Universitat
de les Illes Balears

DOCTORAL THESIS
2023

COMPLEXITY IN LANGUAGE VARIATION:

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

THOMAS LOUF

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

Doctor by the Universitat de les Illes Balears

SUPERVISORS:
José Javier Ramasco
David Sánchez

LOCATION:
Palma, Spain

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

— Lilo & Stitch

Dedicated to the loving memory of Rudolf Miede.

1939 – 2005

ABSTRACT

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

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

RÉSUMÉ

Mon résumé

PUBLICATIONS

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

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

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

[3] Thomas Louf et al. *American Cultural Regions Mapped through the Lexical Analysis of Social Media*. Aug. 2022. DOI: 10.48550/arXiv.2208.07649. arXiv: 2208.07649 [physics].

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

— Donald E. Knuth [33]

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 15

API application programming interface 10

POI point of interest 13

QRT quote retweet 9

RT retweet 9

SES socio-economic status 6

Part I

INTRODUCTION

1

WHERE LANGUAGE DIVERSITY COMES FROM

Language can generally be defined as a structured system that human beings use to communicate. More specifically here, we use the term *language* to refer to natural languages, meaning languages which have evolved naturally, or, said differently, that have not been designed intentionally — as opposed to programming languages, for instance. Our objects of study are thus languages in the common sense of the word, that is coherent systems that define words and how their combinations convey meaning — like English, Mandarin Chinese or Hindi, to cite the three most spoken nowadays. As a primary means of communication, language is ubiquitous in any individual's life and in the workings of any human society. It is so much so that it is considered a "cultural universal", meaning all known human societies have some form of language [9, 28]. And it is so much so that researchers are unable to trace back to the origin of such a structured system of communication [23, 33, 55, 66]. Those who have ventured into this kind of inquiry have estimated that language dates back tens or even hundreds of thousands of years [7, 11, 14, 58]. One fact is for certain though: for what could be colloquially called *a very long time*, human beings have come up with, innovated upon, used, and more generally interacted with languages. It is then safe to say that human history must have seen a huge diversity of languages emerge. What is more ambiguous, though, is how this diversity is shaped through individuals' interactions, as they form societies. This is the central question that defined the whole field of sociolinguistics [10, 46, 47, 70, 75], which is also the broad subject of this thesis. In the following sections, we will touch on the different roles of language in society that may bring about variation, or, on the contrary, reduce existing linguistic diversity.

1.1 LANGUAGE AS A VECTOR FOR COMMUNICATION

The first obvious function that language serves is to facilitate communication between individuals, more specifically the kind of communication called *verbal communication*. To optimize language with regard to that function, there should only be one single language, shared homogeneously among all individuals. This has not been the case historically though, for many reasons, including historical and political ones, but also a very down-to-earth one. It is the very simple fact that, for most of its history, humanity has been spread around the Earth and unable to communicate at long distances. There is one

very well known example that illustrates this. Humans have been in America for thousands of years: according to recently-found evidence, they have for more than 21 000 years [3]. Yet, the first lasting contact between Europeans and indigenous Americans came less than 600 years ago. During all this time, people on the two continents have had ample time to come up with new languages, innovate upon existing ones, and mix within their own continent, at least partially. Thus, on the scale of all these languages' histories, it is only very recently that the two groups came into contact. Since then, things have accelerated extremely fast though. First, transport has allowed long distance communication on the scale of months with boats for roughly the past 500 years, and then on the scale of hours with planes since the start of the last century. In the last two centuries, telecommunication has enabled long distance and near-real-time communication, and it has truly been widely democratized with the Internet in the last two decades. On the technical front, the communication barriers between individuals across the globe seem to have come down. But does this imply a push towards a reorganization of the world in what the philosopher Marshall McLuhan called a *global village* [50]? Does this imply a more interconnected world, and in turn, that we will naturally tend towards the communication-optimal state of homogenous language?

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

> Among all the individuals that are linked together by speech, some sort of average will be set up: all will reproduce — not exactly of course, but approximately — the same signs united with the same concepts. [15]

But this has also been questioned, starting with the central 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, known as Dunbar's number. Its existence was first hypothesized [17, 18], and later demonstrated, not only on real-world social networks [37, 49], but also for a massive, online one [24]. 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 [16, 51,

69, 76]. This idea and the experiments behind it have received some criticism though [43]. 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 [22, 48, 67].

