



**University of
Zurich** ^{UZH}

Modeling spatio-temporal diffusion of complex linguistic structures in Spanish Twitter

GEO 511 Master's Thesis

Author

Patricia Schnidrig
15-713-316

Supervised by

Dr. Peter Ranacher

Prof. Dr. Carlota de Benito Moreno (carlota.debenitomoreno@uzh.ch)

Faculty representative

Prof. Dr. Robert Weibel

25.06.2021

Department of Geography, University of Zurich

Abstract

This study focuses on the spatio-temporal dynamics of diffusion of linguistic innovations through the social media channel Twitter. To analyse the spatial diffusion, I consider over 57 million tweets with geoinformation between 2012 and 2019. The main objective is to understand how Spanish linguistic innovations spread in space and time. Specifically, I aim to understand the diffusion of Spanish linguistic innovations by (1) analyzing the role of geography in the diffusion of Spanish linguistic innovations, and (2) estimating the influence Spanish speaking regions have on each other.

Methodically, the analysis consists of two key parts. On the one hand, the study examines the spatial properties of the diffusion process with descriptive statistics including the measures focus, entropy, spread and impact. On the other hand, the study infers the network of linguistic influence with a Hawkes process. This approach is based on the idea that past adoptions of innovations trigger the occurrence of future adoptions of innovations.

The results show that some linguistic innovations clearly diffuse through different Spanish speaking regions around the globe, which means that there are interactions between Twitter users from different regions. The probability of a global diffusion of a phenomenon increases as the popularity of an innovation increases. Although not all of the innovations spread globally, the large areas over which some innovations spread showed that the large distances between regions on different continents do not seem to act as a linguistic barrier.

The inferred network of linguistic influence concludes that mainly the Argentine and Uruguayan Twitter users influence the European Twitter users. Mexico also seems to be more of a leading or influencing country and at the same time Mexico is the region that is most influenced by itself. Finally, the study demonstrates that a Hawkes process can be used to successfully infer the unknown network of linguistic influence.

Key words: spatial diffusion, innovation propagation, Hawkes process, social media, spatial impact, spatio-temporal analysis, linguistic influence

Acknowledgments

I would like to thank everyone who supported, motivated and inspired me during the process of this thesis.

Especially, I would like to thank my advisors, Peter Ranacher and Carlota de Benito Moreno, for their help and support during each stage of the process.

In addition, I am generally very grateful for the support of members of the linguistic faculty who invested many hours in the time-consuming linguistic preprocessing of the huge Twitter dataset. In particular, I would like to express my gratitude to Sarah Elisabeth Kiener, Maxine Hofstetter and Carlota de Benito Moreno.

Finally, I would like to thank my friends and family for their continuous support.

Contents

List of Figures	vi
List of Tables	viii
1. Introduction	1
1.1. Motivation and goal	1
1.2. Research gap and research questions	2
1.3. Thesis structure	3
2. Theoretical Background	5
2.1. The internet as a linguistic medium	5
2.1.1. Twitter as a data source and linguistic medium	5
2.1.2. The geography on Twitter	6
2.1.3. Challenges of Twitter as a data source	7
2.2. The generation and adoption of innovations	8
2.3. Diffusion of Innovations	9
2.3.1. Theories of linguistic innovation diffusion in time	11
2.3.2. Theories of linguistic innovation diffusion in space	12
2.4. Time-dependent Hawkes process	16
2.4.1. The concept of the Hawkes process	16
2.4.2. The intensity function	17
2.4.3. Historical applications of a Hawkes process	19
3. Data	20
3.1. Data collection	20
3.2. Study area	21
3.3. Tweets in space and time	23
3.4. Analyzed linguistic phenomena	25

Contents

4. Methodology	27
4.1. Preprocessing	27
4.1.1. Spatial binning	27
4.1.2. Innovation filtering	29
4.2. Descriptive statistics	30
4.2.1. Spatial properties of innovation diffusion	30
4.2.2. Relationship between regions	32
4.2.3. Measuring spatial impact	33
4.3. Hawkes Model	35
4.3.1. Multi-dimensional Hawkes process	35
4.3.2. Time transformation	37
4.3.3. The event's timestamp as model input	38
4.3.4. Parameter estimation	39
5. Results	40
5.1. Overview of the innovations	40
5.1.1. Innovations and innovation popularity	40
5.1.2. Origin of innovations	44
5.2. Diffusion patterns of innovation propagation	46
5.3. Relationship between regions and their innovations	52
5.4. Spatial impact	54
5.5. Adjacency matrix of Hawkes process	56
6. Discussion	60
6.1. Spread of innovations	60
6.2. Factors driving the diffusion of innovations	61
6.3. Network of linguistic influence - Leaders and followers	63
6.4. Materials and methods	65
6.4.1. Data-related limitations	65
6.4.2. Critical evaluation of methodologies	66
7. Conclusion	70
7.1. Summary and major findings	70
7.2. Future work	70
References	71

Contents

Appendix	79
A. List of linguistic innovations	79
B. Documentation Twitter corpus	85
C. Point plots	86
D. Adjacency matrix in numbers	87
Personal Declaration	88

List of Figures

2.1.	S-shaped curve of innovation diffusion, adopted from Rogers (2003)	11
2.2.	Bell-shaped curve of innovation diffusion, adopted from Rogers (2003) . .	12
2.3.	The wave model of linguistic diffusion from Wolfram and Schilling-Estes (2003, p. 714)	13
2.4.	Hierarchical structure of linguistic diffusion from Wolfram and Schilling- Estes (2003, p. 724)	13
2.5.	Visualization of multivariate Hawkes process model from Nickel and Le (2020, p. 1)	17
2.6.	Hawkes process with exponential decays from Rizoïu et al. (2017, p. 7) . .	18
3.1.	Geographical distribution of geolocated Spanish tweets in 2012 in prepro- cessed data	21
3.2.	Overview Spanish speaking regions covered in this thesis (Esri, 2015; Eu- rostat, 2020)	22
3.3.	Number of tweets over time in the database	24
4.1.	Procedure of the spatial binning	28
4.2.	Examples illustrating different cases of spatial impact from Kamath et al. (2013, p. 675)	34
4.3.	Visualization of the time transformation proving that times are propor- tional to each other	37
4.4.	Theoretical structure of model input	38
5.1.	Change in frequency (number of tweets with an innovation per 1000 tweets) from 2012 - 2019 in all Spanish speaking regions	40
5.2.	Change in the absolute number of innovations over space and time	42
5.3.	Plot of region and time of the first occurrence of all 33 innovations	45
5.4.	Origin of innovations on a map	46
5.5.	Focus	47
5.6.	Entropy	48
5.7.	Spread	49
5.8.	Boxplot of global and local innovations	50

List of Figures

5.9. Spatial properties compared to the frequency of adoptions of innovations .	51
5.10. Correlation between the spatial properties focus, entropy and spread . . .	52
5.11. Heatmap of Jaccard coefficients	52
5.12. Correlation between distance and Jaccard coefficients	53
5.13. Correlation between distance and adoption lag	54
5.14. Spatial impact plot modeling impacts of all six regions	55
5.15. Spatial impact plot modeling impacts of all regions except Uruguay	56
5.16. Adjacency matrix of first model run including all regions	57
5.17. Adjacency matrix of second model run including all regions except Uruguay	58

List of Tables

3.1. Distance matrix in km	23
3.2. Tweets and users per regions in the database	24
4.1. Sample of final dataset after preprocessing	30
5.1. Adoptions of innovations per regions	43
5.2. Global and local innovations.	50
5.3. Inferred intensity baseline of first model run including all regions	57
5.4. Inferred intensity baseline of second model run including all regions except Uruguay	58

1. Introduction

1.1. Motivation and goal

Language is dynamic and is changing constantly over space and time. Historically, through the change of languages, new languages have been formed. But even in a much shorter period of time, language is constantly changing and new emerging forms can be called linguistic innovations. Especially, in recent years with the rise of online social networks, the internet language became a new research field with two main challenges to linguists who want to explore this medium. On the one hand, there is a huge amount of data in the internet. There has never been such a large language corpus, which is rapidly increasing day by day (Crystal, 2011). On the other hand, the speed of change is immense. New linguistic phenomena are emerging continuously and some of them quickly disappear again. Thus, it is precisely the framework of internet communication that is changing language in new forms in order to better fit the communication demands of social platform users.

Much of social media texts contain geographical information, which opens new windows to an interdisciplinary research field focusing on spatial aspects in connection with linguistic questions. In particular, Twitter offers worldwide data that can be analyzed spatially and linguistically in order to identify diffusion patterns of language change. Thus, in recent research tweets were often used to analyze the role of geography in diffusion of linguistic innovations. Despite the intensive research in recent years, there are still many open questions regarding the diffusion of linguistic innovations. The network of linguistic influence is unknown and unobserved. The main goal of this work is a better understanding of the spatial diffusion of linguistic innovations and reconstruct the spatial pathways of diffusion. From a methodological perspective I propose a time-dependent Hawkes process to track the diffusion of linguistic innovations. The result of the analysis is a network of influence, which might explain the relationship between spatial distance and linguistic influence.

By modeling how innovations propagate over networks, we can gain insight into the diffusion process of language change in online communities and better understand how

1. Introduction

linguistic innovations are transmitted. As Zhao et al. (2015, p. 641) state: "the problem of modeling the influence between people is a vital task for studying social networks. Equally important, the issue has also gained much attention in recent years due to its wide-spread applications in e-commerce, online advertisement and so on [...]". So, by inferring a network of linguistic influence, we gain a deeper insight into the social network of online communities. Once understanding the pathways of the underlying network, advertising and news can be disseminated in a very specific way. Additionally, Kersgaw et al. (2017, p. 1852) point out that "understanding which users and communities have greater influence on a language will allow for foreign language teachers to pre-empt new words entering a language, or for companies to place greater emphasis on communication in the language that has the greatest influence in a given region of network".

1.2. Research gap and research questions

Despite the intensive research in recent years, this work differs from previous research in three main aspects. Firstly, most of the research in the language diffusion field is based on analysis of the English language (Doyle, 2014; Eisenstein et al., 2014; Huang et al., 2016; Jones, 2015). This thesis analyzes how linguistic innovations in Spanish spread in space and time. Spanish is one of the most spoken languages worldwide and is spatially distributed across several countries and continents, which enables a large-scale analysis across the globe.

Secondly, whereas most studies analyze lexical innovations or similar structures such as hashtags (Romero, Meeder and Kleinberg, 2011; Kamath et al., 2013), in this work innovations of several linguistic levels including morphological, syntactic, phraseological and lexical phenomena are analyzed. Innovations are not simply defined by their increasing occurrence, but by a well-considered selection of linguists.

Thirdly, the analysis is mainly based on a time-dependent Hawkes model resulting in a network of influence. Recently, Hawkes processes have been used in research to model the dynamics of earthquakes, crimes, finance and social networks (Bacry et al., 2015; Ogata, 1988; Reinhart and Greenhouse, 2018; Zhao et al., 2015), but a Hawkes process has rarely to never been used to infer a network of diffusion of linguistic innovations. Additionally, recent research lacks the incorporation of geography into a Hawkes process model. In particular, the multidimensionality of the Hawkes process allows the inclusion of space as an additional dimension and thus infers the diffusion of innovations in space.

1. Introduction

The project aims at addressing the following main research question:

How do new linguistic features in Spanish spread in space and time?

To answer this main research question I focus on the following sub-questions:

(RQ1) How does geographical distance influence the diffusion patterns of linguistic innovations on online platforms such as Twitter?

There are different theories of language variation and change which mainly assume that the geographic diffusion of linguistic innovations is characterized by physical distance, population density and cultural patterns. Although interactions on online platforms are in principle not affected by physical distance, geographically proximate individuals will be more likely to be connected on social platforms and thereby adopt changed language (Sadilek, Kautz and Bigham, 2012). But how exactly do the wide distances between two continents influence the diffusion network? Do continental borders act as a linguistic barrier? Are all analyzed phenomena global or also local or regional phenomena? Or in other words, do the given innovations spread over small or large geographical areas?

(RQ2) What role do specific Spanish speaking regions play in the diffusion of linguistic innovations and how influential are they?

Influence is defined by Katz and Lazarsfeld (1955 in Kersgaw et al. (2017, p. 1852)) as "getting people to change their attitudes and behaviours". Probably not all regions will be equally influential in the diffusion process of innovations. Which are "the driving forces" of Spanish linguistic innovations? Where are the origin of the innovations and thus who are the innovators? Which regions seem to have an impact on others, and which are impacted by others?

1.3. Thesis structure

The thesis is structured as follows: In chapter 2 the theoretical background is illuminated. It specifically addresses the role of Twitter in linguistic research, theories of innovation diffusion in space and time and the theoretical framework of a Hawkes process. In chapter 3, the data is presented and described in more detail. In the fourth chapter the methodology is explained. The analysis consists of two key parts: First the study of descriptive statistics and secondly the inference of a network of linguistic influence with a Hawkes process model. The results are presented in chapter 5. The discussion in chapter 6 embeds the results in the existing literature and specifies limitations of the

1. Introduction

data and methods used. Finally, chapter 7 concludes this study with a summary and recommendations for future work.

2. Theoretical Background

2.1. The internet as a linguistic medium

Sornig (1981) already analyzed slang, colloquialisms and casual speech in 1981 by studying emerging linguistic phenomena. He states that "the terms 'slang', 'colloquialism' and 'casual speech' are used rather indiscriminately to denote a type of language usage somewhere between individual speech and standard language norms" (Sornig, 1981, p. 2). It is precisely such newly emerging linguistic forms in everyday language use that are the focus of this work. But since the study of Sornig (1981) the internet has evolved leading to two significant developments. Firstly, new language forms are emerging on the internet to better fit the needs of users. For example on short message platforms, many abbreviations are used that did not previously exist. Secondly, the communication on the internet itself leads to new linguistic phenomena such as idioms.

In the following subsection, the potential of Twitter as a data source is illuminated. Thereafter, the different options providing geographical information on Twitter is explained. Lastly, a few practical concerns of using Twitter in linguistic studies are elaborated.

2.1.1. Twitter as a data source and linguistic medium

Twitter is a social media platform which only allows a short size of text, specifically not more than 280 characters per post (Twitter, 2021). Therefore, it is a so called "microblogging platform", which was created in 2006 and since then rapidly grew (Crystal, 2011). The short posts on the platform are called tweets and can include text, but also images or videos. Diverse Twitter API's allow to download a random sample of all tweets posted in form of text. Consequently, Twitter became a frequently used data source in research. Due to the spatial information of tweets, geographical patterns in different topics such as information diffusion (Eisenstein et al., 2014; Kamath et al., 2013) or language variation and dialects (Gonçalves and Sanchez, 2014) are analyzed with Twitter data.

On Twitter users can follow and be followed by other users. Thus, users share their

2. Theoretical Background

content among their own followers, respectively their own network. As Zhang et al. (2017, p. 2) state: "This generates the network of Twitter, a directed graph where users are connected with each other through explicit relation between them". Zhao et al. (2015) present several key features of such online social networks. For example, they explicate that individuals are influenced by their social network. In other words, the likelihood of the adoption of a behavior by an individual increases if people around him or her have already adopted the behavior. Another key feature of social network mentioned by Zhao et al. (2015) is the uneven network structure of individuals. Some users have a huge number of connections and thereby influence many people, whereas other users only influence a small number of individuals.

2.1.2. The geography on Twitter

Different options exist to provide spatial information within Twitter. Graham et al. (2013) lists the following three options to add geographical information to tweets.

Firstly, the user can give some location information in his or her profile. However, there are no restrictions and Twitter allows users to type anything. Therefore, users also indicate that they are from "Middle-earth" or simply type "here".

Secondly, Twitter users can simply mention their location in the tweet. However, again, a great number of those location names can probably not be located on the globe or are wrongly spelled.

Thirdly, Twitter users can allow Twitter to directly geolocate their tweets by adding either the exact longitude and latitude coordinates or an approximate region specified as a bounding box. This is the most accurate geographic information, but only a small portion of users publish geocoded tweets. The literature varies on how many tweets include geolocation information - the numbers are mainly between 0.6% and 1.6% (Leetaru et al., 2013; Li et al., 2013; Morstatter et al., 2013; Sloan et al., 2013; Takahashi et al., 2011). So, only a small percentage of users explicitly provide their location.

