when someone talks about globalization, most of the time what they are referring to is economic globalization. In this context, good command of a non-native language, like English in most cases, can often be a requirement to apply for a job. As a result, the status of a language, or its perceived value in society, can depend heavily on the value it is given by the market. As the sociologist Pierre Bourdieu put it, individuals, as they speak differently, possess different quantities of *linguistic capital* [19].

Further, the manner with which one speaks a language can identify them as member of a certain socio-economic group. Indeed, all languages have a number of varieties, some with superior status. The standard variety of a language, when one is identified as such by an official institution like a language academy, is often the most prestigious one. A language has then a "correct" way to be spoken and written, which is the one taught in schools. As proven by the PISA reports of the OECD, its latest included [117], in many countries, linguistic proficiency of 15-year-olds strongly varies based on their *socio-economic status* (SES) of origin. Individuals from low socio-economic classes can then be identified by their lack of command of the standard variety of their language, which translates into a lesser linguistic capital. This can be detrimental to these parts of a population, as these differences can entail segregation in several spheres of society, notably on the job market, but also in social interactions.

1.4 LANGUAGE AS A CULTURAL TRAIT

As a vector for communication, language is also necessarily central in cultural acquisition. It is even so intertwined with culture that some aspects of a culture may be embedded directly in a language. It follows that the diffusion of a culture goes hand in hand with the diffusion of a certain language. Here, language is to be understood in the broad sense: it can either be a language like English that is diffused by Hollywood cinema for instance, or a certain jargon within a language, like the (mostly English) vocabulary associated to the

Internet culture. As part of a culture, language may thus contribute to building a sense of group identity to which individuals may adhere. Conversely, rejecting the dominant, or mainstream, culture may also mean rejecting its language, and protecting one's own. It may also mean coming up with one's own language, as part of building a sub or counter-culture. This mechanism is not the only one at work that pushes for language diversity. In 2003, the UNESCO adopted the *Convention for the Safeguarding of the Intangible Cultural Heritage*, which states that language, "as a vehicle of the intangible cultural heritage", is to be safeguarded against the effects of globalization [150]. Also, very often, language preservation policies are implemented based on the argument that cultural diversity embedded in languages needs to be preserved [30, 63, 87].

1.5 LANGUAGE AND SEGREGATION

[92] [95]

1.6 SCOPE AND OUTLINE OF THE THESIS

If there is one chief takeaway from the last pages, it would be that language variation is complex. There is variation between languages, but also within languages which all have variants, as there are as many variants as speakers. There is variation in space and time. And there is variation for many, often entangled, reasons. Linguistics, and especially in its social branch, is therefore at the interface of many interwoven disciplines. We have shown how language and its study fall within the scope of various disciplines of social sciences, as we touched on subjects related to economics, communication science, human geography and politics. Throughout this work, we will also cross the boundaries between those, sometimes lying in-between. Our contribution is humble: we will neither address every aspect of language variation, nor provide the definitive explanation for one aspect of the problem. We will rather provide some further evidence and understanding of some of the phenomena at play.

Up next in Chapter 2, we will present the backbone of this work: the general methodology that has been used throughout the thesis, along with the previous literature that oriented our choices.

After this thorough methodological review, in Chapter 3 we will investigate inter-language competition in space. There, we will consider languages as coherent units that compete for speakers, which leads to geographically-embedded linguistic communities. Our goal is first to observe these, second to measure the differences between different kinds of language competition, and third to try to explain them.

In Chapter 4, we will turn to intra-language variation, still with a geographical component. We will see how socio-economic factors and social mixing can be predictors of the variation between speakers of a single language.

Chapter 5 deals again with intra-language variation in space, but this time investigating how these can reveal different cultural values among individuals sharing the same language.

Finally, Chapter 6 will allow us to take a step back and reflect on the road we have travelled during this thesis, and to envision what could be the next steps to go forward in this general direction.

2 METHODOLOGY

As shown in the previous chapter, language and society are interwoven in so many ways that sociolinguistics should be studied most carefully. As scientists, we would like to establish simple laws that explain reasonably well the interactions between society and language. But to proclaim a law is not enough to establish it: to do so one needs evidence that supports its claims. That is why we will start this chapter with a presentation of the empirical aspect of our methods, before introducing the kind of theoretical modelling that is relevant to this study.

2.1 DATA

2.1.1 *What for*

To understand a phenomenon, one should first observe it carefully. From the observation, we may then be able to make representative measurements of reality and encounter patterns, expected according to our previous knowledge and intuition, or not — the latter being the most interesting case. Indeed, since all models are approximations of reality, it is of utmost importance to be able to find out where they fall short. Observed data also serve as an essential guide to make sensible hypotheses on which to build models.

As we are dealing with language, there is no doubt that the centre of our attention should be the language produced by individuals. It is so omnipresent in our lives that we can find it anywhere there are human beings, in all kinds of context and in very different forms. The amount of information that is available is thus colossal. In comparison, our ability to retrieve it is very limited.

2.1.2 *Traditional sources in linguistics*

This was especially true historically. The field of linguistics is centuries-old, and has relied mostly on written texts and transcriptions of interviews throughout its history. But despite the considerable efforts that have been made to conserve written records, only a tiny proportion of texts survived through centuries. The diachronic study of how language evolves with time has thus very limited empirical resources. These sources also used to be only accessible physically, meaning the researcher would need to travel to collect the data — or the other way around. This has a cost and is a source of bias:

collecting a geographically uniform sample is very challenging in these conditions. Less accessible areas, and countries where there was no institution systematically keeping written records are thus vastly underrepresented.

The texts that were originally in written form present other significant biases. Since only the elites were able to write until a couple of centuries ago, the texts available to us from the distant past are not representative of the societies of the time. Also, for the very nature of the texts that were conserved, colloquial language is almost completely absent from those.

Transcriptions of oral productions may help in this regard. Indeed, they potentially give access to a different kind of language produced by more diverse speakers. They are thus very valuable, but unfortunately also very costly to produce, as they require direct involvement of the researcher in the collection process. This takes considerable time and effort. This direct involvement also calls for careful procedures to collect representative samples of languages, and avoid tainting them with the researchers' own biases.

Using the same kinds of sources to study the languages spoken today has become slightly easier. Travelling to most parts of the world is fast and affordable, and anyway collected texts can be digitized and then analysed and shared on large scales. There are now very large corpora of modern languages, produced in both written [11, 64, 65, 99] and spoken [89, 99, 135, 142] forms, that have been shared among researchers for a vast array of analyses. But the written texts still suffer from a lack of representativity, because, still today, few social classes publish books, articles or letters. As for interviews, although they are still highly relevant for the access they provide to spoken language, conducting and transcribing them faithfully still remains time-consuming. Hence the reduced size of the oral speech corpora, which are almost exclusively in English.

2.1.3 *New sources from online media*

The telecommunication age has brought great promise for the collection of natural language data. As we already mentioned, sharing information and texts has been made much easier. But more importantly, the very nature of the new channels of communication that have been opened allow for systematic collection of natural speech, both in spoken and written forms. Technically, a mobile network operator can record calls, an internet provider or a social media platform can record text sent through them — when not end-to-end encrypted. As these media become more and more globally accessible to people, gone should be the issues of sample representativity, and much easier should be the collection of linguistic data overall.

There are two major caveats to mention here though. First, the systematic collection of such data opens up the potential for serious privacy violations. That is why the majority of countries have extensive legislation to regulate the collection, processing and sharing of telecommunication data. Whether all communication service providers actually abide by these laws internally is doubtful [54, 55], but regardless, they cannot carelessly share private data with researchers. This brings us to the second caveat: these data are owned by private companies. There is little to no incentive for them to share their data, and especially if it requires efforts from them to anonymize the data or make sure that the use that is made of them does not breach privacy laws. That is why, when they do open channels for sharing them, they are almost always paid. Further, even with good will, respecting the privacy of individuals when processing their data is far from trivial. It has been shown repetitively that anonymization is tricky, as researchers managed to de-anonymize some public datasets [51, 110]. In order to conduct research that is both ethical and legal [88, 118], researchers dealing with such data thus have to take very particular care.

All in all, the advent of telecommunication has not proven to be the panacea for linguistic data that some may have hoped for. But there is still more linguistic material to study than ever, as demonstrated by the multiplication of the number of works in the field of computational (socio)-linguistics [114]. These have drawn form a variety of sources, like blogs [111, 134], online forums [9, 53, 111], online reviews [32, 76, 120], or a certain microblogging website called Twitter [2, 29, 96, 106, 158].

2.1.4 *The case of Twitter*

The biggest source of data we have used throughout this thesis is Twitter. Twitter is a microblogging website where people can register to share and view short posts called Tweets. In them, they can write, mention another user, share images, videos or links to other websites. The platform is called a *micro*blogging website because these posts cannot exceed 280 characters (140 before 2017). Tweets can have a public geotag if the user wishes to include one, which is suggested to the user when they tweet with their device's GPS turned on. Tweets can be of four kinds:

- a simple post that appears on the user profile and is shown on the homepage of all the users following this user, which is what people generally refer to with the term *Tweet*;
- a reply to another post, which can be seen by anyone but only shown on the homepage of the users involved in the conversation;

- a repost to one's profile, to share a Tweet that is already posted (which can be one's own), called *retweet* (RT);
- a retweet but with some added text commenting on the quoted post, called *quote retweet* (QRT).

There are many ways to interact with others, and Twitter thus hosts a huge network of inter-user interactions. It is one of the most popular online social media, with hundreds of millions of users worldwide. In the US, for instance, since 2015, more than 20 % of the population use the platform [5]. Other than in the US, it is also popular in many countries globally, although with a slight bias towards developed, western countries [71]. Even though only around 1 % of Tweets are geotagged [108], when only counting users who tweet with a geotag, in around 80 countries there is more than one Twitter user for ten thousand inhabitants [106]. Hence why Twitter has been extensively used for the analysis of geographically-embedded text [4, 15, 16, 59–61, 66, 77, 86, 92, 106, 112], and why do so in this thesis as well. In the following, we will thoroughly present the steps we take to leverage Twitter as a source of geotagged text.