The existence of a process of globalization is undisputed, but the idea that more interconnectedness would create a global village has been challenged [4, 59]. 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. [4]

Nonetheless, there is some evidence of a homogenizing trend. English is on a steady path to become a global language [13]. Most of the estimated 6000 languages that exist in the world today are endangered [12, 29, 45] and getting replaced by a few dominant languages [31, 74].

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

1.2 LANGUAGE AND POLITICS

As Pierre Bourdieu argued, language is not only a means of communication but also a medium of power [8]. Some aspect of this idea has been popularized in the world of fiction with the concept of the *Newspeak* language in George Orwell's *1984* [61]. 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 [21], and not necessarily with dystopian, *Big Brother*, intentions of total control over individuals. When nation-states were built, a common language was seen as a means to unite a nation [78]. This was particularly the case in post-revolution France and imperial Great Britain [31, 36], where respectively French and English were heavily pushed as the languages of higher status. Still today, most constitutions specify one or a maybe a few languages as the languages of the state. In theory, this would be beneficial as it enables the state to build a common ground to guarantee equal opportunity, for instance with public education and the rule of law. It also would not necessarily mean dropping local languages, but only learning a common one, resulting then in a large population of multilinguals. In practice though, these political pushes for a shared language have lead to the near extinction of many regional languages and dialects. It is however not uncommon that a later reaction tried to reverse this process, with policies switching roles completely. Politics can then on the contrary oppose homogenization and protect language

diversity. Current examples include the policies introduced to protect national languages against global English [65], and also the ones to preserve regional languages and dialects within nations [42].

1.3 LANGUAGE AS A COMMODITY

Language can also be seen as part of the set of skills that an individual possesses and may need to perform their job. Knowing a language has thus an economic value, and this is particularly true in a globalized economy [35]. Indeed, the world has not only become more interconnected in terms of communication, but also in terms of trade. It is the aspect of globalization that has had the most impact on our contemporary societies: in fact, when someone talks about globalization, most of the time what they are referring to is economic globalization. In this context, good command of a non-native language, like English in most cases, can often be a requirement to apply for a job. As a result, the status of a language, or its perceived value in society, can depend heavily on the value it is given by the market. As the sociologist Pierre Bourdieu put it, individuals possess different quantities of *linguistic capital* [8]

Further, the manner with which one speaks a language can identify them as member of a certain socio-economic group. Indeed, all languages have a number of varieties, some with superior status. The standard variety of a language, when one is identified as such by an official institution like a language academy, is often the most prestigious one. A language has then a "correct" way to be spoken which is the one taught in schools. As proven by the PISA reports of the OECD, its latest [60] included, in many countries, linguistic proficiency of 15-year-olds strongly varies based on their *socio-economic status* (SES) of origin. The lower socio-economic class This can be detrimental to parts of a population, as these differences can entail segregation in several spheres of society, notably on the job market, but also in social interactions.

1.4 LANGUAGE AS A CULTURAL TRAIT

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

Conversely, rejecting the dominant, or mainstream, culture may also mean rejecting its language, and protecting one's own. It may also mean coming up with one's own language, as part of building a sub or counter-culture. This mechanism is not the only one at work that pushes for language diversity. In 2003, the UNESCO adopted the *Convention for the Safeguarding of the Intangible Cultural Heritage*, which states that language, "as a vehicle of the intangible cultural heritage", is to be safeguarded against the effects of globalization [73]. Also, very often, language preservation policies are implemented based on the argument that cultural diversity embedded in languages needs to be preserved [12, 29, 45].

2 METHODOLOGY

2.1 DATA

2.1.1 *What for*

2.1.2 *Traditional sources in linguistics*

2.1.3 *New sources from online media*

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 three 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 thus many ways to interact with others, and Twitter thus hosts a huge network of inter-user interactions. It is one of the most popular online social media, with hundreds of millions of users worldwide. In the US, for instance, since 2015, more than 20 % of the population use the platform [2]. Other than in the US, it is also popular in many countries globally, although with a slight bias towards developed, western countries [34]. Although only around 1 % of Tweets are geotagged [54], when only counting in users who tweet with a

geotag, in around 80 countries there is more than one Twitter user for ten thousand inhabitants [53].

