Universitat
de les Illes Balears

DOCTORAL THESIS
2023

COMPLEXITY IN COMPUTATIONAL SOCIOLINGUISTICS:
EXPLORING THE INTERPLAY BETWEEN GEOGRAPHY, CULTURE AND THE SOCIAL FABRIC

THOMAS LOUF

Universitat
de les Illes Balears

DOCTORAL THESIS
2023

Doctoral programme in Physics

COMPLEXITY IN COMPUTATIONAL SOCIOLINGUISTICS:
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

SUPERVISORS:
José Javier Ramasco
David Sánchez

LOCATION:
Palma, Spain

Ce qui embellit le désert, dit le petit prince, c'est qu'il cache un puits quelque part...

— *Antoine de Saint-Exupéry,* Le petit prince *[189]*

ABSTRACT

Language has a crucial role in and is greatly influenced by widely different spheres of society, from simple interpersonal communication to the economy and culture. This is what makes sociolinguistics, the study of the interactions of language and society, a complex but decidedly worthwhile endeavour. As a wealth of linguistic data can be retrieved from online social media, the development of new theoretical models aimed at uncovering mechanisms underlying sociolinguistic phenomena can be better guided and tested than ever before. In this thesis, we harness this great potential, and take an interdisciplinary approach to sociolinguistics that is inspired by methods of complexity and data science.

First, we study languages as coherent units that compete with others for speakers, in order to try to identify the drivers of language death and how coexistence of multiple languages in an interconnected society might come to be. Crucially, we take into account the spatial embedding of languages, and first observe it using Twitter data. We find that two languages can coexist with completely separated communities but also with communities mixed in space, featuring a large population of bilinguals. We capture this diversity of coexistence states by introducing a model that considers a potential cultural attachment for one language that may counteract a globally lower prestige, and the relative ease to learn a language knowing the other. Both simulations' and analytical results are used to support our claims.

We then focus on variation within a language to point out the dependence of standard language use with socio-economic status. Focusing on England, we find that there is a slight tendency for English Twitter users to make more grammatical mistakes the lower their income is. This tendency is however very different from one metropolitan area to another, and actually, it seems to be weaker the more socio-economic classes mix together. We propose a model that accounts for potentially different mixing patterns and preferences for a language variety. It reproduces this effect we observed in toy simulations and more realistic ones. We thus find that increased social mixing is crucial to tackle potential social and economic segregation resulting from this language variation.

Lastly, we leverage the interrelationship between language and culture in a case study of the United States to define its major cultural regions. From geotagged tweets written in English, we find the usage hotspots of words found in them to then compute the principal dimensions of lexical variation. With these, we are able to infer coherent cultural regions and the topics that define them.

The strength of the results we obtained across quite diverse areas of sociolinguistics is a mirror of the strength of the approach we took throughout our work. It calls for further developments of this kind, which are most probably only in their infancy.

RÉSUMÉ

Mon résumé

PUBLICATIONS

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

[1] 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.

[2] Thomas Louf, Bruno Gonçalves, José J. Ramasco, David Sánchez and Jack Grieve. 'American Cultural Regions Mapped through the Lexical Analysis of Social Media'. In: *Humanities and Social Sciences Communications* 10.1 (1 30th Mar. 2023), pp. 1–11. ISSN: 2662-9992. DOI: 10.1057/s41599-023-01611-3.

We also had the occasion to explore different topics over the course of this thesis, some leading to other publications not discussed in this manuscript:

[3] 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*. 10th Dec. 2020.

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 23–25, 29, 30, 43, 60

API application programming interface 12

CLD Compact Language Detector 70

EMD earth mover's distance 33, 35, 85, 86

EMR earth mover's ratio 35–37, 47, 50, 51, 85

IPF iterative proportional fitting 19, 46

MSOA middle layer super output area 54, 55

NLP natural language processing 20

PC principal component 73

PCA principal component analysis 20, 73, 74

POI point of interest 15

QRT quote retweet 12

RT retweet 12

SES socio-economic status 7, 53–55, 57, 60, 62, 63

Part I

INTRODUCTION

1 WHAT SHAPES DIVERSITY IN LANGUAGE?

> *[L]anguage [is] partly something originally given, partly that which develops freely. And just as the individual, however freely he may develop, can never reach the point at which he becomes absolutely independent, [...] so too with language [...].*
>
> — *Søren Kierkegaard, 'Journals' [126]*

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 [34, 96]. And it is so much so that researchers are unable to trace back to the origin of such a structured system of communication [91, 106, 163, 204]. Those who have ventured into this kind of inquiry have estimated that language dates back tens or even hundreds of thousands of years [31, 45, 52, 169]. 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. And this, all over the globe. Here, in particular, we will focus on languages as understood in the common sense of the word, that is, coherent systems that define linguistic signs and how their combinations convey meaning — like English, Mandarin Chinese or Hindi, to cite the three most spoken nowadays. Given how long language has existed, and how many human beings have interacted with it, it is then safe to say that human history must have seen a huge diversity in and of language 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 [44, 134, 135, 211, 219], 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 heterogeneity, or, on the contrary, push it towards homogeneity. When not explicitly specified, these effects of heterogenisation or homogenisation will concern both language diversity — that is, differences between languages, when understood as coherent, clearly separated units — and language variation — that is, differences of usage within a language.

1.1 LANGUAGE AS A VECTOR FOR COMMUNICATION

The first obvious function that language serves is to facilitate communication between individuals, and more specifically the kind of

communication called *verbal communication*. To optimise language with regard to that function, there should 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 [21]. 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 democratised 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 reorganisation of the world in what the philosopher Marshall McLuhan called a *global village* [151]? Does this imply a more interconnected world, and in turn, that we will naturally tend towards the communication-optimal state featuring a unique, homogenous language?

A physicist's intuition would say that the more individuals interact with each other, the more language should *thermalise*, 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, homogenisation of language would be both desirable and inevitable. Ferdinand de Saussure, a prominent linguist of the late XIXth - early XXth 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. [54]

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 hypothesised [59, 60], and later demonstrated, not only on real-world social networks [111, 149], but also for a massive, online one [92]. 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 [55, 153, 209, 220]. This idea and the experiments behind it have received some criticism though [127]. 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 [81, 139, 207].

In any case, the existence of a process of globalisation is undisputed, but the idea that more interconnectedness would eventually lead to a global village has been challenged [25, 170]. 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. [25]

Networks of interpersonal communication are therefore complex systems that are hard to summarise with a few simple properties. One can also start to see the limitations of considering language as a neutral means of communication. In fact, as important as the act of communicating itself are the reasons and contexts of this communication. It thus seems that studying language through the lens of communication alone is too limiting to gain a proper understanding of sociolinguistic phenomena.

1.2 LANGUAGE IN THE REALM OF POLITICS

As Pierre Bourdieu argued, language is not only a means of communication, but also a medium of power [32]. Some aspect of this idea has been popularised in the world of fiction with the concept of the *Newspeak* language in George Orwell's *1984* [175]. 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 [75], and not necessarily with dystopian, *Big Brother*-like intentions of total control over individuals. When nation-states were built, pushing forward a common language was seen as a means to unite a nation [226]. This was particularly the case in post-revolution France and imperial Great Britain [102, 110], where respectively French and English were heavily pushed as the languages of higher status. Still today, around the globe most countries' constitutions specify one or a maybe a few languages as the languages of the state. In theory, this would be beneficial as

it allows the state to build a common ground to guarantee equal opportunity, for instance with public education and the rule of law [68]. 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: 90 % of the languages that exist today may be replaced by a handful of dominant languages over the course of this century, according to estimates of the UNESCO [215]. Still, it is not uncommon that a later reaction tried to reverse this process, with policies switching roles completely. Politics can then oppose homogenisation and protect language diversity. Current examples include the policies introduced to protect national languages against global English [203], and the ones to preserve regional languages and dialects within nations [124]. Politics thus very often push for a given language ideology, which impacts whole languages [66, 78, 186], but also varieties within a language. Indeed, numerous languages have a rigorously defined standard variety, sometimes defined by a state institution, like the *Académie Française* for French in France, or the *Real Academia Española* for Spanish in Spain, and often taught in all schools of a country. As such, it is considered as the only "correct" way to speak and write the language. This tends to uniformise language, as it favours the standard variety to the expense of all the others [51, 154], usually called non-standard or vernacular varieties, and which arise naturally within social groups.

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 therefore has an economic value, and this is particularly true in a globalised economy [109]. Indeed, the world has not only become more interconnected in terms of communication, but also in terms of trade. It is the aspect of globalisation that has had the most impact on our contemporary societies: in fact, when someone talks about globalisation, most of the time what they are referring to is economic globalisation. 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* [32].

Further, the manner with which one speaks a language can identify them as member of a certain socio-economic group Indeed, as we mentioned in the previous section, all languages have a number of varieties, some with superior status, such as the standard form. As proven by

the PISA reports of the OECD, its latest included [172], in many countries, linguistic proficiency — in the sense of the standard language — 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, a certain jargon within a language, like the (mostly English) vocabulary associated to the Internet culture, or any language variety. 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. Indeed, different social groups may have different language attitudes [84], meaning they may not value a language or a language variety the same way as other groups. To go back to the opposition between standard and non-standard varieties, different social groups may have different perceptions of what is the normal way to speak [131]. Some may thus oppose the language ideology pushed by society as a whole or the state, which favours the standard form. This mechanism is not the only one at work that pushes against homogenisation. 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 globalisation [214]. Also, very often, language preservation policies are implemented based on the argument that cultural diversity embedded in languages needs to be preserved [47, 97, 130]. Indeed, there is some evidence of a homogenising trend. English is on a steady path to become a global language [48]. Most of the estimated 6000 languages that exist in the world today are endangered [47, 97, 130] and getting replaced by a few dominant languages [102, 218].

1.5 SCOPE AND OUTLINE OF THE THESIS

If there was 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 varieties, as there are as many varieties as speakers. There is variation in space and time. And there is variation for many, often entangled, reasons. Linguistics, and especially its social branch, is 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. This will be used to introduce concepts that permeate this whole thesis along its two complementary streams of data analysis and theoretical modelling.

After this thorough methodological review, in Chapter 3 we will investigate interlanguage 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. Cultural regions of the United States are thus inferred through the analysis of social media posts.

Finally, Chapter 6 will be the occasion for 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

> *New theories have to compete with existing ones, partly on the basis of coherence and generality, but ultimately according to whether they explain existing observations and correctly predict new ones.*
>
> — *Murray Gell-Mann,* The Quark and the Jaguar *[88]*

As shown in the previous chapter, language and society are interwoven in so many ways that sociolinguistic phenomena should be studied most carefully. As scientists, we would like to establish general principles 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 quite 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 pro-

portion of texts survived through centuries. The diachronic study of how language evolves with time has thus 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. The texts available to us from the distant past are not representative of the societies of the time, since only the elites were able to write, and women were essentially barred from publishing until a couple of centuries ago [201]. Also, for the very nature of the texts that were conserved — books, legal texts mostly [50, 114, 185] — colloquial language is almost completely absent from them.

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 [8, 180].

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 digitised and then analysed and shared on large scales. There are now very large corpora of modern languages, produced in both written [23, 98, 99, 150] and spoken [133, 150, 195, 205] forms, that have been shared among researchers for a vast array of analyses. But the written texts still suffer from a lack of representativity, since, 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 also 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, and 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 [86, 87], 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 anonymise 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 repeatedly that anonymisation is tricky, as researchers managed to de-anonymise some public datasets [79, 164]. In order to conduct research that is both ethical and legal, researchers dealing with such data thus have to take very particular care [132, 173].

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 [168]. These have drawn form a variety of sources, like blogs [165, 194], online forums [19, 82, 165], online reviews [49, 113, 176], or a certain microblogging website called Twitter [6, 7, 46, 141, 159].

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 uses the platform [14]. Other than in the US, it is also popular in many countries globally, although with a slight bias towards developed, western countries [107]. Even though only around 1 % of tweets are geotagged [162], when only counting users who tweet with a geotag, in around 80 countries there is more than one Twitter user for ten thousand inhabitants [159]. Hence why Twitter has been extensively used for the analysis of geographically-embedded text [12, 28, 29, 93–95, 101, 115, 128, 136, 159, 166], and why we 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) [212]. 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 [213]. They are all geotagged tweets posted between 2015 and 2021, both years included, without any geographic restriction, which implies that, a priori, they could have originated from anywhere in the world.

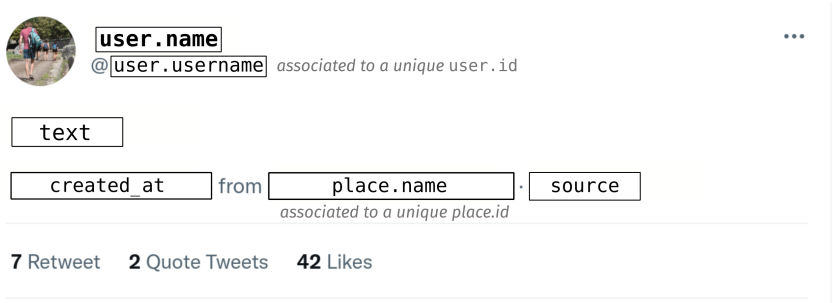
We show in Figure 2.1 an example of the data we can have for each tweet. There are two fields in these data that particularly interest us for the works we will present in the next part and that need careful processing: the textual content of the tweet (in `"text"`) and its geotag (in `"geo"`). Next, 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

(a)

(b)



(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).

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 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 [177]. 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 [147] — 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 [190]. 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 now 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, for which we kept only the text that can be considered the author's natural language production.

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 has enabled their device's GPS but has disabled precise geotagging, this `"coordinates"` field is absent, how do we then 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 [74] (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 sizes 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, as shown in Figure 2.2, 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.

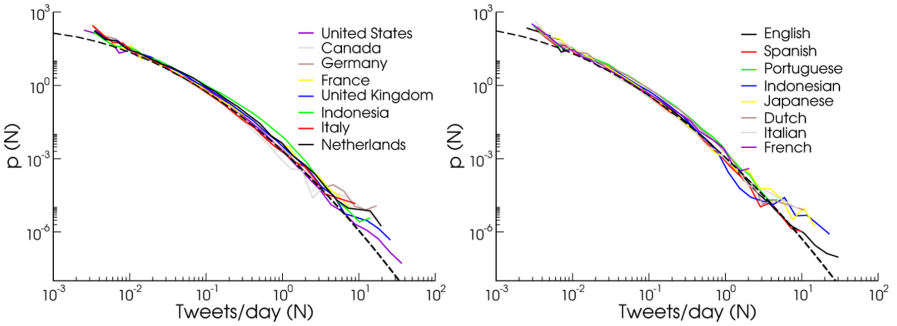


Figure 2.2: Distribution of the activity of Twitter users. The probability density $p(N)$ of user activity (number of daily tweets N) grouped by country (left) and language (right) is plotted. Different curves collapse naturally, which indicates the presence of a seemingly universal distribution of user activity, independent of cultural backgrounds. Dashed lines represent log-normal distributions $p(N) \sim 1/(N\sigma\sqrt{2\pi}) \cdot \exp\left[-(\ln N - \mu)^2/2\sigma^2\right]$ with $\mu = -5.16$ and $\sigma = 1.67$ on the left and $\mu = -5.55$ and $\sigma = 1.70$ on the right. Adapted from [159].

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 formalise 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: $n_{u,c}^{\text{GPS}}$, 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 weighted count of tweets of user u in cell c. It can be decomposed into the contributions of those with GPS coordinates, $n^{\text{GPS}}_{u,c}$, and of those tagged with a place, $r^{\text{P}}_{u,c}$:

$$r_{u,c} = n^{\text{GPS}}_{u,c} + r^{\text{P}}_{u,c}. \tag{2.1}$$

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:

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

We provide an illustration for this cell attribution in Figure 2.3.

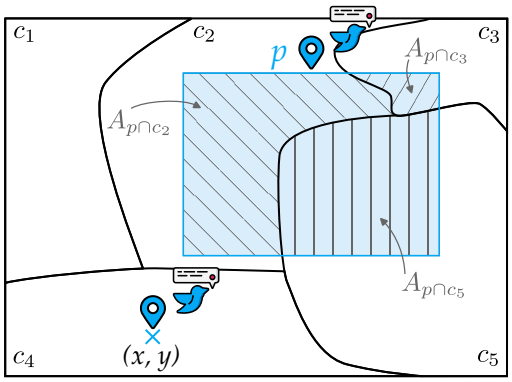


Figure 2.3: Diagram illustrating how geotagged tweets are attributed to cells. A first tweet has GPS coordinates (x, y) attached to it, and can thus be directly attributed to c_4. It thus increments n^{GPS}_{u,c_4} by 1. Another is tagged in place p, defined by a bounding box, shown in blue. It will be attributed partially to c_2, c_3 and c_5, with weights equal to the ratio of overlapping area $A_{p\cap c_k}/A_p$, thus incrementing r^{P}_{u,c_k} by these values for the three cells.

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 $r^{\text{NT}}_{u,c}$. Among those left, the cell of residence c^* is then the one such that $r^{\text{NT}}_{u,c^*}/\sum_{c'} r^{\text{NT}}_{u,c'} \geq 0.5$, 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 were chosen because we believe them to be reasonable, but they may be adjusted to each analysis, and also tweaked for sensitivity analyses. They are summarised along with other criteria in Table 2.1.

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. Again, the values for the criteria given above were set as they were deemed reasonable and allowed us to safely discard problematic users. A summary of these values is given in Table 2.1, while Table 2.2 gives the counts of Twitter users in our dataset that we found to be locals of one of several multilingual regions, using these criteria.

Language detector	Minimum length of tweets after cleaning	4 words
	Minimum confidence of the model	90 %
Relevant places	Maximum number of overlapped cells	$\leq$ 4 cells contain $\geq$ 90 % of its area
Residence cell	Minimum activity in cell	3 tweets
	Minimum proportion of activity in cell	10 %
	Minimum proportion of nighttime activity in cell	50 %
Real users	Maximum tweeting rate	10 tweets per hour
	Maximum speed	$1000\,\mathrm{km\,h^{-1}}$
	Minimum period of activity	Once a month for 3 consecutive months

Table 2.1: Criteria used for Twitter data preprocessing. They are applied over the whole tweeting span of a user in our dataset, which, in this work, is at most from the years 2015 to 2021, both included. This means for example that the tweeting rate is the total number of tweets posted by a user divided by the amount of time that passed between their first and last tweet in the dataset.

Region	Number of local Twitter users
Balearic islands	13 731
Basque country (ES)	22 120
Belgium	41 214
Catalonia	101 688
Cyprus	4227
Estonia	2667
Finland	15 789
Galicia	30 850
Java	840 223
Latvia	15 502
Luxembourg	1656
Malaysia	347 328
Paraguay	45 745
Quebec	16 848
Switzerland	18 552
Valencian Community	59 475

Table 2.2: Number of Twitter users found to be residents and speaking a local language between 2015 and 2019 in several multilingual regions.

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 [107, 159]. In terms of more local geographical biases, densely populated urban areas are usually overrepresented [14, 120, 158]. As for demographics, Twitter users are on average younger [14, 167, 200], with more degrees and income, and more likely to be male [14, 158, 200] than the general population. These biases have to be taken into consideration, and alleviated whenever possible. For instance, some metrics can be rescaled non-uniformly across space, matching marginals obtained from the census using *iterative proportional fitting* (IPF) [53, 71]. For instance, as we will see in Chapter 3, we used it when rescaling cell counts of speakers of different languages in a given region found on Twitter. IPF allowed us to make their sum over cells by language match a given country's official statistics on overall number of speakers by language, and also to match the real number of residents of every cell.

2.2 COMPUTATIONAL MODELS OF NATURAL LANGUAGE

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 give a very brief overview of the field and 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 [36], Google's BERT [56] and its derivatives [191], or the BigScience workshop's BLOOM [225], with an increasing number of pre-trained models made freely accessible [161, 223]. These kinds of models are best known for their derivatives optimised for usage as chatbots, but they have many variants and parts that can serve different functions: lemmatising, that is finding the base forms of words (like removing plurals or conjugation), classifying documents, or finding similarity between words and documents by embedding them in a vector space (with for instance `word2vec` [152] and `tok2vec` [10], respectively). As said above in Section 2.1.4.2, and as we will see in the next chapters, these kinds of models can also be used to detect the language a text is written in. Although they are very powerful tools to carry out some tasks, their training is very costly [9], 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 the one found on Twitter for instance. Some works have strived to normalise informal texts to be able to use such tools, but doing so throws away valuable information [61]. 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 [58] can be good enough to compute document similarities and then cluster documents together. It simply consists in computing a matrix of word frequencies by document, reducing its dimensionality with a singular value decomposition, thus performing *principal component analysis* (PCA) [222] to then cluster documents together based on their cosine similarity. This method can be adapted, by using a matrix containing a different metric than simple frequencies, or clustering based on a different method to compute similarity, as we will show in Chapter 5. Latent Dirichlet allocation [24] 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 [90]. Our results presented in Chapter 4, based on the detection of the use of standard language, are based on a grammar and spell checker, LanguageTool, which simply performs part-of-speech tagging before searching for matches of patterns indicating a broken

rule of the standard language. The tools are thus many, but the context in which they are used greatly conditions their actual power.

2.3 THEORETICAL MODELS

2.3.1 *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 regularisation 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 [198, 206]. 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 to lay a fertile ground for further development, and may help understand other phenomena than the one they were initially intended to explain [64]. 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? This 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 [64]. Theoretical models thus remain invaluable, and rather than a replacement, machine learning models can serve as useful guides in their development by quickly uncovering patterns found in the data.

2.3.2 *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 microscopic and the macroscopic. To illustrate this distinction, we will take the example of the modelling of language competition, which has been approached by many physicists and applied mathematicians with different frameworks [27, 41].