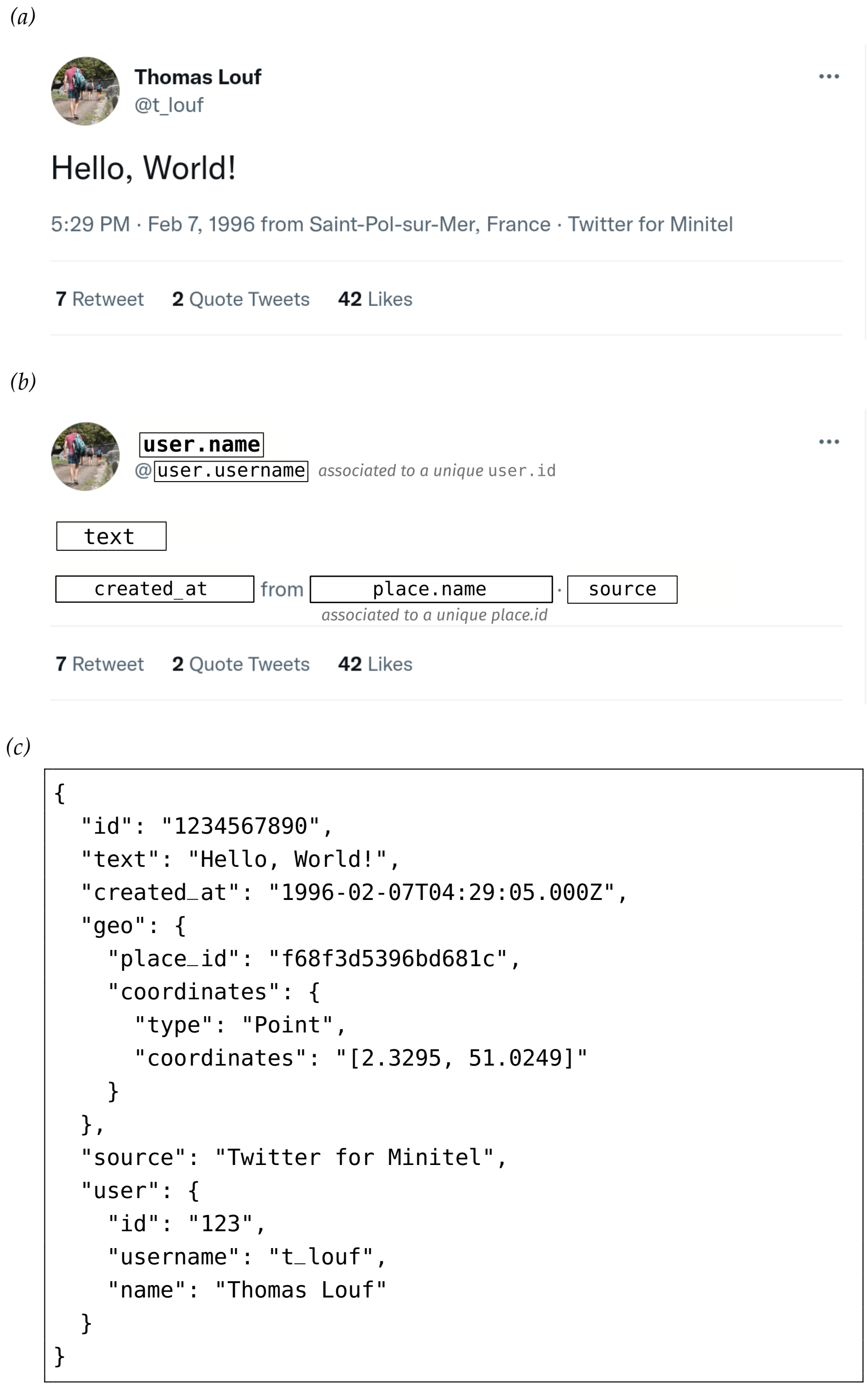
2.1.4.1 *Accessing the data*

A major advantage of Twitter for academic research is how open the platform is to giving access to its data to researchers. One can send automatic queries to Twitter for data through their public *application programming interface* (API) [148]. In these queries, one can specify rules to, for instance, retrieve all the Tweets posted in a given country, in a given time period, or which contain some given text. All the Twitter data we have used throughout this thesis was retrieved from the filtered stream endpoint of the Twitter API [149]. We show in Figure 2.1 an example of the data we can have for each Tweet.

There are two fields in these data that particularly interest us for the works we will present in the next part and that need careful processing: the textual content of the Tweet (in `text`) and its geotag (in `geo`). Next, we detail the usual steps we take to process these.

2.1.4.2 *Text processing*

Since we are interested in the speech produced by users, we need to clean parts of the text which cannot be considered as natural language production. Those are the URLs, mentions of other users (in the form `@username`) and hashtags (in the form `#topic`). It is not completely obvious that the latter should be discarded though. Hashtags are used on Twitter to aggregate Tweets by topics. It is an important feature of the website, whose aim is to enable users to easily find the Tweets of other users discussing similar topics, or inversely to make one's Tweets more discoverable by others, and to see real time trends on the



(a)

Thomas Louf
@t_louf

Hello, World!

5:29 PM · Feb 7, 1996 from Saint-Pol-sur-Mer, France · Twitter for Minitel

7 Retweet **2** Quote Tweets **42** Likes

(b)

user.name
@user.username *associated to a unique* user.id

text

created_at from place.name · source
associated to a unique place.id

7 Retweet **2** Quote Tweets **42** Likes

(c)

```
{
  "id": "1234567890",
  "text": "Hello, World!",
  "created_at": "1996-02-07T04:29:05.000Z",
  "geo": {
    "place_id": "f68f3d5396bd681c",
    "coordinates": {
      "type": "Point",
      "coordinates": "[2.3295, 51.0249]"
    }
  },
  "source": "Twitter for Minitel",
  "user": {
    "id": "123",
    "username": "t_louf",
    "name": "Thomas Louf"
  }
}
```

Figure 2.1: A Tweet data. We show (a) an example Tweet as displayed on Twitter and (b) a version annotated with the name of the fields in (c) the data as it would be sent by the API, which is simply text formatted in a dictionary-like structure (JSON).

platform. Hence, there can be completely different motivations behind writing a hashtag: to actually tag a Tweet with one or more topic, to promote the Tweet, or simply follow a trend. Thus, the content of hashtags can deviate significantly from normal speech [121]. It is therefore safer to discard hashtags entirely, which is no issue as long as we can collect enough textual content without them anyway. We actually made some measurements in our Tweets' database to see if that was the case. We took several random samples of a million Tweets each, stripped them of URLs and mentions, and then computed the ratio of characters within a hashtag compared to the total number of characters left in those Tweets. This proportion was found to be consistently below 5 %. We thus consider the precaution of stripping hashtags off of Tweets worth taking. One last kind of element that we discard are source-dependent. We will not go into details — our text processing code is freely available online anyway — but, for instance, when a Tweet was sent from Foursquare, we strip all location-related content, which can be located after either a “`I'm at`” or “`( @ `” string.

In practice, all the elements cited above are stripped off of Tweets using regular expressions. After this cleaning step, for what follows we then keep only the Tweets still containing at least four words. The next important step that was crucial to all our works was to infer the language the Tweets are written in. To do so, we leverage a trained neural network model for language identification: the Compact Language Detector [131]. It was designed as part of Chromium-based web browsers to detect the language web pages are written in in order to make translation suggestions to users, and it is now openly accessible. Its output is a language prediction along with the confidence of the model. Whenever we focus on a language, we thus keep Tweets which are tagged in that language with a confidence above 90 %.

We have thus described the basic steps of text pre-processing that are recurrent in our works. Out of it we get the Tweets for which we could reliably assign a language, stripped of the parts which are irrelevant to us.

2.1.4.3 *Inferring geolocation*

The steps presented above allow us to measure linguistic features of interest from our Tweets. A next step we usually take is to map those geographically. We are able to do so thanks to the information contained in the `"geo"` field of a Tweet, an example of which is shown in Figure 2.1. This example is actually a particular case, because of the presence of the `"coordinates"` field. This gives precise GPS coordinates of the location of the device used to send out the Tweet: a longitude, latitude pair. It is present in a Tweet's metadata when the device's GPS is enabled *and* when the user opted in for precise location tagging in the parameters of the application. As this setting is opt-in (so off by default), very few users actually have this enabled: from our measurements, roughly between 10 and 20 % of those who posted with their GPS enabled between 2015 and 2019, depending on the country. This setting has been in place since 2015, which is the starting year of the datasets we have used throughout this thesis. So when a user enables their device's GPS with precise geolocation disabled and this `"coordinates"` field is absent, how do we infer geolocation? In this majority case, the geotag we have is the `"place_id"` we show in the example. This identifies a place: a specific, named location, which can be of different scales:

- a country,
- an administrative unit: province, region or department for instance,
- a city,
- a *point of interest* (POI): any kind of public place: restaurant, school, event venue, etc. These are represented by a point, so Tweets tagged with a POI can be considered similarly to the ones with coordinates.

When a user tweets with their device's GPS activated, a place (usually the city they tweet from) is selected by default, and they can switch to another one from a list of close-by places. These places were fed to Twitter by Foursquare [49] (among others), which provides data down to a POI level for more than 190 countries. The geographical extent of places other than POIs is defined by bounding boxes.

To map linguistic features, Tweets must be attributed to the geographical areas of interest of the study. These may be defined by administrative boundaries (US counties, for example) or by us (a regular grid of cells of equal area, for instance). As "area" is an ambiguous term that can also refer to the measure of the size of a surface, in the following we will refer to these areas only by the term *cells*. So, when a Tweet has GPS coordinates or a POI as a geotag, the attribution is straightforward: there can be only one matching cell. When it has a place defined by a bounding box, it is not so trivial. The naive approach would be to take the centroid of the place and attribute to the cell containing it. This is problematic, though. As the cells to match are not necessarily regular, this method does not systematically match the place to the cell with the most overlap. A less naive approach would then consist in computing the area of overlap for every candidate cell and match to the one with biggest such area. This would still be an all-or-nothing attribution though. What if the place has 51 % of its area in one cell and 49 % in another? It would not be reasonable to attribute that Tweet to the first cell only. To account for the uncertainty we have when doing this cell matching, we thus rather do a partial attribution. We attribute the Tweet to possibly more than one cell, with ratios defined by the ones of the place's area that lies within each cell. For the example above, and when computing a basic metric like a count, this means that we attribute 51 % of the count to one cell and 49 % to the other. Because the scales of places span orders of magnitude, some may intersect many cells. There can then be so much uncertainty in the actual geographic origin of the Tweet that it is preferable to discard it. Our criterion here is that when the four cells which contain most of the place's area put together do not contain more than 90 % of its total area, the place, and all the Tweets assigned to it, are discarded.

As the activity of Twitter users was found to follow a log-normal distribution spanning almost four orders of magnitude [106], it can be preferable to compute metrics at the user level. Indeed, at the Tweet level, the linguistic behaviour of the most active users could overshadow the one of the many, less active users. Individuals are mobile but for the vast majority they have a preferred location, namely their place of residence. That is why we often strive to attribute a cell of residence to the users in our datasets. To explain the heuristics we defined for residence attribution, let us first formalize some notation. For each user u, there are two counts we get directly from their Tweets: the number of them with GPS coordinates that fall in cell c: $n_{u,c}^{\mathrm{GPS}}$, and those without coordinates but tagged as being from place p: $n_{u,p}$. We wish to compute $r_{u,c} \in \mathbb{R}^+$, the ratio of Tweets of user u in cell c. It can be decomposed into the contributions of those with GPS coordinates $r_{u,c}^{\mathrm{GPS}}$, and of the others: $r_{u,c}^{\mathrm{P}}$:

$$r_{u,c} = r_{u,c}^{\mathrm{GPS}} + r_{u,c}^{\mathrm{P}} = n_{u,c}^{\mathrm{GPS}} + r_{u,c}^{\mathrm{P}}. \tag{2.1}$$

Denoting A_p and $A_{p\cap c}$ the areas of the place p and of the intersection between p and c, respectively, the partial attribution described above yields:

$$r_{u,c} = n_{u,c}^{\mathrm{GPS}} + \sum_p n_{u,p} \frac{A_{p\cap c}}{A_p}. \tag{2.2}$$

To attribute a cell of residence to each user u, we first only consider cells where $r_{u,c} \geq 3$ and $r_{u,c} / \sum_{c'} r_{u,c'} \geq 0.1$. We also compute $r_{u,c}$ considering only Tweets posted at nighttime (from 6pm to 8am), that we denote $r_{u,c}^{\mathrm{NT}}$. Among those left, the cell of residence c^* is then the one such that $r_{u,c^*}^{\mathrm{NT}} / \sum_{c'} r_{u,c'}^{\mathrm{NT}} \geq 0.5$, if any. This roughly means that we impose that a user must have tweeted at least three times and at least 10 % of the time from that cell, and that at night the majority of their Tweets were from there. All users for whom a cell of residence cannot be attributed are subsequently discarded from the analysis. The three thresholds given above may be adjusted to each analysis, and also tweaked for sensitivity analyses.