Hence why Twitter has been extensively used for the analysis of geographically-embedded text [1, 5, 6, 25–27, 30, 38, 44, 53, 56].

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*

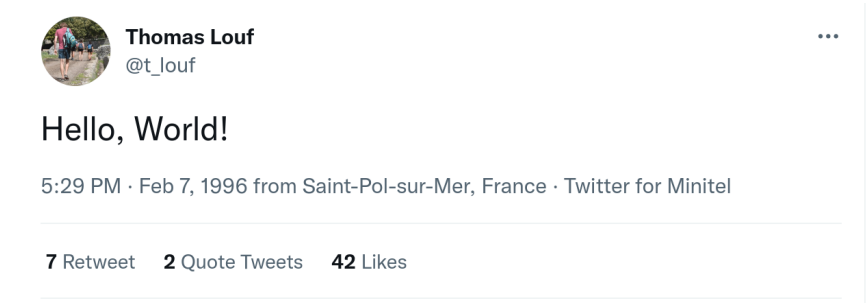
A major advantage of Twitter for academic research is how open the platform is to giving access to its data to researchers. One can send automatic queries to Twitter for data through their public *application programming interface* (API) [71]. In these queries one can specify rules to, for instance, 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 [72]. We show in Figure 2.1 an example of the data we can have for each Tweet.

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

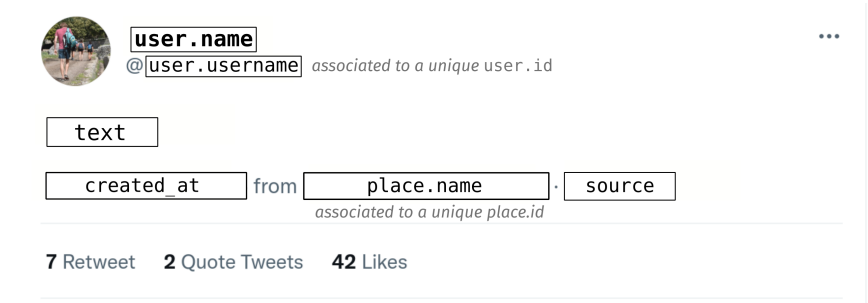
2.1.4.2 *Text processing*

Since we are interested in the speech produced by users, we need to clean parts of the text which cannot be considered as natural language production. Those are the URLs, mentions of other users (in the form `@username`) and hashtags (in the form `#topic`). It is not completely obvious that the latter should be discarded though. Hashtags are used on Twitter to aggregate Tweets by topics. It is an important feature of the website, whose aim is to enable users to easily find the Tweets of other users discussing similar topics, 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 [62]. 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

(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).

discard are source-dependent. We will not go into details — our text processing code is freely available online anyway — but, for instance, when a Tweet was sent from Foursquare, we strip all location-related content, which can be located after either a `"I'm at"` or `"( @ "` string.

In practice, all the elements cited above are stripped off of Tweets using regular expressions. After this cleaning step, for what follows we then keep only the Tweets still containing at least four words. The next important step that was crucial to all our works was to infer the language the Tweets are written in. To do so, we leverage a trained neural network model for language identification: the Compact Language Detector [64]. It was designed as part of Chromium-based web browsers to detect the language web pages are written in in order to make translation suggestions to users, and it is now openly accessible. Its output is a language prediction along with the confidence of the model. Whenever we focus on a language, we thus keep Tweets which are tagged in that language with a confidence above 90 %.

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

2.1.4.3 *Inferring geolocation*

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

- a country,
- an administrative unit: province, region or department for instance,

- a city,
- a *point of interest* (POI): any kind of public place: restaurant, school, event venue, etc. These are represented by a point, so Tweets tagged with a POI can be considered similarly to the ones with coordinates.

When a user tweets with their device's GPS activated, a place (usually the city they tweet from) is selected by default, and they can switch to another one from a list of close-by places. These places were fed to Twitter by Foursquare [20] (among others), which provides data down to a POI level for more than 190 countries. The geographical extent of places other than POIs is defined by bounding boxes.