Many studies rely either on the geocoded tweets with their GPS coordinates or on the location information supplied in the user profiles. The consideration of the location field in the user profile is useful to access a larger set of text messages and individuals than only those who have attached GPS coordinates (Pavalanathan and Eisenstein, 2015). Pavalanathan and Eisenstein (2015) analyze the influence of those different geolocation acquisition methods on datasets and research. They concluded that young people and women write more often GPS-tagged tweets, whereas older people and men tend to mention more geographic-specific names (Pavalanathan and Eisenstein, 2015; Wood-Doughty

2. Theoretical Background

et al., 2017). In addition, the geotagging behavior is dependent on the spoken language of users (Sloan and Morgan, 2015). Tweets in Turkish, Portuguese or Spanish are more likely to contain geoinformation than tweets in Korean, German or Russian (Sloan and Morgan, 2015). In summary, research has shown that demographic characteristics of users influence their behaviour on Twitter - whether or not users provide information about their location.

2.1.3. Challenges of Twitter as a data source

Although Twitter is often used for linguistic pattern analysis over space and time, using Twitter as an object of linguistic studies brings certain difficulties.

Firstly, regarding the demographics, Twitter users are not a representative sample of the population. Research on American Twitter has shown that Twitter users are generally younger, more urban and predominantly male compared to the overall population (Mislove et al., 2011). Similarly, Blank (2017) conclude that British Twitter users are younger, wealthier, and better educated than the overall British population. Therefore, Twitter users are not a representative demographic sample of their regions. However, this work does not aim to analyze the diffusion of linguistic innovations in a representative population. It specifically focuses on social media users, which is why there is no need for representativeness.

Secondly, Twitter users are not distributed homogenously in space. Twitter is significantly overrepresented in densely populated regions (Mislove et al., 2011). In this study, however, very large spatial regions are considered and no distinction is made between urban and rural areas.

Thirdly, Eisenstein et al. (2014) explicate the unaccountable changing Application Programming Interface (API) sampling rate as a major challenge for analysis with tweets. The sampling rate can change in unclear ways over time and can drastically impact the empirical probability of an individual using a specific word (Eisenstein et al., 2014). Therefore, the authors introduce two additional parameters in their model which control for global effects such as changes to the API sampling rate. The dataset used in this work has fewer tweets in more recent years, although the platform has become more popular. The reason for the varying amount of data over the years is probably the changing API sampling rate.

Fourthly, Crystal (2011) raises the question if we should include retweets in linguistic analysis. The phenomenon of retweeting is an important stylistic feature of Twitter where Twitter users can simply repost another tweet with or without adding an additional comment. Those retweets can strongly influence the number of occurrences of

2. Theoretical Background

linguistic phenomena and therefore affect the analysis. This issue is discussed in more detail in chapter 6.4.1.

In addition, Li et al. (2013) states that there is a contribution bias in Twitter data. Most tweets come from a small number of Twitter users. Similarly, Leetaru et al. (2013) analyzed the different engagement of Twitter users and concluded that a small number of users is very active and account for the majority of the tweet volume. In this analysis, however, the contribution bias is not problematic because I am trying to infer the network of linguistic influence of the Twitter community, and the effects of influential individuals are also part of the network.

Another issue that is often mentioned when working with Twitter data is the presence of organizational accounts or bots. Bots are automated programs. Chu et al. (2010) differentiates between legitimate bots which generate tweets about the news or the weather and malicious bots that spread spam. Because both post a large amount of tweets that are not from a human, researchers often tried to distinguish bots from normal accounts (Chu et al., 2010; Grier et al., 2010). Looking only at new adopters of innovations, bots may be included in the data but are not overrepresented or relevant in this study.

2.2. The generation and adoption of innovations

Linguistic innovations can be defined as new forms of language that did not exist before. These phenomena are non-standard in current linguistic norms and thus are somehow innovative. Examples of linguistic innovations are new abbreviations or idioms that have a specific meaning depending on their context. Another example of linguistic innovation would be the use of non-standard orthography. Many new words and constructions on the internet are characterized by strongly informal or spoken language (Crystal, 2011). Eisenstein (2013) analyzed “bad language” on the internet and referred to Jones (2010) to explain the use of non-standard spelling on the internet. Jones (2010) found in her study about spelling on the internet that most people explain non-standard spelling by saying that people are unsure of the correct spellings, it is faster, it has become the norm, or people want to represent their own dialects or accents with non-standard spelling. Another study by Finin et al. (2010) argues that the limit of characters per tweet is often the reason why non-standard acronyms and aberrations are commonly used on Twitter. In Summary, there are several reasons for the generation of linguistic innovation on the internet, but all of them lead to an ever-growing sample of new language forms.

There are several theories that explain the adoption process of innovations, such as

2. Theoretical Background

the threshold model and the social learning theory. Rogers (2003, p. 356) explicates that “the threshold models assume that an individual decision to adopt an innovation depends on the number of other individuals in the system who have already made the behavior change”. The threshold is reached when an individual is convinced to adopt an innovation, because a minimum number of individuals in his or her personal network have adopted the innovation (Rogers, 2003). He explicitly differentiates between the threshold, which occurs at an individual level and the critical mass which operates at the system level. The critical mass refers to the number of adopters before the rate of diffusion begins to accelerate and this “tipping point” is visible in the typical s-shaped curves of diffusion (Maybaum, 2013).

On contrary, the central idea of social learning theory is that one individual observes another individual’s behavior and then does something similar (Rogers, 2003, p. 341). In this respect, the process of the adoption of an innovation is more reflexive.

In both theories however, the adoption of innovations is based on the contacts and interactions in a social network. The language used by individuals is largely influenced by their social network. The more speakers use a linguistic innovation, the higher the probability that an even larger mass will adopt the new language form. Exactly the latter premise is the basic assumption of the Hawkes process described in chapter 2.4.

Regarding the adoption of new linguistic norms, it has been shown that in online communities the probability of the adoption of innovations is related to the longevity of a user in the community (Danescu-Niculescu-Mizil et al., 2013). Danescu-Niculescu-Mizil et al. (2013) identified a two-stage lifecycle. First the users adopt the language of the community, which the authors described as a “linguistically innovative learning phase”. It is followed by a “conservative phase” in which users do not adopt the evolving language norms anymore and stop changing in this regard.

2.3. Diffusion of Innovations

Rogers (2003, p. 11) defines the diffusion of innovations as “the process by which (1) an innovation (2) is communicated through certain channels (3) over time (4) among the members of a social system”. Those four marked elements in the previous definition, can be found in every diffusion research study. The main elements of the diffusion of new innovations is described in the following in more detail:

(1) According to Rogers (2003), an innovation is an idea, practice, or object perceived as new by an individual. Many innovations whose diffusion has been analyzed are technological innovations. But there are also innovations of a different nature, such

2. Theoretical Background

as linguistic innovations, which are analyzed in this study.

(2) Rogers (2003, p. 36) describes the communication channel as “the means by which messages get from one individual to another”. He differentiates between mass media channels and interpersonal channels. Mass media channels involve a mass medium, which enable one or a few individuals to reach an audience of many, such as radio, televisions or newspapers. On the other hand, interpersonal channels involve a face-to-face exchange between two or more individuals. Additionally, he mentions that interactive communication via the internet has become more important for the diffusion of innovations in recent years. This study is based on the micro-blogging platform Twitter, which is a mass media channel.

(3) Rogers (2003) specifies time as the third element in the diffusion process, which allows us to categorize adopters and to draw diffusion curves. For example in the innovation-decision process time can be measured by observing the time dimensions an individual requires from first knowledge of an innovation through its adoption. The inclusion of time in research can also be measured through the innovation’s rate of adoption in a system, which corresponds to the number of members in a system who adopt the innovation in a given time period (Rogers, 2003). Time also plays a central role in inferring the unknown linguistic network with a time-dependent Hawkes process, as it is done in this study.

(4) Finally, diffusion always occurs in a social system, which is defined by Rogers (2003, p. 23) as “ a set of interrelated units that are engaged in joint problem solving to accomplish a common goal”. Individual, groups or organizations can be the members of a social system. The system analyzed in this study consists of all Twitter users tweeting in Spanish and could be located in one of the regions specified in chapter 3.2.

Many innovations whose diffusion has been analyzed are technological innovations, because economists have been interested in understanding diffusion of new technologies across firms and industries. But originally, diffusion theory has its origin in the epidemic approach which was based on theories modeling the spread of diseases (Baptista, 2001). The research of the diffusion of innovation has involved various disciplines, such as sociology, public health, marketing, communication and geography. Although the aspect of geography in the overall research field of innovation diffusion is rather smaller-sized, already in 1995, there were 160 diffusion studies by geographers (Rogers, 2003). Those studies mainly focus on space as a factor affecting the diffusion of innovations, specifically analyzing the effect of spatial distance on diffusion (Rogers, 2003).

In the following two subsections, theories of innovation diffusion in time and in space will be elaborated in more detail.

2.3.1. Theories of linguistic innovation diffusion in time

By analyzing the time element of the diffusion process, one can draw diffusion curves and thereby gain insight into the frequency of innovation over time. Moreover, those temporal diffusion curves allow to categorize adopters in different categories. Innovators can then be identified and differentiated by the adopters of an innovation. Rogers (2003) observes that the cumulative number of adopters of non-linguistic innovations over time forms an S-shaped curve. The adoption of diffusion starts slowly, then the percentage of adopters quickly increases until the number of new adopters finally slows down again at the end. Similar findings are observed by Labov (2001) regarding the temporal diffusion of linguistic change, which also follows a logistic or “s-shaped” progression with respect to time.

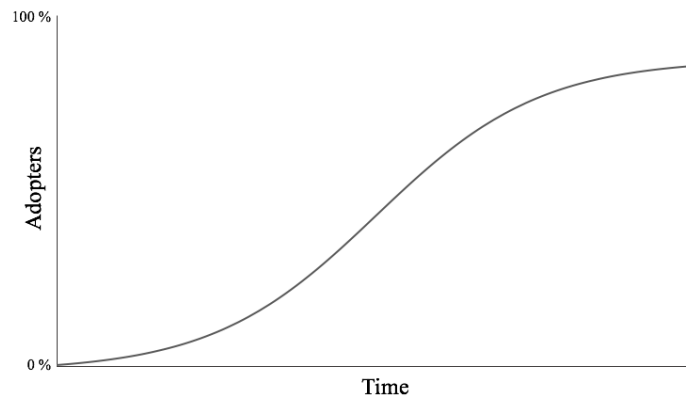


Figure 2.1.: S-shaped curve of innovation diffusion, adopted from Rogers (2003)

When Labov (2001) talks about the s-shaped curve, he refers to the more lasting innovations that end up spreading among the whole population. However, in this work and generally in Twitter, we are looking at colloquial speech, which has on one hand a very fast pace and on the other hand might not spread outside the internet. For example, Danescu-Niculescu-Mizil et al. (2013) show that linguistic innovations might quickly become dated. By analyzing newly emerging language forms in Twitter, it is also possible to observe the decline of the innovations.

In this regard, Tredici and Fernandez (2018) differentiate between successful and unsuccessful innovations. Successful innovations show the described S-shaped curve, whereas unsuccessful innovations can either present a flat dissemination trajectory or have a peak which is followed by a sudden decrease with no stable recovery (Tredici and Fernandez, 2018). To identify successful innovations, Tredici and Fernandez (2018, p. 6) define a slope index which is "based on the dissemination slope of a term, computed as

2. Theoretical Background

the difference between its average dissemination value in the first six months and in the last six months in the dissemination trajectory vector".

The frequency distribution of the number of adopters per time period follows a normal, bell-shaped curve, which is just a different way displaying the same data (figure 2.2). However, this visualization simply allows to categorize adopters into five categories: innovators, early adopters, early majority, late majority and the laggards (Rogers, 2003).

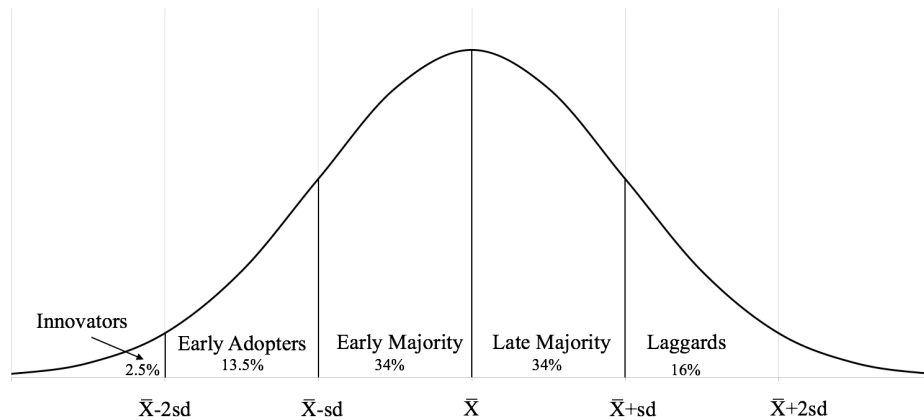


Figure 2.2.: Bell-shaped curve of innovation diffusion, adopted from Rogers (2003)

Rogers (2003, p. 221) defines the rate of adoption as “the relative speed with which an innovation is adopted by members of a social system”. According to Rogers (2003, p. 221) the rate of adoption is measured as “the number of individuals who adopt a new idea in a specified period, such as a year”.

2.3.2. Theories of linguistic innovation diffusion in space

Recent research has also addressed the geographical diffusion and distribution of linguistic innovations. The wave model proposes that the spread of language change behaves as a simple wave, diffusing radially from a central location, which is the point of origin where the source of the innovation has started and the most recent development has occurred (Bailey, 1973).

2. Theoretical Background

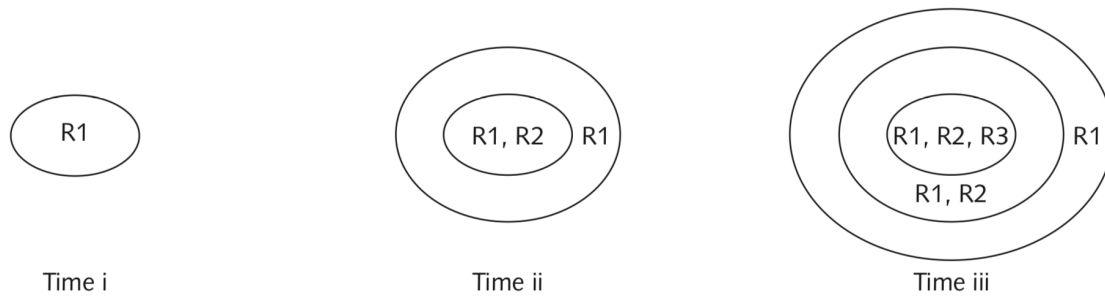


Figure 2.3.: The wave model of linguistic diffusion from Wolfram and Schilling-Estes (2003, p. 714)

However, other studies argue that language change diffuses unevenly over space and affects some communities before others (Boberg, 2000). In these theories, the adoption of linguistic innovation generally follows an urban hierarchy, which is why those theories are named hierarchical models (Boberg, 2000). For example the gravity model is one of them, which not only includes physical distance but also population. This model supposes that the likelihood of contact between individuals from different cities depend on their distance as well as the size of their population (Eisenstein et al., 2014). As a consequence, linguistic innovations spread through large cities first, only later reaching the less urban areas (Trudgill, 1974). This model assumes that language change spreads in a predictable order, where the influence of one city to another is proportional to the population of the city and inversely proportional to the distance between them (Labov, 2003). Figure 2.4 illustrates the gravity model: the larger the circle size, the higher the population density (Wolfram and Schilling-Estes, 2003).

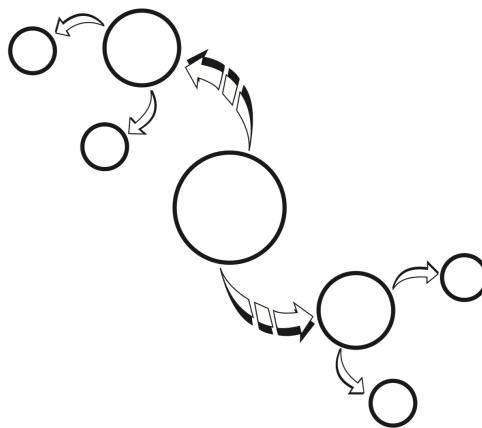


Figure 2.4.: Hierarchical structure of linguistic diffusion from Wolfram and Schilling-Estes (2003, p. 724)

2. Theoretical Background

Wolfram and Schilling-Estes (2003) argue that linguistic innovations often spread in a hierarchical order due to the fact that in denser populated areas more interpersonal contact takes place, which promotes the diffusion of innovations. Another hierarchical model is the closely-related but less specific cascade model, which supposes that language change proceeds from the largest city to the next largest city, passing over sparsely populated regions (Eisenstein et al., 2014; Labov, 2003). The focus is on differences in population rather than on physical distance. Large cities have a great influence on smaller ones and the rural area in between is rarely affected. In this case, linguistic innovations strictly diffuse from larger cities to smaller ones (Wolfram and Schilling-Estes, 2003).