2.1.4.4 *Selecting relevant users*

As we are interested in the natural speech produced by individuals, we actually start our analyses by filtering out users whose behaviour resembles that of a bot. We first eliminate those tweeting at an inhuman rate, set at an average of ten tweets per hour over their whole tweeting period. Then, we only keep those who tweeted either from a Twitter official app, Instagram, Foursquare or Tweetbot (a popular third-party app). These were selected because they are significantly popular among real users. Also, consecutive geolocations implying speeds higher than a plane's ($1000\,\mathrm{km\,h^{-1}}$) are detected to discard users. The final filter is optional: when we wish to only keep residents of the region considered, we impose for a user to have tweeted from there in at least three consecutive months.

2.1.4.5 *Caveats*

No data source is without bias, and Twitter is no exception. First, at the global scale, as we already mentioned, Twitter is more representative of people living in western, developed countries with widespread access to the Internet [71, 106]. In terms of more local geographical biases, densely populated urban areas are usually overrepresented [5, 80, 105]. As for demographics, Twitter users are on average younger [5, 113, 138], with more degrees and income, and more likely to be males [5, 105, 138] than the general population.

2.1.5 *Tools for natural language processing*

The increased availability of natural language data has also been followed by the development of new tools for *natural language processing* (NLP). Rather than an in-depth review, we will here simply mention a few algorithms and tools that were useful to us throughout this thesis. Nowadays NLP has become familiar to the general public through large language models: deep neural networks with billions of parameters trained on billions or soon-to-be trillions of tokens. Popular examples are Open AI's GPT models [21] or Google's BERT [37] and its derivatives [132], with an increasing number of pre-trained models made freely accessible [107, 156]. These kinds of models are best known for their derivatives optimized for usage as chatbots, but they have many variants and parts that can serve different functions: lemmatize, that is to find the base forms of words (like removing plurals or conjugation), classify documents, find similarity between words and documents by embedding them in a vector space (with for instance "word2vec" [101] and "tok2vec" [3], respectively). Although they are very powerful tools to carry out some tasks, their training is very costly, and most of the pre-trained models were trained on rather formal speech, like blog or news articles. They are thus not always well suited to analyse social media speech, as found on Twitter for instance. Some works have strived to normalize more informal texts, but doing so throws away valuable information [42]. Also, NLP has existed for some decades now, and some simpler tools developed further in the past may sometimes suffice to carry out some tasks. Latent semantic analysis [38] can be good enough to compute document similarities and then cluster documents together. Latent Dirichlet allocation [12] can perform good topic modelling in some settings. A recent work has shown that algorithms developed to infer communities in networks can be adapted to provide robust topic modelling and document clustering at the same time [56]. The tools are thus many, but the context in which they are used greatly conditions their actual power.

2.2 THEORETICAL MODELS

2.2.1 *What for?*

We mentioned some models just above, but there is an important distinction to make here with the models we will describe further down. The former are machine learning models: they are algorithms that, after being trained on input data, can predict some output when presented with new data. Hence their name: they are models learnt from data through an algorithm. They can even learn so much from the data that they can end up having learned the training data itself, which makes them unusable to make further predictions. This is what is called over-fitting. But there are many techniques, like regularization or cross-validation, that can be used to avoid this pitfall. They can also be so complex that they become what is known as a black-box model: a model that cannot be interpreted in terms of the influence of the input variables, and whose behaviour is thus unpredictable. Again, this issue can be alleviated with methods such as the computation of Shapley values [136, 143]. But crucially, what even the best-trained algorithms cannot provide is explanation. Models, as those we will mention further down, and as understood by most scientists, try to uncover the mechanisms underlying some observed phenomena. They do not only aim to predict, but also lay a fertile ground for further development, and may help understand other phenomena than the one they were initially intended to explain [43]. Additionally, they are built on explicit assumptions, which allow to clearly set out their scope of validity. Incidentally, a machine learning model needs input data to be trained, but how does one select what variables to look for, what measurements to make, and how? That is again where a theory can help. That is why many times, throughout the history of science, it was the theory that preceded the empirical works [43].

2.2.2 *What kind?*

One of the first things to consider when trying to model a phenomenon is at what level we wish to study it. We will here distinguish between two levels: the micro and the macro. There are two perspectives one can take when trying to understand the mechanisms at play that make individuals vary their language based on their interactions within society:

agent-based model (ABM)

2.3 SOURCE MATERIALS AND TOOLS

Following the principles of open science, throughout my thesis, I have made all source materials for my results openly accessible, whether they are codes[1] or datasets[2], including this very manuscript's[3]. Equally importantly, I believe, I have strived to use almost exclusively free and open source software in my work. I cannot realistically cite here all projects I have relied on to carry out my work, but I can cite a few central ones. I wrote all my code in the Python 3 programming language, using libraries such as NumPy [69], pandas [145] or GeoPandas [81]. In their vast majority, figures presented here were prepared with Matplotlib [78], and sometimes edited, or entirely drawn, with Inkscape[4].

This document was prepared using LaTeX with the `classicthesis` style[5] developed by André Miede and Ivo Pletikosić, and the LaTeX Workshop extension[6] of Visual Studio Code.

[1] Hosted on GitHub at https://github.com/TLouf
[2] Hosted on figshare at https://figshare.com/authors/Thomas_Louf/9441395
[3] Hosted at https://github.com/TLouf/phd-thesis
[4] Available at https://inkscape.org
[5] Hosted at https://www.ctan.org/pkg/classicthesis
[6] Hosted at https://github.com/James-Yu/LaTeX-Workshop

Part II

RESULTS

You can put some informational part preamble text here. Illo principalmente su nos. Non message *occidental* angloromanic da. Debitas effortio simplificate sia se, auxiliar summarios da que, se avantiate publicationes via. Pan in terra summarios, capital interlingua se que. Al via multo esser specimen, campo responder que da. Le usate medical addresses pro, europa origine sanctificate nos se.

3 CAPTURING THE DIVERSITY OF MULTILINGUAL SOCIETIES

TODO

TODO epigraphcite

The following chapter is a reprint of an article, 'Capturing the Diversity of Multilingual Societies', that was previously published by the author of this thesis with José J. Ramasco and David Sánchez [97].

Because major language shifts are bound to the passing of generations, this system has a considerable inertia. Despite this inertia intrinsic to language evolution, these changes are still taking place at dramatic speeds.

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 realised that the world may be facing a mass extinction of languages [30, 63, 87]. 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.

Modelling language shift has been the subject of much research in the last decades [14, 25], which employed various approaches such as the formulation of evolution equations based on ecological models [72, 83, 104, 125, 139], of reaction-diffusion equations [79, 82, 123, 127], or approaches within the framework of agent-based modelling [24, 26, 103, 127]. 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 non-trivial way, and as there is no model that can claim to be the universal solution to build spatial interaction networks [8], 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 [1]. 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 modelling attempts have been made in that direction [26, 122, 123, 151]. In these models, agents can be in a third state AB through which they have to pass to switch from being monolingual in a language to another. Apart from [127] 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 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, [114] 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 modelling. 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 behaviour of bilinguals and the ease of learning a language have their importance for the coexistence of languages. Finally, Sec. IV contains our conclusions.

3.1 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 the Supplementary Information (SI) Sec. II and Figs. S13 and S14 for Quebec [**supp**].

3.1.1 *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 geotagged. It has thus good potential as a data source to extract spatial distributions of language use, as shown in [41, 59, 61, 77, 106, 124]. 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 the SI [**supp**]), and analysed geotagged 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. 3.1). The cell size has to be adapted for each studied region, as explained in Sec. I C of the SI [**supp**]. We have checked the effect of modifying the cell size and made sure that our results are robust (see Supplementary Figs. S10, S11 and S12 [**supp**]). The language of the messages is detected using the *Compact Language Detector* (CLD) that provides the most likely language of a text from the messages along with a confidence. After thoroughly cleaning and analysing the collected Tweets, we obtained a sample of local Twitter users to which a cell of residence and a set of languages were attributed.

3.1.2 *Data access*

The aggregated data giving the counts of local users by language group by cell have been deposited on figshare[1]. The data on commuting patterns at the municipality level in Belgium were obtained from the 2011 census[2]. 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[3].

3.1.3 *Metrics*

Before introducing any metric, let us specify our definition of language groups. First, we focus only on the languages considered to be local in the area under consideration. 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 fulfil the condition in respectively one and 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 [6], 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 the sake of brevity.

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] https://figshare.com/articles/dataset/Spatial_distributions_of_languages_on_Twitter/14339321

[2] https://statbel.fgov.be/en/open-data/census-2011-matrix-commutes-sex

[3] https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dt-td/Index-eng.cfm

- concentration in cell i of L-speakers:

$$c_{L,i} = \frac{N_{L,i}}{N_L}, \tag{3.3}$$

- proportion of L-speakers in i's population:

$$p_{L,i} = \frac{N_{L,i}}{N_i}, \tag{3.4}$$

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 [106], we can define the polarization of a language A for every cell i in a bilingual system with languages A and B as

$$\theta_{A,i} = \frac{1}{2}(1 + p_{A,i} - p_{B,i}). \tag{3.5}$$

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 indicator 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 conceptualised 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 [130], and it was shown to be a proper distance (in the metric sense) between probability distributions [94]. Here, we consider the distributions defined by the signatures $P = \{(i, c_i)\}$ and $Q_L = \{(i, c_{L,i})\}$. We then define EMD_L as

$$\mathrm{EMD}_L \equiv \mathrm{EMD}(P, Q_L) = \sum_{i,j} \hat{f}_{ij} d_{ij}, \tag{3.6}$$

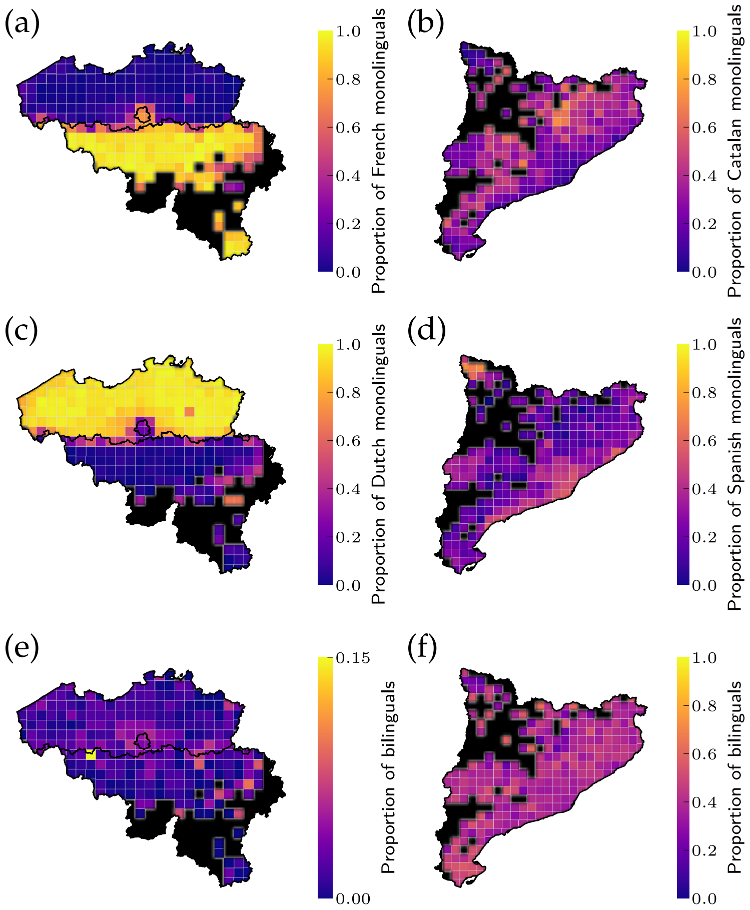


Figure 3.1: Paradigmatic examples illustrating the diversity of multilingual societies. For each cell of $10 \times 10\,\mathrm{km}^2$, the proportions $p_{L,i}$ of monolinguals in (a) French, (b) Catalan, (c) Dutch and (d) Spanish in Belgium (left) and Catalonia (right) are shown. The maps (e) and (f) 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 10 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).

