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

IFISC

DOCTORAL THESIS
2023

Doctoral programme in Physics

COMPLEXITY IN LANGUAGE VARIATION:

EXPLORING THE INTERPLAY BETWEEN GEOGRAPHY, CULTURE AND THE SOCIAL FABRIC

THOMAS LOUF

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

Doctor by the Universitat de les Illes Balears

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

SUPERVISORS:
José Javier Ramasco
David Sánchez

LOCATION:
Palma, Spain

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

— Lilo & Stitch

Dedicated to the loving memory of Rudolf Miede.
1939 – 2005

ABSTRACT

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

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

RÉSUMÉ

Mon résumé

PUBLICATIONS

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

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.

— **HauserMysteryLanguage2014 [HauserMysteryLanguage2014]**

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)

ACRONYMS

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 [**GreenbergLanguageUniversals2020**, **BrownDonaldHumanUniversals1991**]. And it is so much so that researchers are unable to trace back to the origin of such a structured system of communication [**MullerLectureIX1861**, **StamInquiriesOrigin1976**, **GibsonOxfordHandbook2011**, **HauserMysteryLanguage2014**]. Those who have ventured into this kind of inquiry have estimated that language dates back tens or even hundreds of thousands of years [**NicholsOriginDispersal1998**, **ChomskyLanguageMind2004**, **BothaCradleLanguage2009**, **DediuAntiquityLanguage2013**]. 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 defining question of the whole field of sociolinguistics [**LabovSociolinguisticPatterns1973**, **TrudgillSociolinguisticsIntroduction2000**, **ChambersSociolinguisticTheory2007**, **WardhaughIntroductionSociolinguistics2008**, **LabovPrinciplesLinguistic2001**], that is also the core question that we address in 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 be only one homogeneously shared 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 [**BennettEvidenceHumans2021**]. 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, 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. Then, telecommunication has enabled long distance and near-real-time communication in the last two centuries, and it has truly been widely democratized with the internet in the last two decades. Are the new forms of communication brought by these technological shifts then pushing a reorganization of the world towards McLuhan's *global village* [**McLuhanGutenbergGalaxy2008**]? Does this imply 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 maximum entropy. Another view would be that, because it costs energy for humanity to maintain language diversity, homogenization of language would be both desirable and inevitable. But physics' success lies in its ability to provide good models of the world, that is, useful approximations of it. So while the model this intuition is drawn from was applied to the movement of physical bodies and has made the triumph of thermodynamics, is it here of any use, when trying to understand and make predictions about language diversity? 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. [**deSaussureCourseGeneral2011**]

But this view has also been questioned, starting with its central hypothesis that the global society would tend toward complete interconnectedness. This idea was for instance challenged by the anthropologist Robin 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 [**DunbarNeocortexSize1992**, **DunbarSocialBrain1998**], and later demonstrated, not only on real-world social networks [**HillSocialNetwork2003**, **McCartyComparingTwo2005**], but also for a massive, online one [**GoncalvesModelingUsers2011a**]. The existence of a global village is then arguable

> 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. [**BlommaertSociolinguisticsGlobalization2010**]

small world, but that is contested, but anyway no death of distance: . The

Nonetheless, there is some evidence of a homogenising trend. English is on a steady path to become a global language [**CrystalEnglishGlobal2010**]. Most of the estimated 6000 languages that exist in the world today are endangered [**CrystalLanguageDeath2000**, **GrenobleEndangeredLanguages1998**, **KraussWorldLanguages1992**] and getting replaced by a few dominant languages [**GrilloDominantLanguages1989**, **WardhaughLanguagesCompetition1987**].

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. As Pierre Bourdieu argued, language is not only a means of communication but also a medium of power [**BourdieuLanguageSymbolic2009**].

1.2 LANGUAGE AS A POLITICAL WEAPON

As Pierre Bourdieu argued, language is not only a means of communication but also a medium of power [**BourdieuLanguageSymbolic2009**].

1.3 LANGUAGE ON THE MARKET

For different speakers possess different quantities of 'linguistic capital' [**BourdieuLanguageSymbolic2009**]

1.4 LANGUAGE AS A CULTURAL TRAIT

In this thesis, we investigate - inter-language spatio-temporal evolution - intra-language lexical variations in space, highlighting cultural differences - intra-language spatial variations from the standard form, and its interplay with socio-economic status

things that push towards/against diversity (non exhaustive): - need to comm = against - culture: tricky one. extremely popular stuff, ("global culture" that nowadays can spread easily) can bring homogeneity (eg Hollywood). Other hand, this call for counter cultures, push against dominant one. Group identity but at different levels, the closer the stronger - ses: source of diversity. But tricky: this diversity actually brings by segregation

While sometimes considering languages as coherent, clearly-separated units (TODO ref chapter?) to study interlanguage interactions, we will also do away with this simplified view to study intra-language variations.

Linguistics, and especially in its social branch, is thus at the interface of many entangled 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 political science.

[**LabovPrinciplesLinguistic1994**, **LabovPrinciplesLinguistic2001**, **LabovPrinciplesLin**

2 METHODOLOGY

start with recap of previous chapters, into: ok so now what tools can we use. start with "old" way: classic linguistics, transition in ot computational and finish with complex systems stuff. [**NguyenComputationalSociolinguistics2016**]

different media to convey language HERE written language exclusively, WHY: our methods, because that's what computers process better SO we lose a lot of things that cannot be transcribed, or that are simply not because it would require considerable effort to do so (accent, intonation), and access to spoken language is much more limited. also lose all non-verbal communication between human

Enormous amount of information exchanged through language. Data and metadata: . Someone's language tells a lot about them

methods with books "older data"

2.1 DATA

2.1.1 *What for*

2.1.2 *Twitter data*

2.1.2.1 *Accessing*

```
"str"
print('lol')
```

[mathescape, linenos, numbersep=5pt, gobble=2, frame=lines, framesep=2mm]csharp string title = "This is a Unicode in the sky" /* Defined as $\pi = \lim_{n\to\infty} \frac{P_n}{d}$ where P is the perimeter of an n-sided regular polygon circumscribing a circle of diameter d. */ const double pi = 3.1415926535

2.1.2.2 *Text processing*

2.1.2.3 *Infering geolocation*

2.1.2.4 *Caveats*

cover Twitter (biases but also basic technical details, API, what Tweet looks like), why remove HTs, blabla, language IDtion data driven analysis, cite Bruno papers eg

theoretical models: AS, MW

finally computational methods, very general (data, PCA...)

2.2 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 [**HarrisArrayProgramming2020**], pandas [**teamPandasdevPandas2020**] or GeoPandas [**JordahlGeopandasGeopandas2020**]. In their vast majority, figures presented here were prepared with Matplotlib [**HunterMatplotlib2D2007**], 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.

2.3 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 TODO make public

[4] Hosted 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.

— **ProustGuermantesWay1927**, **ProustGuermantesWay1927** [**ProustGuermantesWay1927**]

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.

— **WeilNeedRoots2002**, **WeilNeedRoots2002** *[***WeilNeedRoots2002***]*

Part III

CONCLUSION

6 CONCLUSION

Guido
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.

— **Fellini1963**, **Fellini1963** *[***Fellini1963***]*

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: **??** (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: A floating example (`listings` manual)

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