In several studies a hierarchical diffusion of linguistic innovation was observed. Nevertheless, these models have their limitations. Trudgill (1983 in Wolfram and Schilling-Estes (2003, p. 726)) points out that there is need to include other factors than distance and population into the model. For example, he states that the possibility of the adoption of an innovation is higher if dialects are similar. Therefore, Trudgill (1983 in Wolfram and Schilling-Estes (2003, p. 726)) proposes to include a structural similarity factor into the model. Another weakness of the model was mentioned by Gregory (1985; 2000 in Britain (2018, p. 481)), who explains that the gravity model depends on Euclidean distance and does not consider the spatiality of that distance. For instance, physical objects such as mountains or rivers can increase the walking distance between two apparently close locations. The gravity model assumes a planar space and that the accessibility of locations in space is always equal.

There are also other, less popular patterns of linguistic diffusion which have been observed, such as the contra-hierarchical model of Bailey et al. (1993). The authors discover that whereas some linguistic innovations diffuse hierarchically, others diffuse in a contra-hierarchical way, moving from rural to urban areas. Especially the language features representing the revitalization of traditional norms have their origin in more rural areas and only later spread to large metropolitan areas (Bailey et al., 1993). In addition, Horvath and Horvath (1997) propose a cultural hearths model which assumes that a linguistic innovation first spreads within one cultural region and then moving on to the next.

This effect of “cultural geography” has also been recognized by other studies. Particularly, Eisenstein et al. (2014) present a model capable of identifying the demographic and geographic factors that drive the spread of newly popular words on online platforms. The geographical representation and analysis of such data have shown that cultural factors play an important role in the diffusion of language change (Eisenstein et al., 2014).

2. Theoretical Background

Similarly, Grieve et al. (2018) present evidence that cultural patterns appear to be an important predictor of lexical influence. They propose a method for mapping lexical innovation, which they use to track the origin and spread of new words on Twitter. They found that the origin of lexical innovations is usually in urban areas and African American English is a main source of newly emerging words on American Twitter. Those emerging words then first tend to spread within the cultural region from which they originate, e.g. within the African American culture, before diffusing further (Grieve et al., 2018). Relatedly, Boberg (2000) analyzes transnational linguistic diffusion at the national boundary between the United States and Canada and observes that the linguistic diffusion is slowed down by the border, which also suggest culture being a main driver of linguistic diffusion. In summary, the literature has shown that the diffusion of linguistic innovation is mainly characterized by physical distance, the population density and cultural patterns (Grieve et al., 2018).

But why do those factors of geography, population and culture play a role in the spatial diffusion of linguistic innovations on online platforms? The role of geographical distance in linguistic change has often been explored and although interactions in online media should principally not be affected by space, physical distance plays a role in the diffusion of linguistic changes in social media. Mainly because the probability of interaction and a subsequent friendship generally increases as the distance between individuals decreases (Sadilek et al., 2012). Therefore social networks in online media depend on space as individuals who are geographically proximate will be more likely to be connected on social platforms.

A similar argumentation describes the importance of population size for the diffusion of innovations. Since contacts and interactions between people are a prerequisite for the transmission of linguistic innovations, living in regions with a high population density increases the probability of interactions and accordingly also promotes the diffusion of innovations.

Lastly, the role of culture in the diffusion process can be described by the concept of heterophily and homophily introduced by Rogers (2003). He defines heterophily as “the degree to which two or more individuals who interact are different in certain attributes, such as beliefs, education, social status, and the like” (Rogers, 2003, p. 36). Homophily is described as the opposite, meaning that those individuals are similar in certain attributes. He concludes that individuals tend to be linked to others who are relatively homophilous in social characteristics (Rogers, 2003). Supplementary, socially connected individuals tend to use language in similar ways, which is known as linguistic homophily (Balusu et al., 2018; Yang and Eisenstein, 2017). Hence, culture is important because

2. Theoretical Background

of the assumption that people tend to have more friends who are similar to themselves, also regarding the culture. In summary, contacts and interactions between people are the main condition for the diffusion of linguistic innovations, and these are higher (1) between people who are spatially close, (2) between people who live in densely populated areas, and (3) between people who have similar cultures.

2.4. Time-dependent Hawkes process

2.4.1. The concept of the Hawkes process

Point processes basically describe the timing of events. A homogeneous Poisson point process is the simplest class of point processes and assumes a constant intensity of events over time. Events are independent from each other. Rizoïu et al. (2017) describe them as memoryless, because future events depend not on information from further in the past.

The Hawkes process, however, is a point process indexed by time which counts the number of events over time. Observing an event increases the likelihood of observing similar events in the future. This dependency of future events on past events is the reason why the Hawkes process is considered as a self-exciting point process (Rizoïu et al., 2017). In summary, the Hawkes process is a non-homogeneous Poisson process, in which the intensity depends on previous events (Rizoïu et al., 2017).

In a mathematical formula the Hawkes process can be defined as (Rizoïu et al., 2017):

$$\lambda(t) = \lambda_0(t) + \sum_{i:t>T_i} \phi(t - T_i)$$

where

- $T_i < t$ are all the event time having occurred before current time t .
- $\lambda_0(t)$ is the base intensity function. It describes the arrival of events not triggered by previous events but by external sources. Thus, the baseline intensity measures the level of exogeneity of a node. In other words, Bacry et al. (2020, p. 2) explained it as an intensity that indicates "the spontaneous apparition of an action, with no influence from other nodes of the network".
- ϕ is the memory kernel, which is described in more detail in the next section.

2. Theoretical Background

Figure 2.5 shows a theoretical visualization of a multivariate Hawkes process model by Nickel and Le (2020). The input parameter base intensity is the mutual excitation of a node. The resulting network can be inferred from the observations of events shown in the plot below.

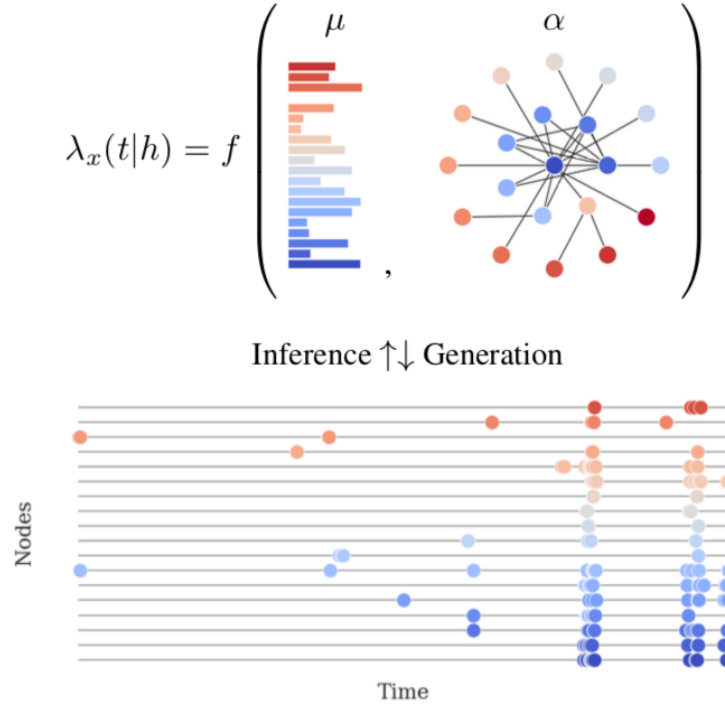


Figure 2.5.: Visualization of multivariate Hawkes process model from Nickel and Le (2020, p. 1)

The result of the Hawkes process is a weighted asymmetrical adjacency matrix, which includes the values indicating the level of interaction between nodes (Bacry et al., 2020). In other words, the matrix indicates the strength of influence each region has on other regions (Alvari and Shakarian, 2019).

2.4.2. The intensity function

Rizoiu et al. (2017, p. 6) define a self-exciting point process as "a point process in which the arrival of an event causes the conditional intensity function to increase". So, the intensity function depends on all previously occurred events and is expressed through the kernel function ϕ .

Often the kernel ϕ is monotonically decreasing after an event occurred. As a consequence, more recent events have a higher influence on the event intensity, whereas events

2. Theoretical Background

having occurred further away in the past have a lower influence on the current event intensity (Rizoïu et al., 2017).

There are different options to model the decay of the intensity function after an event occurred. Although the intensity function does not have to be monotonically decreasing, the influence of an event often decreases over time which is why typically decreasing intensity functions are modelled.

The exponential function is a popular decay function, which is defined as follows:

$$\phi(x) = \alpha e^{-\delta x}$$

Figure 2.6 graphically visualizes a Hawkes process with an exponential kernel. The first graphic (a) shows exemplary nine events that occurred over time. The times T_1, T_2, \dots, T_9 indicate the times an event was observed and $\tau_1, \tau_2, \dots, \tau_9$ represent the corresponding inter-arrival times. The second graphic (b) visualizes the counting process over time. Every time an event occurs the count increases by one unit. On the third graphic (c) the intensity function over time is shown. Visibly, each event provokes a jump in the intensity function, followed by an exponential decay in this example.

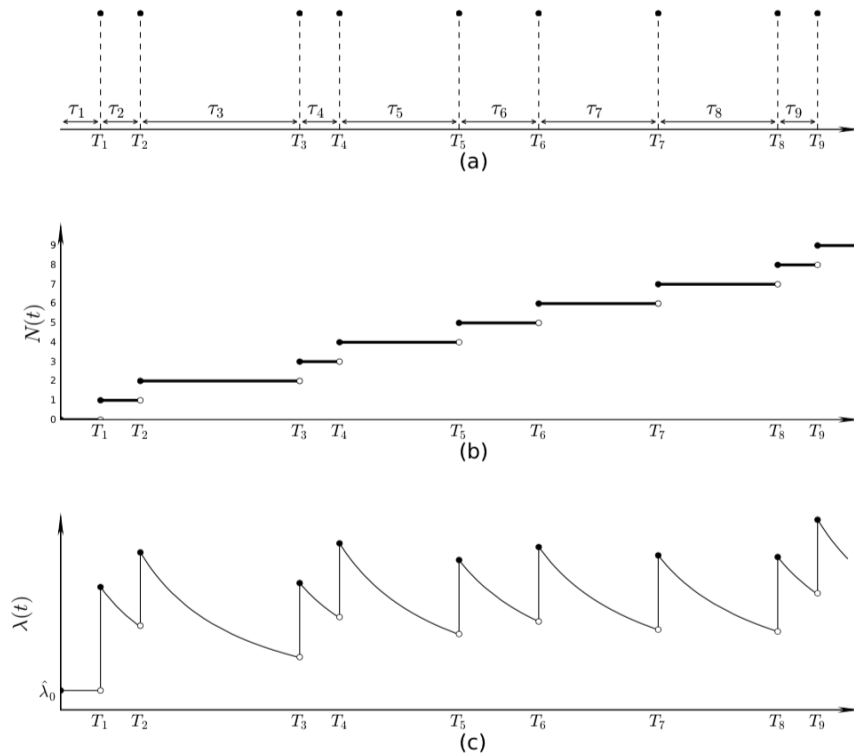


Figure 2.6.: Hawkes process with exponential decays from Rizoïu et al. (2017, p. 7)

2.4.3. Historical applications of a Hawkes process

In 1971, Alan G. Hawkes introduced a self-exciting point process which was named after him (Hawkes, 1971). Hawkes processes have been applied in a wide variety of settings to describe and predict the behaviour of data. Hawkes (1971) himself gave the example of an application as an epidemic model in 1971, because a high number of infections increase the probability of further infections.

The Hawkes process can also be used to describe earthquake dynamics (Ogata, 1988, 1999). Because of geophysical processes, an earthquake in a specific region increases the probability of another earthquake in that region. Thus, for example Helmstetter and Sornette (2003) apply the Hawkes process to forecast where and when aftershocks will occur.

Another application field of Hawkes processes is finance. For instance on the stock market, past buy and sell transactions influence future prices and volumes of such transactions (Bahamou et al., 2019; Rizoïu et al., 2017). Bacry et al. (2015) summarize the applications of Hawkes process in finance based on recent academic literature and give an overview over Hawkes processes used in models of market activity and risk.

Additionally, Hawkes processes were used for crime modeling (Reinhart and Greenhouse, 2018; Stomakhin et al., 2011; Zhu and Xie, 2019). For instance, Stomakhin et al. (2011) assume that gang crimes in Los Angeles follow a temporary dependent point process, so that the occurrence of a crime increases the likelihood of subsequent crimes in the future. Based on this assumption, they predict gang-related crimes in Los Angeles.

Finally, the Hawkes process is applied to model information diffusion and the popularity of online content. For instance, Zhao et al. (2015, p. 1513) argue that the Hawkes process "is ideal for modeling information cascades in networks because every new re-share of a post not only increases its cumulative reshare count by one, but also exposes new followers who may further reshare the post". Kobayashi and Lambiotte (2016) and Zhao et al. (2015) focus on the temporal patterns of retweet activities on Twitter. Both try to predict the number of retweets of original tweets by a Hawkes process. Hawkes processes were also used to model the spread of rumors and fake news on online social networks (Farajtabar et al., 2017; Nie et al., 2020). Using a multivariate Hawkes process, Nie et al. (2020) were able to model the dynamics of user influence in rumor propagation.

3. Data

This chapter describes in more detail the data used to analyse the diffusion of Spanish linguistic innovations. Besides explaining where the data comes from and what regions and linguistic phenomena this study focuses on, the following subsections present some figures on the data collected.

3.1. Data collection

Linguists from the University of Zurich have collected Twitter data from the online archive of the general Twitter stream (Internet Archive, 2020). About 1% of all tweets are available there - the exact rate is unknown. For the analysis, tweets from 2012 to 2019 are used. Due to the enormous amount of tweets and the time-consuming preprocessing, only the first seven days of each month were collected and included in the analysis.

The twitter archive is not complete and there are missing data files in some weeks. In order to minimize missing values, sometimes not only the first seven days were considered, but also other days of the months or even days from another month which are close in time. A detailed table in the appendix B documents all time periods considered in this analysis.

The tweets were filtered by language since only Spanish tweets are considered. For entries dated after April 2013, tweets with the attribute "language" = "es" are selected. For those dated before April 2013, all tweets with the attribute "user_language" = "es" are selected. In addition, tweets that are recognized as Spanish by the Python library langdetect are also included in the analysis. Due to these different selecting methods, the resulting filtered dataset may erroneously contain some non-Spanish tweets. In the end however, only defined Spanish innovations are filtered - a few misclassified tweets do not affect the results of the analysis.

On Twitter, users have the possibility to re-share tweets posted by their social connections. As a consequence, a tweet can be shared by a lot of different users and reach a wide network. Although the dataset has information about whether a tweet is a retweet or not, I do not consider this differentiation for two reasons. On the one hand, some

3. Data

innovations are extremely rare. Ignoring the retweets would result in an uninteresting sample. On the other hand, I argue that even sharing a tweet with an innovation is a use of that innovation. So if a new user shares an already existing tweet with an innovation, it is also an adoption of the innovation. As a result, innovations spread and reach more and more other users.

For each tweet the online archive provides a lot of information. For this work, only the information about the tweet id, the user id, the time of tweet, the content of tweet and some spatial information is relevant.

Note, on Twitter users can follow and be followed by other users. However, the data has no information about the social connections or the number of followers of a user.

3.2. Study area

Figure 3.1 shows the geographical distribution of geolocated Spanish tweets in 2012 on the globe. Not surprisingly, most tweets were posted in Spain, South America and Central America. However, many Spanish tweets occurred also in the United States and the non-Spanish speaking Europe.

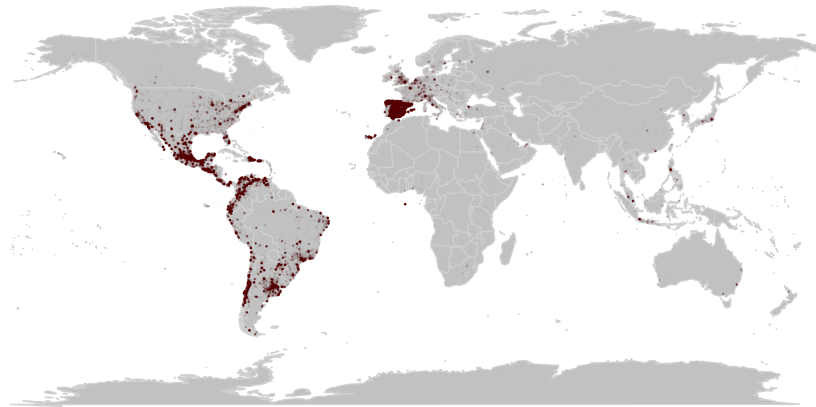


Figure 3.1.: Geographical distribution of geolocated Spanish tweets in 2012 in preprocessed data

The plots and analysis of the geospatial information provided in 2012 helped to identify the regions on which to focus this study. The initial idea was to focus on about 20 metropolitan areas where the most Spanish tweets originated from. In many regions, however, little to no innovations occurred. Consequently, I decided to work with less but larger regions in order to improve the data availability.