At the macro level, one considers the system of interest as a whole and tries to come up with equations that describe the dynamics of some global metrics that summarise the state of the system. In language competition, the system can be a society with many more internal than external interactions, like a country, and the global metrics can be the proportions of people speaking each language. One such model that has been very influential in language competition modelling is the Abrams-Strogatz model [5]. It considers a group where individuals speak either language A or B. It captures the fact that, the more people speak A, and the more prestigious A is in society, the more B speakers will want to switch to A, and inversely. Its proposal is considered a seminal work in this field for the simplicity of its formulation, which inspired many further works. The model is defined by a single equation:

$$\frac{\mathrm{d}p_A}{\mathrm{d}t} = (1-p_A)sp_A^a - p_A(1-s)(1-p_A)^a, \tag{2.3}$$

thus describing the evolution with time of the proportions of A speakers p_A, which depends on an exponent $a > 0$ and a factor $s \in [0,1]$ called the relative status of language A. When this status has a value above 0.5, A is relatively more prestigious than B, and the inverse is true when it is below 0.5. The exponent a, later on called (inverse) volatility, controls the effect of social pressure. For instance, in the first term of Equation (2.3), it controls how much influence the proportion of A speakers p_A will have on the B speakers — present in a proportion equal to $(1-p_A)$ — when considering learning A. The lower a, or the higher the volatility, the more importance this term will have, so the easier the switch. A clear advantage of such a model is its high tractability. First, it is straightforward to fit to existing data, which is what Abrams and Strogatz did with historical data on the proportion of speakers of Scottish Gaelic, Welsh and Quechua, thus predicting the death of these languages [5]. Second, it lends itself to mathematical analysis, which in this instance allowed Vazquez, Castelló and San Miguel to determine that the model only allowed stable coexistence of two languages for $a > 1$, as shown in Figure 2.4. Many other models of the same kind have been proposed to understand language shift, extending previous formulations to consider additional aspects of the problem. Notably, most subsequent models have incorporated bilingualism, as they included a third population of bilinguals AB, which may influence the dynamics in different ways. For instance, including inter-linguistic similarity in a model with bilinguals, the latter may keep a minority language alive when this similarity is high enough [157]. Minett and Wang have stressed the importance of considering different maximum rates between the processes of learning and losing a language, as these happen on different timescales, and argued that individuals cannot transition from being monolingual in a language to monolingual in another, but rather that they should necessarily go

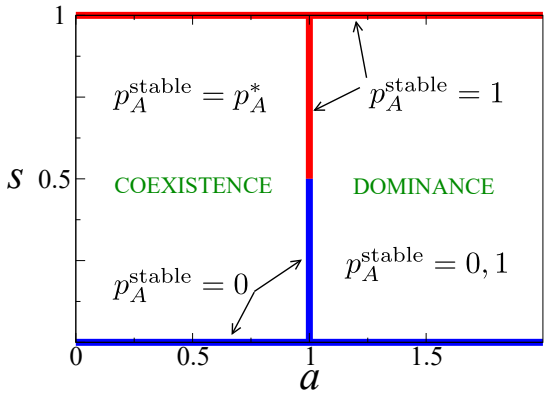


Figure 2.4: Parameter space of the Abrams-Strogatz model. The different stable fixed points are presented, split into two kinds by the $a = 1$ line: coexistence and dominance. In the coexistence region of the parameter space, the stable fixed point is such that $0 < p_A^* < 1$, meaning that the proportions of both A and B speakers are non-null. Adapted from [217].

through bilingualism to make the transition [155]. Heinsalu, Patriarca and Léonard tweaked the model, making the bilingual population also influence the monolinguals to learn the other language [108]. These three models have greatly inspired our own modelling endeavour, that we will present in Chapter 3. Also of note, several have taken inspiration from models of interspecies competition from ecology [123, 181, 202], and others have taken the spatial embedding of languages into account as well using reaction-diffusion-like models [117, 122, 178, 183].

Instead of a single population, one may consider a metapopulation, that is a population of populations that may interact with each other as they move around. This has been used extensively in ecological models to take into account the movements of species for the dynamics of interspecies competition [103]. It was used similarly in epidemiology to try to understand how the propensity of individuals to stay in their home region or to move to others and potentially transmit a disease to another population may contribute to an epidemic [11, 17, 192]. Even closer to our topic, the metapopulation framework has proved useful to study a model of voting behaviours [70]. It can thus surely be adequate to test a sociolinguistic model, and in particular a model of language shift, as we will show in Chapter 3.

While some social aspects can be modelled with some parameters in global equations, or the mobility of the population can be taken into account in a metapopulation framework, the social structure itself can hardly be incorporated into these two kinds of model. That is where the micro approach shines, notably with a class of models called *agent-based model*s (ABMs). In such a model, it is not the evolution of the global population that is modelled, but the evolution of the state of each individual, or agent, in that population. Their state could

be linked to a linguistic variable, or even a set of them. In the case of language competition, it may be the language(s) they speak. The model then defines switching rules between states. For instance, one can rewrite the Abrams-Strogatz model of Equation (2.3) as an ABM, writing the following transition probabilities:

$$\begin{aligned} P(A \to B) &= (1-s)(1-p_A)^a, \\ P(B \to A) &= s p_A^a, \end{aligned} \tag{2.4}$$

from A to B, and B to A, respectively. Each agent, in function of their current state, may switch to another, with potentially their own probability. Indeed, if each agent sees different proportions p_A, because they interact with different people, then the social structure can be reflected in Equation (2.4). And from this transition rule, any kind of social structure can be plugged into the model, from the very basic but mathematically-tractable complete network, to more realistic social networks from synthetic models or even real-world data. ABMs of language competition have thus been tested in 2D lattices, small-world networks [220], or synthetic networks with community structure [39, 42, 43, 155, 217]. They may still allow for analytic study, especially working in the simple case of a complete network of interactions and in mean-field, as these approximations enable to write evolution equations that may provide a good, approximate prediction of the average global trend in the population [217]. We have here cited works focused on language dynamics, from physicists mainly, but there is a long tradition of leveraging ABMs in the social sciences, with the seminal works of Axelrod on the dissemination of culture [15], or Schelling's on the emergence of segregation [193].

Despite their nature, one should not expect these models to offer predictions at the micro, or individual, level. Even though they define rules at this level, they do not — or at least should not — claim that they predict the behaviour of each individual. Rather, what they try to achieve is chiefly to capture the crucial mechanisms that give rise to observed global trends [148]. In that regard they are thus not so different from models defined at the macroscopic level. What they definitely bring to the table is how they enable us to study how some properties of the social structure may matter in sustaining the observed dynamics. We here reference again to the citation of Jan Blommaert we gave in the previous chapter, that the world had become more of "a tremendously complex web of villages towns, neighbourhoods, settlements" than a simple global village, where everyone is perfectly connected. Then to understand the "complexity" that emerges from this, a model that describes dynamics that can happen on any kind of complex network of interaction is invaluable. For example, ABMs can provide insights into questions with a long history in social sciences [138], such as: does this cultural feature persist in part of the population partly because it creates a sense a

group identity that binds people together? Does an increased number of contacts between these populations tend to smooth out differences between them? Naturally, as it is part of the social sciences, these questions are also relevant to sociolinguistics. Language is a pervasive cultural feature, so for instance, the hypothesis that it sustains a sense of group identity deserves very much to be investigated, as we will show in Chapters 1 and 3 within an ABM framework.

2.4 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 [104], pandas [208] or GeoPandas [121]. In their vast majority, figures presented here were prepared with Matplotlib [116], and sometimes edited, or entirely drawn, with Inkscape[4].

This document was prepared using LaTeX with the `classicthesis` style[5], and the LaTeX Workshop extension[6] of Visual Studio Code, and the references were managed using Zotero[7].

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
7 Available at https://www.zotero.org

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

> *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 *[221]*

Much of the work presented in this chapter is included in an article entitled 'Capturing the Diversity of Multilingual Societies', that was previously published by the author of this thesis with José J. Ramasco and David Sánchez [1].

The research presented in this chapter is concerned with the study of languages in contact, that is how different languages interact with one another, which emerged a few decades ago as a hot topic when linguists realised that the world may be facing a mass extinction of languages [47, 97, 130]. As we pointed out in Section 1.3, there is a great cultural wealth embedded in these endangered languages, and its loss would be irreversible. Hence the need to understand what mechanisms drive shifts from one language to another.

Modelling language shift has been the subject of much research in the last decades [27, 41], which employed various approaches such as the formulation of evolution equations based on ecological models [108, 123, 156, 181, 202], of reaction-diffusion equations [117, 122, 178, 183], or approaches within the framework of agent-based modelling [39, 42, 155, 183]. While global evolution equations determine how the proportions of each language group will evolve in a system, *agent-based model*s (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. As explained in Section 2.3.2, ABMs allow to freely define a linguistic environment, and can thus enable us to assess the impact of the social structure on the language dynamics. That is why in this chapter we will formulate existing models and a new proposal of ours in the framework of ABMs, before showing the strength of our proposal in both completely interconnected populations and in a metapopulation framework incorporating the mobility of individuals. We will here turn to slightly more complex models than one as simple as the Abrams-Strogatz model, which simply considers the possibility of two monolingual states.

Indeed, 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 [42, 178, 179, 217]. 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 [183] which relied on census data, none of the aforementioned models have been tested against 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.

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 analyse the spatial distribution of languages in order to evaluate the models. As data on language use with a fine spatial resolution and large sample sizes are hard to come by using the traditional sources mentioned in Section 2.1.2, we rather strive to extract these spatial distributions from Twitter data.

Almost all data analyses, simulations and plots were made using Python code, freely available on GitHub [145].

In this chapter, we therefore present results obtained through a combination of a large-scale empirical study of the spatial distribution of languages with metapopulation modelling. In Section 3.1, we show empirically that multilingual societies are characterised 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 Section 3.2 a model able to capture the diversity seen in the data. The model also shows how the preference of bilinguals and the ease of learning a language have their importance for the coexistence of languages. Finally, we present a brief discussion of all the results presented in this chapter in Section 3.3.

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.

3.1.1 *Twitter data analysis*

As already mentioned in Section 2.1.4, Twitter has good potential as a data source to extract spatial distributions of language use. 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 mentioned in Section 2.1.4.5, 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 Table 2.2), and analysed hundreds of millions of 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 Figure 3.1). The choice of the size of the cells is critical, as it defines the bins of the spatial distributions that we are studying. There are two limits to the cell size. First is an upper one, because if too many users are aggregated, we may smooth out relevant geographical variations in the languages' distributions. The second is a lower one due to the nature of the geolocation data at hand: the Twitter places data described in Section 2.1.4.3. The typical size of places varies between countries, as for instance cities in America are typically more extended than in Europe. Considering these two constraints, we thus have to choose cell sizes carefully for each region considered. Usually, a range of cell sizes is acceptable, and although we choose to show only one in Figure 3.1 for instance, our analysis can also be carried out for other sizes. We will thus show later the robustness of our metrics against cell size changes.

After thoroughly cleaning and analysing the collected tweets, as described in Section 2.1.4, we obtain a sample of local Twitter users to which a cell of residence and their frequency of usage of different languages. As a user may occasionally tweet in a language they do not speak by way of quoting someone else or using a translator, we do not keep all of them. We set that at least 10 % or 5 of the tweets of a user must be in a certain language to consider them a speaker in this language.

Out of hundreds of millions of geotagged tweets posted over a 4-year range, we thus obtain counts of local users by language group

by cell in our 16 regions of interest. These aggregated data have been deposited on figshare [143].

3.1.2 *Defining pertinent 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). An individual can naturally be in a monolingual or in a multilingual group if they fulfil the conditions given above 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 [16], 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:

- concentration in cell i of L-speakers:

$$c_{L,i} = \frac{N_{L,i}}{N_L}, \tag{3.1}$$

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

$$p_{L,i} = \frac{N_{L,i}}{N_i}, \tag{3.2}$$

where $N_L = \sum_i N_{L,i}$ are all the users classified as L-speakers in the country or region considered, and $N_i = \sum_L N_{L,i}$ is the population of Twitter users residing in cell i speaking any of the local languages. As in [159], we can define the polarisation of a language A for every cell i in a bilingual system with languages A and B as

$$\theta_{A,i} = \frac{1}{2}(1 + p_{A,i} - p_{B,i}). \tag{3.3}$$

The polarisation vanishes when there are only B monolinguals and goes to 1 when there are only A monolinguals. It takes the neutral value of 0.5 when there are as many A-speakers as B-speakers. It also takes this value when there are only bilinguals, as the proportion of

bilingual does not impact this metric and we would then have no A-speakers and B-speakers — in the sense given above. 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

$$c_i = N_i / N. \tag{3.4}$$

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 employ 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 [188], and it was shown to be a proper distance (in the metric sense) between probability distributions [140]. 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

$$\mathrm{EMD}_L \equiv \mathrm{EMD}(P, Q_L) = \sum_{i,j} \hat{f}_{ij} d_{ij}, \tag{3.5}$$

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:

$$\begin{cases} f_{ij} \geq 0, \forall\, i, j \\ \sum_j f_{ij} = c_{L,i}, \forall\, i \\ \sum_i f_{ij} = c_j, \forall\, j \\ \sum_i \sum_j f_{ij} = \sum_i c_{L,i} = \sum_j c_j = 1, \end{cases} \tag{3.6}$$

where c_i and $c_{L,i}$ are the concentrations of the population and L-speakers in every cell i, as defined in Equation (3.4) and Equation (3.1). 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 [73], which uses the method of [30]. However, in its raw form, it is dependent on the

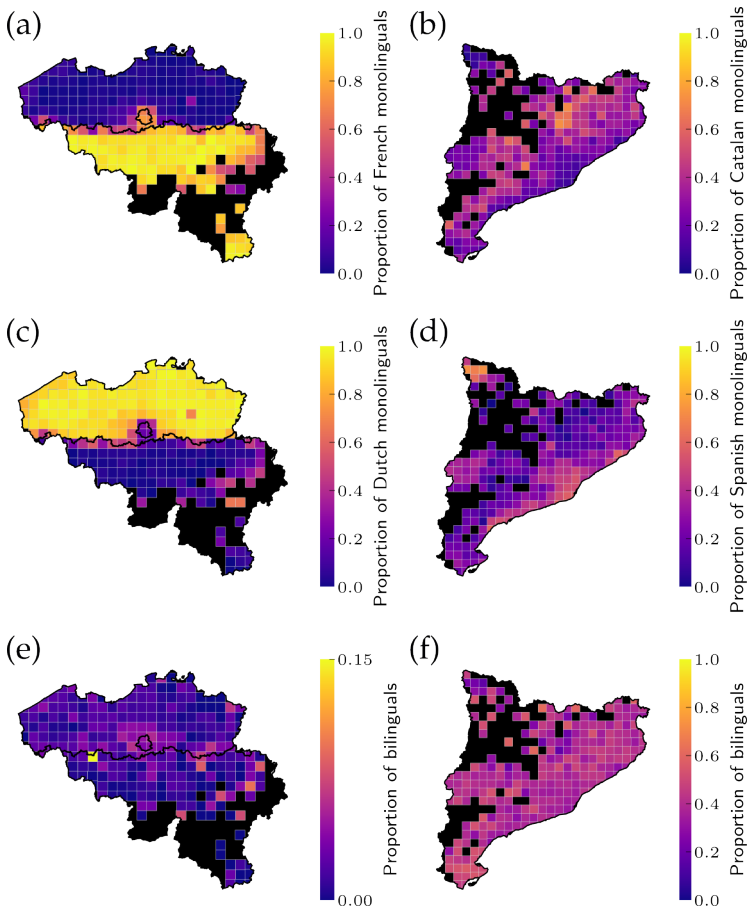


Figure 3.1: Paradigmatic examples illustrating the diversity of multilingual societies. For each cell of $10 \times 10\,\mathrm{km}^2$, the proportions $p_{L,i}$ of monolinguals in (a) French, (b) Catalan, (c) Dutch and (d) Spanish in Belgium (left) and Catalonia (right) are shown. The maps (e) and (f)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

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:

$$k_{\mathrm{EMD}} = \frac{\sum_i \sum_j N_i N_j d_{ij}}{\left(\sum_k N_k\right)^2}. \tag{3.7}$$

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

$$\mathrm{EMR}_L = \frac{\mathrm{EMD}_L}{k_{\mathrm{EMD}}}. \tag{3.8}$$

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. Seeing the relatively low counts of users found in some language groups, one may wonder whether such samples may yield reliable measures of segregation through the EMR. Therefore, to determine whether the measure of the EMR for a group can be deemed reliable, we first set a hard minimum of 50 users detected in that group. Then, we order the cells by descending concentration of the whole population, and take as group count threshold the inverse of the concentration in the cell corresponding to the 90^{th} percentile of the cumulative distribution. This way, the sample we have can be expected to have been sufficient to populate significant cells. However, this threshold may not be passed for a minority group localised in low-density areas, while we may still have a sufficient sample relatively to its actual size. For this reason, we also test whether the EMR calculation is robust to bootstrap resampling, as we generate 50 samples from the concentration distribution of this group, calculate the EMR each time, and if the relative standard deviation of these is below 10 %, we consider the measured EMR of the group to be reliable. We have thus determined that the EMR calculated for the bilinguals in Cyprus, all the multilingual groups in Luxembourg and the trilinguals in Switzerland were not reliable.

3.1.3 *Empirical results*

We propose a first visualisation of the collected data in Figure 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 $10 \times 10\,\mathrm{km}^2$. 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 (Figure 3.1(e)-(f)).

The results are summarised in Figure 3.2(a)-(b), 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. As mentioned previously, we filtered out regions where we deem not sufficient the statistics gathered from Twitter. A wide diversity of situations can be observed. Multilingual societies may have rather balanced monolin-

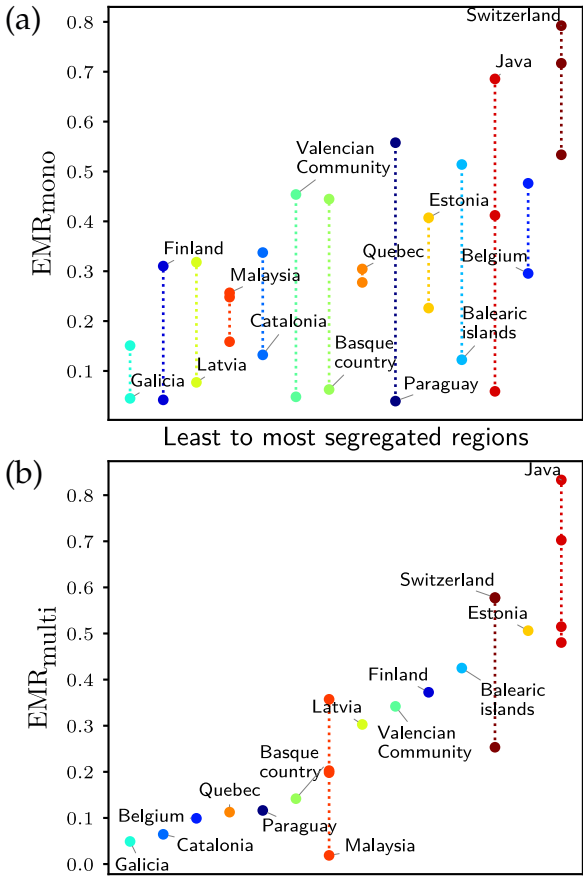


Figure 3.2: EMRs of the (a) monolingual and (b) multilingual groups of multilingual regions of interest, ranked left to right by increasing average of the y-axis values. In (b), the point for trilinguals in Switzerland is not displayed because its value was deemed unreliable. All values of the EMR shown here, as well as those that we discarded are given in Table A.1. A rich diversity of mixing patterns is shown, beyond the two paradigmatic cases of Catalonia and Belgium.

gual 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 spoken by the majority, 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. To demonstrate the robustness of the EMR as a metric on spatial distributions and the soundness of our choices of cell sizes, we show in Figure 3.3 that the EMR is cell-size-invariant for reasonable choices of cell sizes.

All metrics introduced in this section can be computed using similar data taken from other sources to evaluate the spatial mixing of languages.

3.2 MODELS CAPTURING DIVERSITY

As language use in a society only sees significant changes on a timescale of generations [134], the maps obtained from Twitter are only snapshots of the situation around the years 2015 to 2019. In other words, it gives a synchronic view. We do not have access to data providing the longitudinal evolution (diachronic framework), while the models at hand describe the dynamics of the system. Still, 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. Figure 3.4 displays the states: monolingual in A and B, and bilingual AB, with the associated transition probabilities in two previous models and in our proposal. We denote p_A and p_B the proportions of monolinguals in A and B, respect-

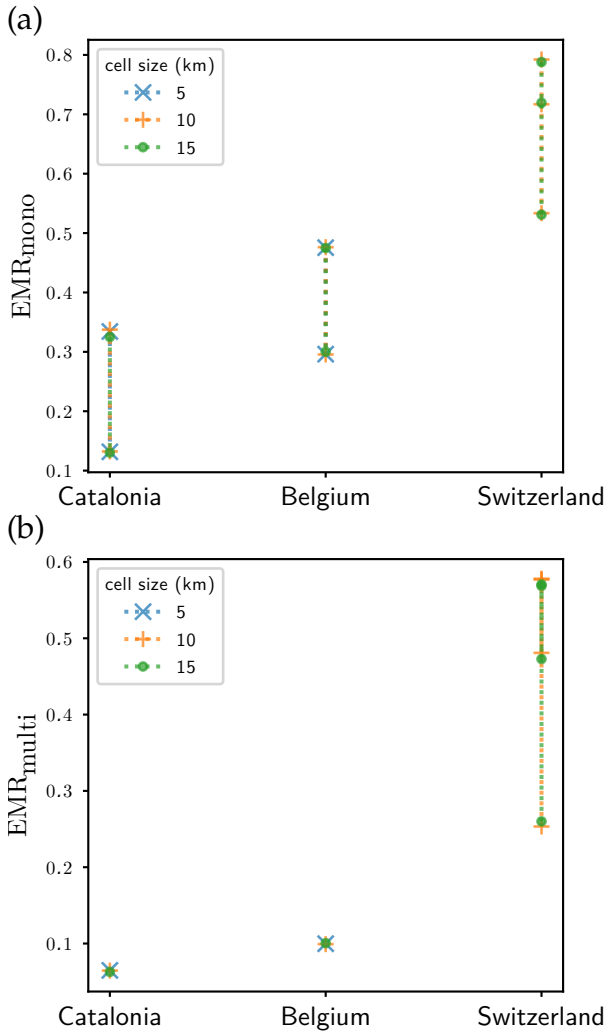


Figure 3.3: Robustness of the EMR with regard to the choice of cell size. For cell sizes ranging from $5 \times 5\,\mathrm{km}^2$ to $15 \times 15\,\mathrm{km}^2$ in Catalonia, Belgium and Switzerland, the values of the EMR (a) between the monolinguals in each language and the whole population, and (b) between the multilinguals and the whole population are shown.

ively, and p_{AB} the proportion of bilinguals. These satisfy the equality $p_A + p_B + p_{AB} = 1$. Within this notation, a state of coexistence is a state in which the two languages remain spoken, which corresponds to either $p_{AB} > 0$, or $p_A > 0$ and $p_B > 0$. Extinction of A (B), for instance, corresponds to $p_A = p_{AB} = 0$ ($p_B = p_{AB} = 0$).