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 $\sum_{i,j} f_{ij} d_{ij}$ under the following constraints:

$$\begin{cases} f_{ij} \geq 0, \forall\, i, j \\ \sum_j f_{ij} = c_{L,i}, \forall\, i \\ \sum_i f_{ij} = c_j, \forall\, j \\ \sum_i \sum_j f_{ij} = \sum_i c_{L,i} = \sum_j c_j = 1, \end{cases} \tag{3.7}$$

where c_i and $c_{L,i}$ are the concentrations of the population and L-speakers in every cell i, as defined above. 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 [48], which uses the method of [17]. However, in its raw form, it is dependent on the spatial scale of the system considered. Hence the need for a normalisation factor k_{EMD} in order to enable comparisons between regions of different sizes. The first, obvious choice for k_{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:

$$k_{\text{EMD}} = \frac{\sum_i \sum_j N_i N_j d_{ij}}{\left(\sum_k N_k\right)^2}. \tag{3.8}$$

Our final metric is then the normalised version of the EMD, the *earth mover's ratio* (EMR), defined as:

$$\text{EMR}_L = \frac{\text{EMD}_L}{k_{\text{EMD}}}. \tag{3.9}$$

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 the SI Fig. S13 and Sec. S14 **[supp]**, the EMR is cell size invariant and, quite generally, a reliable metric when a careful statistical analysis is made.

3.1.4 *Empirical results*

We propose a first visualisation of the collected data in Fig. 3.1(a)-(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 Supplementary

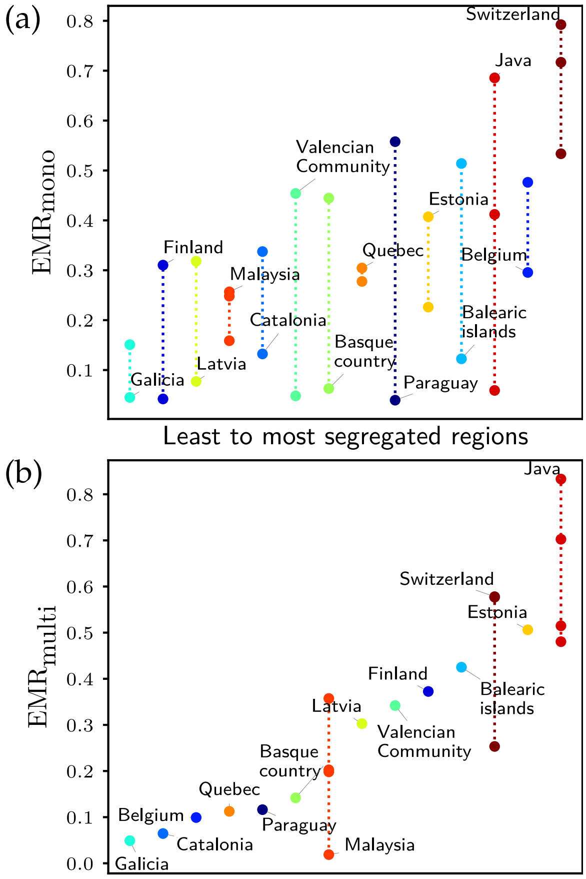


Figure 3.2: EMRs of the (a) monolingual and (b) multilingual groups of multilingual regions of interest, ranked left to right by increasing average of the y-axis values. In (b), the point for trilinguals in Switzerland is not displayed because its value was deemed unreliable (for more details see SI Sec. IF [**supp**]). A rich diversity of mixing patterns is shown, beyond the two paradigmatic cases of Catalonia and Belgium.

Figs. S10 and S11 [**supp**] for equivalent maps with cells of $5 \times 5\,\mathrm{km}^2$ and $15 \times 15\,\mathrm{km}^2$). 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 concentrates 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 (Fig. 3.1(e)-(f)). Results for the other multilingual regions listed in the SI Table S1 are shown in the SI Figs. S1-S14 [**supp**]. These findings are summarized in Fig. 3.2(g)-(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 the SI Table S2 for all measured metrics and cell sizes used [**supp**]). 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 localised. 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. 3.1 are shown using both data from the census and from Twitter for Quebec in the SI Figs. S13 and S14 [**supp**]. Similar mixing patterns can be observed from both data sources.

3.2 MODELS CAPTURING DIVERSITY

As language use in a society only sees significant changes on a timescale of generations [90], 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 fulfil it, the underlying mechanisms of language shift are not therein fully captured, missing a significant element that could be key to language preservation.

3.2.1 *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. Fig. 3.3 displays the states: monolingual in A and B, and bilingual AB, with the associated trans-

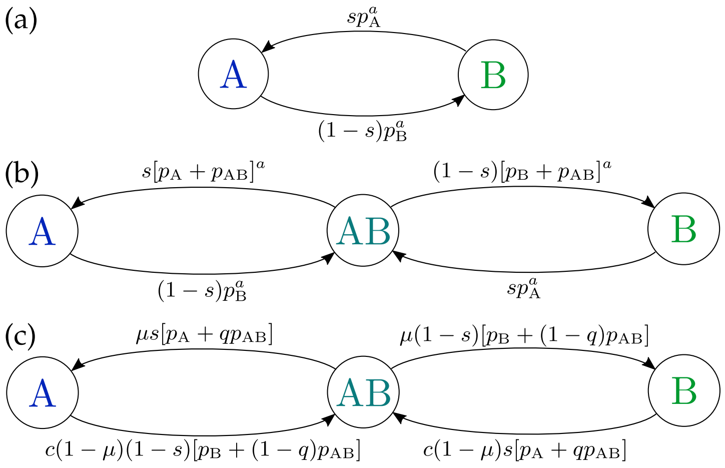

(a)
sp_{A}^{a}
A
B
$(1-s)p_{\mathrm{B}}^{a}$
(b)
$s[p_{\mathrm{A}}+p_{\mathrm{AB}}]^{a}$
$(1-s)[p_{\mathrm{B}}+p_{\mathrm{AB}}]^{a}$
A
AB
B
$(1-s)p_{\mathrm{B}}^{a}$
sp_{A}^{a}
(c)
$\mu s[p_{\mathrm{A}}+qp_{\mathrm{AB}}]$
$\mu(1-s)[p_{\mathrm{B}}+(1-q)p_{\mathrm{AB}}]$
A
AB
B
$c(1-\mu)(1-s)[p_{\mathrm{B}}+(1-q)p_{\mathrm{AB}}]$
$c(1-\mu)s[p_{\mathrm{A}}+qp_{\mathrm{AB}}]$

Figure 3.3: Diagrams of the models presented in the text, showing the transition probabilities from one state to another. (a) Abrams-Strogatz model from [1]. (b) Bilinguals model from [26]. (c) Our model of bilinguals including both their preference and the ease to learn the other language (see equation (3.12)).

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

The first model to mention is the one introduced in Ref. [1] by Abrams and Strogatz (Fig. 3.3A). 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 [1]. It was thoroughly analysed in [151], 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. [26] (see Fig. 3.3(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. [151] 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 [45] with a linguistic functional specialisation. A difference in prestige favours 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 [30, 126]. 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 favoured 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. 3.3(b) is not able to produce a stable solution in which the bilinguals coexist with monolinguals of a single language.

3.2.2 *Our model*

Our proposal stems from the realisation 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. [103] 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:

$$P(\mathrm{AB} \to \mathrm{X}) = \mu \, s_{\mathrm{X}} \, \left[p_{\mathrm{X}} + q_{\mathrm{X}} \, p_{\mathrm{AB}} \right], \tag{3.10}$$

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, $q_{\mathrm{X}} \in [0,1]$ 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:

$$P(\mathrm{X} \to \mathrm{AB}) = c \, (1 - \mu) \, s_{\mathrm{Y}} \, \left[p_{\mathrm{Y}} + q_{\mathrm{Y}} \, p_{\mathrm{AB}} \right], \tag{3.11}$$

where Y is the language other than X, and, critically, $c \in [0,1]$ is a factor adjusting the learning rate. The time scales 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 $s = s_{\mathrm{A}} = 1 - s_{\mathrm{B}}$ and $q = q_{\mathrm{A}} = 1 - q_{\mathrm{B}}$. The transitions in our model are illustrated in Fig. 3.3(c) and we write here below the transition probabilities that define it:

$$\begin{cases} P(\mathrm{A} \to \mathrm{AB}) = c \, (1 - \mu) \, (1 - s) \, \left[p_{\mathrm{B}} + (1 - q) \, p_{\mathrm{AB}} \right] \\ P(\mathrm{B} \to \mathrm{AB}) = c \, (1 - \mu) \, s \, \left[p_{\mathrm{A}} + q \, p_{\mathrm{AB}} \right] \\ P(\mathrm{AB} \to \mathrm{A}) = \mu \, s \, \left[p_{\mathrm{A}} + q \, p_{\mathrm{AB}} \right] \\ P(\mathrm{AB} \to \mathrm{B}) = \mu \, (1 - s) \, \left[p_{\mathrm{B}} + (1 - q) \, p_{\mathrm{AB}} \right] \end{cases} . \tag{3.12}$$

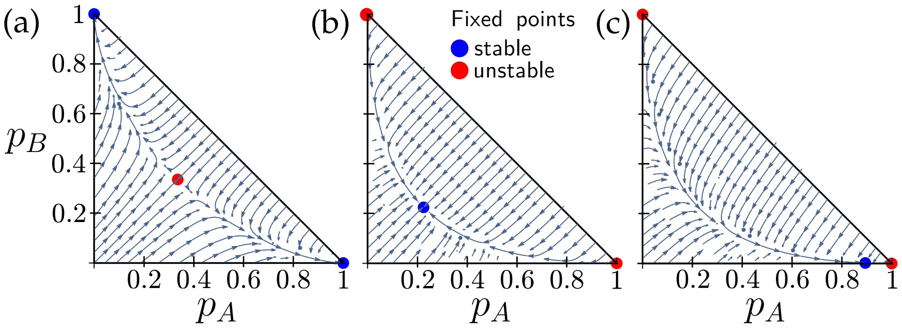


Figure 3.4: Flow diagrams for the dynamics of two languages according to our model described in equation (3.12) set in a well-mixed population. p_A and p_B denote the proportions of monolinguals in A and B, respectively, and the proportion of bilinguals p_{AB} is such that $p_A + p_B + p_{AB} = 1$. The 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.

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 [129]. 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 non-linearities 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.

3.2.3 *A single population*

We first analyse the model in the simplest setting of a single well-mixed population to determine the typology of possible solutions. Given the normalisation 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 the SI Sec. III **[supp]**). 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. 3.4 show where the model converges to in three characteristic examples, depending on the model parameters. In the first one (Fig. 3.4(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. 3.4(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. 3.4(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.