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

As the activity of Twitter users was found to follow a log-normal distribution spanning almost four orders of magnitude [53], it can be preferable to compute metrics at the user level. Indeed, at the Tweet level, the linguistic behaviour of the most active users could overshadow the one of the many, less active users. Individuals are mobile but for the vast majority they have a preferred location, namely

their place of residence. That is why we often strive to attribute a cell of residence to the users in our datasets. To explain the heuristics we defined for residence attribution, let us first formalize some notation. For each user u, there are two counts we get directly from their Tweets: the number of them with GPS coordinates that fall in cell c: $n_{u,c}^{\mathrm{GPS}}$, and those without coordinates but tagged as being from place p: $n_{u,p}$. We wish to compute $r_{u,c} \in \mathbb{R}^+$, the ratio of Tweets of user u in cell c. It can be decomposed into the contributions of those with GPS coordinates $r_{u,c}^{\mathrm{GPS}}$, and of the others: $r_{u,c}^{\mathrm{P}}$:

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

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

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

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

2.1.4.4 *Selecting relevant users*

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

2.1.4.5 *Caveats*

No data source is without bias, and Twitter is no exception. First, at the global scale, as we already mentioned, Twitter is more representative

of people living in western, developed countries with widespread access to the Internet [34]. In terms of more local geographical biases, densely populated urban areas are usually overrepresented [2, 40, 52]. As for demographics, Twitter users are on average younger [2, 57], with more degrees and income, and are more likely to be a male than the overall population [2, 52].

2.2 MODELS

2.2.1 *What for*

2.2.2 *What kind*

agent-based model (ABM)

2.3 SOURCE MATERIALS AND TOOLS

Following the principles of open science, throughout my thesis, I have made all source materials for my results openly accessible, whether they are codes[1] or datasets[2], including this very manuscript's[3]. Equally importantly, I believe, I have strived to use almost exclusively free and open source software in my work. I cannot realistically cite here all projects I have relied on to carry out my work, but I can cite a few central ones. I wrote all my code in the Python 3 programming language, using libraries such as NumPy [32], pandas [68] or GeoPandas [41]. In their vast majority, figures presented here were prepared with Matplotlib [39], and sometimes edited, or entirely drawn, with Inkscape[4].

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

2.4 OUTLINE

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

Part II

RESULTS

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

3

MULTILING

TODO

TODO epigraphcite

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

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 *[63]*

5

ACR

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

— *Simone Weil,* The Need for Roots *[77]*

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*, $8\frac{1}{2}$ [19]

Part IV

APPENDIX

A

APPENDIX TEST

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

More dummy text.

A.1 APPENDIX SECTION TEST

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

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

Table A.1: Autem usu id.

A.2 ANOTHER APPENDIX SECTION TEST

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

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

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

BIBLIOGRAPHY

[1] 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.

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

[3] 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.

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

[5] 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. arXiv: 1903.04329.

[6] 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. arXiv: 1605.02951.

[7] Rudolf Botha and Chris Knight. *The Cradle of Language*. OUP Oxford, 30th Apr. 2009. 418 pp. ISBN: 978-0-19-156767-4. Google Books: IVRkzK1VX1oC.

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

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

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

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

[22] 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.

[23] 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.

[24] 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.

[25] Bruno Gonçalves and David Sanchez. '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. pmid: 25409174.

[26] 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. arXiv: 1511.04970.

[27] 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. pmid: 29799872.

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

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

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

[40] 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.

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

[42] 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.

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

[44] 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.

[45] 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.

[46] William Labov. *Sociolinguistic Patterns*. University of Pennsylvania Press, Sept. 1973. 374 pp. ISBN: 978-0-8122-1052-1. Google Books: hD0PNMu8CfQC.

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

[48] 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.

[49] 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.

[50] 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.

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

[52] 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.

[53] 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. pmid: 23637940.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

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

[59] 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. Google Books: 5q0hAwAAQBAJ.

[60] OECD. *Where All Students Can Succeed*. Vol. II. PISA 2018 Results. Paris: Organisation for Economic Co-operation and Development, 2019.

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

[62] 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.

[63] 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.

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

[65] 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. Google Books: reRuAAAAQBAJ.

[66] 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.

[67] 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.

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

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

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

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

[72] *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).

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

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

[75] 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.

[76] 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.

[77] 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.

[78] 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. Google Books: Gd5d1CwtI7cC.