The first model to mention is the one introduced in [5] by Abrams and Strogatz (Figure 3.4(a)). We already briefly introduced the model in Section 2.3.2, but let us give a brief reminder of notation and features of the model. 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

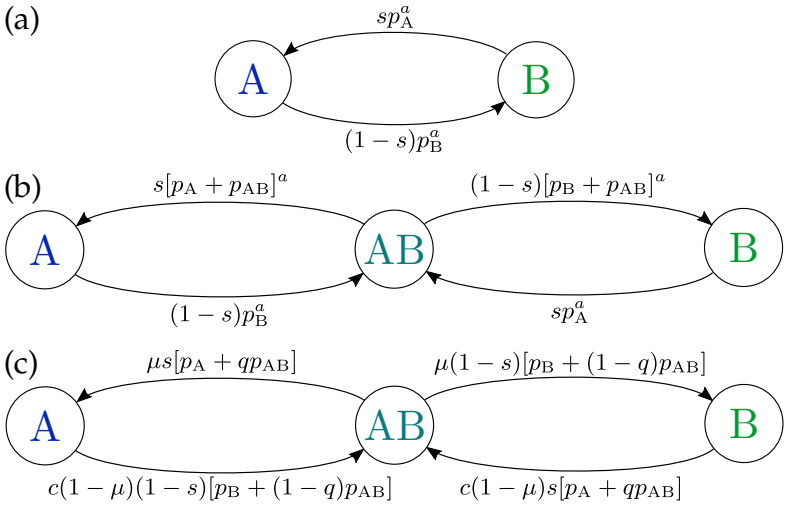


Figure 3.4: Diagrams of the models presented in the text, showing the transition probabilities from one state to another. (a) Abrams-Strogatz model from [5]. (b) Bilinguals model from [42]. (c) Our model of bilinguals including both their preference and the ease to learn the other language (see Equation (3.11)).

($a > 1$). 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. The model was thoroughly analysed in [217], 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, as shown previously in Figure 2.4. 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. [42] (see Figure 3.4(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 [217] 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 situation of diglos-

sia [69] 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 [47, 182]. 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. It is however hard to argue that French speakers in Quebec or Catalan speakers in Catalonia are loosely attached to these 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 Figure 3.4(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 [155] 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:

$$P(\mathrm{AB} \to \mathrm{X}) = \mu\, s_{\mathrm{X}}\, [p_{\mathrm{X}} + q_{\mathrm{X}}\, p_{\mathrm{AB}}]\,, \tag{3.9}$$

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, present in the model of [155], and $q_{\mathrm{X}} \in [0, 1]$, that reflects the preference of bilinguals to speak X, which has not been considered in previous works. 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:

$$P(\mathrm{X} \to \mathrm{AB}) = c\,(1-\mu)\,s_{\mathrm{Y}}\,\left[p_{\mathrm{Y}} + q_{\mathrm{Y}}\,p_{\mathrm{AB}}\right], \tag{3.10}$$

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 Figure 3.4(c) and we write here below the transition probabilities that define it:

$$\begin{cases} P(\mathrm{A} \to \mathrm{AB}) = c\,(1-\mu)\,(1-s)\,\left[p_{\mathrm{B}} + (1-q)\,p_{\mathrm{AB}}\right] \\ P(\mathrm{B} \to \mathrm{AB}) = c\,(1-\mu)\,s\,\left[p_{\mathrm{A}} + q\,p_{\mathrm{AB}}\right] \\ P(\mathrm{AB} \to \mathrm{A}) = \mu\,s\,\left[p_{\mathrm{A}} + q\,p_{\mathrm{AB}}\right] \\ P(\mathrm{AB} \to \mathrm{B}) = \mu\,(1-s)\,\left[p_{\mathrm{B}} + (1-q)\,p_{\mathrm{AB}}\right] \end{cases}. \tag{3.11}$$

Given the normalisation condition $p_{\mathrm{A}} + p_{\mathrm{B}} + p_{\mathrm{AB}} = 1$, these transition probabilities can actually be rewritten in terms of, let us say, p_{A} and p_{B} only. 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 [186]. 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 ($p_{\mathrm{B}} = 0$). Still, the use of B by the bilinguals keeps attracting monolinguals of the group A proportionally to $(1-q)\,p_{\mathrm{AB}}$.

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*

Figures shown in this section were generated with Mathematica code, freely available on GitHub [144].

We first analyse the model in the simplest setting of a single completely interconnected population to determine the typology of possible solutions. A mean field approximation yields:

$$\begin{cases} \dfrac{\mathrm{d}p_\mathrm{A}}{\mathrm{d}t} = (1 - p_\mathrm{A} - p_\mathrm{B})P(\mathrm{AB} \to \mathrm{A}) - p_\mathrm{A}P(\mathrm{A} \to \mathrm{AB}) \\ \dfrac{\mathrm{d}p_\mathrm{B}}{\mathrm{d}t} = (1 - p_\mathrm{A} - p_\mathrm{B})P(\mathrm{AB} \to \mathrm{B}) - p_\mathrm{B}P(\mathrm{B} \to \mathrm{AB}) \end{cases} . \tag{3.12}$$

Fixed points are the solutions for which $\frac{\mathrm{d}p_\mathrm{A}}{\mathrm{d}t} = \frac{\mathrm{d}p_\mathrm{B}}{\mathrm{d}t} = 0$. The stability of these points is studied by performing a linear perturbation analysis around them, which requires the calculation of the Jacobian of the linearised 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 Figure 3.5 show where the model converges to in three characteristic examples, depending on the model parameters. In the first one (Figure 3.5(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 Figure 3.5(b), the stable fixed point falls in the middle of the diagram and, therefore, the solution is

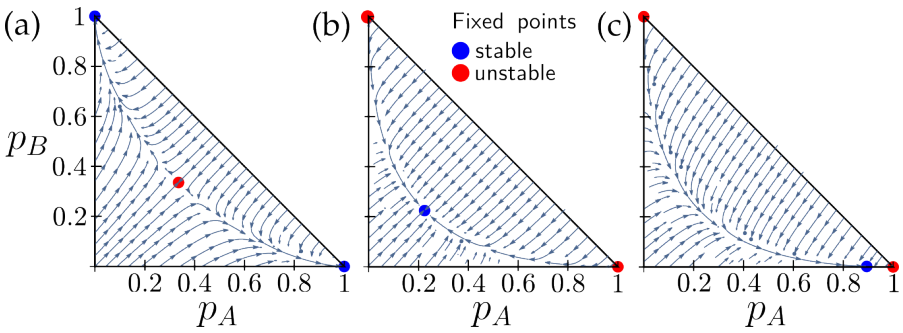


Figure 3.5: Flow diagrams for the dynamics of two languages according to our model described in Equation (3.11) set in a well-mixed population. p_A and p_B denote the proportions of monolinguals in A and B, respectively, and the proportion of bilinguals p_AB is such that $p_\mathrm{A} + p_\mathrm{B} + p_\mathrm{AB} = 1$. 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 a lower prestige, B survives in a small community of bilinguals as it is the preferred language among them.

symmetric coexistence with a majority ($\sim 1/2$) of bilinguals. Finally, in Figure 3.5(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. Remarkably, it does so without introducing extra parameters to fit. Indeed, μ and q are both quantities that can be measured in real-world scenarios, as opposed, for instance, to the inter-linguistic similarity introduced by [156] that is, as said explicitly in this work, not straightforward at all to calculate, and is thus an extra parameter that needs to be fit to data. Moreover, although it is out of the scope of this work, the parameter q can naturally be measured on an individual level, which can be used to initialise more realistic simulations within an ABM framework.

We change now the viewpoint from the phase space to the parameter space. In Figure 3.6, 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

$$r = \frac{\mu}{c\,(1-\mu)}, \tag{3.13}$$

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, $q^{\text{opt}} > 1 - s$. For $q < q^{\text{opt}}$, A is more at risk of extinction whereas for $q > q^{\text{opt}}$, 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 will now analyse the model in a metapopulation framework to uncover the effect of including space and check whether this pattern can arise.

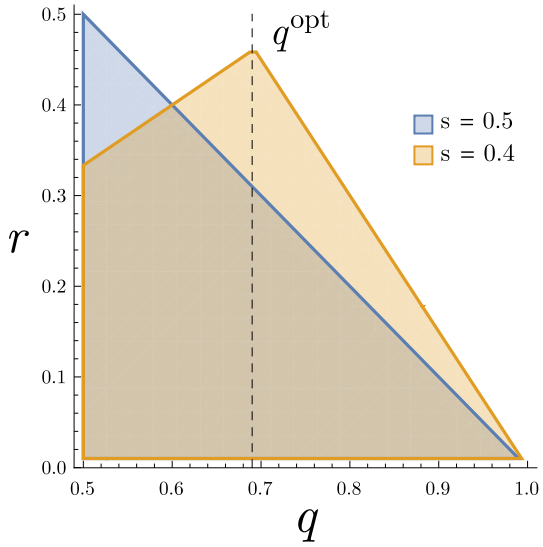


Figure 3.6: Region of the parameter space where the dynamics of our model in a single population converge to stable coexistence of languages. We show two 2D cuts of the coexistence region in the (q, r) space for fixed values of $s = 0.5, 0.4$, with r as defined in Equation (3.13). Lower values of r favour coexistence, as well as a neutral prestige and bilingual preference q. When $s < 0.5$, coexistence is favoured for an optimal value $q^{\text{opt}} > 1 - s$.

3.2.4 *The model in space*

In our context, we would need some information to build the extended model within this metapopulation framework. 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 Belgium as a benchmark, 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 [18, 37, 38].

The populations and commuting patterns at the municipality level are thus obtained from the national census of 2011[1]. We implement

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

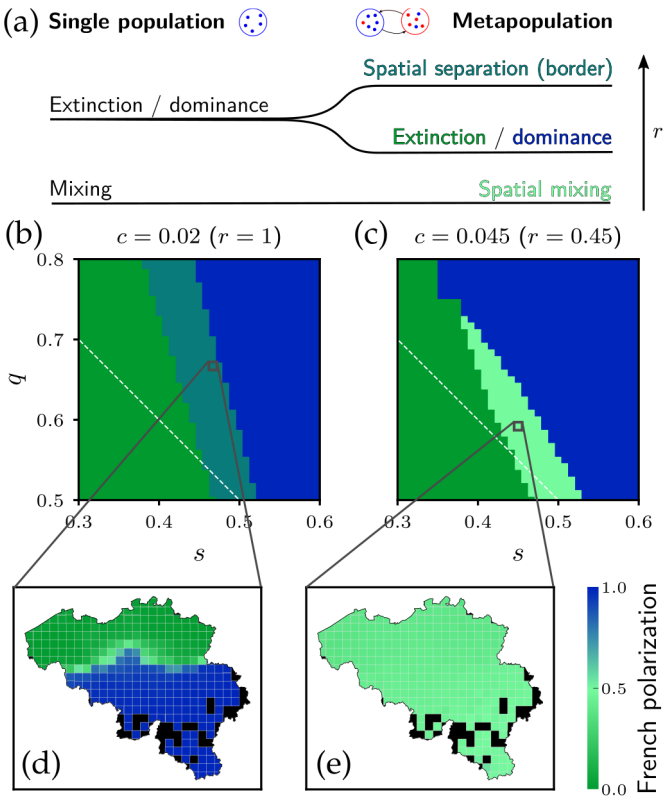


Figure 3.7: Types of stable states of convergence of our model in a metapopulation set for Belgium. (a) Diagram illustrating the effect of adding metapopulations in the stable states of a single population: the former extinction state bifurcates in full extinction and in a boundary-like state with monolinguals separated in space. Larger values of r favour homogeneity, either by full extinction or by separation states. 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 maps of the French polarisation (A = French and B = Dutch in Equation (3.3)) 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 $r = 0.45$, $s = 0.45$, $q = 0.592$. s and q are defined as the prestige of and preference for French.

a mapping process from municipalities to our cells based on area overlap, similarly to what we described in Section 2.1.4.3.

Regarding the language groups, since we are mixing commuting data from the census with data from Twitter, we first have to up-scale the number of individuals we found from Twitter data to match the number of commuters. Instead of up-scaling uniformly by the ratio between these two numbers, we choose here to alleviate the socio-demographic biases that we mentioned in Section 2.1.4.5 through re-scaling factors $k_{L,i}$, which are different between cells and language groups. The first bias to address is how users are more urban than average [158], which leads us to impose the constraint $\sum_L k_{L,i} N_{L,i} = N_i^{(\text{census})}$. The second one is how language usage is different on Twitter, because one's audience is wider, potentially more cosmopolitan than offline [166], and with different demographics [158]. Hence we want to also re-scale such that $\sum_i k_{L,i} N_{L,i} = N_L^{(\text{census})}$, with $N_L^{(\text{census})}$ the global numbers of L-speakers, data which are much more widely available than the counts for each census tract. By imposing these two constraints, it is assumed that the two biases are independent: the bias in the choice of language used online is space-independent in a specific multilingual society, and the over-representation of urban people is L-group-independent. In practice, we fit the values of the upscaling factors $k_{L,i}$ via *iterative proportional fitting* (IPF) [53, 71].

Once the metapopulation has been initialised, the model can be simulated. As in [70], 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 (3.11), and then move to their work cells where again they interact with the local population. The agents encounter thus different environments characterised 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 [17, 192] for epidemiological models. Our derivations and the resulting equations are given in Appendix B. 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 Figure 3.7 and Figure 3.8. 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 (Figure 3.7(e)), global extinction and local extinction of a language in part of the territory leading to a boundary-like state (Figure 3.7(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, as shown in Figure 3.8. 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 Figure 3.6 may not have been reached.

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 Figure 3.9. 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 Figure 3.9(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 Figure 3.7(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 Figure 3.9(b) shows that the final stable state exhibits more segregation for both monolinguals and bilinguals, since the boundary between communities lies in the

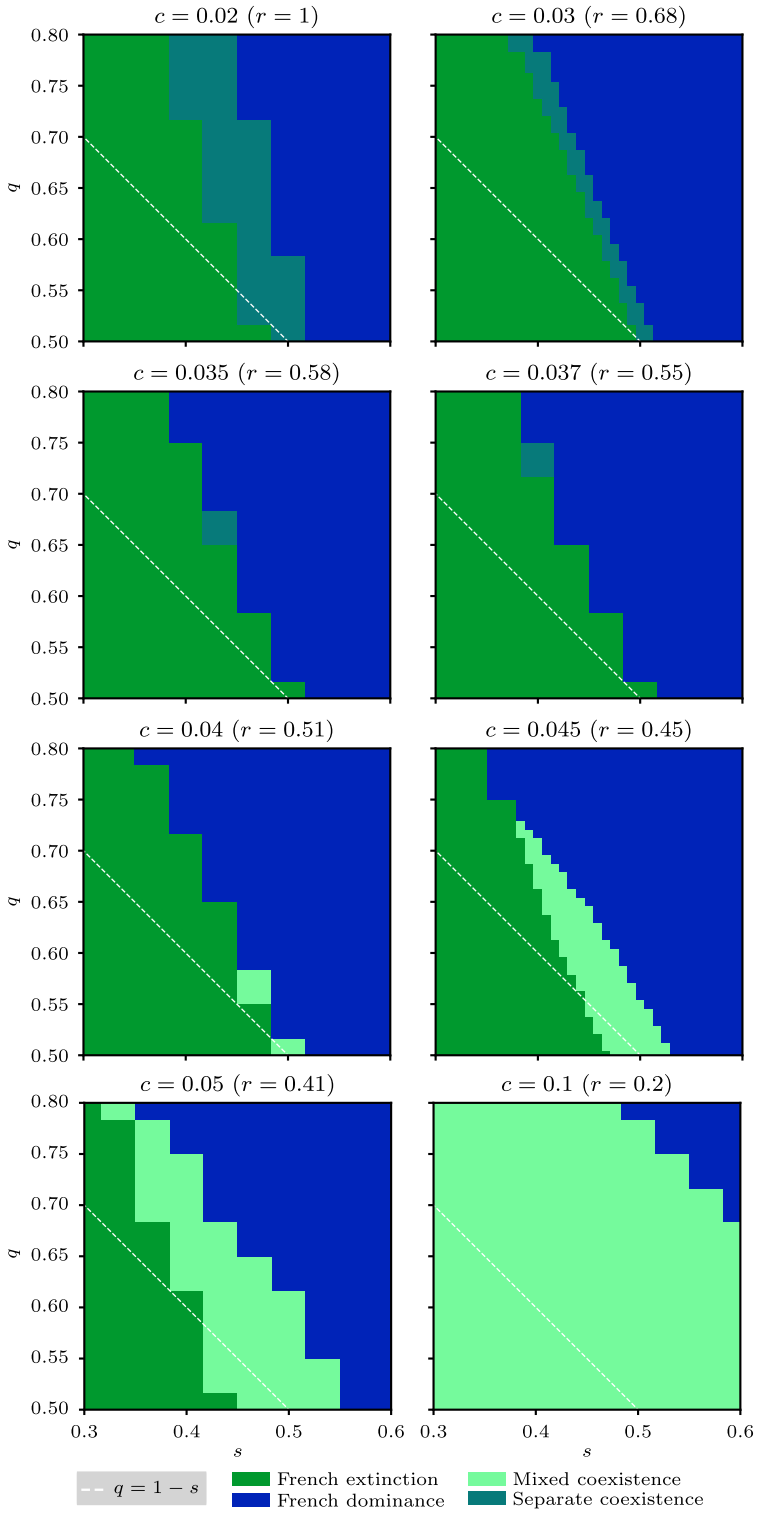


Figure 3.8: Phase space of the stable states of convergence of our model when iterated within our metapopulation framework in Belgium. Colours show the kind of convergence state reached for the corresponding set of parameters. A lower learning rate, that is a lower value of c, favours the separate coexistence of French and Flemish, while a higher one favours mixed coexistence. In the transition between the two however, coexistence is almost impossible, as the corresponding region in the (s, q) space becomes very narrow.

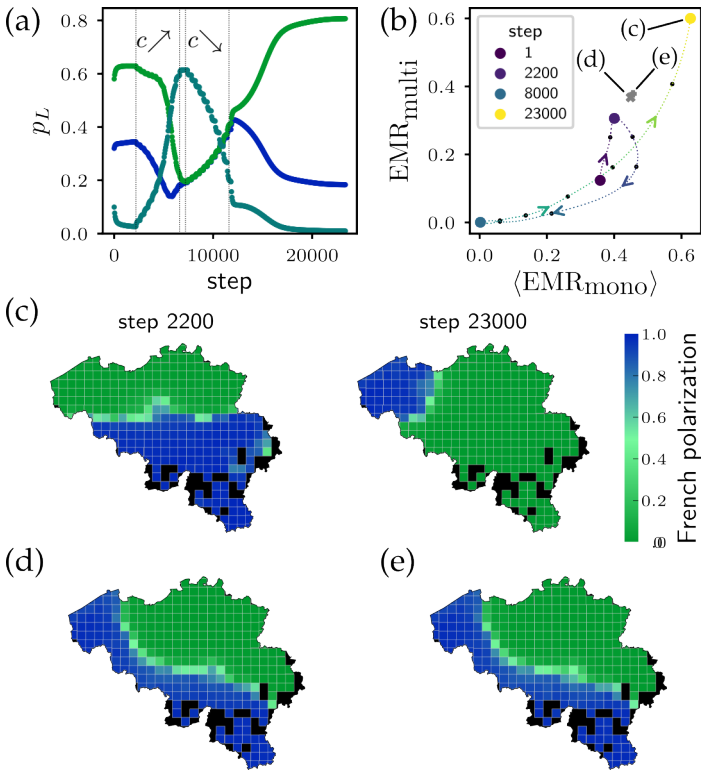


Figure 3.9: Evolution of the state of the metapopulation model in Belgium when c varies, first slowly increased and then decreased to recover the original value. We fixed $s = q = 1/2$ and $\mu = 0.02$. (a) Evolution of the global proportions p_L of individuals belonging to each L-group. The blue curve 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) Polarisation 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) Polarisation 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).

countryside, and not around Brussels as in the original scenario. The importance of the history of languages is hence clearly shown by this experiment.

The seemingly random placement of the boundary may be owed to the absence of constraints on the system, which is completely closed. In reality a country is an open system with exterior influences, notably from its direct 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. We use the population censuses of these two countries at a municipality level[2] to determine how many commuters will come from each municipality close to the border. On each side of the two borders, we only keep cells and municipalities located less than 50 km away from the border. We take a fixed proportion p_{TB} of the population of each of these municipalities as our transborder commuters. Then, to spread the commuters of each municipality to the cells in Belgium, we use a very simple gravity law [196], stating that the number of commuters going from municipality γ to cell j is such that

$$T_{\gamma j} \propto \frac{P_\gamma P_j}{d_{\gamma j}^2}, \tag{3.14}$$

with P denoting the population and d the inter-centroid distance. The normalisation for every γ is given by

$$\sum_{j \in \mathcal{B}} T_{\gamma j} = p_{\mathrm{TB}} P_\gamma, \tag{3.15}$$

with $\mathcal{B}$ the set of border cells. These commuters $T_{\gamma j}$ then appear as an additional fixed population of monolinguals at every work step in the cell j. 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 Figure 3.9(d-e)). This positioning is a clear improvement over the closed-system simulation, albeit still not quite the one we observed in Figure 3.1. In Figure 3.9(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 contemplated 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 [199]. However, in

2 Available at https://www.insee.fr/fr/statistiques/4989724 for France and at https://opendata.cbs.nl/statline/#/CBS/nl/dataset/70072ned/table?dl=3B993 for the Netherlands.