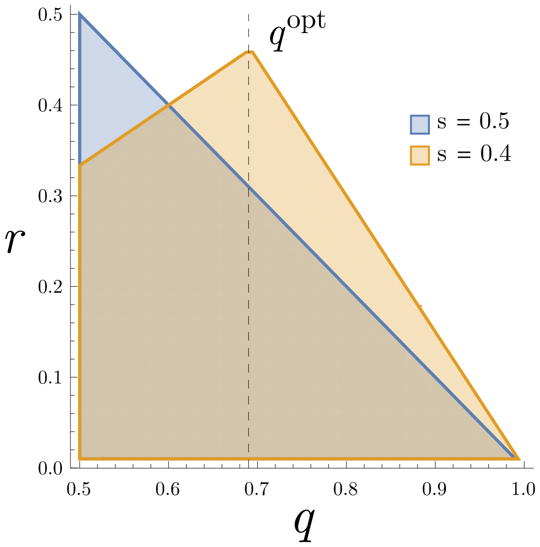


Figure 3.5: Region of the parameter space where the dynamics of our model in a single population converge to stable coexistence of languages. We show two 2D cuts of the coexistence region in the (q, r) space for fixed values of $s = 0.5, 0.4$, with $r = \mu/(c\,(1-\mu))$. Lower values of r favour coexistence, as well as a neutral prestige and bilingual preference q. When $s < 0.5$, coexistence is favoured for an optimal value $q^{\mathrm{opt}} > 1 - s$.

We change now the viewpoint from the phase space to the parameter space. In Fig. 3.5, we plot the region of parameters where the model converges to stable coexistence. Since c and μ 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, $q^{\mathrm{opt}} > 1 - s$. For $q < q^{\mathrm{opt}}$, A is more at risk of extinction whereas for $q > q^{\mathrm{opt}}$, 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 favours 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 analyse the model in a metapopulation framework to uncover the effect of including space and check whether this pattern can arise.

3.2.4 *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 [68] and epidemiology [7, 133]. 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 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 using a model of human mobility, such as gravity, radiation or distance-kernel-based models [8, 22, 23].

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 the SI Sec. IV A [**supp**]). Regarding

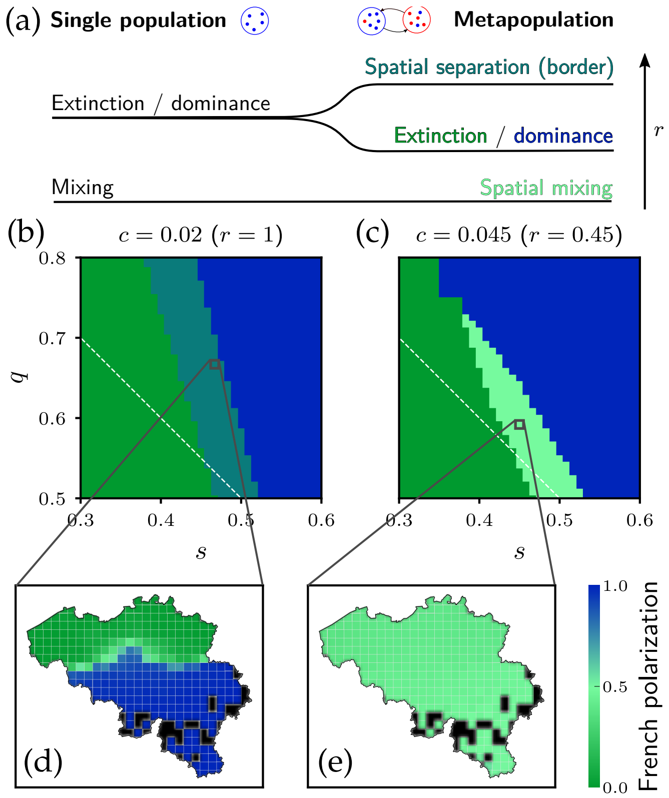


Figure 3.6: Types of stable states of convergence of our model in a metapopulation set for Belgium. (a) Diagram illustrating the effect of adding metapopulations in the stable states of a single population: the former extinction state bifurcates in full extinction and in a boundary-like state with monolinguals separated in space. Larger values of r favour homogeneity, either by full extinction or by separation states (see the SI Fig. S16 for more r values [**supp**]). 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 $r = 0.45$, $s = 0.45$, $q = 0.592$.

the language groups, Twitter data may suffer from different sociodemographic biases [105, 112], 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) [34, 47].

Once the metapopulation has been initialised, the model can be simulated. As in Ref. [46], 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 equation (3.12), 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 analyse 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 [7, 133] for epidemiological models (see the SI Sec. IV B for details [**supp**]). 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 state. We show the results for the metapopulation setting of Belgium in Fig. 3.6. 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. 3.6(e)), global extinction and local extinction of a language in part of the territory leading to a boundary-like state (Fig. 3.6(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 decreasing r, or in other words making it easier or more common to learn the other language, does not necessarily favour coexistence. Indeed, as r decreases, at one point boundary states become unstable and this may not necessarily lead to fully mixed states. When r shrinks 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 low 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. 3.5 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 SI Fig. S7 [**supp**]). We initialised 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 SI Fig. S15 [**supp**]).

3.2.5 *Dynamics in the parameters*

The effect of multilingual education or, in general, policies favouring 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 23000 steps and present the results in Fig. 3.7. 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. 3.7(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. 3.6(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. The resulting trajectory in the EMR space in Fig. 3.7(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, not-

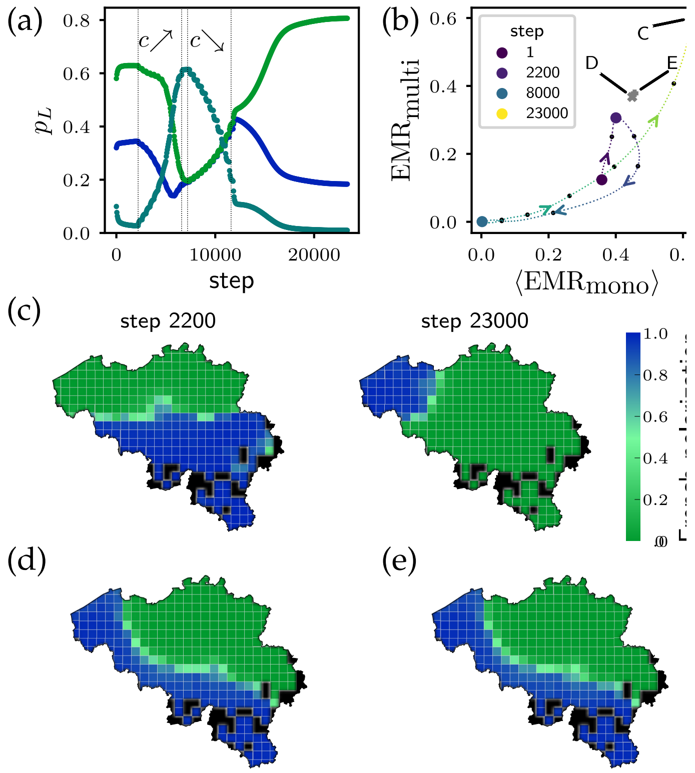


Figure 3.7: Evolution of the state of the metapopulation model in Belgium when c varies, first slowly increased and then decreased to recover the original value. We fixed $s = q = 1/2$ and $\mu = 0.02$. (a) Evolution of the global proportions p_L of individuals belonging to each L-group. The blue curve 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 coloured 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)-(e) Polarization maps of French in the final states of simulations including transborder 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 panel (b).

ably from its direct neighbours. Thus, we ran the same simulation with transborder 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 SI Sec. IV D [**supp**]). 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 Fig. 3.7(d-e)). This positioning is a clear improvement over the closed-system simulation, albeit still not quite the one we observed in Fig. 3.1. In Fig. 3.7(b), the positions of these two states in the EMR space 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 [137]. 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.

3.3 DISCUSSION

This chapter has presented our exploration of the spatial distribution patterns of language competition and coexistence in multilingual societies. It consisted in first 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 defence 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 counterintuitively, 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 [84] in many countries where accelerated changes are threatening cultural diversity.

4

SES X LANGUAGE

This was not to say that Albertine had not already possessed [...] a quite adequate assortment of those expressions which reveal at once that one's people are in easy circumstances, and which, year by year, a mother passes on to her daughter just as she bestows on her [...] her own jewels.

— *Marcel Proust,* The Guermantes Way *[128]*

5 ACR

To be rooted is perhaps the most important and least recognized need of the human soul. It is one of the hardest to define.

— *Simone Weil,* The Need for Roots [155]

Part III

CONCLUSION

6

CONCLUSION

Guido (Marcello Mastroianni)
I'm not afraid anymore of telling the truth, of the things I don't know, what I'm looking for and what I haven't found yet. This is the only way I can feel alive and I can look into your faithful eyes without shame.

— *Federico Fellini,* $8\frac{1}{2}$ *[44]*

Part IV

APPENDIX

A APPENDIX TEST

Lorem ipsum at nusquam appellantur his, ut eos erant homero concludaturque. Albucius appellantur deterruisset id eam, vivendum partiendo dissentiet ei ius. Vis melius facilisis ea, sea id convenire referrentur, takimata adolescens ex duo. Ei harum argumentum per. Eam vidit exerci appetere ad, ut vel zzril intellegam interpretaris.

More dummy text.

A.1 APPENDIX SECTION TEST

Test: Table A.1 (This reference should have a lowercase, small caps A if the option `floatperchapter` is activated, just as in the table itself → however, this does not work at the moment.)

LABITUR BONORUM PRI NO	QUE VISTA	HUMAN
fastidii ea ius	germano	demonstratea
suscipit instructior	titulo	personas
quaestio philosophia	facto	demonstrated

Table A.1: Autem usu id.

A.2 ANOTHER APPENDIX SECTION TEST

Equidem detraxit cu nam, vix eu delenit periculis. Eos ut vero constituto, no vidit propriae complectitur sea. Diceret nonummy in has, no qui eligendi recteque consetetur. Mel eu dictas suscipiantur, et sed placerat oporteat. At ipsum electram mei, ad aeque atomorum mea. There is also a useless Pascal listing below: Listing A.1.

Listing A.1: A floating example (`listings` manual)

```
for i:=maxint downto 0 do
begin
{ do nothing }
end;
```

BIBLIOGRAPHY

[1] Daniel M. Abrams and Steven H. Strogatz. 'Modelling the Dynamics of Language Death'. In: *Nature* 424.6951 (Aug. 2003), p. 900. ISSN: 00280836. DOI: 10.1038/424900a.

[2] Thayer Alshaabi et al. 'Storywrangler: A Massive Exploratorium for Sociolinguistic, Cultural, Socioeconomic, and Political Timelines Using Twitter'. In: *Science Advances* 7.29 (16th July 2021), eabe6534. DOI: 10.1126/sciadv.abe6534.

[3] Dimo Angelov. *Top2Vec: Distributed Representations of Topics*. 19th Aug. 2020. DOI: 10.48550/arXiv.2008.09470.

[4] Rudy Arthur and Hywel T. P. Williams. 'The Human Geography of Twitter: Quantifying Regional Identity and Inter-Region Communication in England and Wales'. In: *PLOS ONE* 14.4 (15th Apr. 2019), e0214466. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0214466.

[5] Brooke Auxier and Monica Anderson. *Social Media Use in 2021*. Pew Research Center, 2021.

[6] Colin Baker. *Foundations of Bilingual Education and Bilingualism*. Vol. 31. Bilingual Education and Bilingualism 79. Bristol, UK ; Tonawanda, NY: Multilingual Matters, 1997. 378 pp. ISBN: 978-1-84769-356-3.

[7] Duygu Balcan et al. 'Modeling the Spatial Spread of Infectious Diseases: The Global Epidemic and Mobility Computational Model'. In: *Journal of Computational Science* 1.3 (Aug. 2010), pp. 132–145. ISSN: 18777503. DOI: 10.1016/j.jocs.2010.07.002.

[8] Hugo Barbosa et al. 'Human Mobility: Models and Applications'. In: *Physics Reports* 734 (2018), pp. 1–74. DOI: 10.1016/j.physrep.2018.01.001.