3. Data

Finally, I focused on six regions in Spain and Latin America shown in figure 3.2. In particular, the Spanish regions Andalucia, Catalunya and Madrid, as well as the countries Mexico, Argentina and Uruguay were used for the spatial aggregation. Only tweets which are located in these regions have been considered for the analysis.

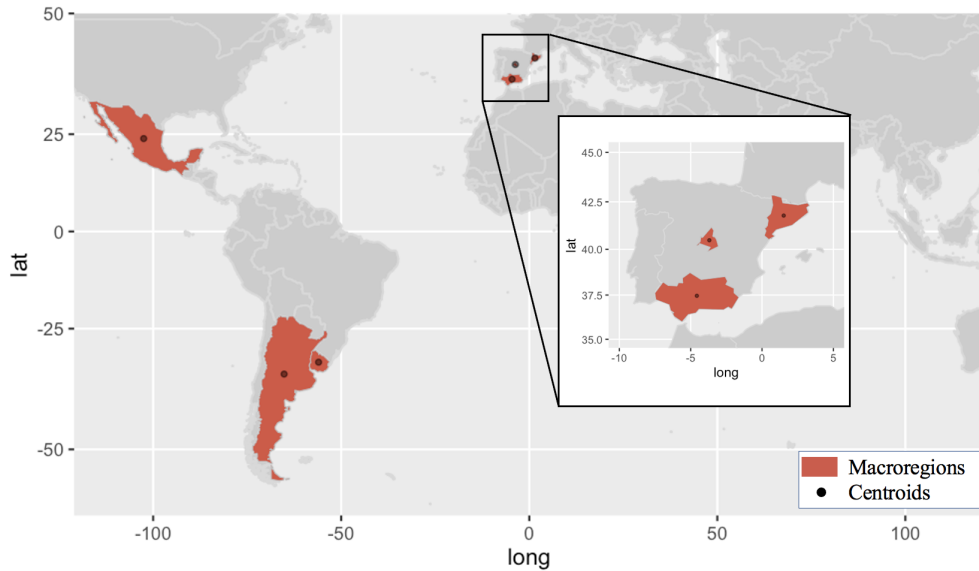


Figure 3.2.: Overview Spanish speaking regions covered in this thesis (Esri, 2015; Eurostat, 2020)

Andalucia, Catalunya and Madrid are the most densely populated regions in Spain and the majority of active Twitter users are based in these regions. Spain was deliberately not grouped into one spatial bin, but split up into these three regions in order to study the linguistic influence of the country in more detail. Mexico is the biggest Spanish speaking country in the world. However, Argentina has more active Twitter users. Uruguay consists of a much smaller area and population than Argentina but both countries have very engaged Twitter users; Their average user posts about the same amount of tweets.

For the spatial data of Argentina, Mexico and Uruguay, I used a polygon dataset from all countries in the world (Esri, 2015). The Spanish regions Andalucia, Catalunya and Madrid were extracted from a dataset of the European Union which contains the administrative units of Spain (Eurostat, 2020). The two sub-datasets were joined into one dataset containing spatial boundaries of all six regions. The polygons describing the boundaries were slightly simplified since less detailed polygons speed up the spatial

3. Data

binning process.

For all regions, the location of the centroids were also calculated for the following two reasons. Firstly, many spatial visualizations can be created more easily if spatial information is provided with points rather than polygons. Secondly, the spatial distance between the regions is needed for some descriptive statistics in this study.

Table 3.1 lists the Haversine distance between the centroids of all regions in kilometer. The Haversine distance function determines distances between two points and accounts for the effects of the Earth’s spherical space (Kamath et al., 2013).

Table 3.1.: Distance matrix in km

	Andalucia	Argentina	Catalunya	Madrid	Mexico
Argentina	10'207				
Catalunya	709	10'916			
Madrid	344	10'496	460		
Mexico	9'087	7'657	9'353	9'020	
Uruguay	9'431	893	10'139	9'732	7'987

The greatest distance is between Argentina and Catalunya and the smallest distance is between Andalucia and Madrid. The selection of the regions leads to the fact that there are some geographically close regions with small distances but then there are also very large distances between the regions on different continents.

3.3. Tweets in space and time

All tweets that could be located in one of the six regions were stored in a database. Finally, in total the database consists of 57'779'055 tweets from 6'664'596 unique users.

For first visualizations, the data was weekly aggregated in 96 temporal bins. Weekly binning is the same as monthly binning in this study because only the first seven days of each month are included in the dataset. The number of tweets per temporal bin is visualized in figure 3.3.

3. Data

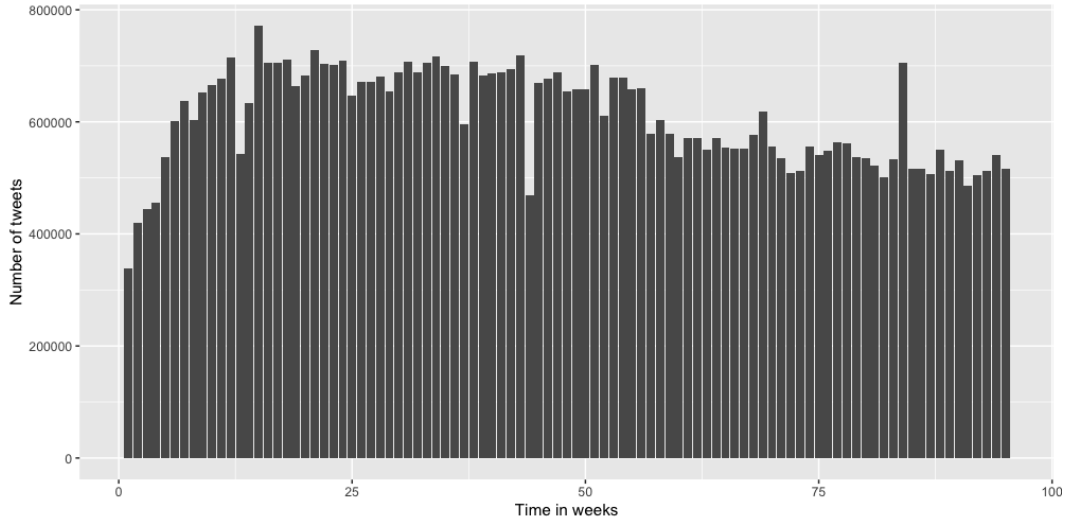


Figure 3.3.: Number of tweets over time in the database

In the beginning of 2012, slightly less data is available. This is probably because of less and inaccurate geoinformation. In the years 2013 and 2014, the API sampling rate provided more tweets than in the following years. Therefore, there is a tendency for a slightly higher number of tweets in the early years. Over all years there is an average of 608'200 tweets per week in the dataset.

Some of the outliers were examined more closely, but no clear reason was found why in the given time periods less or more tweets occurred. It is either due to the varying sampling rate of the API or because of the varying availability of geoinformation.

Table 3.2 summarizes the number and percentage of tweets as well as the number and percentage of users per region. In addition, the ratio "tweets per user" is listed in the table indicating how actively Twitter is used in a region.

Table 3.2.: Tweets and users per regions in the database

Regions	Number of Tweets	Percentage of Tweets	Number of Users	Percentage of Users	Tweets per User
Andalucia	5'431'520	9.4%	689'551	10.2%	7.9
Argentina	26'352'076	45.6%	2'517'240	37.4%	10.5
Catalunya	2'624'313	4.5%	381'045	5.6%	6.9
Madrid	4'395'978	7.6%	523'652	7.8%	8.4
Mexico	16'914'356	29.3%	2'425'820	36.0%	7.0
Uruguay	2'060'812	3.6%	201'156	3.0%	10.2

Most tweets in the dataset are located in Argentina, followed by Mexico with the

3. Data

second most tweets. Around 85% of all tweets are located in the latter two regions. The number of users per regions show that 73% of all users can be located either in Argentina or Mexico, slightly less but still a very large proportion. Interestingly, the last column in the table shows that the ratio of tweets per user looks more similar across all regions. The ratio indicates the average number of tweets by a user in a given region, which is between 7 and 10 tweets in all regions. Argentina and Uruguay have the highest ratio of tweets per user. Twitter users are extremely active there - the average Twitter user posts around 10 Tweets in my dataset. In the Spanish regions Andalucia, Catalunya and Madrid the average user posts around 7.7 tweets.

3.4. Analyzed linguistic phenomena

The tweets were thematically binned by looking for all tweets which include given linguistic innovations. Linguists from the University of Zurich had processed the text of the tweets to identify innovations and spatial attributes. They compiled a list of innovations, which contains different newly popular constructions in colloquial Spanish.

The number of innovations considered in total is 33. All innovations were selected and defined by linguists according to their own assumptions. A survey in the linguist's own Twitter network has led to a large selection of new language forms in colloquial Spanish. However, since a large part of the linguist's network is based in Spain, there is a bias in the selection of innovations. Thus, the data is more likely to include innovations which originated or successfully spread in Spain. There are a few innovations which only occurred in the three Spanish regions on the European continent. In contrast, there are only two innovations that can only be found in Latin America or Central America.

A detailed list describing all analyzed innovations can be found in appendix A. In the following chapters, the type of innovation is mainly referred by its identification number, the meaning of which is given in the list in appendix A. Besides explaining what they mean, what they are used for and why they have become popular (if relevant or known), the innovations are classified in different types of linguistic innovations:

- Morphosyntactic innovations: They do not follow the grammatical system of Standard Spanish.
- Non-standard orthography: They are characterized by innovative spelling.
- Phraseology: They are about longer utterances that affect more than one word and are instances of repeated phrases.

3. Data

They might be

- a) fixed: They are always the same (or with minimal variations, either related to their spelling or the inflection of nouns, verbs, etc.);
- b) productive: They have a “free slot” that the speaker fills with new material;
- c) productive multimedia: The free slot is filled by multimedia material, typically pictures or gifs, but might also be other tweets (links).

The innovations analyzed in this work propagate within a relatively short time. The focus lies on a time span of a few years rather than historical time periods. Thus, it is not always obvious whether or not I have the start point and blooming period of the innovations covered in the dataset. Nevertheless, when selecting the innovations, the linguists focused on innovations which mainly became popular after 2012.

4. Methodology

To answer the question of how linguistic features in Spanish spread in space and time, different approaches were used. Section 4.1 contains the documentation of how the Twitter dataset was preprocessed. Section 4.2 mainly introduces the measures proposed by Kamath et al. (2013) to study the spreading of hashtags, which were adopted in this thesis to investigate the role of geography in the diffusion process. The final section 4.3 describes the Hawkes process model that was used to infer the network of linguistic influence.

4.1. Preprocessing

Once the tweets and innovations were linguistically preprocessed by the linguists, my part was to spatially bin the data. The subsection thereafter describes how the geoinformation was extracted from the tweets. Then, the next subsection specifies how all innovations were filtered from the entire dataset, resulting in the final input dataset for the analysis.

4.1.1. Spatial binning

As described in chapter 2.1.2, different options exist to provide spatial information within Twitter. In this work, I use the georeferenced tweets which contain either exact coordinates or a bounding box of the locations. Unfortunately, only few tweets provide specific geographical information such as the explicit coordinates. Thus, I also use the information provided in the “user_location”-field in the user profile. It is a free text attribute, which is non-mandatory and user-provided. While some user provide their home country, state or city, other users provide nonsensical information (e.g. "between heaven and hell", "Bailando bajo la lluvia"). To distinguish between useful and nonsensical information, the geo_tweets were further processed: “user_location” has first been cleaned by removing emojis, smileys, and unnecessary punctuation.

All six regions were defined by regular expressions (Regex) resulting in a long string of the most important city names and abbreviations of the given region. These resulting six Regex strings then allow to search for relevant geoinformation in text format.

4. Methodology

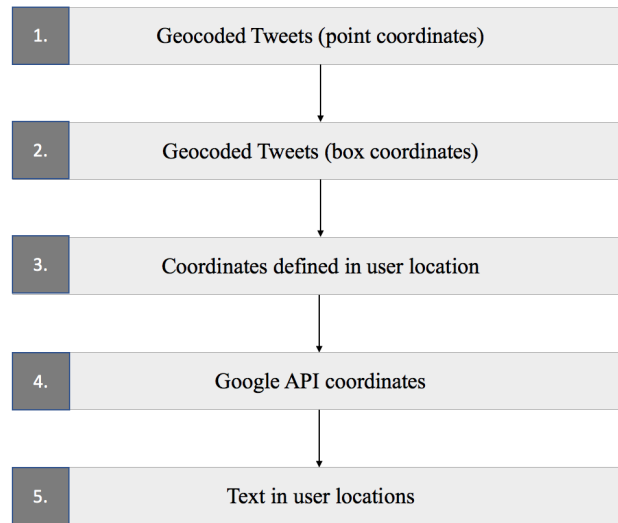


Figure 4.1.: Procedure of the spatial binning

The procedure of spatial binning is visualized in diagram 4.1. The binning is structured according to the priority of the geoinformation. If exact longitude and latitude coordinates were added directly to the tweet, they are considered with priority while all content of the user profile is ignored. So the more precise the geoinformation, the higher its priority in the spatial binning. The lowest priority has the text information in the user profile (user location).

First, when point coordinates were assigned to tweets, it was checked if they are located within the six considered regions. Therefore, the simple point-in-polygon-algorithm "over" of R was used, which tests if a point is inside a polygon.

Second, the processing of the bounding box coordinates was done in a similar way. To simplify the process, the focus was only on the lower left corner of the bounding box, so only the minimum latitude and the minimum longitude coordinate were taken into account. As before, it was tested whether this coordinate is located in one of the regions with the point-in-polygon-algorithm "over".

Third, some users define their location in their user profile, not only by specifying the location in text form, but by indicating the location with coordinates. These coordinates were also taken into account and checked in the same way as before whether they are located in one of the relevant regions.

Fourth, the Google API was applied to place names whose ambiguity could not be resolved by additional references. Some location names appear more than once on the globe, making it difficult to identify them geographically. This is for example the case

4. Methodology

for "Puebla" which often stands for the city Puebla in Mexico, but can also refer to numerous places with the name "Puebla de XXX". Finally, the preprocessed regions from the user location field were taken into account.

By including the defined location in text format of the user profile, I make the simplifying assumption that all tweets from a user can also be located at the location defined in his or her user profile. This assumption will certainly lead to some miss-located tweets, for example if a user actively tweets during his or her holidays in a foreign place. Nevertheless, most tweets of a user are probably posted in the user's home location.

By far the most tweets could actually be localized through this option. About 93 percent of all tweets with geoinformation were localized by the text in the user location field, whereas only about 5 percent were localized by either point or bounding box coordinates.

4.1.2. Innovation filtering

Once all tweets were spatially binned and stored in the database, I simply filtered all tweets with an innovation. The result is a dataset with 9'840 rows, where each row is a tweet with an innovation. As mentioned before, in total over 57 million tweets are stored in the database and thus could be located in one of the six regions. Comparing these numbers, it turns out that only one out of 5'871 tweets contains an innovation analyzed in this work.

However, in this study not the number of tweets with an innovation is relevant, but the number of unique users who adopt an innovation. The tweets with innovations were further filtered in a way that only innovations which had been adopted by new users were selected. In other words, if a Twitter user tweeted a specific innovation more than once, only the first usage of that given innovation and user is relevant. In numbers this means that the dataset was reduced from 9'840 to 9'618 rows. 222 innovations were used multiple times by the same user.

A sample of the final dataset used for the following analysis is listed in table 4.1. The table shows how the data is structured and which information it contains. The column "Tweet ID" contains a unique id for each tweet. The column "User ID" provides information about the user who posted or shared and thereby adopted an innovation. The third column, named "Time created at", notes the exact time when a tweet was posted. This information is important for the time-dependent Hawkes process. Since the data was binned temporally, spatially and thematically, each row contains the information in which month the tweet occurred, in which region it occurred and which innovation it is.