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 EMR, 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 non-linearities. 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 counter-intuitively, increasing this ease to learn the other language may break the existing boundary and lead to extinction, and not to the desired coexistence with mixing of the languages. This calls for caution when designing policies since the final state is strongly history-dependent.

Overall, our findings shed light on the role of heterogeneous speech communities in multilingual societies, and they may help shape the objectives and nature of language planning [124] 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 *[184]*

Language bounds individuals together just as much as it divides them. Upon reading the last sentence, one may first think of interlanguage boundaries as the ones we have seen in the previous chapter: if two individuals cannot understand each other's language, communication is effectively very limited. This chapter, on the other hand, is dedicated to intra-language differences: when people speak the same language, but differently, which is also a strong marker of social divides. As we already emphasized in Chapter 1, language variation in a population is driven by many factors, among which the socio-economic background of individuals is essential. And as we also said previously, one's ability to use the standard form of their language can be associated to a higher linguistic capital, and has thus economic values. It is not uncommon that some classes of the population, particularly those of relatively low *socio-economic status* (SES), prefer to use the non-standard form. This has been shown by prominent sociolinguists in the past, as they conducted field studies and analysed differences between socio-economic classes [133, 210]. Although this phenomenon has long been known [44], it has been observed on very limited scales, and, more importantly, no real explanation of how it might emerge has been proposed. These are the two shortcomings we wish to address in this chapter. As groups of lower SES have less linguistic capital, their economic opportunities will also be affected. SES and this linguistic variation are thus mutually sustaining: one inherits a SES along which comes a linguistic capital that contributes — among other things — to constrain them to their status of origin. Understanding the mechanisms that lead to this linguistic segregation is therefore of great importance.

First, as we have already shown in previous chapters, a first step toward understanding is to measure the phenomenon. It has already been quantified by the PISA reports of the OECD [172], which consistently show that students with lower socio-economic background tend to have a lower reading proficiency. While these confirm there is an issue to tackle, they are not extensive enough. They do not test language production, and not the whole population but only a sample

of students of a specific age. Alternative empirical works are thus needed. Data from social media have repeatedly proven useful to link SES and different social behaviours [80]. In particular, Twitter data were used in the past to study geographical variation of linguistic variables [62]. Closer to our topic of interest still, some past work has investigated the dependence on SES of the frequency of a few markers of non-standard language in France, as seen on Twitter [4], showing the potential of this data source for such an analysis. Here, we investigate these dependencies in the UK, by way of measuring how Twitter users abide by the rules of the standard variety of their language.

4.1 AN INFLUENCE OF SOCIO-ECONOMIC STATUS?

4.1.1 *Twitter data analysis*

To do so, we analyse the tweets written in English of users identified as UK residents, following the methodology described in Section 2.1.4. To assign a SES to these users, we also determine their area of residence from the geotags attached to their tweets. Our unit areas for the study are the 7201 *middle layer super output area*s (MSOAs), which are areas created by the Office for National Statistics of the UK for the output of the census estimates. They host at least 5000 and at most 15 000 inhabitants, with a typical population of 10 000. Their boundaries can be downloaded from the Open Geography portal of the Office for National Statistics[1]. Importantly, the average annual net income of their residents can be obtained from the census[2], which gives us a proxy for the SES of our Twitter users.

The second ingredient we need for our analysis is their propensity to deviate from the rules defining standard English. The key to this measurement is LanguageTool, an open-source grammar, style and spell checker. The first advantage of using such a tool over a pre-defined set of rules, as in [4], is that the tool is implemented in 15 languages. Our study can thus quite easily be replicated in other countries. The second advantage is that it covers a very wide spectrum of potential mistakes: it has more than 5500 rules defined for the English language. These rules are categorised in 11 categories, among which are grammar mistakes, confused words or typos. In this work we focus on grammar mistakes, as they are among the most common and are the most characteristic of non-standard language. LanguageTool therefore enables us to compute the frequency at which Twitter users identified as UK residents make grammar of mistakes, according to

1 https://geoportal.statistics.gov.uk/maps/msoa-dec-2011-boundaries-generalised-clipped-bgc-ew-v3

2 In particular the 2018 net annual income estimates, available at https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/smallareaincomeestimatesformiddlelayersuperoutputareasenglandandwales

standard rules. From this raw frequency we then compute the frequency of mistakes per word written by the user. As in the previous chapter we compute this frequency on a user level, to try and be more representative of the general population, and not only of the few, very active users. At the cell level, we then compute the average of these relative frequencies for all residents. This implies that we cannot keep all users, as some have been so little active that they did not write enough to make mistakes, so we keep only users who have written at least 100 words. Also, to exclude cells with too little statistics, in the following we only keep cells with at least 15 residents left after this previous filter. This leaves us with 131 402 users spread across 4879 MSOAs. In each of these cells, our analysis yields estimates of the income of its residents and of the frequency at which they make grammar mistakes.

4.1.2 *Correlations with the frequency of mistakes*

We find a significant, but weak correlation (Pearson r equal to -0.25) between the net income and the average frequency of mistakes in all cells of the country. We then focus on metropolitan areas because we expect them to host more linguistic and socio-economic heterogeneity than their rural counterparts. We also happen to have more data in these on both mobility and language production, due to the fact that Twitter users tend to be more urban [158]. We therefore consider the 8 largest metropolitan areas in England, and find large differences between these areas, with correlations ranging from -0.07 in Sheffield to -0.5 in Bristol. The average frequency of grammar mistakes is plotted against the average of net annual income of their residents for all MSOAs of England and Wales, London, Manchester and Sheffield in Figure 4.1.

4.1.3 *The role of assortativity*

To find out what could make these cities so different in that regard, we measure the assortativity in the mobility patterns of their residents [112]. We thus determine how likely people from different socio-economic classes are to interact with each other.

SES classes are defined such that each class has roughly the same population. Since our proxy for SES is the average income in every MSOA, every resident of a cell will necessarily be assigned to the same class. Considering the cells in a given metropolitan area, we rank them by increasing average net income. We get their population from the census[3], denoted N_c for each cell c in the following. Denoting

3 https://www.nomisweb.co.uk/census/2011/ks101uk

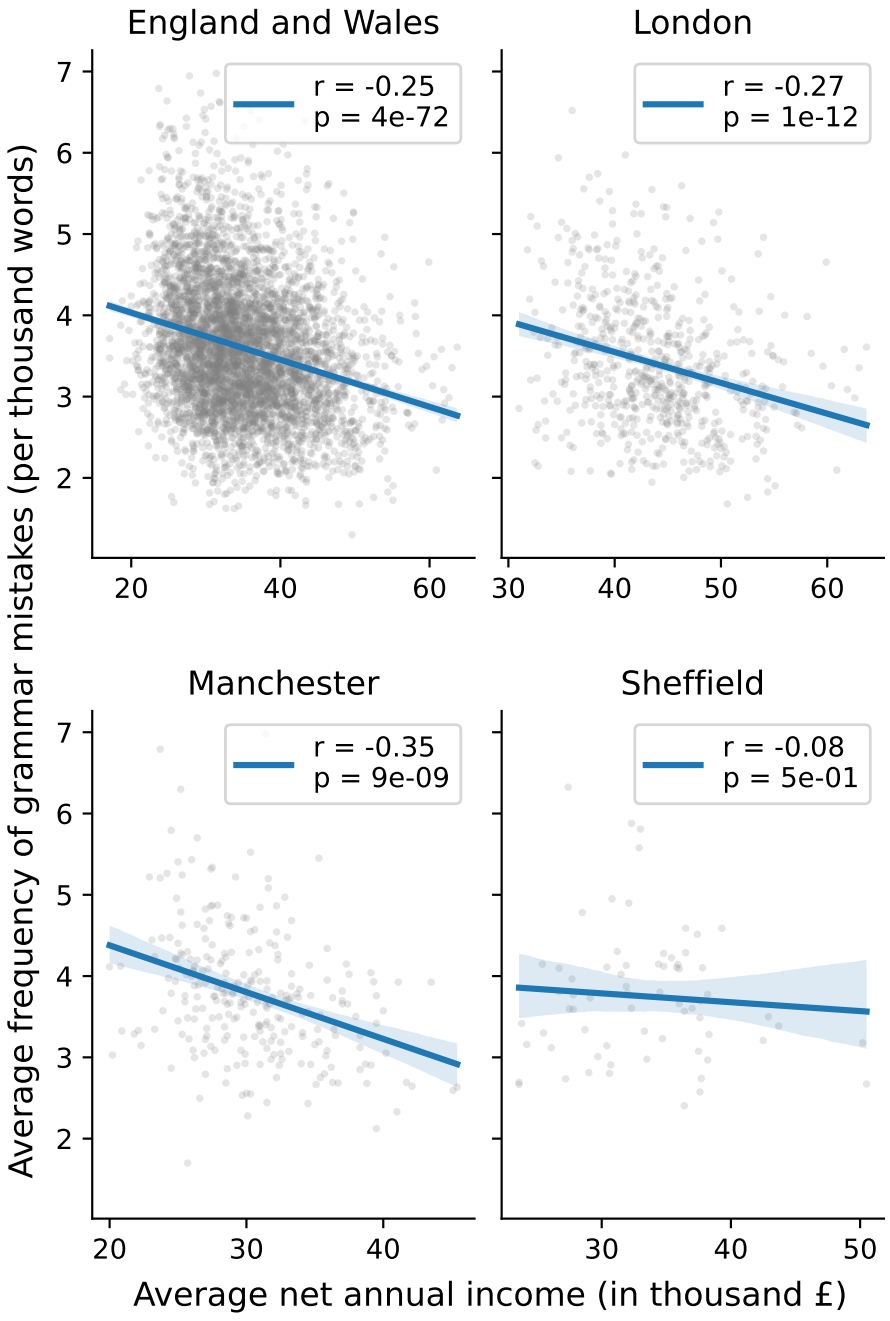


Figure 4.1: Correlation between frequency of grammar mistakes and net income. We show them for all MSOAs of England and Wales, and then of the metropolitan areas of London, Manchester and Sheffield. Each grey dot corresponds to an MSOA, while blue lines indicate the result of a linear regression, with the corresponding Pearson r and p-value given in each legend. The areas coloured in light blue indicate the 95 % confidence interval for the regression estimates.

P_c the set of cells less populated than c, formally defined as $P_c = \{c' \in C \mid N_{c'} < N_c\}$, we define the SES class of c as follows:

$$S_c = n_S \left\lceil \frac{\sum_{c' \in P_c} N_{c'}}{\sum_{c' \in C} N_{c'}} \right\rceil, \tag{4.1}$$

with $\lceil \cdot \rceil$ representing the ceiling function and n_S the number of classes we wish to define. We denote $t_{u,c}$ the number of trips made by user u to cell c, and introduce the probability for an individual residing in a cell of class i to visit another of class j:

$$M_{i,j} = \frac{\sum_{u \in S_i} \sum_{c \in S_j} t_{u,c}}{\sum_{j \in S_j} \sum_{u \in S_i} \sum_{c \in S_j} t_{u,c}}, \tag{4.2}$$

which is normalised row-wise, meaning $\sum_j M_{i,j} = 1$. Since we are interested in how much individuals mix when they move, we exclude the tweets sent from their cell of residence to compute $M_{i,j}$. Such matrices computed for our eight metropolitan areas are represented in Figure 4.2. They show completely different patterns, from cities where the diagonal elements slightly stand out as in London or Manchester, to cities in which one specific class of cells is the destination of most outward mobility, as in Newcastle or Bristol.

These patterns can be summarised by a measure of how strongly diagonal these matrices are: their Pearson r value, denoted r_M in our case. It is defined as follows:

$$r_M = \frac{\sum_{i,j} M_{i,j} \cdot \sum_{i,j} ijM_{i,j} - \sum_{i,j} iM_{i,j} \cdot \sum_{i,j} jM_{i,j}}{\sqrt{\sum_{i,j} i^2 M_{i,j} - \left(\sum_{i,j} iM_{i,j}\right)^2} \cdot \sqrt{\sum_{i,j} j^2 M_{i,j} - \left(\sum_{i,j} jM_{i,j}\right)^2}}. \tag{4.3}$$

This gives us a measure of assortativity, meaning that this metric gives us a sense of how much people of similar classes tend to stay together. The closer to 1, the more individuals tend to go to areas of similar SES (assortative mixing), and the closer to -1, the more they will stay with people of opposite SES (disassortative mixing). Values close to 0 indicate no assortativity, meaning no preference in mixing. We then measure this assortativity for our metropolitan areas from the matrices shown in Figure 4.2, and show the results in Figure 4.3 against the correlation we computed previously between socio-economic status and the average frequency of grammar mistakes. What we find is a very clear correlation between this assortativity measured at the city level, and the correlation between SES and the frequency of grammar mistakes, as shown in Figure Figure 4.3. This indicates that the more mixing in the population, the less the frequency of mistakes is determined by the SES of origin. Further, as shown in the bottom panel of Figure 4.3, more mixing tends to imply more frequent mistakes. It then appears to favour the spread in popularity of non-standard English.

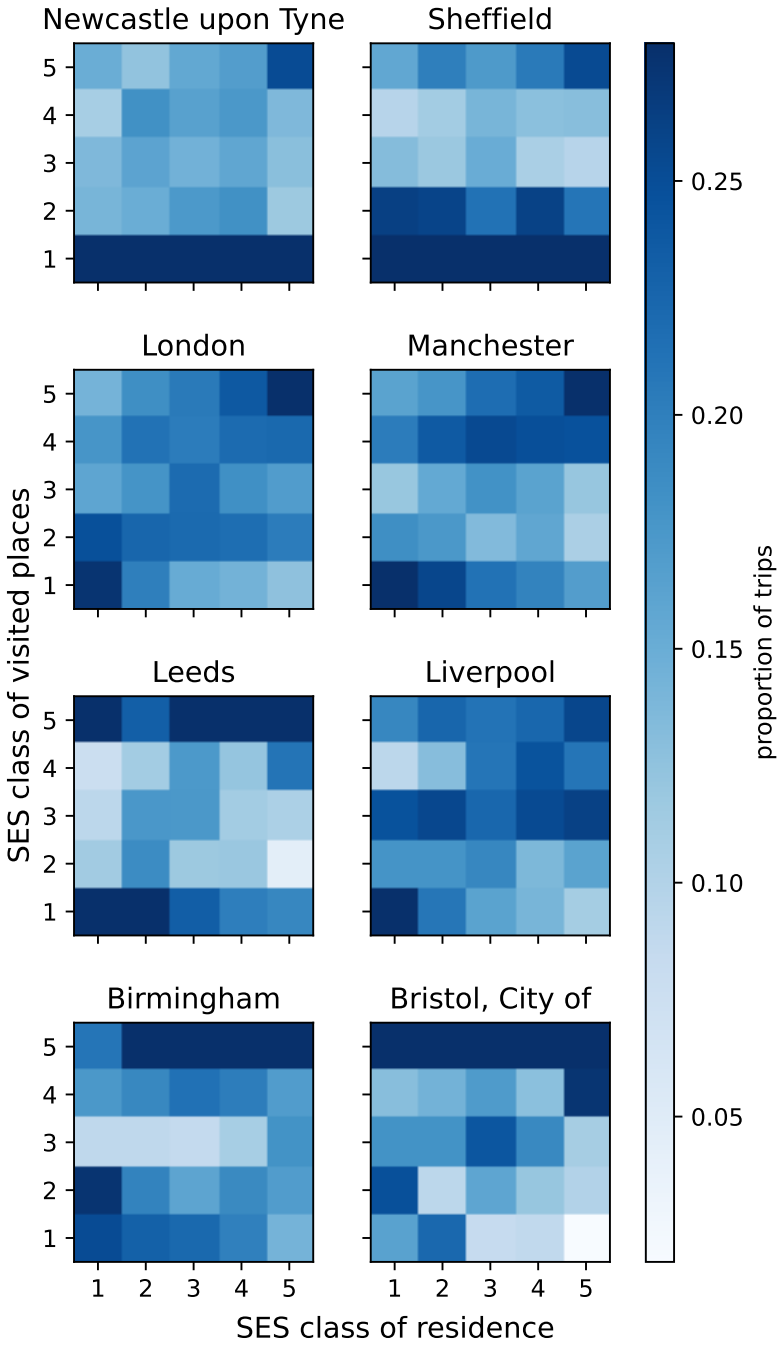


Figure 4.2: Matrices of stratification of SES classes in their mobility in eight metropolitan areas of England. For every pair of SES classes for the cell of residence and the cell visited by Twitter users, this shows the proportion of trips that were made from the latter to the former. A square located at $(x, y) = (i, j)$ corresponds to $M_{j,i}$, as it is defined in Equation (4.2). The normalization is thus column-wise on this representation.

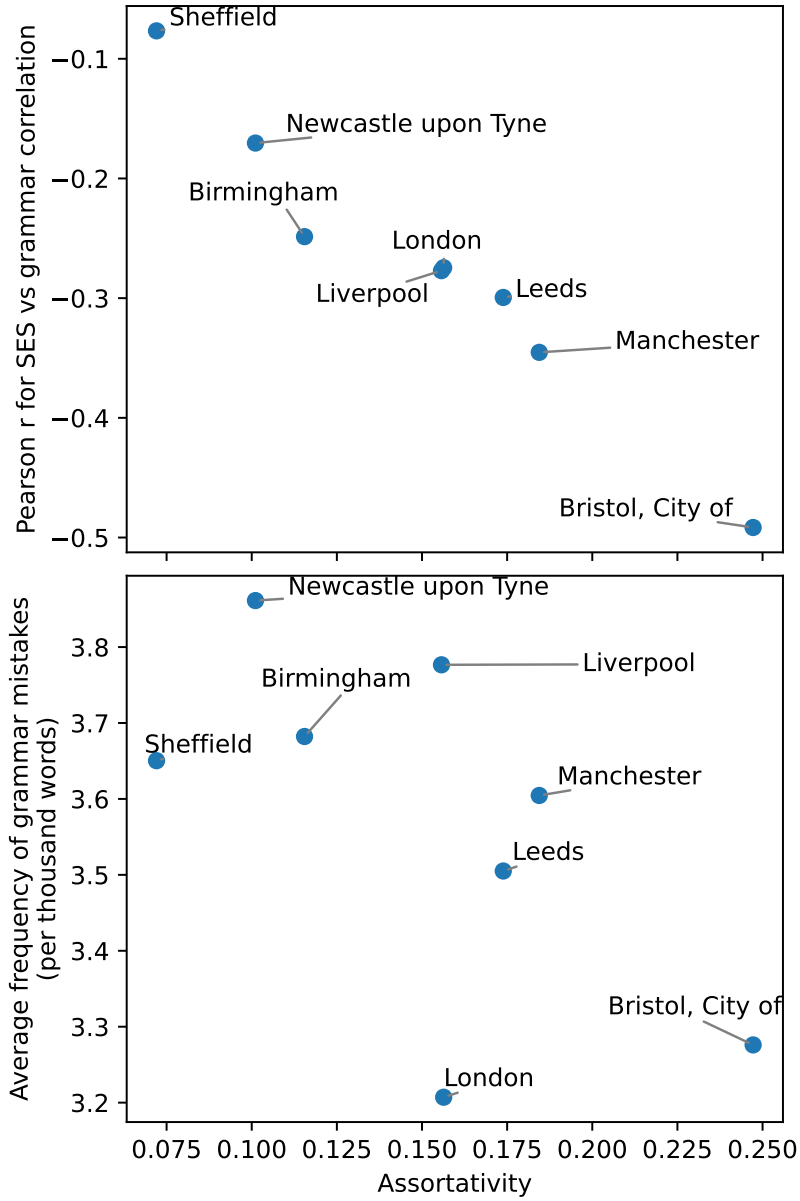


Figure 4.3: The influence of assortativity on the frequency of grammar mistakes (bottom) and their dependence on SES of origin (top). The latter clearly decreases as assortativity increases. This means that the more SES classes mix with one another, the less the usage of the standard form of individuals will depend on their own SES. As for the frequency of mistakes, it seems to rather increase with mixing.

4.2 MODEL

4.2.1 *Definition*

Having identified the importance of social mixing, we wish to understand the mechanisms behind this observation with a simple model. There are several effects we wish to consider, that we describe below.

(i) One of the two varieties of the language may be more prestigious than the other. This is for example the case of the standard form: it is taught at schools and spoken by mainstream media and public institutions [51].

(ii) Even though one variety is less prestigious, it might still be preferred by a part of the population that has some kind of cultural attachment to it. For instance, slang can be preferred by members of lower SES, as using it might give them a sense of group identity [133, 210].

(iii) We previously observed very different mixing patterns in UK metropolises. Indeed, mobility can be very heterogeneous, so it should be possible to plug in any mobility pattern into the model to be able to understand how different mixing may affect the dynamics of the linguistic varieties.

With these considerations in mind, we propose an ABM, that we describe next. It considers agent who can have one of two SES classes, who can either speak standard, or not. The standard form has an intrinsic prestige l_v, the value of which belongs to the unit interval, but that we will always set above 0.5 in the following, to account for (i). Now turning to (ii), each SES class has a preference for one form: the lower class 1 is attached to the non-standard form with a factor q_1, and inversely the higher class 2 is attached to the standard one with q_2. They are also comprised between 0 and 1, and when one of these two parameters has a value above 0.5, it means there is a preference for the respective form. Then for instance, when an agent of low SES speaking non-standard interacts with another agent speaking standard, they have a probability $l_v(1 - q_1)$ to start using the standard form as well. The agents can move from their residence cell with a probability $M_{i,j}$ at each step, which thus controls the mixing of the two populations, as discussed in (iii). A summary diagram of the model is provided in Figure 4.4.