[9] Anton Barua, Stephen W. Thomas and Ahmed E. Hassan. 'What Are Developers Talking about? An Analysis of Topics and Trends in Stack Overflow'. In: *Empirical Software Engineering* 19.3 (1st June 2014), pp. 619–654. ISSN: 1573-7616. DOI: 10.1007/s10664-012-9231-y.

[10] Matthew R. Bennett et al. 'Evidence of Humans in North America during the Last Glacial Maximum'. In: *Science* 373.6562 (24th Sept. 2021), pp. 1528–1531. DOI: 10.1126/science.abg7586.

[11] Douglas Biber, Susan Conrad and Randi Reppen. 'Corpus-Based Investigations of Language Use'. In: *Annual Review of Applied Linguistics* 16 (Mar. 1996), pp. 115–136. ISSN: 1471-6356, 0267-1905. DOI: 10.1017/S0267190500001471.

[12] David M. Blei, Andrew Y. Ng and Michael I. Jordan. 'Latent Dirichlet Allocation'. In: *The Journal of Machine Learning Research* 3 (null 1st Mar. 2003), pp. 993–1022. ISSN: 1532-4435.

[13] Jan Blommaert. *The Sociolinguistics of Globalization*. Cambridge University Press, 2010.

[14] Michael Boissonneault and Paul Vogt. 'A Systematic and Interdisciplinary Review of Mathematical Models of Language Competition'. In: *Humanities and Social Sciences Communications 2021 8:1* 8.1 (22nd Jan. 2021), p. 21. ISSN: 2662-9992. DOI: 10.1057/s41599-020-00683-9.

[15] Eszter Bokányi, Dániel Kondor and Gábor Vattay. 'Scaling in Words on Twitter'. In: *Royal Society Open Science* 6.10 (2nd Oct. 2019), p. 190027. ISSN: 20545703. DOI: 10.1098/rsos.190027.

[16] Eszter Bokányi et al. 'Race, Religion and the City: Twitter Word Frequency Patterns Reveal Dominant Demographic Dimensions in the United States'. In: *Palgrave Communications* 2.1 (26th Apr. 2016), pp. 1–9. ISSN: 2055-1045. DOI: 10.1057/palcomms.2016.10.

[17] Nicolas Bonneel et al. 'Displacement Interpolation Using Lagrangian Mass Transport'. In: *ACM Transactions on Graphics* 30.6 (12th Dec. 2011), pp. 1–12. ISSN: 0730-0301. DOI: 10.1145/2070781.2024192.

[18] Rudolf Botha and Chris Knight. *The Cradle of Language*. OUP Oxford, 30th Apr. 2009. 418 pp. ISBN: 978-0-19-156767-4.

[19] Pierre Bourdieu. *Language and Symbolic Power*. Ed. by John B. Thompson. Trans. by Gino Raymond and Matthew Adamson. Cambridge: Polity Press, 2009. 302 pp. ISBN: 0-7456-0097-2.

[20] Donald Brown. *Human Universals*. New York: McGraw-Hill, 1991. x, 220. ISBN: 978-0-07-008209-0.

[21] Tom B. Brown et al. *Language Models Are Few-Shot Learners*. 22nd July 2020. DOI: 10.48550/arXiv.2005.14165.

[22] James Burridge. 'Spatial Evolution of Human Dialects'. In: *Physical Review X* 7.3 (17th July 2017), p. 031008. ISSN: 21603308. DOI: 10.1103/PhysRevX.7.031008.

[23] James Burridge and Tamsin Blaxter. 'Inferring the Drivers of Language Change Using Spatial Models'. In: *Journal of Physics: Complexity* 2.3 (20th July 2021), p. 035018. ISSN: 2632072X. DOI: 10.1088/2632-072X/abfa82.

[24] Inés Caridi et al. 'Schelling-Voter Model: An Application to Language Competition'. In: *Chaos, Solitons and Fractals* 56 (1st Nov. 2013), pp. 216–221. ISSN: 09600779. DOI: 10.1016/j.chaos.2013.08.013.

[25] Claudio Castellano, Santo Fortunato and Vittorio Loreto. 'Statistical Physics of Social Dynamics'. In: *Reviews of Modern Physics* 81.2 (2009), pp. 591–646. ISSN: 00346861. DOI: 10.1103/RevModPhys.81.591.

[26] Xavier Castelló, Víctor M. Eguíluz and Maxi San Miguel. 'Ordering Dynamics with Two Non-Excluding Options: Bilingualism in Language Competition'. In: *New Journal of Physics* 8.12 (6th Dec. 2006), pp. 308–308. ISSN: 13672630. DOI: 10.1088/1367-2630/8/12/308.

[27] Jack K. Chambers. *Sociolinguistic Theory: Linguistic Variation and Its Social Significance*. 2. ed., [reprinted]. Language in Society 22. Malden, Mass.: Blackwell, 2007. 320 pp. ISBN: 978-0-631-22882-0 978-0-631-22881-3.

[28] Noam Chomsky. *Language and Mind: Current Thoughts on Ancient Problems*. Brill, 1st Jan. 2004, pp. 379–405. ISBN: 978-0-08-047474-8. DOI: 10.1163/9780080474748_018.

[29] Emily M. Cody et al. 'Climate Change Sentiment on Twitter: An Unsolicited Public Opinion Poll'. In: *PLOS ONE* 10.8 (20th Aug. 2015), e0136092. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0136092.

[30] David Crystal. *Language Death*. Cambridge University Press, 26th June 2000. DOI: 10.1017/cbo9781139106856.

[31] David Crystal. *English as a Global Language*. Cambridge: Cambridge Univ. Press, 2010. 212 pp. ISBN: 978-0-521-53032-3 978-0-521-82347-0.

[32] Cristian Danescu-Niculescu-Mizil et al. 'No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities'. In: *Proceedings of the 22nd International Conference on World Wide Web*. WWW '13. New York, NY, USA: Association for Computing Machinery, 13th May 2013, pp. 307–318. ISBN: 978-1-4503-2035-1. DOI: 10.1145/2488388.2488416.

[33] Dan Dediu and Stephen Levinson. 'On the Antiquity of Language: The Reinterpretation of Neandertal Linguistic Capacities and Its Consequences'. In: *Frontiers in Psychology* 4 (2013). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2013.00397.

[34] W. Edwards Deming and Frederick F. Stephan. 'On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known'. In: *The Annals of Mathematical Statistics* 11.4 (Dec. 1940), pp. 427–444. DOI: 10.1214/aoms/1177731829.

[35] Ferdinand de Saussure. *Course in General Linguistics*. Ed. by Perry Meisel. Trans. by Wade Baskin Edited by Perry Meisel and Haun Saussy. Columbia University Press, June 2011, 336 Pages. ISBN: 978-0-231-52795-8.

[36] Ithiel de Sola Pool and Manfred Kochen. 'Contacts and Influence'. In: *Social Networks* 1.1 (1st Jan. 1978), pp. 5–51. ISSN: 0378-8733. DOI: 10.1016/0378-8733(78)90011-4.

[37] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 24th May 2019. DOI: 10.48550/arXiv.1810.04805.

[38] Susan T. Dumais. 'Latent Semantic Analysis'. In: *Annual Review of Information Science and Technology* 38.1 (2004), pp. 188–230. ISSN: 1550-8382. DOI: 10.1002/aris.1440380105.

[39] Robin I. M. Dunbar. 'Neocortex Size as a Constraint on Group Size in Primates'. In: *Journal of Human Evolution* 22.6 (1st June 1992), pp. 469–493. ISSN: 0047-2484. DOI: 10.1016/0047-2484(92)90081-J.

[40] Robin I. M. Dunbar. 'The Social Brain Hypothesis'. In: *Evolutionary Anthropology: Issues, News, and Reviews* 6.5 (1998), pp. 178–190. ISSN: 1520-6505. DOI: 10.1002/(SICI)1520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-8.

[41] Jonathan Dunn. 'Mapping Languages: The Corpus of Global Language Use'. In: *Language Resources and Evaluation* 54.4 (1st Dec. 2020), pp. 999–1018. ISSN: 15728412. DOI: 10.1007/s10579-020-09489-2.

[42] Jacob Eisenstein. 'What to Do about Bad Language on the Internet'. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL-HLT 2013. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 359–369.

[43] Joshua M. Epstein. 'Why Model?' In: *Journal of Artificial Societies and Social Simulation* 11.4 (2008), p. 12. ISSN: 1460-7425.

[44] Federico Fellini, director. *8$\frac{1}{2}$*. scriptwriter Federico Fellini et al. Drama. 24th June 1963.

[45] Charles A. Ferguson. 'Diglossia'. In: *Word* 15.2 (Jan. 1959), pp. 325–340. ISSN: 0043-7956. DOI: 10.1080/00437956.1959.11659702.

[46] Juan Fernández-Gracia et al. 'Is the Voter Model a Model for Voters?' In: *Physical Review Letters* 112.15 (18th Apr. 2014), p. 158701. ISSN: 10797114. DOI: 10.1103/PhysRevLett.112.158701.

[47] Stephen E. Fienberg. 'An Iterative Procedure for Estimation in Contingency Tables'. In: *The Annals of Mathematical Statistics* 41.3 (June 1970), pp. 907–917. ISSN: 0003-4851. DOI: 10.1214/aoms/1177696968.

[48] Rémi Flamary et al. 'POT: Python Optimal Transport'. In: *Journal of Machine Learning Research* 22.78 (2021), pp. 1–8. ISSN: 1533-7928.

[49] *Foursquare Places Service*. 2019. URL: https://developer.foursquare.com/places.

[50] Roger Fowler et al. *Language and Control*. 1st ed. London: Routledge, 1979. 232 pp. ISBN: 978-0-429-43621-5. DOI: 10.4324/9780429436215.

[51] Sébastien Gambs, Marc-Olivier Killijian and Miguel Núñez del Prado Cortez. 'De-Anonymization Attack on Geolocated Data'. In: *Journal of Computer and System Sciences*. Special Issue on Theory and Applications in Parallel and Distributed Computing Systems 80.8 (1st Dec. 2014), pp. 1597–1614. ISSN: 0022-0000. DOI: 10.1016/j.jcss.2014.04.024.

[52] Ruth García-Gavilanes, Yelena Mejova and Daniele Quercia. 'Twitter Ain't without Frontiers: Economic, Social, and Cultural Boundaries in International Communication'. In: *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW'14: Computer Supported Cooperative Work. Baltimore Maryland USA: ACM, 15th Feb. 2014, pp. 1511–1522. ISBN: 978-1-4503-2540-0. DOI: 10.1145/2531602.2531725.

[53] Matt Garley and Julia Hockenmaier. 'Beefmoves: Dissemination, Diversity, and Dynamics of English Borrowings in a German Hip Hop Forum'. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. ACL 2012. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 135–139.

[54] *GDPR Enforcement Tracker - List of GDPR Fines*. URL: https://www.enforcementtracker.com (visited on 16/02/2023).

[55] *GDPR Fines and Notices*. In: *Wikipedia*. 8th Feb. 2023.

[56] Martin Gerlach, Tiago P. Peixoto and Eduardo G. Altmann. 'A Network Approach to Topic Models'. In: *Science Advances* 4.7 (18th July 2018), eaaq1360. DOI: 10.1126/sciadv.aaq1360.

[57] Kathleen R. Gibson and Maggie Tallerman. *The Oxford Handbook of Language Evolution*. Oxford University Press, Nov. 2011. ISBN: 978-0-19-954111-9. DOI: 10.1093/oxfordhb/9780199541119.001.0001.

[58] Bruno Gonçalves, Nicola Perra and Alessandro Vespignani. 'Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number'. In: *PLOS ONE* 6.8 (3rd Aug. 2011), e22656. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0022656.