4. Methodology

Table 4.1.: Sample of final dataset after preprocessing

Tweet ID	User ID	Time created at	Temporal Bin	Spatial Bin	Innovation Bin
201201051224_0141	127554363	Thu Jan 05 19:24:40 +0000 2012	1201	Mexico	inno_44
201506071239_0163	519655844	Sun Jun 07 18:39:21 +0000 2015	1506	Andalucia	inno_02
201811020509_0087	310830497	Fri Nov 02 11:09:43 +0000 2018	1811	Madrid	inno_05

4.2. Descriptive statistics

After roughly exploring the data by means of plots and simple analysis, I have adopted some methods from Kamath et al. (2013) who examined the spatio-temporal propagation of Twitter hashtags. Some measures used by them are adopted in this study as they enable a deeper understanding of geographical factors in the diffusion of innovations. The measures are explained in the following three subsections in more detail.

4.2.1. Spatial properties of innovation diffusion

Kamath et al. (2013) studied three spatial properties of hashtag propagation, namely the focus, entropy and spread of hashtags. The first two measures, focus and entropy, were adopted from a study investigating the relationship between popularity and locality of online YouTube videos (Brodersen et al., 2012). But there were also previous attempts to measure spatial properties of web resources. For example, Ding et al. (2000) examined the geographic scope of web resources and focused on measuring the uniformity of the distribution of web resources. But in this work, the measures from Kamath et al. (2013) are adopted.

Focus

Kamath et al. (2013, p. 671) define the focus as "the maximum probability of observing the hashtag at a single location". The set of all occurrences of an innovation i in region r is defined as O_r^i . So in a formula, the probability of observing an adoption of an innovation i in region r is:

4. Methodology

$$P_r^i = \frac{O_r^i}{\sum_{r \in R} O_r^i}$$

The value of the focus lies between 0 and 1. The focus can be interpreted as a measure indicating if and how adoptions of innovations are spatially clustered. The higher the focus, the more local and spatially clustered the innovation is.

The innovation focus region is the location in which the most adoptions of a given innovation occur. It can simply be determined by assigning the region with the maximum focus to each innovation.

Entropy

The entropy is defined by Kamath et al. (2013, p. 671) as "the randomness in spatial distribution of a hashtag and determines the minimum number of bits required to represent the spread".

$$\varepsilon^i = - \sum_{r \in R} P_r^i \log_2 P_r^i$$

The entropy shows how dispersed an innovation is. If an innovation appears in many regions, the entropy increases.

Spread

Kamath et al. (2013, p. 671) define the spread as "the mean distance for all occurrences of a hashtag from its geographic midpoint". The geographical midpoint G is calculated by weighting all centroids of the regions with the respective number of adoptions of an innovation of this region. The geomidpoint-Calculator then determines the center of gravity resulting in latitude and longitude coordinates, which describe the geographical midpoint (GeoMidpoint.com, 2021). Once having geographical midpoints for all innovations, Kamath et al. (2013) measure the mean distance for all adoptions of an innovation from its center of gravity.

$$S^i = \frac{1}{|O^i|} \sum_{o \in O^i} D(o, G(O^i))$$

The result is the spread value in form of a number of kilometer for each innovation. If the value is high, the innovation has diffused wide and probably became a global phenomena. On contrary, a low value indicates a small spread from its geographical midpoint and thus can be classified as a more local innovation.

4.2.2. Relationship between regions

Jaccard similarity index

Kamath et al. (2013) also study the global footprint of hashtags by examining the hashtag similarity of a location versus the distance between locations. They use the Jaccard coefficient to identify how similar two samples of hashtags are. The Jaccard similarity coefficient is a method popularly used to compare the similarity of samples. The coefficient is measured by dividing the number of features that are in common in two samples by the number of features that occur either in one or in both samples (Niwattanakul et al., 2013).

Thus, the innovation similarity (or Jaccard similarity coefficient) can be defined as:

$$InnovationSimilarity(r_a, r_b) = \frac{I_{r_a} \cap I_{r_b}}{I_{r_a} \cup I_{r_b}}$$

The sample of innovations used in one region is compared with the sample of innovations used in another region. The resulting coefficient is between 0 and 1. A Jaccard coefficient of 1 indicates that regions have all innovations in common, whereas a similarity score of 0 indicates that those locations share no innovations.

In a second step, this index can be used to determine correlation patterns between the similarity of used innovations and the geographical distance between regions. Technically, the correlation between the distance matrix in table 3.1 and the Jaccard index values in figure 5.11 was measured.

Adoption lag

In addition, Kamath et al. (2013) examine if regions that are spatially close are more likely to adopt hashtags at the same time. They define a so called "adoption lag", which compares the time of the first observation of a given hashtag in a region with the time of the first observation of that hashtag in another region. By accumulating all these time differences of the common hashtag between two regions and then divide it by the number of hashtags that occur in both regions, the result is the mean temporal lag between two regions. Thus, similarly to Kamath et al. (2013), the innovation adoption lag for two locations can be defined as:

$$AdoptionLag(r_a, r_b) = \frac{1}{|I_{r_a} \cap I_{r_b}|} \sum_{i \in I_{r_a} \cap I_{r_b}} |t_{r_a}^i - t_{r_b}^i|$$

In order to calculate the difference between two times, all time units were transformed into numerical values. This is done by counting the seconds from a defined start time,

4. Methodology

also called reference time, to the desired time. High numbers are therefore later on the time axis than low values.

The values of the adoption lag can be interpreted as follows: A high adoption lag indicates there is a large time gap between the occurrences of the innovation in the regions. And conversely, a low value indicates that the innovations occurred about the same time in the regions. The adoption lag can also be compared with the distance matrix to find out whether the two patterns correlate.

4.2.3. Measuring spatial impact

Finally, for the last descriptive methodology I adopt the technique to evaluate the impact a location has on other locations proposed by Kamath et al. (2013).

They define the spatial impact as follows:

"The spatial impact $I_{l_i-l_j}$ of location l_i on l_j is a score in the range $[-1, 1]$, such that -1 indicates l_i adopts a hashtag only after l_j has adopted it, $+1$ indicates l_j adopts a hashtag only after l_i adopts it and 0 indicates the locations are independent of each other and adopt hashtags simultaneously."

(Kamath et al., 2013, p. 675)

The spatial impact measure is based on the cartesian product. The cartesian product of two sets, X and Y, is defined as the set of all ordered pairs (x,y), where x is an element of X and y is an element of Y (Dwyer, 2016). Note that the order of terms within pairs is important. Thus, the pair (x,y) is not the same pair as (y,x).

In this study, an occurrence of innovation i in region r at time interval t can be represented as O_r^i . Using the cartesian product, I define two subsets. First, the preceding innovations which is a set of all occurrences of i in r_a that precede r_b in the cartesian product of their occurrences. Borrowing the notation from Kamath et al. (2013), the preceding set can be defined as:

$$O_{r_a}^i \prec O_{r_b}^i = \{o_{r_a}^i(t) | t_a < t_b \forall (o_{r_a}^i(t_1), o_{r_a}^i(t_2)) \in O_{r_a}^i \times O_{r_b}^i\}$$

Second, the subset of succeeding innovations is a sample of all occurrences of i in r_a that succeed r_b in the cartesian product of their occurrences. It is defined as:

$$O_{r_a}^i \succ O_{r_b}^i = \{o_{r_a}^i(t) | t_a > t_b \forall (o_{r_a}^i(t_1), o_{r_a}^i(t_2)) \in O_{r_a}^i \times O_{r_b}^i\}$$

4. Methodology

In a formula, the spatial impact is defined as:

$$ImpactValue(r^a, r^b) = \frac{\sum_{i \in I_{r_a} \cup I_{r_b}} ImpactValue^i(r^a, r^b)}{|I_{r_a} \cup I_{r_b}|}$$

The spatial impact *ImpactValue* from one region to another is measured for all innovations that occur in both regions.

So first, the *ImpactValue* needs to be calculated for each innovation individually using the following equation:

$$ImpactValue^i(r^a, r^b) = \begin{cases} \frac{|O_{r_a}^i < O_{r_b}^i| - |O_{r_a}^i > O_{r_b}^i|}{|O_{r_a}^i \times O_{r_b}^i|} & \text{if } i \in I_{r_a} \text{ and } i \in I_{r_b} \\ 1 & \text{if } i \in I_{r_a} \text{ only} \\ -1 & \text{if } i \in I_{r_b} \text{ only} \end{cases}$$

For a better understanding Kamath et al. (2013) visualize three examples graphically shown in figure 4.2.

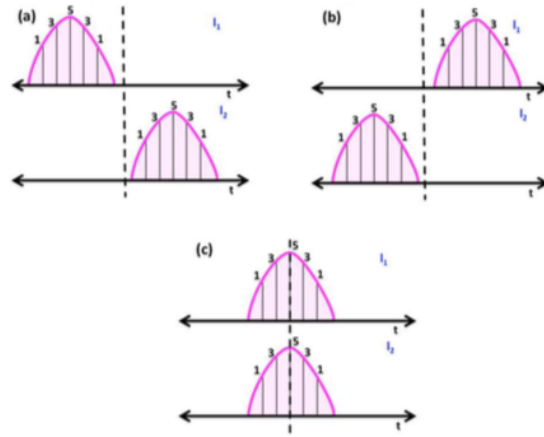


Figure 4.2.: Examples illustrating different cases of spatial impact from Kamath et al. (2013, p. 675)

The first case (a) in figure 4.2 shows the case where location 1 completely influence location 2. The impact value would be 1. In the second case (b) location 2 completely influences location 1 resulting in an impact value of -1. If neither location affects the other, this would be presented as shown in the diagram (c) in figure 4.2.

4. Methodology

In numbers, the cartesian product of the examples in figure 4.2 would be represented as follows:

$$\begin{aligned}
 (a) \quad |O_{r_a}^i < O_{r_b}^i| = 169 \text{ and } |O_{r_a}^i > O_{r_b}^i| = 0 & \quad \text{ImpactValue}^i(r^a, r^b) = \frac{169 - 0}{169} = 1 \\
 (b) \quad |O_{r_a}^i < O_{r_b}^i| = 0 \text{ and } |O_{r_a}^i > O_{r_b}^i| = 169 & \quad \text{ImpactValue}^i(r^a, r^b) = \frac{0 - 169}{169} = -1 \\
 (c) \quad |O_{r_a}^i < O_{r_b}^i| = 62 \text{ and } |O_{r_a}^i > O_{r_b}^i| = 62 & \quad \text{ImpactValue}^i(r^a, r^b) = \frac{62 - 62}{169} = 0
 \end{aligned}$$

Finally, the spatial impacts of a given region are visualized in a one dimensional plot, where the x-axis is dimensioned from -1 to 1. The results of this analysis allows ultimately an evaluation of the results of the Hawkes model.

4.3. Hawkes Model

4.3.1. Multi-dimensional Hawkes process

To model the Hawkes process I used tick, which is a machine learning library for Python 3. The library involves a large set of tools for statistical learning and is currently the most comprehensive library that deals with Hawkes processes (Bacry et al., 2017b). So far, only a few open source packages such as the library *pyhawkes*², the R-based library *hawkes R*³, and the C++ library *PtPack*⁴ were available (Bacry et al., 2017b). However, Bacry et al. (2017b) showed that the computational timings of tick strongly outperform the existing libraries when it comes to simulation and fitting.

In particular, I used the ADM4-Model designed by Zhou et al. (2013) to infer the hidden linguistic network. This is a model that implements parametric inference for Hawkes processes with an exponential parametrisation of the kernels and a mix of Lasso and nuclear regularization (Bacry et al., 2017a).

The intensity of the Hawkes process in ADM4 is defined as (Bacry et al., 2017a):

$$\lambda_i(t) = \mu_i + \sum_{j=1}^D \sum_{t_k^j < t} \phi_{ij}(t - t_k^j)$$

where

- D is the number of nodes

4. Methodology

- μ_i are the baseline intensities
- ϕ_{ij} are the kernels
- t_k^j are the timestamps of all events of node j

The exponential paramertisation of the kernels is defined as (Bacry et al., 2017a):

$$\phi_{ij}(t) = \alpha^{ij} \beta e^{-\beta t} \mathbf{1}_{t>0}$$

where

- Parameter β is the decay and is given to the model.
- Matrix α^{ij} is the inferred adjacency matrix.

Zhou et al. (2013) design the algorithm ADM4 to efficiently infer the network of social influence. They consider several key features in their model. First, Zhou et al. (2013) state that actions are recurrent. Individuals can participate in an event multiple times. Second, Zhou et al. (2013) suggest that actions between interacting individuals are often self-exciting. If an individual or many of his or her neighbors already participated in an event, the likelihood of future participation increases. Third, Zhou et al. (2013) argue that network of social influence have certain topological structures. On the one hand the network is usually sparse, because the majority of individuals probably influence only a small number of others and only a few influence many others. On the other hand, the network have low-rank structures indicating that individuals tend to form communities, with an increasing likelihood of participating in an event under the influence of other members of the same community.

The first and second feature are captured by the multi-dimensional Hawkes process which specifically models recurrent and self-exciting processes. The network topology (sparse and low-rank nature of the network) are considered by introducing nuclear norm and lasso norm regularization on the infectivity matrix (Zhou et al., 2013). The ADM4 algorithm on tick allows to regulate the level of penalization of these regularizations (Bacry et al., 2017a). Since the inferred matrix in this study is neither sparse nor low-rank, the level of penalization is set to the smallest possible value of python to minimize the penalization, so there is actually no penalization. The smallest possible value of Python is used, because the algorithm does not allow to set the parameter C, which represents the level of penalization, to 0.

4.3.2. Time transformation

The used time-dependent Hawkes process is based on the occurrences of events over time. As described in chapter 4.1.2, In this work an event is defined as the adoption of an innovation by a new user. So, the main input to the Hawkes model is a list of times when new users adopt an innovation.

Importantly however, the Hawkes process observes the occurrence of events over *continuous* time. Thus, a major challenge was that my input data was not continuous in time. Since only the first seven days of each month are in the dataset, for each month 3 weeks of data are missing. So the data is in an ordinal scale, in which the numbers can be ranked, but the respective distances are not proportional. In the model, the missing data would indicate that simply no innovations were used during this time - but this is a fatally wrong assumption. Therefore, it was necessary to transform the time so that the missing time segments do not influence the results of the model. In other words, the data needed to be transformed from an ordinal scale into a metric scale, where the times can not only be ranked but are also proportional to each other. As a result, the new time scale is continuous and in addition the interval between time 1 and 10 is equal to the interval between 10 and 20.

In order to make this transformation, I have sorted and numbered all unique time records that occurred in the dataset. This then results in a list with values between 1 and 33'861'534. Since these values are difficult to interpret, they were re-projected between 1 and 1000. Eventually, the time of each tweet in the dataset can be found on this new time scale between 0 and 1000 - at time 0, the first tweet in the dataset occurred and at time 1000 the last one was observed.

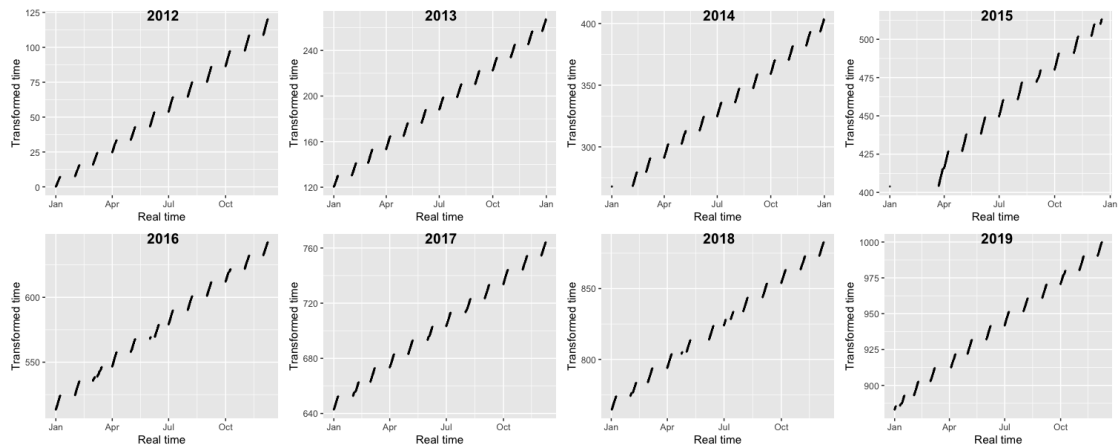


Figure 4.3.: Visualization of the time transformation proving that times are proportional to each other

4. Methodology

In figure 4.3 the relationship between the original real time and the resulting transformed time is plotted. The linearity of the respective points (or lines) show that the proportions are correct. Bent lines would indicate a distorted time - which is not the case here. An irregularity is to be noted in the beginning of 2015, because of missing data in the Twitter archive at that time.

The end time is given to the model, which is in this case simply 1000.