4.2.2 *Properties and behaviour in mean-field*

The process we just described above can be simulated for any number of cells and arrangements of the populations of different SES classes. Here, however, we will consider a rather simple case in order

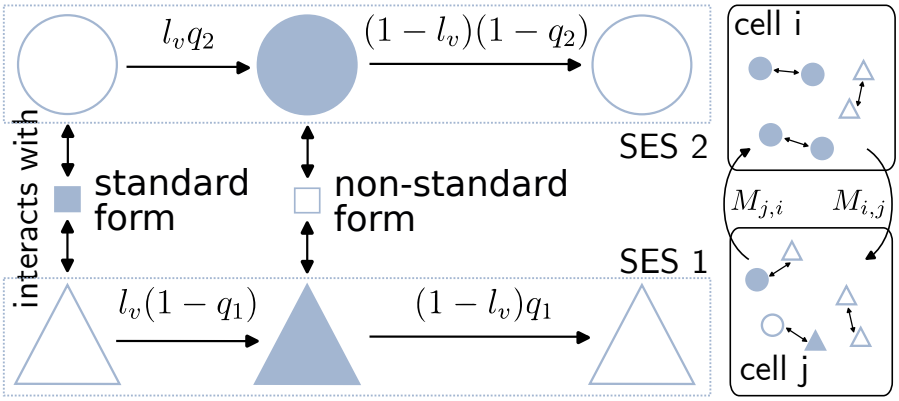


Figure 4.4: Diagram summarizing our agent-based model for the adoption of a linguistic variety. It features two SES classes, represented by circles and triangles, which can speak one of two varieties of their language: the standard form (filled shapes) and the non-standard (empty shapes). Each agent has a cell of residence, and a mobility matrix with elements $M_{i,j}$ defines the probabilities for a resident of i to be at different cells at each time step. After they have moved, for every agent, another agent is picked randomly for them to interact with, and if they happen to interact with an agent using the other variety, in function of their SES class of origin they have different probabilities to adopt this variety. For instance, when a triangle (SES 1) speaking non-standard interacts with another agent speaking standard, they have a probability equal to $l_v(1-q_1)$ to adopt the standard form (transition represented in the bottom left).

to present succinct equations that lend themselves to interpretation and mathematical analysis. We will consider only two cells, with N_1 individuals of class 1 residing in cell 1, and N_2 individuals of class 2 residing in cell 2. This is a situation of complete socio-economic segregation. We will denote p_1 the proportion of individuals of class 1 speaking non-standard (variety 1), and p_2 the proportion of individuals of class 2 speaking standard (variety 2). Individuals have to speak either 1 or 2, which implies that for instance a proportion $(1 - p_1)$ of individuals of class 1 speak the variety 2. The two variables therefore summarise the linguistic state of the system at any given time. Regarding the mobility, we introduce a new variable $m_{i,j}$, such that:

$$m_{i,j} \equiv \frac{N_i M_{i,j}}{\sum_{i'} N_{s_{i'}} M_{i',j}}, \tag{4.4}$$

which can be understood as the probability to pick an individual with residence in i from j. They satisfy $\sum_i m_{i,j} = 1$. Further, since there are only two cells, people either stay at their residence or move to the other cell. As a result, to account for different mobility patterns, out of the 2×2 mobility matrix with elements $M_{i,j}$, we only retain the parameters $M_1 \equiv M_{1,2}$ and $M_2 \equiv M_{2,1}$, the probabilities for people of each class to move away from their residence cell. Likewise, the matrix of $m_{i,j}$ can be summarised with the corresponding m_1 and m_2. Under all these assumptions and new notation, we can then write the probability to pick an individual speaking a variety in each cell:

$$\begin{aligned}
P(2 \mid c = c_1) &= (1 - m_2)(1 - p_1) + m_2 p_2 \\
P(2 \mid c = c_2) &= m_1(1 - p_1) + (1 - m_1)p_2 \\
P(1 \mid c = c_1) &= (1 - m_2)p_1 + m_2(1 - p_2) \\
P(1 \mid c = c_2) &= m_1 p_1 + (1 - m_1)(1 - p_2)
\end{aligned} \tag{4.5}$$

This further enables us to determine the probabilities to switch from one variant to another, depending on one's SES class:

$$\begin{aligned}
P(1 \to 2 \mid s = s_1) &= l_v(1 - q_1)[P(2 \mid c = c_1)(1 - M_1) \\
&\qquad + P(2 \mid c = c_2)M_1] \\
P(1 \to 2 \mid s = s_2) &= l_v q_2[P(2 \mid c = c_1)M_2 \\
&\qquad + P(2 \mid c = c_2)(1 - M_2)] \\
P(2 \to 1 \mid s = s_1) &= (1 - l_v)q_1[P(1 \mid c = c_1)(1 - M_1) \\
&\qquad + P(1 \mid c = c_2)M_1] \\
P(2 \to 1 \mid s = s_2) &= (1 - l_v)(1 - q_2)[P(1 \mid c = c_1)M_2 \\
&\qquad + P(1 \mid c = c_2)(1 - M_2)]
\end{aligned} \tag{4.6}$$

Working in mean field, we can write the following differential equations:

$$\begin{aligned}\frac{\mathrm{d}p_1}{\mathrm{d}t} &= (1-p_1)P(2\to 1 \mid s=s_1) - p_1 P(1\to 2 \mid s=s_1)\\ \frac{\mathrm{d}p_2}{\mathrm{d}t} &= (1-p_2)P(1\to 2 \mid s=s_2) - p_2 P(2\to 1 \mid s=s_2)\end{aligned} \tag{4.7}$$

Now if we simplify further and assume that the two SES classes have the same population, and the same mobility $M = M_1 = M_2$, then we get that $m_1 = m_2 = M$, and it follows that

$$\begin{aligned}\frac{\mathrm{d}p_1}{\mathrm{d}t} &= 2M(1-M)(1-p_1-p_2)[q_1(1-l_v) - p_1(q_1-l_v)]\\ &\quad + p_1(1-p_1)(q_1-l_v)\\ \frac{\mathrm{d}p_2}{\mathrm{d}t} &= 2M(1-M)(1-p_1-p_2)[q_2 l_v - p_2(l_v+q_2-1)]\\ &\quad + p_2(1-p_2)(l_v+q_2-1)\end{aligned} \tag{4.8}$$

One can first notice how each equation features a first term linked to the groups' mobility, maximum for $M = \frac{1}{2}$, which is the situation of maximum mixing. This term is null when $p_1 = 1 - p_2$, and thus tends to smooth out differences in variety usage between SES classes. Indeed, in the first equation, given that $q_1 - l_v = q_1(1 - l_v/q_1)$, and that, as these parameters belong to the unit interval, $1 - l_v/q_1 > 1 - l_v$, then this term's sign follows the one of $(1 - p_1 - p_2)$. It is therefore negative if $p_1 > 1 - p_2$ and positive otherwise, meaning it pushes p_1 towards $1 - p_2$. Similarly in the second equation, since $q_2 l_v - q_2 - l_v > -1$, the mobility term pushes p_2 towards $1 - p_1$.

The second term of each evolution equation is one of "self-growth", independent of mobility and maximum for $p_k = \frac{1}{2}$. This term thus pushes a given class k away from a state of maximum entropy, and rather towards homogeneity. For $k = 1$ for instance, it does it either towards $p_1 = 0$ for $q_1 < l_v$, or towards $p_1 = 1$ for $q_1 > l_v$. When there is no mixing, for $M = 0$ or $M = 1$, as expected the two populations become completely independent, and one can directly see that the only stable fixed points of the system are homogeneous populations in terms of the variety they use. Which one they will end up using depends on the sign of $(q_1 - l_v)$ and $(l_v + q_2 - 1)$ for classes 1 and 2, respectively.

4.2.3 *Simulations on a toy example*

We ran simulations with 200 agents spread evenly between two SES classes, with each class in its own residence cell. We set the standard form to be the more prestigious one, agents with higher SES to be indifferent and ones with lower SES to be more attached to the

non-standard form. In Figure **??**(c), we show the result of this simulation when we set a very low inter-cell mobility. For Figure **??**(d), the mobility was greatly increased. We can see that the more the two populations mix, the closer they are to using the two forms in the same proportions, reflecting what we observed in the data.

4.2.4 *Simulations in real metropolitan areas*

This promising result calls for further investigation of the model, and particularly how it would fare when initialised with the actual data that we have in the metropolitan areas of England.

4.3 DISCUSSION

5

INFERRING AMERICAN CULTURAL REGIONS THROUGH THE LEXICAL ANALYSIS OF SOCIAL MEDIA DATA

> *Language is the road map of a culture. It tells you where its people come from and where they are going.*
>
> — *Rita Mae Brown,* Starting from Scratch *[35]*

Much of the work presented in this chapter is included in an article entitled 'American Cultural Regions Mapped through the Lexical Analysis of Social Media', that was previously published by the author of this thesis with Bruno Gonçalves, José J. Ramasco, David Sánchez and Jack Grieve [2].

Cultural identity is an elusive notion because it depends on a wide range of different cultural factors — including politics, religion, ethnicity, economics, and art, among countless other examples — which will generally differ across individuals, with the cultural background of every individual ultimately being unique. Nevertheless, individuals from the same region can generally be expected to share some cultural traits, reflecting the shared cultural values and practices associated with the region [33]. Identifying the cultural regions of a nation — regions whose populations are characterized by relative cultural homogeneity compared to the populations of other regions within the nation — is very valuable information across a wide range of domains. For example, it is important for governments to understand geographical variation in the values of their population so as to better meet their educational, social, and welfare needs. Similarly, from an economic standpoint, it is important to identify where certain services and products are most required and how best to engage with populations in different regions of the country. In general, defining the cultural regions of a nation is therefore a crucial part of understanding the complex landscape of human behaviour that nation encompasses, providing an accessible and broad classification of the populations of a country [137].

5.1 PREVIOUS WORKS

Mapping cultural regions has been a particularly active area of research in the United States, where there has long been debate over the cultural geography of the country, with a wide range of theories of American cultural regions having been proposed. Seven of the

most prominent theories [63, 83, 85, 142, 171, 224, 227] are mapped in Figure 5.1, showing considerable disagreement. For example, in [227] the geographer Wilbur Zelinsky identified 5 major cultural regions — New England, the Midland, the South, the Middle West, and the West — based on a synthesis of regional patterns in a wide range of cultural factors, including ethnicity, religion, economics, and settlement history. Alternatively, in [85] drawing on a similar but more extensive range of cultural factors, the social scientist Raymond Gastil identified 13 major cultural regions, offering a more complex theory than Zelinsky, including by dividing Zelinsky's Midland, Middle West, and West regions. The two studies illustrate two basic limitations with these types of approaches that subjectively synthesize a range of data to infer cultural regions. First, it is unclear exactly how relevant cultural factors should be identified. Zelinsky considers fewer factors than Gastil, which may explain his simpler proposal. Second, it is unclear how these different factors should be synthesized to produce a single overall map of cultural regions. Zelinsky places greater emphasis on the importance of initial settlement, which may also explain his simpler proposal.

Given the subjectivity underlying these studies, the lack of agreement over the number and location of American cultural regions (as illustrated in Figure 5.1) is not surprising. Only a distinction between the North and South, reflecting the Union-Confederacy border, and a distinction between the East and West, reflecting the path of the Rocky Mountains, are common to these most influential theories of American cultural regions [63, 72, 83, 85, 142, 171, 224, 227]. Otherwise, between 4 and 12 primary cultural areas have been mapped, typically including the Northeast [63, 72, 83, 85, 142, 171, 227], the South [63, 72, 83, 85, 142, 171, 224, 227], the West [63, 83, 85, 171, 224, 227], and the Midwest [63, 83, 85, 171, 227].

In large part, the debate over the geography of American cultural regions has been about which types of cultural factors should be given precedence, and how these factors should be combined. Crucially, these decisions have generally been left entirely to the judgment of the analyst. Quantitative data from the census and elections have sometimes been taken into consideration (e.g. [85, 142, 224, 227]), but less often subjected to statistical analysis (e.g. [142]), while the selection and weighting of these factors has always been subjective. For example, religion and politics are undoubtedly important cultural factors, but they can be measured in various ways, and it is unclear how important they are relatively speaking and whether their importance varies across the United States.

A basic question is therefore how can we infer general American cultural regions in an objective way? In particular, how can we both identify a complete or at least representative range of relevant cultural factors and somehow combine these factors in so as to map American

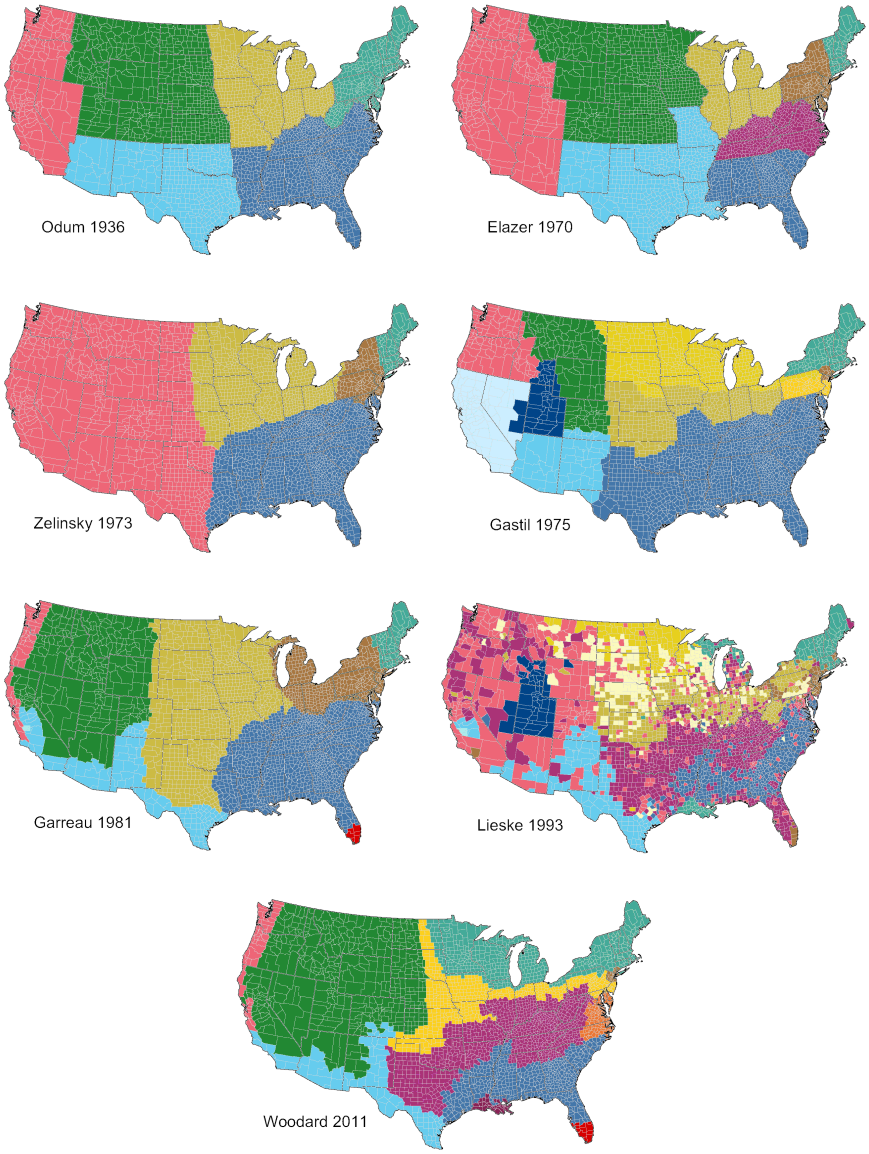


Figure 5.1: Maps showing the primary American cultural regions as identified in eight previous studies at the county level [63, 83, 85, 142, 171, 224, 227].

cultural regions? Defining such regions does not mean that they do not contain internal variation or that they are separated by hard borders — culture is dynamic and complex and humans are highly mobile — but that we can find areas where the cultural practice and values of the people who live within that region are more similar to each other than to those of people who live outside that region.

The goals of this work are therefore to address these issues, by (i) proposing a novel method for discovering cultural regions by identifying regional patterns in topics of conversation, and by then (ii) proposing a theory of American cultural regions derived from the application of this method to a large corpus of geolocated social media data.

5.2 OUR APPROACH

Our starting premise is that cultural regions will necessarily be reflected by regional variation in the topics that people choose to discuss in their everyday lives. If the cultural geography of the US was broadly homogenous, we would expect topics of conversation to be largely the same across the country, aside from different uses of place names and other such relatively superficial and necessarily regionalized vocabulary items. However, if people from different regions exhibit distinct and systematic cultural characteristics—for example, in politics, religion, music, sport, fashion—as research on American cultural geography has consistently shown, then these patterns of cultural variation will necessarily manifest themselves as patterns of topical variation in the language used by people from these regions [129]. For example, if hip hop music, baseball, tattoos, or some other cultural practice is especially popular in some part of the country, we would expect more discussion on that topic in large samples of everyday language use originating from that region, including on social media. Furthermore, previous works show that cultural factors often show related regional patterns, owing presumably to interrelationships between these different factors. For example, regional settlement patterns can help explain differences in ethnicity and religion which can have long term effects on voting patterns. Consequently, analysing these regional topical patterns in the aggregate can be used to infer broader cultural regions. Crucially, there is no need to predefine what these topical patterns are or how much they matter: the topics themselves and their relative importance can be inferred through the analysis of everyday language as well. We therefore introduce an automated method for identifying cultural regions based on the automated identification of patterns of regional variation in topics of discussion in very large corpora of geotagged everyday language use. Our method is especially intended to take advantage of the incredibly large amount of geotagged social media data that online communic-

ation now provides us with for the first time, although our method could be used to identify cultural regions within any area based on any substantial source of regionalized everyday language use.

Specifically, to map modern American cultural regions, we identify regional patterns in the topics that Americans tend to discuss on social media through a quantitative analysis of ten thousand lexical items in over 3.3 billion geotagged tweets from across the US, collected between 2015 and 2021. Large corpora of geotagged Twitter data have been used frequently in computational sociolinguistics [168], and also in particular to map patterns of dialect variation [4, 57, 62, 93, 95, 99–101, 115], while others leveraged methods such as Latent Dirichlet Allocation to identify regional topical patterns [77, 128]. Despite this wealth of research that has used large corpora of social media to identify regional patterns in language use, we are aware of no research that has used this type of information to infer the location of general cultural regions.

Of course, social media or any other form of language can only provide a partial picture of regional patterns in overall topics of discussion in a region. In general, big data corpora generated from microblogging platforms certainly present a number of biases that we cited in Section 2.1.4.5. However, if cultural regions are real and pervasive, then we should expect these regions to manifest themselves in any large sample of everyday language that encompass a large proportion of the population, even if the specific topics of interest vary across these different domains. Furthermore, right now, Twitter is the only variety of geotagged natural language data available in sufficient amounts to allow reliable automatic analyses, and is a very popular social media platform used regularly by millions of people from across the US, mostly in interactive contexts [14], serving as a perfect domain to apply our data-driven approach for automatically mapping cultural regions.

Our main finding is that the modern US can be divided into five primary cultural areas, each defined by its own topical patterns. We emphasize that this result stems from a quantitative analysis in contrast to previous proposals based on more or less informative (qualitative) approaches. Further, beyond the specific number of regions it is most relevant to note that our method yields the list of words and topics that define those regions, which highlights the differences in interests, habits and backgrounds that distinguishes each cultural region from the others. Crucially, by means of a dynamic analysis we show that cultural regions of the US are relatively stable over the past few years, offering further evidence that cultural areas are real phenomena that pervade American society.

The rest of the chapter is structured as follows. The results of the work are first introduced by a description of the dataset collection and pre-processing methodology. Regional variations of words usage

All code used for this work is hosted on GitHub [147].

observed from this dataset are then explored, before obtaining the principal dimensions of these variations. The main result of the work, the cultural regions of the US and the main topics of discussion that define them, is then presented in detail. The possibility of a variation with time of the results is then explored. Finally, a discussion of the insights brought by the analysis and also of where future works could build on it comes to conclude the chapter.

5.3 MEASURING REGIONAL VARIATION

We analyse 3.3 billion geotagged tweets from the contiguous US, posted from January 1st, 2015 to December 31st, 2021. Importantly, we discard users according to the criteria given in Section 2.1.4.4. In our dataset, we thus retain 17 million users. We clean the tweets' text and use the *Compact Language Detector* (CLD) to eliminate tweets written in a language other than English, following the process described in Section 2.1.4.2. We also remove tweets with a geotag that did not allow for reliable assignment to our unit areas, which are the US counties and county equivalents (3108 in total), according to the criterion given in Table 2.1. Certainly, counties vary in both size and population but most of them form a useful division sufficiently large to show a sizeable amount of tweets and sufficiently small to allow for a careful delimitation of cultural areas (states would be too big units whereas towns would be too small).