[59] Bruno Gonçalves and David Sánchez. 'Crowdsourcing Dialect Characterization through Twitter'. In: *PLOS ONE* 9.11 (19th Nov. 2014). Ed. by Tobias Preis, e112074. ISSN: 19326203. DOI: 10.1371/journal.pone.0112074.

[60] Bruno Gonçalves and David Sánchez. 'Learning about Spanish Dialects through Twitter'. In: *Revista Internacional de Linguistica Iberoamericana* 14.2 (16th Nov. 2016), pp. 65–75. ISSN: 15799425.

[61] Bruno Gonçalves et al. 'Mapping the Americanization of English in Space and Time'. In: *PLOS ONE* 13.5 (25th May 2018). Ed. by Tobias Preis, e0197741. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0197741.

[62] Joseph Greenberg. *Language Universals*. De Gruyter Mouton, 20th Jan. 2020. ISBN: 978-3-11-080252-8. DOI: 10.1515/9783110802528.

[63] Lenore A. Grenoble and Lindsay J. Whaley. *Endangered Languages: Language Loss and Community Response*. Cambridge University Press, 1998. 384 pp. ISBN: 978-0-521-59712-8.

[64] Jack Grieve. 'A corpus-based regional dialect survey of grammatical variation in written Standard American English'. PhD thesis. Ann Arbor, United States, 2009. 340 pp. ISBN: 9781109318388.

[65] Jack Grieve. *Regional Variation in Written American English*. Cambridge University Press, 2016. ISBN: 978-1-139-50613-7. DOI: 10.1017/CBO9781139506137.

[66] Jack Grieve et al. 'Mapping Lexical Dialect Variation in British English Using Twitter'. In: *Frontiers in Artificial Intelligence* 2 (12th July 2019), p. 11. ISSN: 2624-8212. DOI: 10.3389/frai.2019.00011.

[67] R. D. Grillo. *Dominant Languages : Language and Hierarchy in Britain and France*. In collab. with Internet Archive. Cambridge ; New York : Cambridge University Press, 1989. 282 pp. ISBN: 978-0-521-36540-6.

[68] Ilkka Hanski. 'Metapopulation Dynamics'. In: *Nature* 396.6706 (5th Nov. 1998), pp. 41–49. ISSN: 00280836. DOI: 10.1038/23876.

[69] Charles R. Harris et al. 'Array Programming with NumPy'. In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

[70] Marc D. Hauser et al. 'The Mystery of Language Evolution'. In: *Frontiers in Psychology* 5 (2014). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2014.00401.

[71] Bartosz Hawelka et al. 'Geo-Located Twitter as Proxy for Global Mobility Patterns'. In: *Cartography and Geographic Information Science* 41.3 (27th May 2014), pp. 260–271. ISSN: 1523-0406. DOI: 10.1080/15230406.2014.890072.

[72] E. Heinsalu, M. Patriarca and J. L. Léonard. 'The Role of Bilinguals in Language Competition'. In: *Advances in Complex Systems* 17.1 (20th Apr. 2014). ISSN: 02195259. DOI: 10.1142/S0219525914500039.

[73] Monica Heller. 'The Commodification of Language'. In: *Annual Review of Anthropology* 39.1 (21st Oct. 2010), pp. 101–114. ISSN: 0084-6570, 1545-4290. DOI: 10.1146/annurev.anthro.012809.104951.

[74] Patrice L.-R. Higonnet. 'The Politics of Linguistic Terrorism and Grammatical Hegemony during the French Revolution'. In: *Social History* 5.1 (1st Jan. 1980), pp. 41–69. ISSN: 0307-1022. DOI: 10.1080/03071028008567470.

[75] R. A. Hill and Robin I. M. Dunbar. 'Social Network Size in Humans'. In: *Human Nature* 14.1 (1st Mar. 2003), pp. 53–72. ISSN: 1936-4776. DOI: 10.1007/s12110-003-1016-y.

[76] Dirk Hovy, Anders Johannsen and Anders Søgaard. 'User Review Sites as a Resource for Large-Scale Sociolinguistic Studies'. In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 18th May 2015, pp. 452–461. ISBN: 978-1-4503-3469-3. DOI: 10.1145/2736277.2741141.

[77] Yuan Huang et al. 'Understanding U.S. Regional Linguistic Variation with Twitter Data Analysis'. In: *Computers, Environment and Urban Systems* 59 (2016), pp. 244–255. ISSN: 01989715. DOI: 10.1016/j.compenvurbsys.2015.12.003.

[78] J. D. Hunter. 'Matplotlib: A 2D Graphics Environment'. In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[79] Neus Isern and Joaquim Fort. 'Language Extinction and Linguistic Fronts'. In: *Journal of the Royal Society Interface* 11.94 (6th May 2014), p. 20140028. ISSN: 17425662. DOI: 10.1098/rsif.2014.0028.

[80] Yuqin Jiang, Zhenlong Li and Xinyue Ye. 'Understanding Demographic and Socioeconomic Biases of Geotagged Twitter Users at the County Level'. In: *Cartography and Geographic Information Science* 46.3 (4th May 2019), pp. 228–242. ISSN: 15450465. DOI: 10.1080/15230406.2018.1434834.

[81] Kelsey Jordahl et al. *geopandas/geopandas: GeoPandas*. Version v0.8.1. Zenodo, July 2020. DOI: 10.5281/zenodo.2585848.

[82] Anne Kandler. 'Demography and Language Competition'. In: *Human Biology* 81.2-3 (1st Apr. 2009), pp. 181–210. ISSN: 0018-7143. DOI: 10.3378/027.081.0305.

[83] Anne Kandler and James Steele. 'Ecological Models of Language Competition'. In: *Biological Theory* 3.2 (20th June 2008), pp. 164–173. ISSN: 1555-5542. DOI: 10.1162/biot.2008.3.2.164.

[84] Robert B Kaplan and Richard B Baldauf. *Language Planning: From Practice to Theory*. Bristol, UK ; Tonawanda, NY: Multilingual Matters, 1997. ISBN: 1-85359-372-9.

[85] Judith S Kleinfeld. 'The Small World Problem'. In: *Society* 39.2 (2002), pp. 61–66. DOI: 10.1007/BF02717530.

[86] Caglar Koylu. 'Uncovering Geo-Social Semantics from the Twitter Mention Network: An Integrated Approach Using Spatial Network Smoothing and Topic Modeling'. In: *Human Dynamics Research in Smart and Connected Communities*. Ed. by Shih-Lung Shaw and Daniel Sui. Human Dynamics in Smart Cities. Cham: Springer International Publishing, 2018, pp. 163–179. ISBN: 978-3-319-73247-3. DOI: 10.1007/978-3-319-73247-3_9.

[87] Michael Krauss. 'The World's Languages in Crisis'. In: *Language* 68.1 (1992), pp. 4–10. ISSN: 1535-0665. DOI: 10.1353/lan.1992.0075.

[88] Stefan Kulk and Bastiaan van Loenen. 'Brave New Open Data World?' In: *International Journal of Spatial Data Infrastructures Research* 7.0 (0 4th May 2012), pp. 196–206. ISSN: 1725-0463. DOI: 10.2902/ijsdir.v7i0.285.

[89] William Labov. *The Social Stratification of English in New York City*. Cambridge University Press, 1966.

[90] William Labov. *Sociolinguistic Patterns*. University of Pennsylvania Press, Sept. 1973. 374 pp. ISBN: 978-0-8122-1052-1.

[91] William Labov. *Principles of Linguistic Change, Vol. 2: Social Factors*. Chichester: Blackwell Publishers, 22nd Mar. 2001. 592 pp. ISBN: 978-0-631-17916-0.

[92] Fabio Lamanna et al. 'Immigrant Community Integration in World Cities'. In: *PLoS ONE* 13.3 (2018), pp. 1–19. ISSN: 19326203. DOI: 10.1371/journal.pone.0191612.

[93] Jure Leskovec and Eric Horvitz. 'Planetary-Scale Views on a Large Instant-Messaging Network'. In: *Proceedings of the 17th International Conference on World Wide Web*. WWW '08. New York, NY, USA: Association for Computing Machinery, 21st Apr. 2008, pp. 915–924. ISBN: 978-1-60558-085-2. DOI: 10.1145/1367497.1367620.

[94] Elizaveta Levina and Peter Bickel. 'The Earth Mover's Distance Is the Mallows Distance: Some Insights from Statistics'. In: *Proceedings of the IEEE International Conference on Computer Vision*. Vol. 2. Vancouver, Canada: IEEE, 2001, pp. 251–256. DOI: 10.1109/ICCV.2001.937632.

[95] Jacob Levy Abitbol et al. 'Socioeconomic Dependencies of Linguistic Patterns in Twitter: A Multivariate Analysis'. In: *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*. 2018, pp. 1125–1134. ISBN: 978-1-4503-5639-8. DOI: 10.1145/3178876.3186011.

[96] Lizi Liao et al. 'Lifetime Lexical Variation in Social Media'. In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence: 27-31 July 2014, Québec*. 1st July 2014, pp. 1643–1649.

[97] Thomas Louf, David Sánchez and José J. Ramasco. 'Capturing the Diversity of Multilingual Societies'. In: *Physical Review Research* 3.4 (30th Nov. 2021), p. 043146. ISSN: 2643-1564. DOI: 10.1103/PhysRevResearch.3.043146.

[98] Christopher McCarty et al. 'Comparing Two Methods for Estimating Network Size'. In: *Human Organization* 60.1 (8th Nov. 2005), pp. 28–39. ISSN: 0018-7259. DOI: 10.17730/humo.60.1.efx5t9gjtgmga73y.

[99] Anthony McEnery and Zhonghua Xiao. 'Swearing in Modern British English: The Case of Fuck in the BNC'. In: *Language and Literature: International Journal of Stylistics* 13.3 (Aug. 2004), pp. 235–268. ISSN: 0963-9470, 1461-7293. DOI: 10.1177/0963947004044873.

[100] Marshall McLuhan. *The Gutenberg Galaxy: The Making of Typographic Man*. Repr. Toronto: University of Toronto Press, 2008. 293 pp. ISBN: 978-0-8020-6041-9.

[101] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space*. 6th Sept. 2013. DOI: 10.48550/arXiv.1301.3781.

[102] Stanley Milgram. 'The Small World Problem'. In: *Psychology today* 2.1 (1967), pp. 60–67.

[103] James W. Minett and William S.Y. Wang. 'Modelling Endangered Languages: The Effects of Bilingualism and Social Structure'. In: *Lingua* 118.1 (1st Jan. 2008), pp. 19–45. ISSN: 00243841. DOI: 10.1016/j.lingua.2007.04.001.

[104] J. Mira and Á. Paredes. 'Interlinguistic Similarity and Language Death Dynamics'. In: *EPL (Europhysics Letters)* 69.6 (2nd Feb. 2005), p. 1031. ISSN: 0295-5075. DOI: 10.1209/EPL/I2004-10438-4.

[105] Alan Mislove et al. 'Understanding the Demographics of Twitter Users'. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 5. 1. Barcelona: AAAI Press, 2011, pp. 554–557.

[106] Delia Mocanu et al. 'The Twitter of Babel: Mapping World Languages through Microblogging Platforms'. In: *PLoS ONE* 8.4 (18th Apr. 2013). Ed. by Yamir Moreno, e61981. ISSN: 19326203. DOI: 10.1371/journal.pone.0061981.

[107] Ines Montani et al. *explosion/spaCy: Industrial-strength Natural Language Processing*. Version v3.5.0. explosion, 20th Jan. 2023. DOI: 10.5281/ZENODO.1212303.