4.3.3. The event's timestamp as model input

Once a new time was assigned to each tweet, I could simply extract all the innovations from the dataset resulting in a list of all innovations with the associated new time. As described the Hawkes process tracks events over time.

Due to the multidimensionality of the model, this list was further divided and separated according to the type of innovation and the region in which it appeared. As a result, I had multiple small subsets each consisting of one type of innovation in one region. The input to the model is a numpy array with timestamps. In my case, the data is wrapped in a list of lists of numpy arrays with timestamps in order to track the behavior of different innovations and regions. Figure 4.4 shows exemplary the structure of the input data to the model.

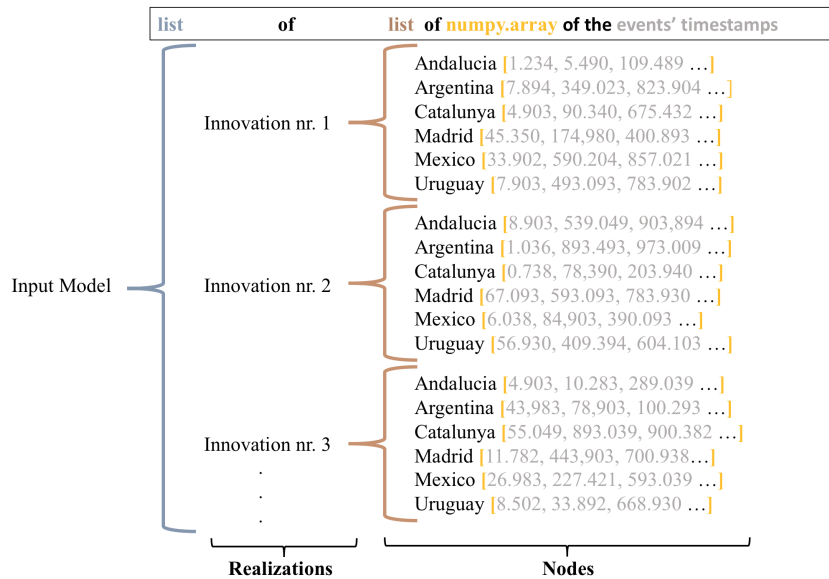


Figure 4.4.: Theoretical structure of model input

The ADM4-algorithm only allows the modeling of realizations with presence in all nodes. In my case, this means that only the modeling of innovations that occur in all

4. Methodology

regions is possible. Consequently, in a first model run, I focus on a subset of popular innovations with at least one occurrence in all regions. Only the eleven innovations with the number 1, 2, 3, 5, 8, 15, 20, 21, 23, 25, 32 are used to infer the first influence network.

In a second step, I reduced the regions and excluded Uruguay, because it is the region with the fewest innovations. So, I only focused on the five regions Andalucia, Argentina, Catalunya, Madrid and Mexico, which allows to take into account all the innovations that occur in these five remaining regions (18). One can easily detect which 18 innovations have been used in table 5.1, by counting only the innovations that occur in all regions except Uruguay.

4.3.4. Parameter estimation

A major challenge when modeling a Hawkes process with exponential decays is the estimation of parameters from observed data. The parameters of the decaying kernel ϕ can be estimated by maximizing the likelihood over the observed data. Rizoïu et al. (2017) list two reasons to not maximize the likelihood itself but the log of the likelihood. Firstly, maximizing the likelihood is more complex from the computational and numerical perspective because when it comes to maximizing, summing is less expensive than multiplication. Secondly, maximizing the likelihoods themselves would result in very small numbers that might then run out of floating point precision.

Borrowing the notation from Rizoïu et al. (2017), I define θ as the set of parameters of the Hawkes process. Then, the log of the likelihood function is defined as:

$$l(\theta) = \log L(\theta) = -\int_0^T \lambda(t) dt + \sum_{i=1}^{N(T)} \log \lambda(T_i)$$

Maximizing the natural logarithm automatically implies maximizing the likelihood function (Rizoïu et al., 2017).

The function `score` in the Python library `tick` computes the log likelihood (Bacry et al., 2017a). Since the higher the log likelihood, the better, one can simply maximize the score.

The estimated decay is 0.7 for the first model run and 0.5 for the second model run.

5. Results

All results presented in this chapter address the main research question, which aims to understand how new linguistic features in Spanish spread in space and time. In particular, the results in the sections 5.1, 5.2, 5.3 aim to answer RQ1, whereas sections 5.4, 5.5 aim to answer RQ2.

5.1. Overview of the innovations

5.1.1. Innovations and innovation popularity

In total, there are 9618 users who adopt a new innovation in the database. Figure 5.1 shows the change of frequency of innovations over time. There are six tiles to make the plots readable, otherwise the occurrence of rare innovations over time could not be observed in plots with popular innovations.

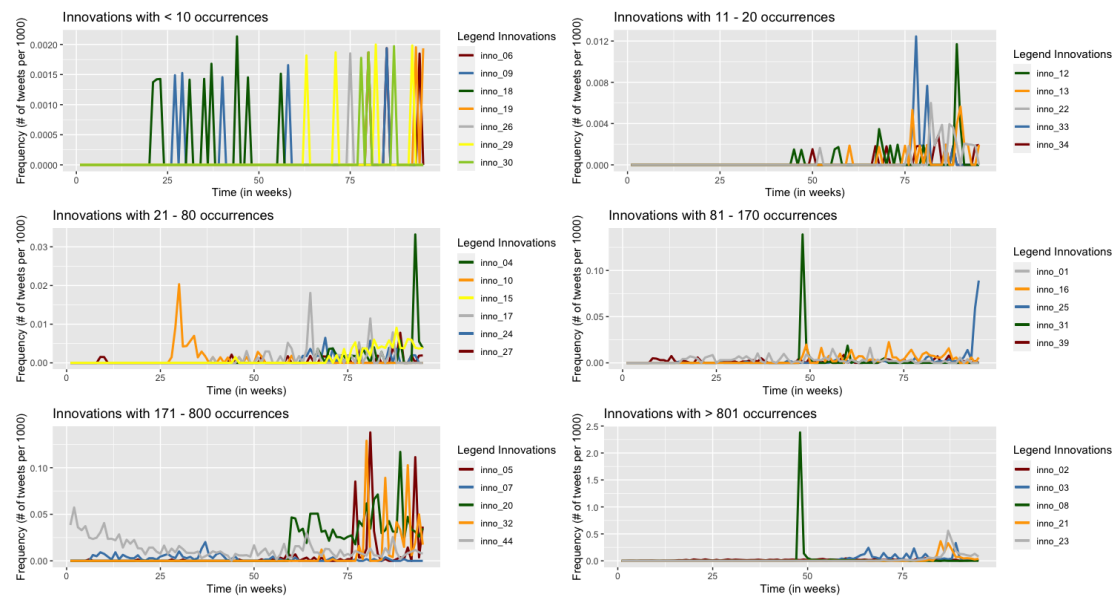


Figure 5.1.: Change in frequency (number of tweets with an innovation per 1000 tweets) from 2012 - 2019 in all Spanish speaking regions

5. Results

The innovations in this work propagate within a relatively short time. For many innovations, the typical increase in frequency can be observed in the last years of the observation period. Some innovations, on the other hand, do not show a distinctive change in their frequency. For example, the intensifier suffix "-érrimo" (innovation nr. 44) or the idiom "... que enamoró a Spielberg" (innovation nr. 39) have no significant changes on their frequency and consistently appear over the whole observation period. Innovation nr. 39 is an example of a phraseological phenomena. The sentence "la serie que emocionó a Spielberg" ("The show that moves Spielberg") has its origin in an advertisement for a TV show. Many Twitter users made fun of the slogan and it is nowadays often used with a sarcastic tone. In addition, the decrease of frequency of some innovations can also be observed. Some innovations are very short-lived forms and do not succeed in becoming language norms.

Furthermore, there is a very conspicuously high peak in the use of a few innovations. For example, innovation nr. 31 frequently appeared in January 2016 because a politician tweeted it on Three King's Day. In contrast, explaining the enormous peak of the innovation nr. 8 in January 2016 is not that trivial. The sentence "jaja para k kieres saber eso" ("hahaha why do you want to know that") is a common reply to a rhetorical question that is trying to make a point. It's sarcastic and has it's origin on the "yahoo answer" platform, where this sentence was an answer to a question. Interestingly, the original appearance of this innovation is much older, but in the first week of January 2016 the sentence occurred in 1594 tweets by 1559 Twitter users. Most of the tweets, around 1130, are classified as retweets. The innovation was used in all six regions, however conspicuously many users tweeted it in Argentina (954), Mexico (344) and Uruguay (120). So presumably the innovation spreads to Central and South America at this time.

Figure 5.2 shows the increasing occurrences of the innovations. However, not all innovations in the data spread successfully and some are only used by a relatively low number of users. So even though all those considered innovations are perceived as innovative and popular, some of them do not spread.

5. Results

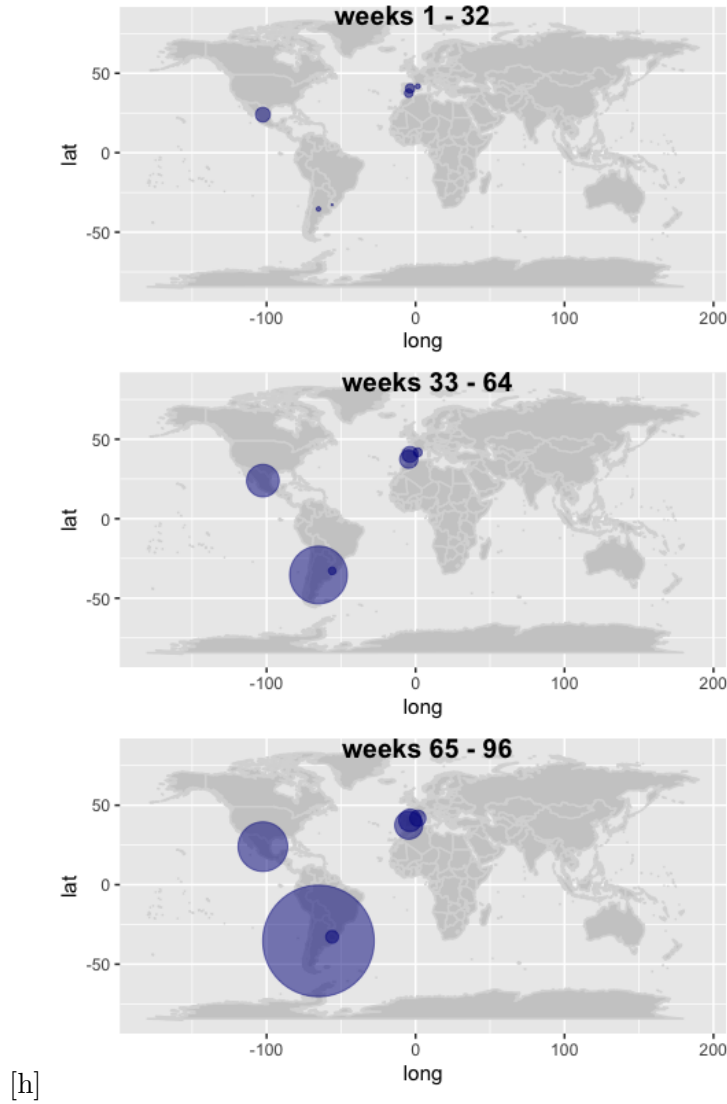


Figure 5.2.: Change in the absolute number of innovations over space and time

Table 5.1 lists the number of users who adopt an innovation per type of innovation and region. In other words, the table shows the rate of adoption defined in chapter 2.3.1. Here, the rate of adoption counts the number of individuals who adopted an innovation in the time period from 2012 to 2019. The column on the right summarizes how many users adopted a specific innovation. The last row in the table summarizes the number of users who adopted innovations in a specific region. Most of the innovations occurred in five or more region (57%). But not all innovations have appeared in all regions. Only 11 innovations occurred in all six regions. 18 innovations occurred in all regions except in Uruguay.

5. Results

Table 5.1.: Adoptions of innovations per regions

<i>Innovation</i>	<i>Andalucia</i>	<i>Argentina</i>	<i>Catalunya</i>	<i>Madrid</i>	<i>Mexico</i>	<i>Uruguay</i>	Σ
Nr. 1	12	70	7	8	63	5	165
Nr. 2	298	204	124	215	120	17	978
Nr. 3	111	710	52	76	508	61	1518
Nr. 4	2	30	0	6	3	3	44
Nr. 5	85	42	53	79	5	8	272
Nr. 6	0	1	0	0	2	0	3
Nr. 7	54	7	33	53	3	0	150
Nr. 8	109	1109	29	70	384	135	1836
Nr. 9	1	0	1	3	0	0	5
Nr. 10	15	0	8	23	2	1	49
Nr. 12	4	1	8	5	0	0	18
Nr. 13	4	1	4	4	4	0	17
Nr. 15	9	29	5	4	1	3	51
Nr. 16	62	6	46	42	6	0	162
Nr. 17	14	6	16	15	0	0	51
Nr. 18	1	0	0	7	1	0	9
Nr. 19	1	0	1	1	0	0	3
Nr. 20	18	445	15	17	187	66	748
Nr. 21	70	426	31	64	198	35	824
Nr. 22	6	7	3	2	1	0	19
Nr. 23	56	734	11	42	322	79	1244
Nr. 24	12	0	1	14	2	1	30
Nr. 25	3	107	1	6	1	5	123
Nr. 26	0	1	0	0	0	0	1
Nr. 27	5	0	4	15	0	0	24
Nr. 29	1	1	1	1	0	0	4
Nr. 30	1	1	0	0	1	0	3
Nr. 31	48	1	25	47	1	0	122
Nr. 32	71	114	32	45	18	6	286
Nr. 33	6	1	4	4	1	0	16
Nr. 34	3	5	0	2	2	1	13
Nr. 39	36	3	24	23	2	0	88
Nr. 44	102	35	67	138	400	0	742
Σ	1220	4097	606	1031	2238	426	9618

5. Results

The right column in the table shows which innovations are adopted often and which are adopted rarely. Innovation nr. 6 is the rarest which only occurred once in the dataset. There are a few other innovations that occurred less than 10 times during the entire observation period, which is also extremely rare.

On the contrary, some innovations were adopted by over a thousand users, such as innovation nr. 3, innovation nr. 8 and innovation nr. 23. Almost half of all innovations in the dataset are one of these three innovations. In summary, there are some very rare innovations and some very popular ones - the differences are substantial.

The last row shows the differences of the number of users who adopt an innovation per region. In Argentina, by far the most innovations occurred. Mexico also shows a high number of innovations. The other regions have a significantly smaller number. Least innovations were adopted in Uruguay.

The analysis of the innovations demonstrates that most Twitter users use some innovations only once. Nevertheless, there are some users who use either multiple innovations or one innovation multiple times. For example, one user used the innovation "Inyustisia" (innovation nr. 7) 17 times in different months over the years 2014 and 2015. This innovation is a phonetic spelling which is due to the conspicuous pronunciation of the word "injustice" by Cristiano Ronaldo.

5.1.2. Origin of innovations

In order to answer the question when and where innovations first appeared, I focused on the first observation of an innovation. Figure 5.3 shows that innovations have emerged throughout almost the entire observation period. The earliest occurrence of an innovation was the innovation nr. 44 on the 1st January 2012. Innovation nr. 6 was observed for the first time at the 5th September 2018 and is thus the newest innovation in the dataset. Note, there may be unobserved innovations before and after the time considered.