From the remaining tweets, we extract and count the tokens in their text, and assign them to counties. Counties that accumulate fewer than 50 000 tokens are not taken into account, leaving us with $N_c = 2576$ counties which define our sub-corpora. We thus keep 83 % of the total number of counties. After this filtering, the full dataset contains 9.1 billion tokens. We subsequently convert the remaining word forms to lowercase and aggregate the token counts on these forms. We then remove all function words (like *the, and*) and interjections (like *um, oh*), and consider the 10 000 most common remaining word forms. Note that this list of word forms emerges from the data, and is not imposed by any previous topical or dialect classification. All aggregated data generated by our Twitter data analyses as well as our list of excluded words are available for download from a figshare repository [146]

We then measure and map the relative frequencies $f_{c,w}$ for every word w in every county c. We illustrate our raw results by plotting in Figure 5.2 the relative frequency in each county of four representative words: (a) *today*, (c) *mountain*, (e) *traffic* and (g) *bruh* (cells that appear greyed out do not reach a minimum number of tweets as explained in the paragraph above). In the first case, *today* appears at relatively stable rates in most of the counties, as expected. Alternatively, *mountain* is a regionally-dependent word as clearly seen. The item *traffic* appears more frequently in urban areas. Finally, *bruh* is an African-American

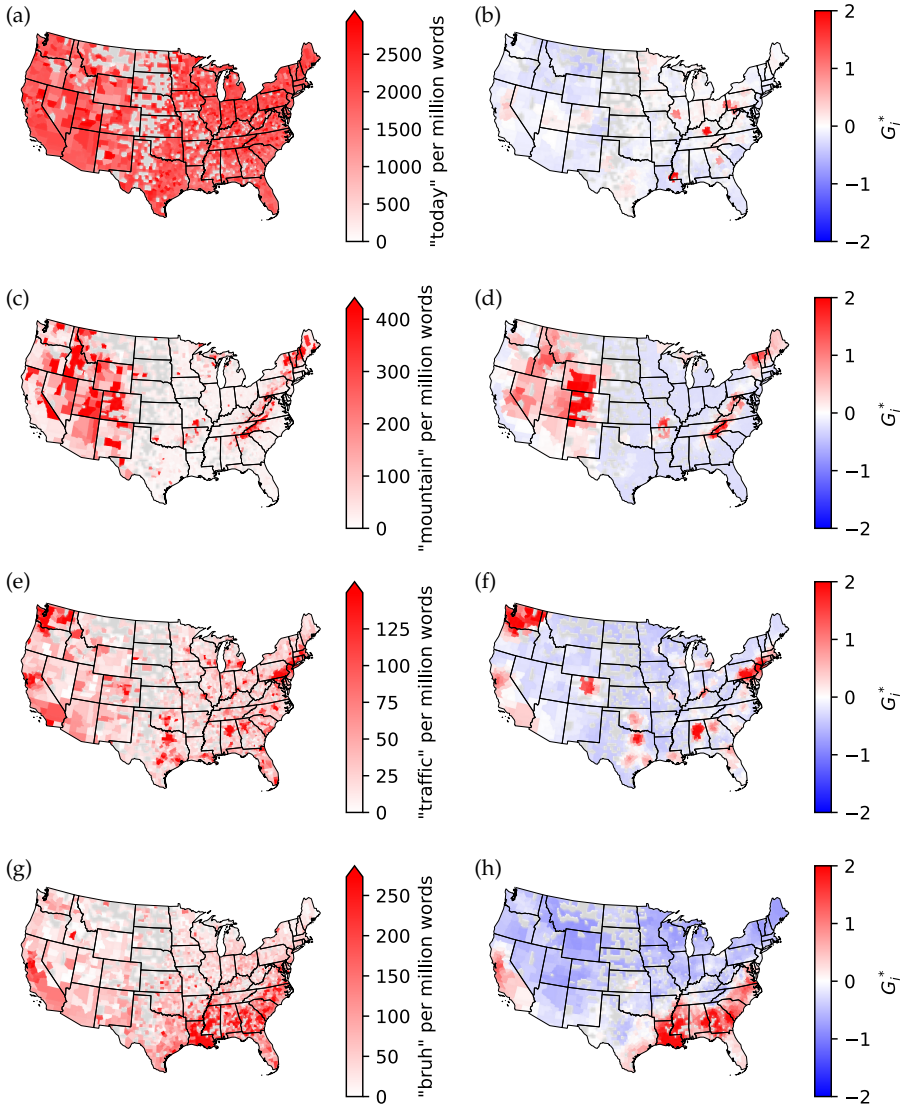


Figure 5.2: Maps showing the (a-c-e-g) relative frequency and (b-d-f-h) Getis-Ord G_i^* z-score for the words *today*, *mountain*, *traffic* and *bruh*, respectively. One can note how the latter metric enables to reveal word usage hotspots, smoothing out the raw noisy signal from the data.

English variant that appears to be especially common in southern counties, where there are large African-American populations.

A word of caution is now required. A relative frequency map alone is not able to fully reveal regional variations due to the wide range of different factors besides regional variation that affect word use and add noise to the signal. To extract the underlying regional signal from each word map, we conduct a multivariate spatial analysis [99, 100] of the relative frequencies of our 10 000 word forms. In order to identify geographical hotspots in the usage of each word (Figure 5.2), we compute Getis-Ord's z-scores (G_i^* [174]) for each county c and word w, which are defined as:

$$G^*_{c,w} = \frac{\sum_{c'} W_{c,c'}(f_{c',w} - \bar{f_w})}{\sigma_w \sqrt{\frac{N_c \sum_{c'} W^2_{c,c'} - \left(\sum_{c'} W_{c,c'}\right)^2}{N_c - 1}}}, \tag{5.1}$$

with $\bar{f_w}$ the average frequency of w over the whole dataset, σ_w the standard deviation in w's frequencies, and $W_{c,c'}$ are the elements of a proximity matrix, which we take as equal to 1 if $c' = c$ or c' belonging to c's 10 nearest neighbours, and equal to 0 otherwise.

The metric given by Equation (5.1) ultimately diminishes spurious data variation and smooths spatial patterns, allowing us to discern a regional pattern in a word's usage. In Figure 5.2(b), (d), (f) and (h) we show, respectively, the G_i^* z-scores for the previous words *today*, *mountain*, *traffic* and *bruh*. White, light blue or light red counties do not depart significantly from an average utilization, whereas a bright red or blue respectively mean that the word is relatively frequently or infrequently used in that region. Since *today* is a rather generic word, we do not find any strong regional pattern, whereas the others do. The usage hotspots of *mountain* display the main mountain ranges of the country. While the map for *traffic* is correlated with large urban areas (and can be interpreted as a topical word), the dialect word *bruh* seems to be significantly more used in counties pertaining to the Deep South. We see here that different attributes that define a culture (interests, behaviour, dialect) are captured within our scheme and, notably, are treated on equal footing.

5.4 OBTAINING THE PRINCIPAL DIMENSIONS OF REGIONAL VARIATION

The G_i^* distributions for all 10 000 top words by usage thus hold valuable information. However, a considerable part of this information can be analysed more efficiently, since some words may belong to the same semantic field (*mountain* and *peak*) or characterize the same particular dialect (*bruh* and *aight*). Furthermore, a few variations may simply be uninformative noise, intrinsic to real individuals' behaviours, but also potentially resulting from an imperfect filtering of Twitter data,

as aforementioned. The most important dimensions of regional lexical variation are then found by subjecting the hotspot maps for the complete set of words to a *principal component analysis* (PCA) [142, 222]. Another possible approach would have consisted in performing topic modelling, for instance by ways of a latent Dirichlet allocation on the word frequency matrix, to then infer a distribution of topics for every county. It is however more computationally intensive, and poses the questions of the selection of the number of topics, their interpretability and their internal coherence [13, 105]. In a case like ours where documents are so large (aggregating all tweets in a county), it is far from obvious to select a number of topics such that there is little overlap between them, and to know that these topics are actually representative of the dataset as a whole. This is much more clear when selecting components in PCA, as we show below.

From the $N_w = 10\,000$ dimensions of our dataset, we thus project to a *principal component* (PC) space of $N_{\mathrm{PC}} = 326$ dimensions. It turns out that these 326 components explain 92 % of the observed variance (see Figure 5.3(f)). We do not set this number of components arbitrarily, by choosing one directly or by setting a percentage of variance we wish to explain using these components. Instead, we use the broken-stick rule to fix the number of components [76, 118]. This heuristic compares the decrease of the variance explained by each successive component to the one expected from a random partition of the whole variance in N_w parts. Components, sorted by decreasing explained variance, are kept until they do not explain more variance than their corresponding random part would. With this method, we do not make any assumption about the amount of variance in our data that is simply due to random fluctuations.

We show the projected data along the first four PCs in Figure 5.3(a-d), which displays a neat visualization of the spatial patterns. The map for each dimension shows two opposing regions (red and blue) which can be linked to their characteristic words, the ones with the highest (positive, in red) and lowest (negative, in blue) loading. For an illustration, in Figure 5.3(e) we show in a word cloud the most characteristic words for each of the two regions in Figure 5.3(b), which corresponds to the second component.

5.5 INFERRING CULTURAL REGIONS

We are now in a position to generate a single overall taxonomy of American cultural regions by clustering together counties with similar lexical signature. To do so, we subject the previous PC maps to a hierarchical clustering, using the Euclidean distance and the Ward variance minimization algorithm [65]. This is how we define the cultural regions from our corpus, as depicted in Figure 5.4. From the dendrogram and the evolution of the average silhouette score for different

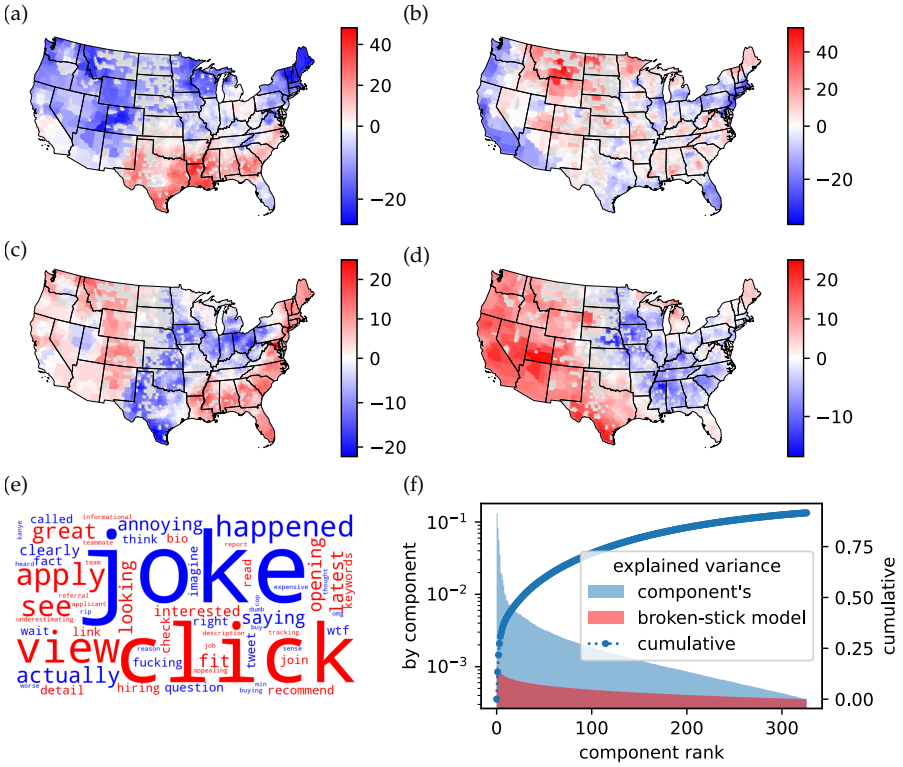


Figure 5.3: Result of the PCA carried out on our whole dataset. (a-d) Four maps show the projection of the data along the first four components, highlighting regional lexical variations. Note that the scale on the divergent colour scales are not symmetrical around zero in order to utilize the full range of colours of the colour map. (e) Word cloud showing the words with the strongest positive (red) and negative (blue) loadings for the second component, with each word's font size depending on its loading's absolute value. (f) Explained variance of the PCs compared to the broken-stick model on a logarithmic scale, which shows how the number of components to keep is selected at the first intersection of the two curves. The cumulative proportion of the variance explained by the components is also plotted, showing that our dimension reduction explains around 92% of the observed variance. The first four components shown in panels (a-d) capture alone 31% of the variance.

levels of clustering, we select a meaningful number of clusters n_{clusters} [187]. The hierarchical nature of the clustering is useful to see how regions are grouped together at different levels of clustering, indicating which regions are closer together. Importantly, applying hierarchical clustering to the principal dimensions of variation of the data obtained through PCA allows us to focus on the main regional patterns of variation. Applying the algorithm directly to the 10 000-word distance matrix would yield highly noisy results.

We plot the main divisions in Figure 5.4(a). This is the main result of the chapter. In the map, we present the division into five clusters

since it is one of the two best options as characterized by the Silhouette score analysis in Figure 5.4(b), and at a clear-cut on the dendrogram in Figure 5.4(d). The optimal choices correspond to the two significant drops in the score: the first (second) corresponds to a cluster number equal to 2 (5).

Indeed, the dendrogram in Figure 5.4(d) shows that the counties can be initially classified into two large-scale subgroups representing a North vs South divide. The North is then further fragmented into the clusters 2, 3 and 4 shown in Figure 5.4(a), whereas the South group splits into the clusters 1 and 5. For the most part, our map in Figure 5.4(a) is consistent with standard theories of American cultural regions, with all five of our regions finding analogues in existing systems. Yet, taken as a whole our clusters do not match any previous system and reveal non-contiguous culture regions such as the clusters 3 and 4. Moreover, in contrast to previous proposals our results have the advantage of being data-driven, based on variation in the topics people care to discuss as opposed to factors selected by hand by the researcher (and consequently subjected to many more, uncontrolled biases than our Twitter data).

Further, to be able to better interpret the obtained regions, it is insightful to know which words characterize each cluster the most. To infer them, we start by taking the centre of each cluster in words-G_i^*-space. Hence, for each cluster we take the average G_i^* score over its counties for all words. From these n_{clusters} vectors of N_w elements, we calculate the minimum absolute difference between each cluster centre's word's score and the ones of all other clusters, i.e., we take the distance to the closest cluster's centre along the word's dimension. More formally, we define the specificity $S_{C,w}$ of word w for cluster C as:

$$S_{C,w} = \min_{C' \in \mathcal{C} \setminus C} \left(\frac{1}{N_C} \sum_{c \in C} G^*_{c,w} - \frac{1}{N_{C'}} \sum_{c \in C'} G^*_{c,w} \right)^2 , \tag{5.2}$$

where $\mathcal{C}$ denotes the set of clusters, N_C the number of counties belonging to cluster C, and $G^*_{c,w}$ the G_i^* score of word w in county c. For each cluster C, we thus define the most characteristic words as the ones with highest $S_{C,w}$ values. In the case of the division in five clusters, the top 5 most characteristic words per cluster are shown in Figure 5.4(c), according to the specificity metric defined in Equation (5.2). In all cases, the five cultural regions are linked to clear and distinct topical patterns (see the Supplementary Information for a more exhaustive list). We stress that these characteristic words are automatically identified based on the quantitative analysis presented above. Notably, for each cluster we see three basic type of lexical patterns.

First, we see words associated directly with those locations, most commonly the names of cities, states, and sports teams. This is basic evidence that the method works: we would expect these words to be

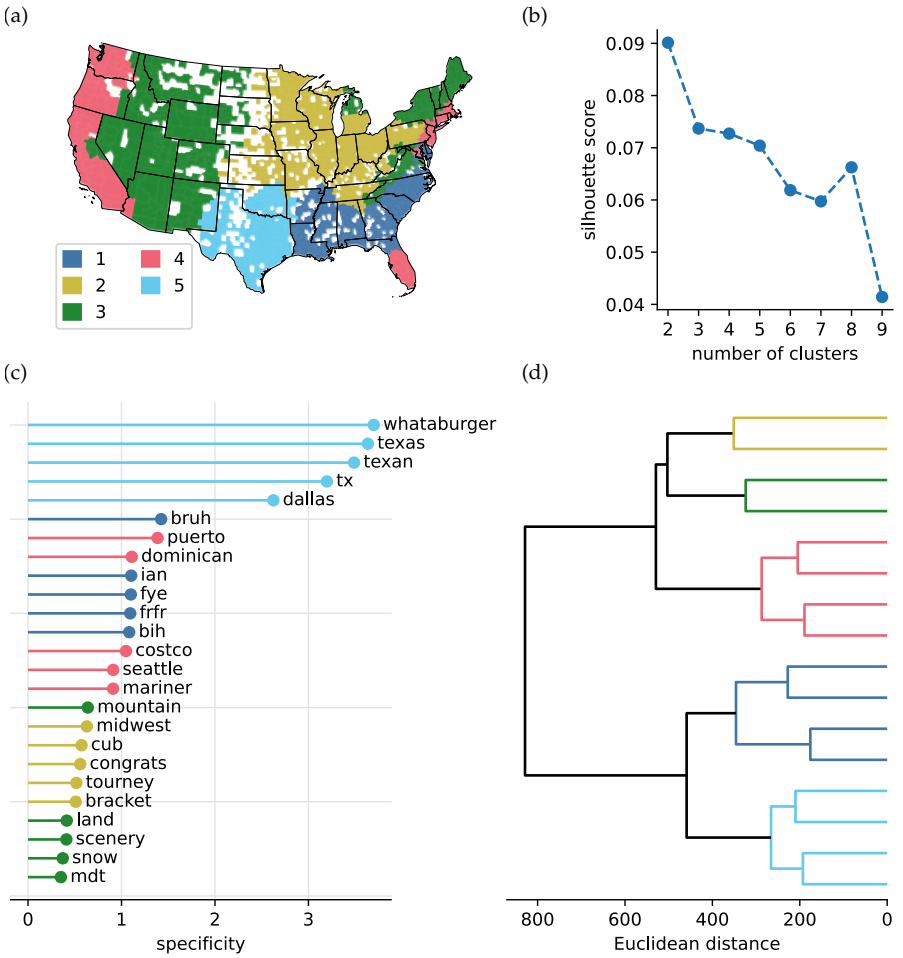


Figure 5.4: Cultural regions obtained from our whole dataset. (a) Map of the five clusters obtained through hierarchical clustering, selected from a high value of (b) the mean Silhouette score. A significant drop of the Silhouette score after 5 levels indicates that further splitting counties in more region does not yield coherent cultural regions. (c) Five most specific words for the five clusters shown in (a), along with their specificity values. (d) The dendrogram allows seeing which clusters are first joined if going to a higher level of clustering, and thus which ones are closer together. It clearly shows that the strongest division is the one between the North and the Southeast (excluding Florida), further splitting as the cluster distance increases.

associated with the cultural regions that contain them. However, these results also reflect how often people from different cities and states refer to each other. For example, the fifth cluster which is centred on Texas also includes Oklahoma, which contributes various place names to the list of words most strongly associated with this region. This means not only that people in Texas and Oklahoma talk more about place names in their own states, as would be expected, but that they talk more about place names in each other's states. This is one type of

regional topical patterns that our approach draws to identify cultural regions.

Second, we observe words connected with non-regional topics, which nonetheless show regional differences. In this case, our approach can be seen as discovering topical patterns and by extension cultural patterns that distinguish between different regions of the US. For example, cluster 2 is strongly associated with the discussion of a range of American sports, as well as the names of the states that fall within this region. Although we would expect that a region centred around the Midwest would be associated with the names for Midwestern states, their preoccupation with the discussion of sports on Twitter is not so easy to predict.

Third, we find words that are dialect items, i.e., alternative ways of referring to a given concept. This type pattern is especially apparent for cluster 1, which aligns closely with the region of African American population density and is therefore associated with numerous lexical items from African-American English (e.g. *bruh, lawd, turnt*). Although dialectologists do not usually focus on the frequencies of individual words, this results is to be expected: dialect regions, which can be seen as a type of cultural region, have been found to generally align with broader cultural regions [99].

We can now examine each of the five cultural regions we have identified in turn and consider what the words that are mostly strongly associated with each tell us about the culture of that region, as well as the factors that drive cultural variation in the US more generally.

The first cluster [blue in Figure 5.4(a)], which identifies a southeastern region, largely reflects African American culture, as can be predicted based on the close correlation between our map and the distribution of counties with relatively large African American populations. Most notably, tweets from the South are more likely to contain words related to African American culture, including, for example, cuisine (e.g. *grits, cookout*), fashion (e.g. *braids, dreads*), and music (e.g. *rappers, rapping*). As noted above, this cluster is also strongly characterized by many vocabulary items associated with African American English, especially for referring to people (e.g. *bruh, dawg*), as well as many acronyms (e.g. *frfr, stg*). Place names associated most strongly with this cluster primarily include southern states (e.g. *Georgia, Carolina*), despite the fact that, in general, references to place names are relatively rare compared to other clusters.

The second cluster [yellow in Figure 5.4(a)] has its core in the Midwest and is clearly characterized by more frequent references to sports. American team sports especially stand out, with 40 words of the top 50 most strongly associated with this cluster being directly linked to this topic. In particular, these are words associated with basketball (e.g. *basketball, rebound*) and baseball (e.g. *baseball, innings*), although football, wrestling, and cheering are also referenced, as well

as various more generic sporting terms (e.g. *teams, tourney*). Similarly, many place names are associated with local sports teams (e.g. *Cubs, Chiefs*), although various state names are also strongly associated with this cluster (e.g. *Ohio, Illinois*), as well as the word *Midwest* itself. A smaller number of lexical items are also associated with school (e.g. *locker, choir*). Overall, this cluster therefore shows that sports is a central part of this region.

The third cluster [green in Figure 5.4(a)] can be identified with a discontinuous region that mostly aligns with rural areas of the US, as well as areas that focus on outdoor activities, especially in mountainous regions (e.g., the Rocky or Appalachian Mountains). This cluster is relatively hard to interpret topically, in part because, unlike the other regions, it is characterized by the relative infrequent use of a number of words. In terms of words that are relatively common in this region, the clearest pattern is a relatively large number of words associated with nature (e.g. *mountains, tree*), weather (e.g. *snow, seasonal*), and outdoor activities (e.g. *adventures, trail*). Clearly, people in this region tend to focus more of their natural surroundings. In addition, there are a number of words related to work (e.g. *hiring, jobs*), as well as a numerous place names (e.g. *Colorado, Montana*) that are strongly associated with this region. In terms of words that are uncommon within the cluster, there exist many verbs, especially verbs associated with human actions like communication (e.g. *said, told*), thought (e.g. *understand, confused*), and physical actions (e.g. *put, hit*), which implies overall less focus on the individual. This region is also associated with relatively infrequent use of a wide range of negative words (e.g. *wrong, bad*), which largely hints at a more positive outlook.

The fourth cluster [red in Figure 5.4(a)] also identifies a discontinuous region that primarily encompasses large urban areas in the coasts (Northeast and West). Unsurprisingly, this region is characterized by a wide range of words associated with more urban life (e.g. *homeless, traffic*), especially terms related to different nationalities and immigration (e.g. *Latino, Asian*). We also find a relatively large number of place names (e.g. *California, NYC*). Strikingly, this cluster is associated with a very large number of words with negative connotations, including relating to violence (e.g. *violence, attack*), danger (e.g. *dangerous, crime*), cursing (e.g. *asshole, fucking*), political unrest (e.g. *protests, indicted*), racism (e.g. *Nazi, supremacist*), and general negative adjectives (e.g. *disgusting, abusive*). Quite generally, people from this cluster are more likely to discuss negative topics than other parts of the US, at least on social media. Taken together, the third and fourth clusters suggest an opposition in the culture of more rural and urban areas in the US, which appear to engage in more positive and negative discourse respectively [216].