[108] Fred Morstatter et al. 'Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose'. In: *Proceedings of the International AAAI Conference on Web and Social Media* 7.1 (3rd Aug. 2021), pp. 400–408. ISSN: 2334-0770, 2162-3449. DOI: 10.1609/icwsm.v7i1.14401.

[109] Max Müller. 'Lecture IX. The Theoretical Stage, and the Origin of Language'. In: *Lectures on the Science of Language: Delivered at the Royal Institution of Great Britain in April, May, and June 1861*. London, England: Longman, Green, Longman, and Roberts, 1861, pp. 329–378. DOI: 10.1037/14263-009.

[110] Arvind Narayanan and Vitaly Shmatikov. 'Robust De-anonymization of Large Sparse Datasets'. In: *2008 IEEE Symposium on Security and Privacy (Sp 2008)*. 2008 IEEE Symposium on Security and Privacy (Sp 2008). May 2008, pp. 111–125. DOI: 10.1109/SP.2008.33.

[111] Dong Nguyen, Noah A. Smith and Carolyn P. Rosé. 'Author Age Prediction from Text Using Linear Regression'. In: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. LaTeCH-HLT 2011. Portland, OR, USA: Association for Computational Linguistics, June 2011, pp. 115–123.

[112] Dong Nguyen, Dolf Trieschnigg and Leonie Cornips. 'Audience and the Use of Minority Languages on Twitter'. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 9. 1. Oxford: AAAI Press, 2015, pp. 666–669. ISBN: 978-1-57735-733-9.

[113] Dong Nguyen et al. '"How Old Do You Think I Am?": A Study of Language and Age in Twitter'. In: *Proceedings of the International Conference on Weblogs and Social Media*. Vol. 7. 2013, pp. 439–448.

[114] Dong Nguyen et al. 'Computational Sociolinguistics: A Survey'. In: *Computational Linguistics* 42.3 (21st Sept. 2016), pp. 537–593. ISSN: 15309312. DOI: 10.1162/COLI_a_00258.

[115] Johanna Nichols. 'The Origin and Dispersal of Languages: Linguistic Evidence'. In: *The origin and diversification of language* 24 (1998), pp. 127–170.

[116] Pippa Norris and Ronald Inglehart. *Cosmopolitan Communications: Cultural Diversity in a Globalized World*. Cambridge University Press, 31st Aug. 2009. 447 pp. ISBN: 978-1-139-47961-5.

[117] OECD. *Where All Students Can Succeed*. Vol. II. PISA 2018 Results. Paris: OECD Publishing, 2019. DOI: 10.1787/b5fd1b8f-en.

[118] Paul Ohm. 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization'. In: *UCLA Law Review* 57 (2009–2010), p. 1701.

[119] George Orwell. *1984*. New American Library, 1950.

[120] Jahna Otterbacher. 'Inferring Gender of Movie Reviewers: Exploiting Writing Style, Content and Metadata'. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. CIKM '10. New York, NY, USA: Association for Computing Machinery, 26th Oct. 2010, pp. 369–378. ISBN: 978-1-4503-0099-5. DOI: 10.1145/1871437.1871487.

[121] Ruth Page. 'The Linguistics of Self-Branding and Micro-Celebrity in Twitter: The Role of Hashtags'. In: *Discourse & Communication* 6.2 (1st May 2012), pp. 181–201. ISSN: 1750-4813. DOI: 10.1177/1750481312437441.

[122] M. Patriarca et al. 'Modeling Two-Language Competition Dynamics'. In: *Advances in Complex Systems* 15.3-4 (13th June 2012). ISSN: 02195259. DOI: 10.1142/S0219525912500488.

[123] Marco Patriarca and Els Heinsalu. 'Influence of Geography on Language Competition'. In: *Physica A: Statistical Mechanics and its Applications* 388.2-3 (15th Jan. 2009), pp. 174–186. ISSN: 03784371. DOI: 10.1016/j.physa.2008.09.034.

[124] Umashanthi Pavalanathan and Jacob Eisenstein. 'Confounds and Consequences in Geotagged Twitter Data'. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon: Association for Computational Linguistics (ACL), 2015, pp. 2138–2148. ISBN: 978-1-941643-32-7. DOI: 10.18653/v1/d15-1256.

[125] J. P. Pinasco and L. Romanelli. 'Coexistence of Languages Is Possible'. In: *Physica A: Statistical Mechanics and its Applications* 361.1 (15th Feb. 2006), pp. 355–360. ISSN: 03784371. DOI: 10.1016/j.physa.2005.06.068.

[126] Alejandro Portes and Lingxin Hao. 'E Pluribus Unum: Bilingualism and Loss of Language in the Second Generation'. In: *Sociology of Education* 71.4 (1998), pp. 269–294. ISSN: 00380407. DOI: 10.2307/2673171.

[127] Katharina Prochazka and Gero Vogl. 'Quantifying the Driving Factors for Language Shift in a Bilingual Region'. In: *Proceedings of the National Academy of Sciences of the United States of America* 114.17 (25th Apr. 2017), pp. 4365–4369. ISSN: 10916490. DOI: 10.1073/pnas.1617252114.

[128] Marcel Proust. *The Guermantes Way*. In collab. with Internet Archive. Trans. by C. K. (Charles Kenneth) Scott-Moncrieff. In Search of Lost Time. New York, Random House, 1927.

[129] Suzanne Romaine. 'The Bilingual and Multilingual Community'. In: *The Handbook of Bilingualism and Multilingualism: Second Edition*. Chichester, UK: John Wiley & Sons, Ltd, 3rd Oct. 2012, pp. 443–465. ISBN: 978-1-4443-3490-6. DOI: 10.1002/9781118332382.ch18.

[130] Yossi Rubner, Carlo Tomasi and Leonidas 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. DOI: 10.1109/ICCV.1998.710701.

[131] Alex Salcianu et al. *Compact Language Detector v3 (CLD3)*. 2023.

[132] Victor Sanh et al. *DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter*. 29th Feb. 2020. DOI: 10.48550/arXiv.1910.01108.

[133] Lisa Sattenspiel and Klaus Dietz. 'A Structured Epidemic Model Incorporating Geographic Mobility among Regions'. In: *Mathematical Biosciences* 128.1-2 (1st July 1995), pp. 71–91. ISSN: 00255564. DOI: 10.1016/0025-5564(94)00068-B.

[134] Jonathan Schler et al. 'Effects of Age and Gender on Blogging'. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. 2006.

[135] Martin Schweinberger. 'Swearing in Irish English – A Corpus-Based Quantitative Analysis of the Sociolinguistics of Swearing'. In: *Lingua* 209 (1st July 2018), pp. 1–20. ISSN: 0024-3841. DOI: 10.1016/j.lingua.2018.03.008.

[136] Lloyd S. Shapley. *Notes on the N-Person Game — II: The Value of an N-Person Game*. RAND Corporation, 21st Aug. 1951.

[137] Sherry Simon. *Cities in Translation: Intersections of Language and Memory*. London: Routledge, 1st Jan. 2011. 1-204. ISBN: 978-0-203-80288-5. DOI: 10.4324/9780203802885.

[138] Luke Sloan. 'Who Tweets in the United Kingdom? Profiling the Twitter Population Using the British Social Attitudes Survey 2015'. In: *Social Media + Society* 3.1 (1st Jan. 2017). ISSN: 2056-3051. DOI: 10.1177/2056305117698981.

[139] Ricard V. Solé, Bernat Corominas-Murtra and Jordi Fortuny. 'Diversity, Competition, Extinction: The Ecophysics of Language Change'. In: *Journal of the Royal Society Interface* 7.53 (6th Dec. 2010), pp. 1647–1664. ISSN: 17425662. DOI: 10.1098/rsif.2010.0110.

[140] Selma K. Sonntag. *The Local Politics of Global English: Case Studies in Linguistic Globalization*. Lexington Books, 28th Oct. 2003. 167 pp. ISBN: 978-0-7391-5728-2.

[141] James H. Stam. *Inquiries into the Origin of Language: The Fate of a Question*. New York: Harper & Row, 1976. xii, 307. ISBN: 978-0-06-046403-5.

[142] Anna-Brita Stenström, Gisle Andersen and Ingrid Kristine Hasund. *Trends in Teenage Talk: Corpus Compilation, Analysis and Findings*. John Benjamins Publishing, 27th Sept. 2002. 243 pp. ISBN: 978-90-272-9733-4.

[143] Erik Štrumbelj and Igor Kononenko. 'Explaining Prediction Models and Individual Predictions with Feature Contributions'. In: *Knowledge and Information Systems* 41.3 (1st Dec. 2014), pp. 647–665. ISSN: 0219-3116. DOI: 10.1007/s10115-013-0679-x.

[144] Yuri Takhteyev, Anatoliy Gruzd and Barry Wellman. 'Geography of Twitter Networks'. In: *Social Networks*. Capturing Context: Integrating Spatial and Social Network Analyses 34.1 (1st Jan. 2012), pp. 73–81. ISSN: 0378-8733. DOI: 10.1016/j.socnet.2011.05.006.

[145] The pandas development team. *pandas-dev/pandas: Pandas*. Version latest. Zenodo, Feb. 2020. DOI: 10.5281/zenodo.3509134.

[146] Jeffrey Travers and Stanley Milgram. 'An Experimental Study of the Small World Problem'. In: *Social Networks*. Elsevier, 1977, pp. 179–197.

[147] Peter Trudgill. *Sociolinguistics: An Introduction to Language and Society*. Penguin UK, 2000.

[148] *Twitter API Documentation*. URL: https://developer.twitter.com/en/docs/twitter-api (visited on 24/01/2023).

[149] *Twitter API: Filtered Stream Endpoint*. URL: https://developer.twitter.com/en/docs/twitter-api/tweets/filtered-stream/api-reference/get-tweets-search-stream (visited on 24/01/2023).

[150] UNESCO. *Convention for the Safeguarding of the Intangible Cultural Heritage*. 17th Oct. 2003.

[151] F. Vazquez, X. Castelló and M. San Miguel. 'Agent Based Models of Language Competition: Macroscopic Descriptions and Order-Disorder Transitions'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2010.4 (8th Apr. 2010), P04007. ISSN: 17425468. DOI: 2010040903131700.

[152] Ronald Wardhaugh. *Languages in Competition: Dominance, Diversity, and Decline*. Wiley-Blackwell, 1987.

[153] Ronald Wardhaugh. *An Introduction to Sociolinguistics*. 5. ed., repr. Blackwell Textbooks in Linguistics 4. Malden, Mass.: Blackwell, 2008. 418 pp. ISBN: 978-1-4051-3559-7.

[154] Duncan J. Watts and Steven H. Strogatz. 'Collective Dynamics of 'Small-World' Networks'. In: *Nature* 393.6684 (6684 June 1998), pp. 440–442. ISSN: 1476-4687. DOI: 10.1038/30918.

[155] Simone Weil. *The Need for Roots: Prelude to a Declaration of Duties towards Mankind*. Trans. by Arthur Wills. Routledge Classics. London ; New York: Routledge, 2002. 298 pp. ISBN: 978-0-415-27101-1 978-0-415-27102-8.

[156] Thomas Wolf et al. 'Transformers: State-of-the-Art Natural Language Processing'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6.

[157] Sue Wright. *Community and Communication: The Role of Language in Nation State Building and European Integration*. Multilingual Matters, 1st Jan. 2000. 292 pp. ISBN: 978-1-85359-484-7.

[158] Faiyaz Al Zamal, Wendy Liu and Derek Ruths. 'Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors'. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 6. 1. 2012, pp. 387–390. DOI: 10.1609/icwsm.v6i1.14340.