5. Results

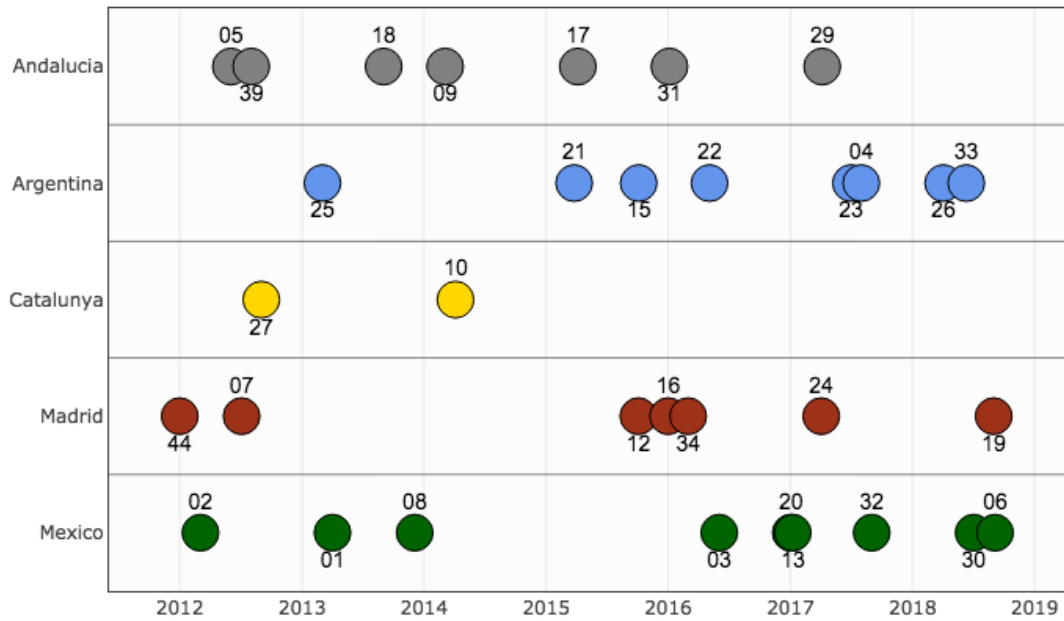


Figure 5.3.: Plot of region and time of the first occurrence of all 33 innovations

Unfortunately, the figure 5.3 does only allow to draw limited conclusions about the innovators, who introduce a new linguistic term. The innovators can be defined as those users who use a linguistic term for the very first time. However, I only consider the years between 2012 and 2019 and I do not know and consider what was before, which makes statements about the origin of older innovations difficult. Conclusions about the origin of the newer innovations are challenging, since the first appearance of a given innovation can be quite random in the sample of tweets provided by Twitter.

For this reason, not only the first occurrence but the first few occurrences of an innovation are considered to identify the innovators. Following the diffusion of innovation theory, I defined the innovators according to when they adopted an innovation relative to all other adopters (Rogers, 2003; Toole et al., 2012). As shown in figure 2.2, Rogers (2003) defines the first 2.5% of individuals who use an innovation as the innovators. Averaged across all innovations, 2.5% correspond to about ten adopters. Correspondingly, only innovations that are adopted at least ten times are taken into account on the following map. From these first ten adoptions of an innovation, also called innovators, the modal value of the regions of these tweets is computed. The modal value of the regions is defined as the region where the most innovators occurred.

Figure 5.4 shows which innovation is most likely to have originated in which region.

5. Results

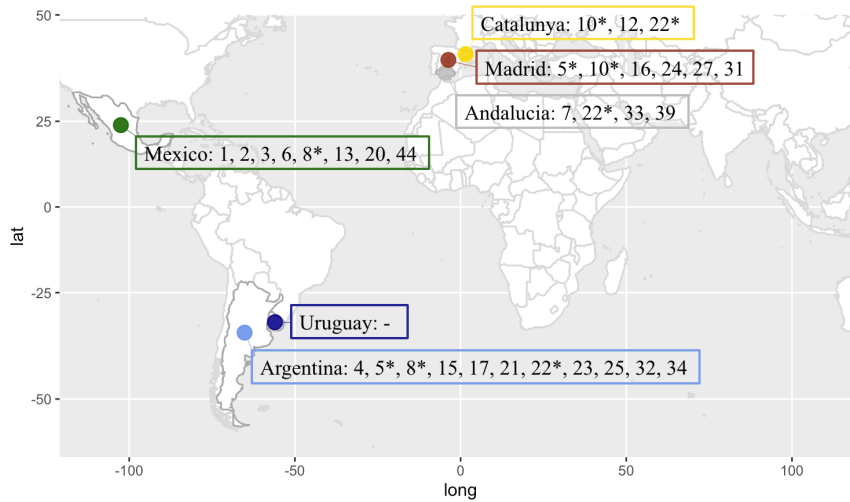


Figure 5.4.: Origin of innovations on a map

Innovation that have the first appearances in two regions with equal frequency and therefore have no clear modal value are marked with a * symbol.

The modal region is thus a more robust measure for identifying the origin of an innovation. However note that we must be careful about concluding from these results how innovative a region is. If I normalize the number of originated innovations per region with the number of users per region, the European regions seem to be more innovative than Argentina and Mexico in terms of the analyzed phenomena in this study. But as mentioned earlier, there is a bias in the innovation selection as it is mainly based on the assumption of Europeans.

5.2. Diffusion patterns of innovation propagation

In this section, three spatial properties of innovation propagation are examined, namely the focus, the entropy and the spread of innovations.

Focus

The focus is a measure describing the maximum probability of observing an adopter of an innovation in one region. The focus value describes how important a region is for an innovation. The higher the focus, the higher the proportion of all adopters that occurred

5. Results

in the given region. Figure 5.5 visualizes the focus value on the x-axis and indicates the focus region of each innovation by color.

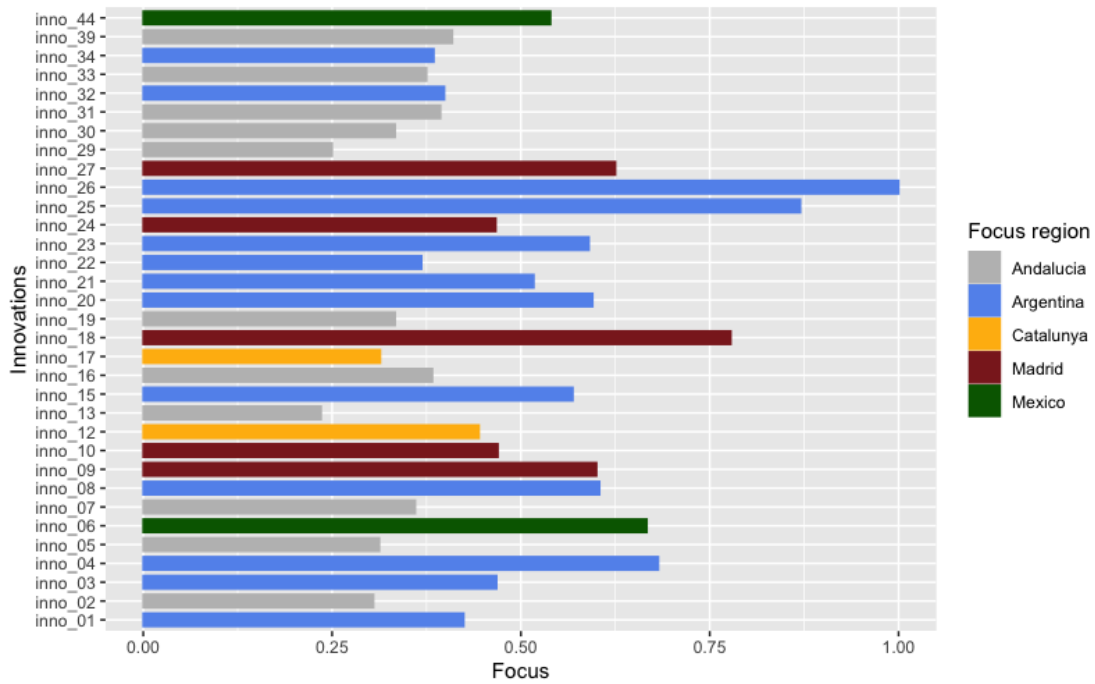


Figure 5.5.: Focus

Innovation nr. 26 has a focus of one, because the innovation only occurred once over the observed time period. No other innovations have only remained in one region. The other extreme is innovation nr. 29 which was adopted once in four different regions resulting in a focus value of 0.25. The mean focus value of all adoptions is nearly 0.5. So on average 50% of all adopters of an innovation occur in a single location.

The focus region can be defined as the region with the highest focus. 13 innovations have their innovation focus region in Argentina, 8 in Andalusia, 7 in Madrid, 3 in Mexico and 2 in Catalunya. For the innovations the focus region are mainly in Argentina but also Madrid and Andalusia. Uruguay is never a focus region.

In figure 5.5, it is striking that innovations which have their focus region in Argentina tend to have a higher focus values than, for example, innovations which have their focus in Andalusia. This means that innovations with their focus region in Argentina are less dispersed than it is the case for innovations with focus region in Andalusia. So, I already assume that innovations with their focus region in Argentina have a smaller entropy than innovations with their focus region in Andalusia.

5. Results

Entropy

The entropy is a measure of the randomness in spatial distribution of the adopters of an innovation. The meaning of the entropy values can be illustrated with the two extreme examples mentioned above. Innovation nr. 26 only occurs in a single location resulting in an entropy value of 0, whereas innovation nr. 29 is adopted equally in four different regions and has an entropy value of two. So the higher the entropy, the more randomly distributed are the adopters. The mean entropy over all innovations is around 1.7, so most innovations have been adopted in several regions.

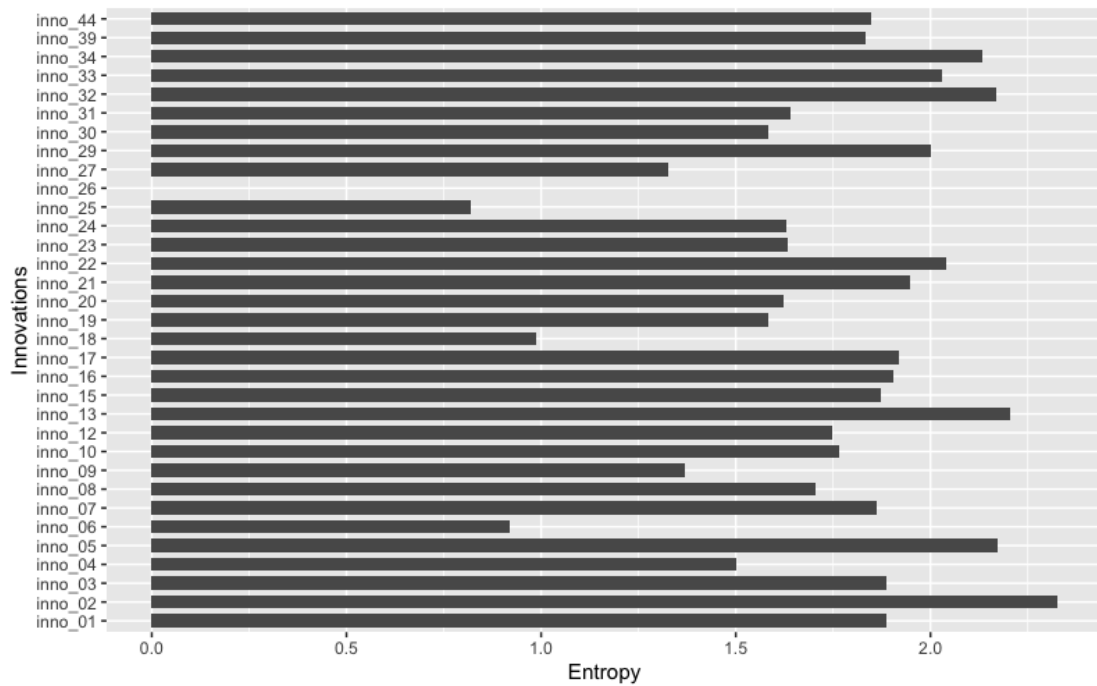


Figure 5.6.: Entropy

Confirming my assumption from before, adopters of innovations with their focus region in Argentina have a mean entropy of 1.6, whereas adopters of innovations with their focus region in Andalucia show a mean entropy of 1.9. The latter adopters are therefore more spatially distributed and less clustered.

Spread

The spread measures the mean geographic distance over which an innovation diffuses. Interestingly, through this measure, one can easily observe which innovations have managed to cross the Atlantic and gain popularity on both sides of the ocean. Innovation

5. Results

nr. 30 is the innovation with the largest average spread of 5'375 km, whereas innovation nr. 26 has the minimum spread of 0 km, occuring only once in a region.

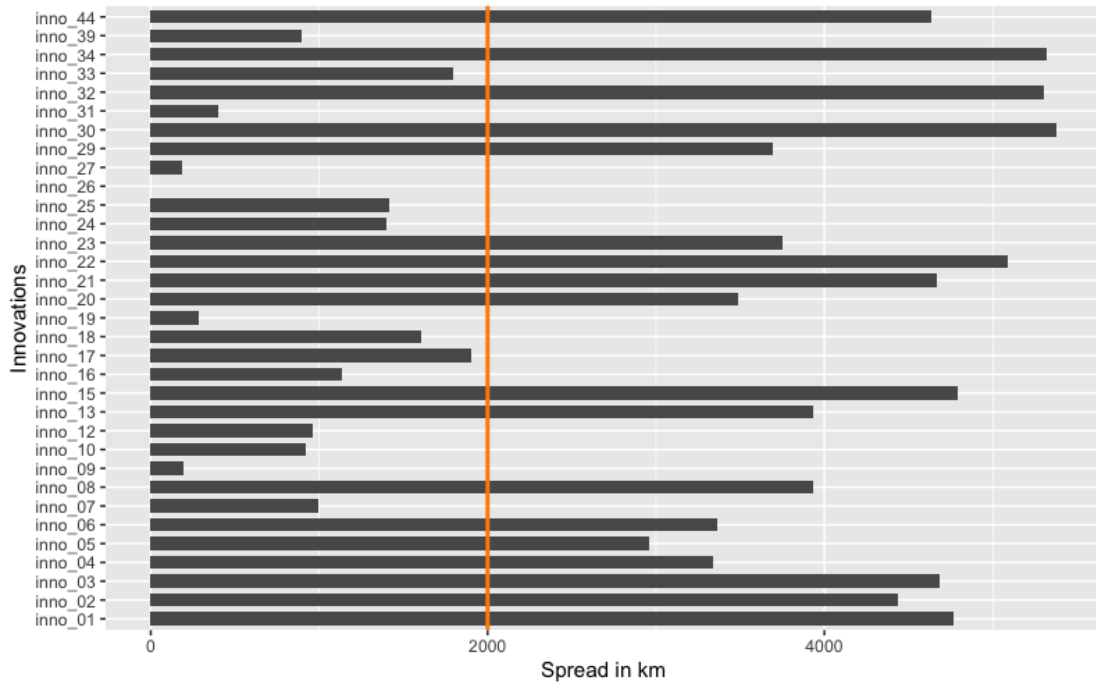


Figure 5.7.: Spread

The plot shows that not all innovations spread globally. Some innovations seem to occur only locally or regionally. The orange line in figure 5.7 symbolizes the division of the two variants - global and local. The threshold was set at 2000 km, because this is a high average spread, which the analysis in this section indicates an innovation can only reach if it spreads globally. The innovations that were categorized, either global or local, are listed in table 5.2. On one side, the local innovations that exhibit a spread between 0 and 2000 km. They have a mean focus of 0.53 and a mean entropy of 1.48. On the other side, the global innovation that have a spread between 2000 km and 5500 km. They have a mean focus of 0.46 and a mean entropy of 1.85.

The number of adopters of each innovation is also given in brackets in table 5.2, because it is assumed that the probability of a global diffusion increases as the popularity of an innovation increases. This correlation was examined in more detail and the distributions of the two groups were visualized in boxplots in figure 5.8.

5. Results

Table 5.2.: Global and local innovations.

Global innovations	Local innovations
Innovation nr. 1 [165]	Innovation nr. 7 [150]
Innovation nr. 2 [978]	Innovation nr. 9 [5]
Innovation nr. 3 [1518]	Innovation nr. 10 [49]
Innovation nr. 4 [44]	Innovation nr. 12 [18]
Innovation nr. 5 [272]	Innovation nr. 16 [162]
Innovation nr. 6 [3]	Innovation nr. 17 [51]
Innovation nr. 8 [1836]	Innovation nr. 18 [9]
Innovation nr. 13 [17]	Innovation nr. 19 [3]
Innovation nr. 15 [51]	Innovation nr. 24 [30]
Innovation nr. 20 [748]	Innovation nr. 25 [123]
Innovation nr. 21 [824]	Innovation nr. 26 [1]
Innovation nr. 22 [19]	Innovation nr. 27 [24]
Innovation nr. 23 [1244]	Innovation nr. 31 [122]
Innovation nr. 29 [4]	Innovation nr. 33 [16]
Innovation nr. 30 [3]	Innovation nr. 39 [88]
Innovation nr. 32 [286]	
Innovation nr. 34 [13]	
Innovation nr. 44 [742]	

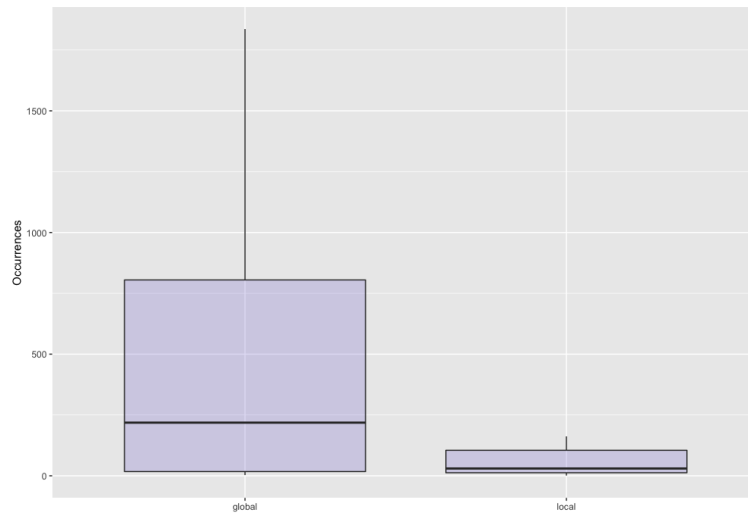


Figure 5.8.: Boxplot of global and local innovations

Innovations categorized as global phenomena have a mean frequency of 487 adopters, while innovations categorized as local phenomena have a mean frequency of 57 adopters. The median values are smaller, with 218 adopters of global phenomena and 30 adopters of local phenomena, but still considerably different. Although it is already fairly obvious

5. Results

from the boxplots visualized in figure 4, a t-test was computed to determine if the two groups are significantly different. With a p-value of 0.006581 the null hypothesis can be clearly rejected, indicating that the two means of the sample are significantly different. In summary, when the popularity of an innovation increases, the phenomenon tends to spread globally.