Finally, the fifth cluster [cyan in Figure 5.4(a)], which is centred around South Central States, especially Texas and Oklahoma, is char-

acterized by frequent reference to place names, relative to the other clusters, especially in these two states, as has already been noted. For example, the first five most strongly associated words are *Whataburger* (a fast food chain from Texas), followed by *Texas, TX, Texan, and Dallas*. This not only shows that people in this region tend to discuss place more on Twitter, but implies that this cultural region is characterized by a relatively high amount of local pride. Correspondingly, this region is also associated with a relatively large number of dialect terms, both of Anglo (e.g. *yalls, fixing*) and Hispanic (e.g. *queso, taco*) origins, reflecting the diverse makeup of this region.

Given the lack of consensus in previous research, our results can help resolve long-standing debates relating to the distribution of American cultural regions. We find that the division between the Southeast and the rest of the US is the strongest. This result attests to the importance of the cultural divide between White and Black America and between the North and the South. Although all previous major theories of American cultural regions has identified a distinction between the North and the South, our southern region is especially similar to relatively recent theories, which identify a southern region that closely aligns with the part of the south with an especially high proportion of African Americans [142, 224]. Another key finding that emerged from our analysis is a broad opposition between coastal and internal areas, which has not previously been identified as important sources of distinction of American cultural regions [63, 72, 83, 85, 142, 171, 224, 227] but reflects a modern political trend of undeniable significance [89] that is currently reconfiguring the nation. The discontinuous nature of these regions, which is not required by our definition of a cultural region, is also notable. It demonstrates how patterns in American culture can be distributed across very wide areas, reflecting complex patterns in physical and human geography, and the underlying complexity and dynamic nature of American society. This result is broadly in line with other recent theories of American cultural regions which have also identified discontinuous cultural regions [142, 224].

Our analysis is further useful for understanding the relationship between these regions. It divides the South into two regions, splitting Texas off from the rest of the Southeast, and splits the Midwest off from the rest of the North, divided into discontinuous countryside/coastal regions, rather than contiguous cultural regions. However, on the question of the number of primary American cultural regions, we can only safely say that with our data and methodology, at least 5 distinct regions can be discerned. We do not see it here, but we still cannot discard recent theories that claim that America is fundamentally far more culturally fragmented [83, 142, 224].

5.6 TEMPORAL ASPECT OF THE RESULTS

Given the success of our analysis, it would be interesting to see how the cultural regions found in Figure 5.4 change with time, as has been done in other research analysing diachronic corpora [7, 22, 26, 125, 160]. Although we would not expect significant changes due to the short timescale imposed by our Twitter dataset, we can still carry out a diachronic study to validate the very existence and meaningfulness of the cultural regions. To do so, we split our corpus into three datasets corresponding to different year ranges: 2015-2016, 2017-2018 and 2019-2021. These periods have a similar amount of tokens and can be then subjected to comparison. We show their maps in Figure 5.5(a-c). We obtain similar patterns, despite the variety of topics and forms employed on Twitter along the years and the heuristic nature of the clustering method that introduces a small amount of noise in the results. The North-South division is stable over time with small variations that can be due to either fluctuations or incipient structural changes. The latter cannot be conclusive due to the short time period considered in this work.

Next, we take the hierarchical clustering in Figure 5.4(d) and select the county-to-cluster assignment corresponding to the highest level of the hierarchy. This is represented by the two-way division between North and South. For each year in our dataset, we then measure the pairwise distances between counties belonging to both clusters. The distances are calculated as Euclidean distances between rows of the matrix $G^*_{c,w}$ (see Equation (5.1)). We thus obtain the evolution with time of the inter-cluster distances distribution as shown in Figure 5.5(d). The box plots demonstrate that (i) the median distance is roughly constant over the years, and (ii) the distance distribution shows little variation. Both findings suggest that the detected cultural regions are no artefact of the method, but a genuine data structure that exists within our corpus.

5.7 DISCUSSION

Overall, our analysis has identified regional patterns of lexical variation of clear cultural importance. Furthermore, the themes associated with each of these patterns provide a new perspective on American cultural geography. For example, although our analysis has confirmed that factors such as ethnicity and religion are important for defining American cultural regions, we found substantial variation in the relevance of these factors across the US. Our analysis has also identified other subtler cultural patterns — such as a focus on social interaction, the outdoors, family, and leisure — which have been overlooked in previous research, in part because they cannot be easily studied through the analysis of traditional sources of secondary data. Our

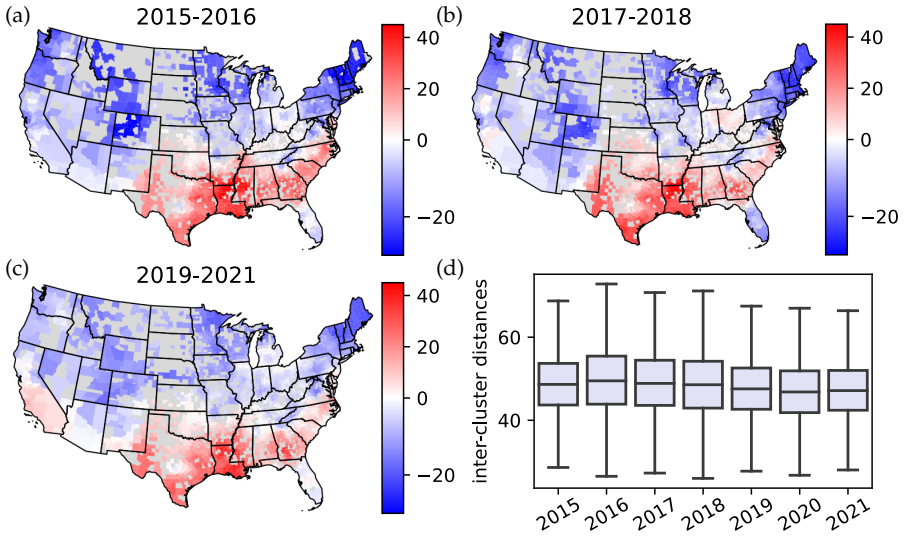


Figure 5.5: Effect of temporal segmentation for the data on the obtained divisions. (a-c) Maps of the data projected along the first PC obtained for the years 2015-2016, 2017-2018 and 2019-2021, respectively. Apart from slight variations in California and Florida, the first component translates the same division between the Southeast and the rest of the US. Note that the scale on the divergent colour scales are not symmetrical around zero in order to utilize the full range of colours of the colour map. (d) Evolution of the distributions of inter-cluster distances along the years spanned by our dataset. The box plots show the median, first and third quartile, and the boundaries of the whiskers are within the 1.5 interquartile range value. We use the cluster assignment obtained with the whole dataset and measure the Euclidean distance in G_i^* space between counties belonging to different counties. The distribution is thus shown to vary little from year to year, which demonstrates the stability of the two-way division we found.

method has therefore not only allowed us to map cultural regions, but it has also allowed us to identify cultural factors that are important for defining these regions, at least in this communicative context, providing a foundation for a more complete picture of the American cultural landscape.

Clearly, this study has only analysed one genre of American English. The specific topical patterns on Twitter would not be exactly replicated in other genres, especially given the communicative purpose and user base associated with microblogging platforms. Nevertheless, assuming that American cultural regions are important and pervasive forces, similar regional patterns should be reflected across all popular genres. This issue could be further clarified when more richly annotated natural language data becomes available in a near future. Our methods, however, will remain valid for any such dataset. Crucially, we expect that our main idea of inferring cultural regions and the topics defining

them from people's speech will be applicable to any big data resource with linguistic value.

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*, $8\frac{1}{2}$ [67]

6.1 SUMMARY OF OUR FINDINGS

6.2 WHAT WE LEARNED ALONG THE WAY

6.2.1 *Crossing disciplinary boundaries*

Given this approach that I took throughout these years of PhD, it is only natural that I ended up crossing disciplinary boundaries. In practice, this is reflected first in the different backgrounds of the researchers I had the chance to collaborate with. For instance, I had the opportunity to work with Jack Grieve, whose background is in linguistics. While I was motly focused on the quantitative aspect of the work, he brought a level of interpretation of our results that I did not see at first (I am here referring to the results of Chapter 5). Further, this made me aware of a whole new range of literature and research areas I had never heard of. This is mostly because this literature was from people outside of the field of complexity science, and consequently published in different journals.

Second, and as just said above, this is partly due to the first point: this cross-disciplinarity is apparent from the diversity of literature cited throughout this thesis. I can illustrate how twisted navigating this diversity can be, but how rewarding it is with the example of how I stumbled upon a metric such as the *earth mover's distance* (EMD), that I use in Chapter 3. First finding out the shortcomings of spatial entropies, widely accepted in the field of spatial complex systems [20] It has mainly been used within the field of computer vision [188], and it was shown to be a proper distance (in the metric sense) between probability distributions [140]. Nothing but a particular version of the Wasserstein or Mallows distance, well known in transport theory. Only to find out a year after writing this article that the geographer John F. Jakubs had defined a metric equivalent to the *earth mover's ratio* (EMR) in 1981 [119], much before computer vision had the occasion to exploit it. Natural that different fields end up using different terms. In this

case, mathematicians have defined a general distance, which is used in other applied fields within a particular case. So the mathematicians need this more general name. And would anyone really think that a geographer or engineer in computer vision should give up the very evocative EMD in favour of "the Wasserstein 1-distance on discrete empirical distributions"? So what can be done to bridge these gaps? Mostly curiosity that drove me to find all of this out. One should not be afraid to search for literature outside of their field, and to simply acknowledge them. Citing a work outside of your field is much more valuable than citing another everyone within your field, who are probably going to be most of your readers, already know.

All in all, the people and the research you interact with are highly interdependent. And as I have shown, more diversity in those can be the foundation for more throrough and further-reaching research. While desirable, increasing this diversity in one's research work is far from straightforward. Beyond the point made above that looking for and citing other fields' literature can help, or the obvious that one should try to go to different conferences, make project proposals with an actually interdisciplinary team, I would like to highlight the lack of interdisciplinary reviews. Review articles are the first literature one seeks to enter into a field. Yet, the notion of field itself limits the scope of reviews such as [41], which is a very good review of physics-inspired models of social phenomena, but exactly as such, it targets almost exclusively researchers with a background in physics. While there is no doubt this kind of review adds great value, another kind that would focus rather on an issue that is tackled by multiple fields, and that would review their different approaches and confront or relate them, would also be beneficial. For instance, instead of answering the question "what physics-inspired models have been used to study social dynamics?", they would answer one such as "what is the state of research on language competition and multilingualism, from the (mostly) qualitative models of linguists to the quantitative ones of physicists, also considering empirical works?". This would provide scientists from different fields insights into what others have to say about a question, and who they might collaborate with to further everyone's understanding. The more recent [27] is the closest to being this kind of interdisciplinary reviews, although more is necessary as it does not confront the conclusions of the research coming from different fields.

6.2.2 *Science and trends*

The number of researchers interested in language competition models like the one we presented in Chapter 3 is very low. The hype has passed, it was 20 years ago when Abrams and Strogatz published a

piece in *Nature*. Yet the problem has not been solved at all, as we have shown ourselves.

6.3 UNCHARTED DIRECTIONS WORTH EXPLORING

While I believe the results summarised above provide valuable insights to sociolinguistic matters, they leave many questions unanswered, and maybe even raise more questions than answers. In the following I will list some potential directions for future research that could build on top of the work we have presented here. First are the directions that naturally emerge from the following question: what else could the approaches or methods we have developed be applied to?

For instance, can the model of language shift we presented in Chapter 3 be equally applied to diachronic data? A dataset with historical data spanning centuries was for example used recently in [197]. One could try to see if our model can be fit to these data, and, if not, what ingredients may be missing. A comparison with the results obtained in [197] with a different model can also bring valuable insights to the field.

The central idea behind the inference of American cultural regions we presented in Chapter 5 can also lead to many further developments. One could look at countries that share a language and uncover their cultural differences. Interesting questions can then be investigated. Can some areas of a country be more similar to another country's than their own? It would also be interesting to see how the colonial past of some countries is reflected in those. Did more conflictual decolonisation processes lead to starker cultural differences? Or, does a more recent split imply more similarity?

The second kind of directions are not based on methodology, but rather on gaps in understanding that remain open.

An idea about modelling language competition that is worth looking into and that we have not mentioned yet is to consider language as a property not only of the individuals but also of their interactions, as suggested in [40]. Indeed, as shown in this article, a language might be preserved within a tight-knit community that uses it internally, while being able to switch to another to communicate with others. This applies to language shift models of Chapter 3, but also to models of usage of a language's varieties, as the one of Chapter 4. What is missing here is more empirical work on the question. This would imply a considerable effort, as obtaining a network of interactions along with the language in which they take place from real-world data is very challenging. But this direction shows great promise, so these efforts to put together such a dataset may well be worthwhile.

Third are simply interesting questions that arose while I worked on this thesis, as, while investigating matters of sociolinguistics, we

stumbled upon interesting issues in social science that are not confined to the realm of linguistics.

One is segregation. This theme permeates Chapter 3, Chapter 4, and also Chapter 5 to a lesser extent. But these are segregations of different kinds: linguistic, cultural and socio-economic. A question that remains very much open is how these types of segregation are interrelated, also taking into account the segregation in social contacts, also known as assortativity or homophily.

Both models presented in Chapter 3 and Chapter 4 integrate effects of social pressure. They simply assume that the most people around you speaks a language or a variety, the more likely you are to adopt it too.

Another is the impact of cultural differences on social behaviours in general. What features matter the most when building social groups? How does this change from one area of the world to another?

There are thus many more directions worth exploring, and, for now, to most questions the most sensible answer I can provide is: I don't know. But while one could consider knowledge as comforting some faith in the universe, I believe uncertainty to be much more entertaining. Let us then keep asking questions to keep things that way. At last, it may seem that the quest for understanding paradoxically leads to less understanding, but not to worry, it's the quest itself that matters, not its destination, as the poet would say.

Part IV

APPENDIX

A

CHARACTERISATION OF MULTILINGUAL REGIONS

Table A.1: Summary metrics for 16 regions of interest. For each group of monolinguals and multilinguals, we give the number of residents of that region found to belong to this group based on their tweets, and the group' EMR (see Equation (3.8)).

Region	L group	Count	EMR
Balearic islands	Catalan	685	0.514
	Spanish	9,892	0.123
	Catalan-Spanish	3,154	0.425
Basque country	Spanish	17,451	0.063
	Basque	674	0.445
	Spanish-Basque	3,995	0.142
Belgium	French	13,556	0.476
	Dutch	26,644	0.296
	French-Dutch	1,014	0.099
Catalonia	Catalan	17,760	0.337
	Spanish	45,844	0.132
	Catalan-Spanish	38,084	0.065
Cyprus	Greek	531	0.448
	Turkish	3,693	0.253
	Greek-Turkish	3	0.607
Estonia	Estonian	1,931	0.226
	Russian	679	0.407
	Estonian-Russian	55	0.506
Finland	Finnish	15,320	0.042
	Swedish	207	0.310
	Finnish-Swedish	262	0.373
Galicia	Spanish	17,820	0.045
	Galician	1,109	0.151
	Spanish-Galician	11,921	0.049
Latvia	Latvian	12,994	0.077
	Russian	2,071	0.318
	Latvian-Russian	437	0.303

Java	Indonesian	726,246	0.059
	Javanese	1,461	0.686
	Sundanese	1,333	0.412
	Indonesian-Javanese	66,597	0.833
	Indonesian-Sundanese	38,424	0.480
	Javanese-Sundanese	93	0.702
	Indonesian-Javanese-Sundanese	6,067	0.515
Luxembourg	German	178	0.411
	French	1,262	0.183
	Luxembourgish	56	0.276
	German-French	31	0.214
	German-Luxembourgish	41	0.466
	French-Luxembourgish	44	0.163
	German-French-Luxembourgish	41	0.351
Malaysia	English	37,919	0.257
	Malay	41,302	0.159
	Chinese	1,225	0.248
	English-Malay	259,853	0.019
	English-Chinese	5,786	0.203
	Malay-Chinese	77	0.357
	English-Malay-Chinese	1,165	0.198
Paraguay	Spanish	41,768	0.040
	Guarani	110	0.558
	Spanish-Guarani	3,866	0.116
Quebec	English	6,877	0.278
	French	4,314	0.304
	English-French	5,657	0.113
Switzerland	German	8,697	0.534
	French	7,637	0.717
	Italian	1,628	0.792
	German-French	287	0.253
	German-Italian	131	0.578
	French-Italian	142	0.577
	German-French-Italian	28	0.481
Valencian	Catalan	1,061	0.454
Community	Spanish	50,908	0.048
	Catalan-Spanish	7,504	0.342

B

APPROXIMATE EVOLUTION EQUATIONS FOR OUR LANGUAGE COMPETITION MODEL IN A METAPOPULATION FRAMEWORK

Here we wish to derive a system of equations describing the evolution of the metapopulation under reasonable assumptions. Let us first rewrite Equation (3.12), the global equations of our model, in terms of counts instead of proportions:

$$\begin{aligned}\frac{\mathrm{d}N_{A,i}}{\mathrm{d}t} &= \mu s N_{AB,i}\frac{\sum_k (N_{A,k}+qN_{AB,k})}{\sum_k N_k} \\ &\quad - c(1-\mu)(1-s)N_{A,i}\frac{\sum_k (N_{B,k}+(1-q)N_{AB,k})}{\sum_k N_k}, \\ \frac{\mathrm{d}N_{B,i}}{\mathrm{d}t} &= \mu(1-s)N_{AB,i}\frac{\sum_k (N_{B,k}+(1-q)N_{AB,k})}{\sum_k N_k} \\ &\quad - c(1-\mu)sN_{B,i}\frac{\sum_k (N_{A,k}+qN_{AB,k})}{\sum_k N_k},\end{aligned} \tag{B.1}$$

for every cell i. Let us translate these to a metapopulation level, for which the equations hold for the subpopulations of each cell i. We will follow what was done in [192], and divide every population $N_{L,i}$ according to their work destination j. We thus introduce the notation $N_{L,ij}(t)$ which is the number of L-speakers who are residents in i and are at j for work at time t. It is such that

$$N_{L,i}(t) = \sum_j N_{L,ij}(t). \tag{B.2}$$

Then, the equations we want to solve to get the equilibrium points are, for every i,

$$\begin{aligned}\frac{\mathrm{d}N_{A,i}}{\mathrm{d}t} &\equiv \sum_j \frac{\mathrm{d}N_{A,ij}}{\mathrm{d}t} = 0 \\ \frac{\mathrm{d}N_{B,i}}{\mathrm{d}t} &\equiv \sum_j \frac{\mathrm{d}N_{B,ij}}{\mathrm{d}t} = 0.\end{aligned} \tag{B.3}$$

Regarding commuting, we will here use the notations from [17], and introduce first σ_{ij}, the commuting rate between the subpopulation i and every other cell j. The return rate of commuting individuals, that is the inverse of the timescale of their stay at work, is denoted τ.

The subpopulation size evolution (summing over all languages) due to commuting is then given by

$$\begin{aligned}\frac{\mathrm{d}N_{ii}}{\mathrm{d}t} &= \tau \sum_j N_{ij}(t) - \sum_j \sigma_{ij} N_{ii}(t) \\ \frac{\mathrm{d}N_{ij}}{\mathrm{d}t} &= \sigma_{ij} N_{ii}(t) - \tau N_{ij}(t).\end{aligned} \tag{B.4}$$

Then, for monolinguals A, we can write the following:

$$\begin{aligned}\frac{\mathrm{d}N_{A,ii}}{\mathrm{d}t} &= \text{A from every destination } j \text{ returning to their residence } i \\ &\quad - \text{A from } i \text{ leaving } i \text{ for work} \\ &\quad + \text{AB from } i \text{ and currently at } i \text{ turning A} \\ &\quad - \text{A from } i \text{ and currently at } i \text{ turning AB},\end{aligned} \tag{B.5}$$

which gives, using both the commuting part from Equation (B.4) and the language competition part from Equation (B.1),

$$\begin{aligned}\frac{\mathrm{d}N_{A,ii}}{\mathrm{d}t} &= \tau \sum_j N_{A,ij} \\ &\quad - \sum_j \sigma_{ij} N_{A,ii} \\ &\quad + \mu s (N_{ii} - N_{A,ii} - N_{B,ii}) \left(\frac{\sum_k (N_{A,ki} + q N_{AB,ki})}{\sum_k N_{ki}} \right) \\ &\quad - c(1-\mu)(1-s) N_{A,ii} \left(\frac{\sum_k (N_{B,ki} + (1-q) N_{AB,ki})}{\sum_k N_{ki}} \right).\end{aligned} \tag{B.6}$$

and similarly for every $j \neq i$:

$$\begin{aligned}\frac{\mathrm{d}N_{A,ij}}{\mathrm{d}t} &= \text{A arriving at } j \text{ coming from their residence } i \\ &\quad - \text{A currently at } j \text{ returning to their residence } i \\ &\quad + \text{AB from } i \text{ and currently at } j \text{ turning A} \\ &\quad - \text{A from } i \text{ and currently at } j \text{ turning AB},\end{aligned} \tag{B.7}$$

which gives

$$\begin{aligned}\frac{\mathrm{d}N_{A,ij}}{\mathrm{d}t} &= \sigma_{ij} N_{A,ii} \\ &\quad - \tau N_{A,ij} \\ &\quad + \mu s (N_{ij} - N_{A,ij} - N_{B,ij}) \left(\frac{\sum_k (N_{A,kj} + q N_{AB,kj})}{\sum_k N_{kj}} \right) \\ &\quad - c(1-\mu)(1-s) N_{A,ij} \left(\frac{\sum_k (N_{B,kj} + (1-q) N_{AB,kj})}{\sum_k N_{kj}} \right).\end{aligned} \tag{B.8}$$

Now when we sum Equation (B.8) over j and add Equation (B.6) to try to solve for the system Equation (B.3), we first see, as in [192], that

the commuting terms simplify. For the language competition terms, there remains to estimate the $N_{A,ij}$. We will use another result from [17], where they show that under the assumption that $\forall\ i, \tau \gg \sigma_i$, we can make the following approximation:

$$\begin{aligned} N_{ii} &= \frac{N_i}{1+\sigma_i/\tau} \\ N_{ij} &= \frac{N_i \sigma_{ij}/\tau}{1+\sigma_i/\tau}. \end{aligned} \tag{B.9}$$

Let us introduce a matrix $\underline{\underline{\nu}}$ such that

$$\begin{aligned} \forall\ i,\ \nu_{ii} &= \frac{1}{1+\sigma_i/\tau}\ , \\ \forall\ i,j \text{ such that } i \neq j,\ \nu_{ij} &= \frac{\sigma_{ij}/\tau}{1+\sigma_i/\tau}\ , \end{aligned} \tag{B.10}$$

so we can rewrite

$$\forall\ i,j,\ N_{ij} = N_i \nu_{ij}. \tag{B.11}$$

These counts, summed over all languages, are then constant. We can also use this approximation for each language, by identification in the equation below:

$$N_{A,ij}(t) + N_{B,ij}(t) + N_{AB,ij}(t) = (N_{A,i}(t) + N_{B,i}(t) + N_{AB,i}(t))\nu_{ij}. \tag{B.12}$$

We thus obtain the following equivalent equations under our assumptions:

$$\begin{aligned} \frac{\mathrm{d}N_{A,i}}{\mathrm{d}t} &= \mu s(N_i - N_{A,i} - N_{B,i}) \sum_j \nu_{ij}[q(1-\gamma_{B,j}) + (1-q)\gamma_{A,j}] \\ &\quad - c(1-\mu)(1-s)N_{A,i} \sum_j \nu_{ij}[(1-q)(1-\gamma_{A,j}) + q\gamma_{B,j}] \\ \frac{\mathrm{d}N_{B,i}}{\mathrm{d}t} &= \mu(1-s)(N_i - N_{A,i} - N_{B,i}) \sum_j \nu_{ij}[(1-q)(1-\gamma_{A,j}) + q\gamma_{B,j}] \\ &\quad - c(1-\mu)sN_{B,i} \sum_j \nu_{ij}[q(1-\gamma_{B,j}) + (1-q)\gamma_{A,j}], \end{aligned} \tag{B.13}$$

where

$$\begin{aligned} \gamma_{A,j} &= \frac{\sum_k N_{A,k}\nu_{kj}}{\sum_k N_k \nu_{kj}}, \\ \gamma_{B,j} &= \frac{\sum_k N_{B,k}\nu_{kj}}{\sum_k N_k \nu_{kj}}. \end{aligned} \tag{B.14}$$

BIBLIOGRAPHY

[4] Jacob Levy Abitbol et al. 'Socioeconomic Dependencies of Linguistic Patterns in Twitter: A Multivariate Analysis'. In: *Proceedings of the 2018 World Wide Web Conference*. WWW '18. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2018, pp. 1125–1134. ISBN: 978-1-4503-5639-8. DOI: 10.1145/3178876.3186011.

[5] 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.

[6] 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.

[7] 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.

[8] Hamza Alshenqeeti. 'Interviewing as a Data Collection Method: A Critical Review'. In: *English Linguistics Research* 3.1 (29th Mar. 2014), p39. ISSN: 1927-6036, 1927-6028. DOI: 10.5430/elr.v3n1p39.

[9] Anil Ananthaswamy. 'In AI, Is Bigger Always Better?' In: *Nature* 615.7951 (7951 8th Mar. 2023), pp. 202–205. DOI: 10.1038/d41586-023-00641-w.

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

[11] Andrea Apolloni et al. 'Metapopulation Epidemic Models with Heterogeneous Mixing and Travel Behaviour'. In: *Theoretical Biology and Medical Modelling* 11.1 (13th Jan. 2014), p. 3. ISSN: 1742-4682. DOI: 10.1186/1742-4682-11-3.

[12] 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.

[13] R. Arun et al. 'On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations'. In: *Proceedings of the 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - Volume Part I*. PAKDD'10. Berlin, Heidelberg: Springer-Verlag, 21st June 2010, pp. 391–402. ISBN: 978-3-642-13656-6. DOI: 10.1007/978-3-642-13657-3_43.

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

[15] Robert Axelrod. 'The Dissemination of Culture: A Model with Local Convergence and Global Polarization'. In: *The Journal of Conflict Resolution* 41.2 (1997), pp. 203–226. ISSN: 0022-0027. DOI: 10.1177/0022002797041002001.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Michael Batty et al. 'Entropy, Complexity, and Spatial Information'. In: *Journal of Geographical Systems* 16.4 (24th Oct. 2014), pp. 363–385. ISSN: 14355949. DOI: 10.1007/s10109-014-0202-2.

[21] 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.

[22] R. Alexander Bentley et al. 'Books Average Previous Decade of Economic Misery'. In: *PLOS ONE* 9.1 (8th Jan. 2014), e83147. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0083147.

[23] 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.

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

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

[26] Vladimir V. Bochkarev, Anna V. Shevlyakova and Valery D. Solovyev. 'The Average Word Length Dynamics as an Indicator of Cultural Changes in Society'. In: *Social Evolution & History* 14.2 (Sept. 2015), pp. 153–175.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] Jan Otto Marius Broek, John Winter Webb and Mei-Ling Hsu. *A Geography of Mankind*. McGraw-Hill New York, 1973.

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

[35] Rita Mae Brown. *Starting from Scratch: A Different Kind of Writers' Manual*. Reissue edition. Toronto: Bantam, 1st Mar. 1989. 272 pp. ISBN: 978-0-553-34630-5.

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

[37] 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.

[38] 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.

[39] 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.

[40] Adrián Carro, Raúl Toral and Maxi San Miguel. 'Coupled Dynamics of Node and Link States in Complex Networks: A Model for Language Competition'. In: *New Journal of Physics* 18.11 (29th Nov. 2016), p. 113056. ISSN: 13672630. DOI: 10.1088/1367-2630/18/11/113056.

[41] 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.

[42] 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.

[43] Xavier Castelló, Lucía Loureiro-Porto and Maxi San Miguel. 'Agent-Based Models of Language Competition'. In: *International Journal of the Sociology of Language* 2013.221 (22nd Jan. 2013), pp. 21–51. ISSN: 16133668. DOI: 10.1515/ijsl-2013-0022.

[44] 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.

[45] 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.

[46] 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.

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

[48] 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.

[49] 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.

[50] Mark Davies. 'Expanding Horizons in Historical Linguistics with the 400-Million Word Corpus of Historical American English'. In: *Corpora* 7.2 (1st Nov. 2012), pp. 121–157. ISSN: 1749-5032. DOI: 10.3366/cor.2012.0024.

[51] Bethany Davila. 'The Inevitability of "Standard" English: Discursive Constructions of Standard Language Ideologies'. In: *Written Communication* 33.2 (Apr. 2016), pp. 127–148. ISSN: 0741-0883, 1552-8472. DOI: 10.1177/0741088316632186.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

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

[57] Gonzalo Donoso and David Sánchez. 'Dialectometric Analysis of Language Variation in Twitter'. In: *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Valencia, Spain: Association for Computational Linguistics (ACL), 27th July 2017, pp. 16–25. DOI: 10.18653/v1/w17-1202.

[58] 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.

[59] 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.

[60] Robin I. M. Dunbar. 'The Social Brain Hypothesis'. In: *Evolutionary Anthropology: Issues, News, and Reviews* 6.5 (1998), pp. 178–190. ISSN: 1520-6505.

[61] 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.

[62] Jacob Eisenstein et al. 'Diffusion of Lexical Change in Social Media'. In: *PLoS ONE* 9.11 (19th Nov. 2014). Ed. by Robert C. Berwick, e113114. ISSN: 19326203. DOI: 10.1371/journal.pone.0113114.

[63] Daniel Judah Elazar. *Cities of the Prairie: The Metropolitan Frontier and American Politics*. New York: Basic Books, 1970. 514 pp. ISBN: 978-0-465-01137-7.

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

[65] Brian S. Everitt et al. *Cluster Analysis*. Wiley, Chichester, UK: John Wiley & Sons, 2011. XII, 330 p. ill. ISBN: 978-0-470-74991-3.

[66] Ralph W. Fasold. *The Sociolinguistics of Society*. In collab. with Internet Archive. Oxford, England ; New York, NY, USA : B. Blackwell, 1984. 358 pp. ISBN: 978-0-631-13462-6 978-0-631-13385-8.

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

[68] Charles A Ferguson. 'The Language Factor in National Development'. In: *Anthropological Linguistics* 4.1 (1962), pp. 23–27.

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

[70] 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.

[71] 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.

[72] David Hackett Fischer. *Albion's Seed*. Oxford, UK: Oxford University Press, 1989. ISBN: 978-1-5226-6831-2.

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

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

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

[76] Serge Frontier. 'Étude de La Décroissance Des Valeurs Propres Dans Une Analyse En Composantes Principales: Comparaison Avec Le Modèle Du Bâton Brisé'. In: *Journal of Experimental Marine Biology and Ecology* 25.1 (15th Nov. 1976), pp. 67–75. ISSN: 0022-0981. DOI: 10.1016/0022-0981(76)90076-9.

[77] Anastasia A. Funkner et al. 'Geographical Topic Modelling on Spatial Social Network Data'. In: *Procedia Computer Science*. 10th International Young Scientists Conference in Computational Science, YSC2021, 28 June – 2 July, 2021 193 (1st Jan. 2021), pp. 22–31. ISSN: 1877-0509. DOI: 10.1016/j.procs.2021.10.003.

[78] Susan Gal. 'Multilingualism'. In: *The Routledge Companion to Sociolinguistics*. Routledge, 2006. ISBN: 978-0-203-44149-7.

[79] 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.

[3] Emir Ganić et al. 'Dynamic Noise Maps for Ljubljana Airport'. In: *10th SESAR Innovation Days*. 10th Dec. 2020.

[80] Jian Gao, Yi-Cheng Zhang and Tao Zhou. 'Computational Socioeconomics'. In: *Physics Reports*. Computational Socioeconomics 817 (10th July 2019), pp. 1–104. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2019.05.002.

[81] 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.

[82] 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.

[83] Joel Garreau. *The Nine Nations of North America*. Boston: Houghton Mifflin Company, 1996. 427 pp. ISBN: 978-0-395-29124-5.

[84] Peter Garrett. 'Language Attitudes'. In: *The Routledge Companion to Sociolinguistics*. Routledge, 2006. ISBN: 978-0-203-44149-7.

[85] Raymond Duncan Gastil. *Cultural Regions of the United States*. Seattle: University of Washington Press, 1975. ISBN: 978-0-295-95651-0.

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

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

[88] Murray Gell-Mann. *The Quark and the Jaguar: Adventures in the Simple and the Complex*. Illustrated edition. St. Martin's Griffin, 15th Sept. 1995. ISBN: 978-0-8050-7253-2.

[89] Andrew Gelman. *Red State, Blue State, Rich State, Poor State: Why Americans Vote the Way They Do*. Princeton: Princeton University Press, 31st Dec. 2009. ISBN: 978-1-4008-3211-8. DOI: 10.1515/9781400832118.

[90] 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.

[91] 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.

[92] 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.

[93] 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.

[94] 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.

[95] 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.

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

[97] 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.

[98] 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.

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

[100] Jack Grieve, Dirk Speelman and Dirk Geeraerts. 'A Statistical Method for the Identification and Aggregation of Regional Linguistic Variation'. In: *Language Variation and Change* 23.2 (2011), pp. 193–221. ISSN: 09543945. DOI: 10.1017/S095439451100007X.

[101] 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.

[102] Ralph 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.

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

[104] 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.

[105] Mahedi Hasan et al. 'Normalized Approach to Find Optimal Number of Topics in Latent Dirichlet Allocation (LDA)'. In: *Proceedings of International Conference on Trends in Computational and Cognitive Engineering*. Ed. by M. Shamim Kaiser et al. Advances in Intelligent Systems and Computing. Singapore: Springer, 2021, pp. 341–354. ISBN: 978-981-334-673-4. DOI: 10.1007/978-981-33-4673-4_27.

[106] 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.

[107] 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.

[108] Els Heinsalu, Marco Patriarca and Jean Leo Léonard. 'The Role of Bilinguals in Language Competition'. In: *Advances in Complex Systems* 17.1 (20th Apr. 2014). ISSN: 02195259. DOI: 10.1142/S0219525914500039.

[109] 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.

[110] 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.

[111] Russell 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.

[112] Rafiazka Millanida Hilman, Gerardo Iñiguez and Márton Karsai. 'Socioeconomic Biases in Urban Mixing Patterns of US Metropolitan Areas'. In: *EPJ Data Science* 11.1 (1 Dec. 2022), pp. 1–18. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-022-00341-x.

[113] 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.

[114] Xiaoling Hu, Nigel Williamson and Jamie McLaughlin. 'Sheffield Corpus of Chinese for Diachronic Linguistic Study1'. In: *Literary and Linguistic Computing* 20.3 (1st Sept. 2005), pp. 281–293. ISSN: 0268-1145. DOI: 10.1093/llc/fqi034.

[115] 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.

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

[117] 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.

[118] Donald A. Jackson. 'Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches'. In: *Ecology* 74.8 (1993), pp. 2204–2214. ISSN: 00129658. DOI: 10.2307/1939574.

[119] John F. Jakubs. 'A Distance-Based Segregation Index'. In: *Socio-Economic Planning Sciences* 15.3 (1981), pp. 129–136. ISSN: 00380121. DOI: 10.1016/0038-0121(81)90028-8.

[120] 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.

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

[122] 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.

[123] 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.

[124] 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.

[125] Andres Karjus et al. 'Quantifying the Dynamics of Topical Fluctuations in Language'. In: *Language Dynamics and Change* 10.1 (10th Feb. 2020), pp. 86–125. ISSN: 2210-5832, 2210-5824. DOI: 10.1163/22105832-01001200.

[126] Søren Kierkegaard. 'Journals'. In: *A Kierkegaard Anthology*. Ed. by Robert W. Bretall. New York, Modern Library, 1936, p. 12.

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

[128] 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.

[129] Claire Kramsch. 'Language and Culture'. In: *AILA review* 27.1 (2014), pp. 30–55. DOI: 10.1075/aila.27.02kra.

[130] 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.

[131] William A. Kretzschmar Jr. 'Language Variation and Complex Systems'. In: *American Speech* 85.3 (1st Aug. 2010), pp. 263–286. ISSN: 0003-1283. DOI: 10.1215/00031283-2010-016.

[132] 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.

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

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

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

[136] 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.

[137] Jan-Erik Lane and Svante Ersson. *Culture and Politics: A Comparative Approach*. 2nd ed. London: Routledge, 18th May 2016. 392 pp. ISBN: 978-1-315-57545-2. DOI: 10.4324/9781315575452.

[138] Bibb Latané. 'The Psychology of Social Impact'. In: *American Psychologist* 36.4 (Apr. 1981), pp. 343–356. ISSN: 0003-066X. DOI: 10.1037/0003-066X.36.4.343.

[139] 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.

[140] 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.

[141] 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.

[142] Joel Lieske. 'Regional Subcultures of the United States'. In: *The Journal of Politics* 55.4 (21st Nov. 1993), pp. 888–913. ISSN: 0022-3816. DOI: 10.2307/2131941.

[143] Thomas Louf. *Spatial Distributions of Languages on Twitter*. 2021. DOI: 10.6084/m9.figshare.14339321.

[144] Thomas Louf. *multiling-analytical*. 2022. DOI: 10.6084/m9.figshare.20627235. URL: https://github.com/TLouf/multiling-analytical (visited on 28/02/2023).

[145] Thomas Louf. *multiling-twitter*. 2022. DOI: 10.6084/m9.figshare.20627238. URL: https://github.com/TLouf/multiling-twitter (visited on 28/02/2023).

[146] Thomas Louf. *Word Counts per US County in Geo-Tagged Tweets Posted between 2015 and 2021*. figshare, 2023. DOI: 10.6084/m9.figshare.20630919.v1.

[147] Thomas Louf. *words-use*. 2023. DOI: 10.6084/m9.figshare.20627034.v1. URL: https://github.com/TLouf/words-use (visited on 28/02/2023).

[1] 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.

[2] Thomas Louf et al. 'American Cultural Regions Mapped through the Lexical Analysis of Social Media'. In: *Humanities and Social Sciences Communications* 10.1 (1 30th Mar. 2023), pp. 1–11. ISSN: 2662-9992. DOI: 10.1057/s41599-023-01611-3.

[148] Michael W. Macy and Robert Willer. 'From Factors to Actors: Computational Sociology and Agent-Based Modeling'. In: *Annual Review of Sociology* 28.1 (2002), pp. 143–166. DOI: 10.1146/annurev.soc.28.110601.141117.

[149] 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.

[150] 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.

[151] 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.

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

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

[154] James Milroy. 'The Ideology of the Standard Language'. In: *The Routledge Companion to Sociolinguistics*. Routledge, 2006. ISBN: 978-0-203-44149-7.

[155] 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.

[156] Jorge Mira and Ángel 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.

[157] Jorge Mira, Luís F. Seoane and Juan J. Nieto. 'The Importance of Interlinguistic Similarity and Stable Bilingualism When Two Languages Compete'. In: *New Journal of Physics* 13.3 (3rd Mar. 2011), p. 033007. ISSN: 13672630. DOI: 10.1088/1367-2630/13/3/033007.

[158] 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.

[159] 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.

[160] Elaheh Momeni et al. 'Modeling Evolution of Topics in Large-Scale Temporal Text Corpora'. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Twelfth International AAAI Conference on Web and Social Media. Vol. 12. 15th June 2018, pp. 656–659. DOI: 10.1609/icwsm.v12i1.15068.

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

[162] 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.

[163] 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.

[164] 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.

[165] 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.

[166] 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.

[167] 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.

[168] 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.

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

[170] 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.

[171] Howard Washington Odum. *Southern Regions of the United States*. Chapel Hill, NC: Univ. North Carolina Press, 1936. ISBN: 978-0-87586-015-2.

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

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

[174] J. K. Ord and Arthur Getis. 'Local Spatial Autocorrelation Statistics: Distributional Issues and an Application'. In: *Geographical Analysis* 27.4 (1st Oct. 1995), pp. 286–306. ISSN: 1538-4632. DOI: 10.1111/J.1538-4632.1995.TB00912.X.

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

[176] 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.

[177] 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.

[178] 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.

[179] Marco 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.

[180] Glenna Ruth Pickford. 'American Linguistic Geography: A Sociological Appraisal'. In: *WORD* 12.2 (Aug. 1956), pp. 211–233. ISSN: 0043-7956, 2373-5112. DOI: 10.1080/00437956.1956.11659600.

[181] Juan Pablo Pinasco and Lilia 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.

[182] 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.

[183] 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.

[184] 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.

[185] Matti Rissanen. 'The Helsinki Corpus of English Texts'. In: *Corpora Across the Centuries*. First International Colloquium on English Diachronic Corpora. Cambridge: Rodopi, 31st Dec. 1993. ISBN: 978-90-5183-615-8.

[186] 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.

[187] Peter J. Rousseeuw. 'Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis'. In: *Journal of Computational and Applied Mathematics* 20.C (1st Nov. 1987), pp. 53–65. ISSN: 0377-0427. DOI: 10.1016/0377-0427(87)90125-7.

[188] 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.

[189] Antoine de Saint-Exupéry. *Le petit prince*. 1st ed. Paris: Gallimard, 18th Mar. 2016. ISBN: 978-2-07-061275-8.

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

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

[192] 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.

[193] Thomas C. Schelling. 'Dynamic Models of Segregation'. In: *The Journal of Mathematical Sociology* 1.2 (1st July 1971), pp. 143–186. ISSN: 0022-250X. DOI: 10.1080/0022250X.1971.9989794.

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

[195] 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.

[196] Ashish Sen and Tony Smith. *Gravity Models of Spatial Interaction Behavior*. 1995. ISBN: 3-642-79880-2. DOI: 10.1007/978-3-642-79880-1.

[197] Luís F. Seoane and Jorge Mira. 'Are Dutch and French Languages Miscible?' In: *The European Physical Journal Plus* 137.7 (20th July 2022), p. 836. ISSN: 2190-5444. DOI: 10.1140/epjp/s13360-022-03020-y.

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

[199] 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.

[200] 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.

[201] Erik Smitterberg and Merja Kytö. 'English Genres in Diachronic Corpus Linguistics'. In: Stockholm University, 2015, pp. 117–133.

[202] 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.

[203] 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.

[204] 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.

[205] 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.

[206] 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.

[207] 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.

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

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

[210] Peter Trudgill. *The Social Differentiation of English in Norwich*. Cambridge Studies in Linguistics 13. Cambridge: Univ. Pr, 1974. 211 pp. ISBN: 978-0-521-29745-5 978-0-521-20264-0.

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

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

[213] *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).

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

[215] Ad Hoc Expert Group on Endangered Languages UNESCO. 'Language Vitality and Endangerment'. In: International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages. Paris, 2003.

[216] Robert M. Vanderbeck and Cheryl Morse Dunkley. 'Young People's Narratives of Rural-Urban Difference'. In: *Children's geographies* 1.2 (2003), pp. 241–259. DOI: 10.1080/14733280302192.

[217] Federico Vazquez, Xavier Castelló and Maxi 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.

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

[219] 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.

[220] 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.

[221] 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.

[222] Svante Wold, Kim Esbensen and Paul Geladi. 'Principal Component Analysis'. In: *Chemometrics and Intelligent Laboratory Systems* 2.1-3 (1st Aug. 1987), pp. 37–52. ISSN: 0169-7439. DOI: 10.1016/0169-7439(87)80084-9.

[223] 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.

[224] Colin Woodard. *American Nations: A History of the Eleven Rival Regional Cultures of North America*. New York NY: Penguin Books, 2012. 384 pp. ISBN: 978-0-14-312202-9.

[225] BigScience Workshop et al. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. 10th Dec. 2022. DOI: 10.48550/arXiv.2211.05100. preprint.

[226] 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.

[227] Wilbur Zelinsky. *The Cultural Geography of the United States*. Englewood Cliffs: Prentice Hall, 1992. 226 pp. ISBN: 978-0-13-194424-4.