Comparison of spatial properties

Figure 5.9 shows the correlation of the spatial properties with the number of adopters occurring. All measures tend to correlate positively with the number of adopters. In other words, the more Twitter users adopt an innovation, the higher their focus, entropy and spread. For the entropy and the spread this relationship is not surprising. Higher number of adopters lead intuitively to more dispersed and more widely spread innovations. For the focus, however, one would intuitively assume that it becomes smaller the more adopters exist. Two points should be noted. First, although the correlation in the plot might be slightly positive, it is almost zero. Secondly, many of the innovations in the dataset are only adopted rarely. In the plot there are many points on the left side which indicates only a few occurrences and a few points on the right side with many occurrences. The spatial properties of rare innovations are rather random. For example, if only two Twitter users adopt an innovation, the measures focus, entropy and spread have less significance than if there are hundreds of adopters of that given innovation.

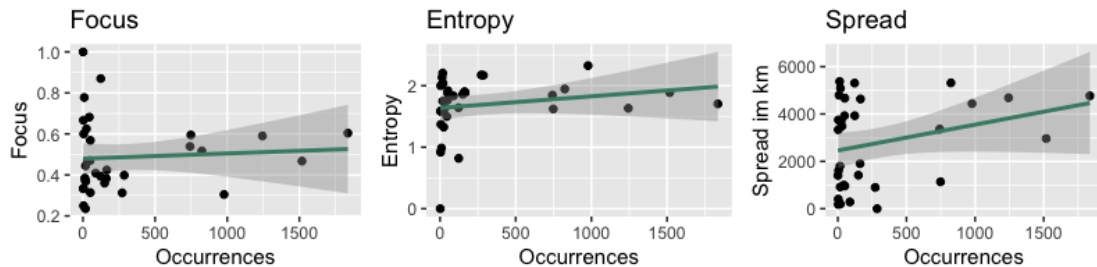


Figure 5.9.: Spatial properties compared to the frequency of adoptions of innovations

In a next step, I compared the three geospatial properties. The first plot in figure 5.10 shows that the focus and the entropy exhibit strong proportionality. The higher the entropy, the lower the focus. While the randomness of distribution increases, the focus of that given innovation decreases. Similarly, figure 5.10 shows that the focus decreases as the spread increases. Not surprisingly, the more widespread an innovation is, the more concentrated is its focus. Eventually, when the spread of an innovation increases, the entropy also increases.

5. Results

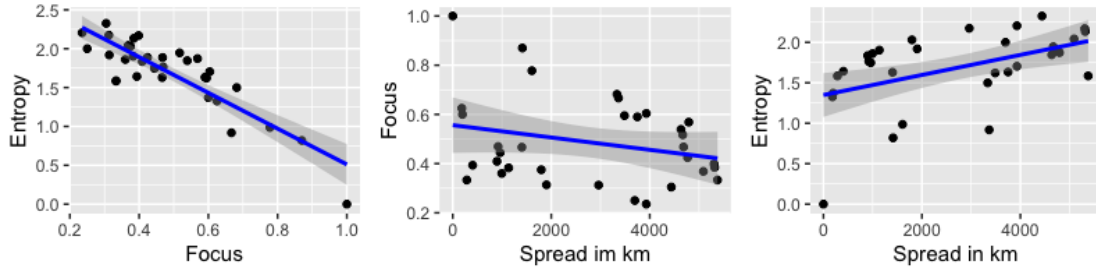


Figure 5.10.: Correlation between the spatial properties focus, entropy and spread

5.3. Relationship between regions and their innovations

Jaccard Similarity Coefficient

The innovation similarity, measured with the Jaccard coefficient, describes the similarity of two regions regarding the innovation which are used there.

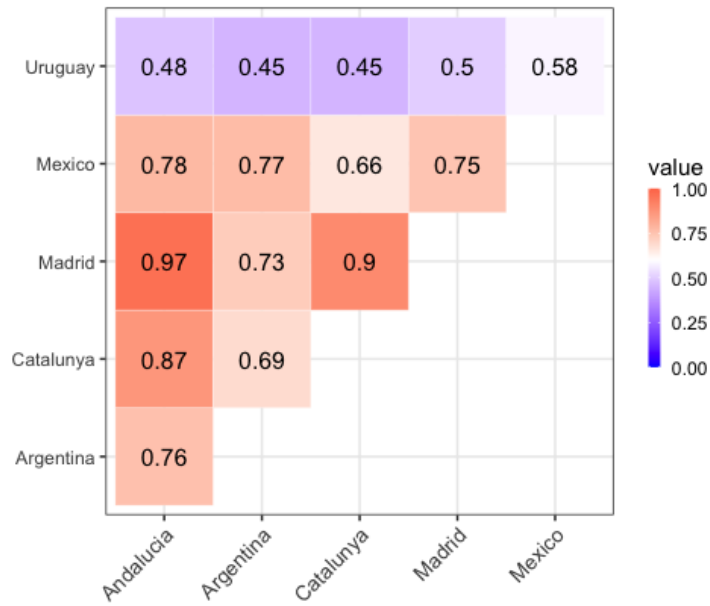


Figure 5.11.: Heatmap of Jaccard coefficients

The heat-map in figure 5.11 shows the upper triangle of the Jaccard index between all regions (the lower triangle is symmetric). The three regions of Spain share the highest coefficients for each other and are consequently most similar in terms of the

5. Results

used innovations. They have a 95 percent match of used innovations, meaning that for example almost all innovations that were used in Madrid were also used in Andalucia and Catalunya. Uruguay is the region where the fewest innovations occurred among those analyzed. Correspondingly, the region has a low Jaccard coefficient indicating little similarity with all other regions. Interestingly, however, Uruguay and Argentina have the lowest Jaccard coefficient, although they are spatially close to each other. In other words, the two Latin American regions share the smallest sample of common innovations, which is rather surprising. The reason is probably that Uruguay has the smallest amount of tweets and, therefore, a smaller amount of innovations that occurred there.

In a next step, I studied the relationship of the Jaccard coefficient and distance: Are regions that are spatially close, more likely to adopt the same innovations?

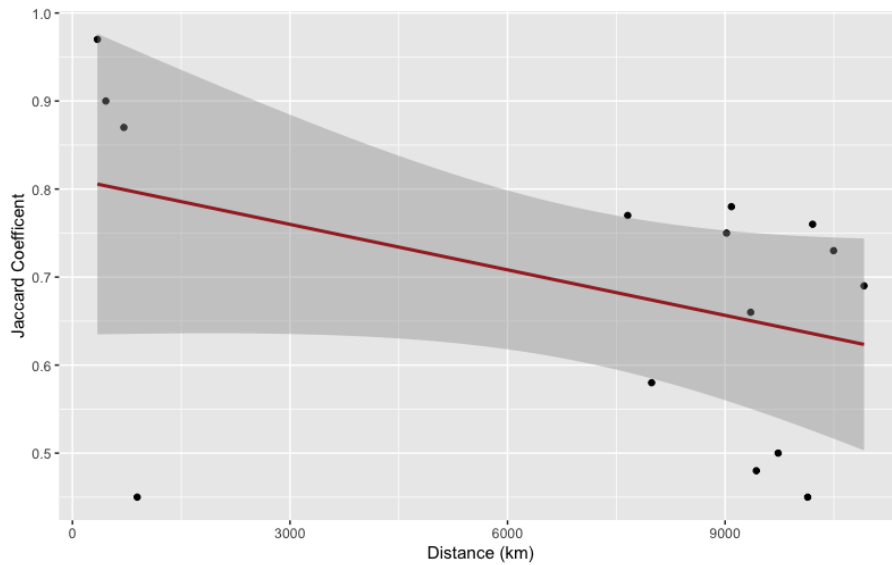


Figure 5.12.: Correlation between distance and Jaccard coefficients

The innovation similarity of locations versus their distance between locations can be measured resulting in a correlation value. Over all regions and innovations, a correlation value of -0.64 was measured indicating that the Jaccard coefficient is negatively correlated with distance. In other words, the smaller the distance between two regions, the more similar is the sample of innovations which is used there. Vice versa, if regions are far apart, they also tend to have a different sample of innovations.

Adoption Lag

The adoption lag answers the question if regions that are spatially close also are more likely to adopt innovations at the same time. Figure 5.13 shows the correlation between distance and the mean adoption lag between regions in days. The correlation is weaker than before, indicating that the time at which an innovation is adopted does not correlate with space. Or in other words, geographical close regions do not adopt innovations at the same time.

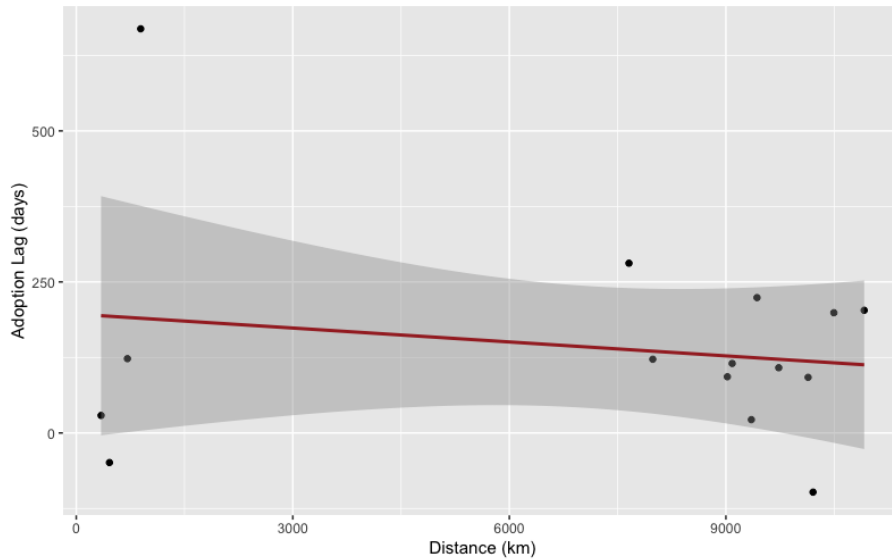


Figure 5.13.: Correlation between distance and adoption lag

The values of the adoption lag are rather high, because of the given structure of the dataset which only includes seven days per month. The many missing days in the dataset are counted as days when no innovation occurred, so care must be taken when interpreting adoption lag values.

5.4. Spatial impact

The spatial impact is a measure to evaluate the impact that regions have on each other. The results are visualized in a one-dimensional plot in figure 5.14. On the x-axis the spatial impact is shown, which is between -1 and 1.

5. Results

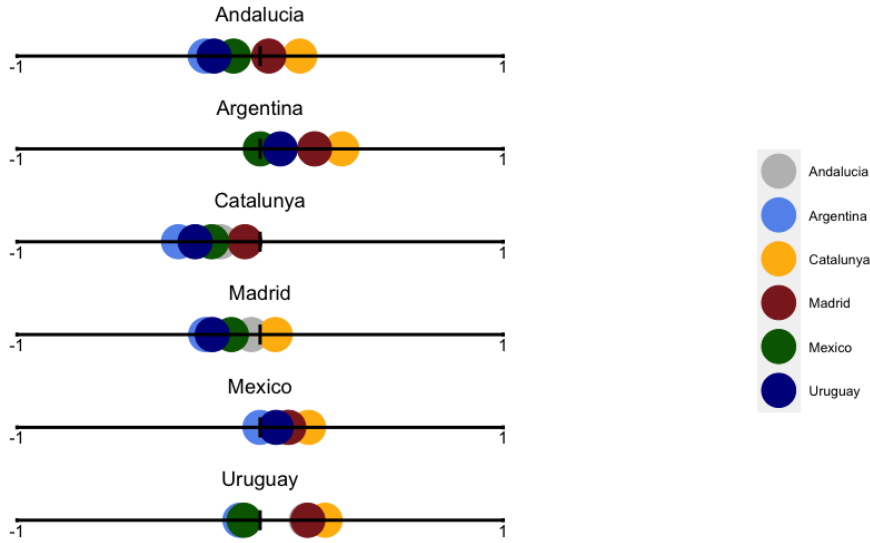


Figure 5.14.: Spatial impact plot modeling impacts of all six regions

Each of these one-dimensional plots in figure 5.14 is about one specific region. The region itself is not visualized in its own plot, but one can imagine that it would be located at the zero point. Regions which impact the given location, are positioned on the left half of the plot. They have a negative *ImpactValue* and are so called impacting regions. On the right half of the plot are the regions with a positive *ImpactValue* which are the impacted regions.

Andalucia does both, it influences and it is influenced. Regions in South America and Latin America have an influence on Andalucia, which in turn influences the other Spanish regions Catalunya and Madrid. A similar picture can be seen with Uruguay - it is influenced by regions west of the Atlantic and influences regions east of the Atlantic.

The plot shows that, in particular, the two regions Catalunya and Madrid are influenced by others. Especially, Catalunya does not seem to affect any other region, but is affected by all others.

According to these results, Mexico can also be called an impacting region, since it is only slightly influenced by Argentina and influences most of the other regions. Argentina is the region that influences most other regions and is itself the least influenced.

5. Results

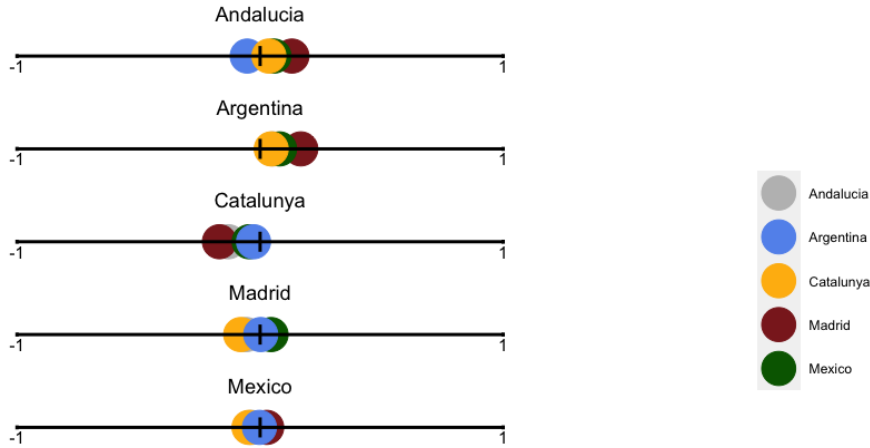


Figure 5.15.: Spatial impact plot modeling impacts of all regions except Uruguay

In a second step, the spatial impact was calculated with only five regions (all regions except Uruguay). As mentioned in chapter 4.3.3, I model the Hawkes process twice and exclude Uruguay once because it has the fewest innovations. For the spatial impact, I model the same data as in the Hawkes process in the next chapter, so that the two results can be compared. The results of the second run are visualized in figure 5.15, which are similar to the previous results. However, the measured *Impact Values* seem to be smaller indicating that the modeling with additional innovations have rather weakened the tendencies.

5.5. Adjacency matrix of Hawkes process

A Hawkes process model identifies the influence of a specific region on all other regions, including the region itself. The output of the model is an adjacency matrix, which shows the inferred network of influence. In other words, the matrix reflects the estimated influence on the adoption of innovations of each region on others and itself. Figure 5.16 visualizes the adjacency matrix of the first model run which includes all six regions and the corresponding eleven innovations that occurred in all of them.

It is worth noting that the model is very efficient - the entire script has a running time of only about one minute. Loading and structuring the data is included in this time, so the running time of the actual inference takes only a few seconds.

5. Results

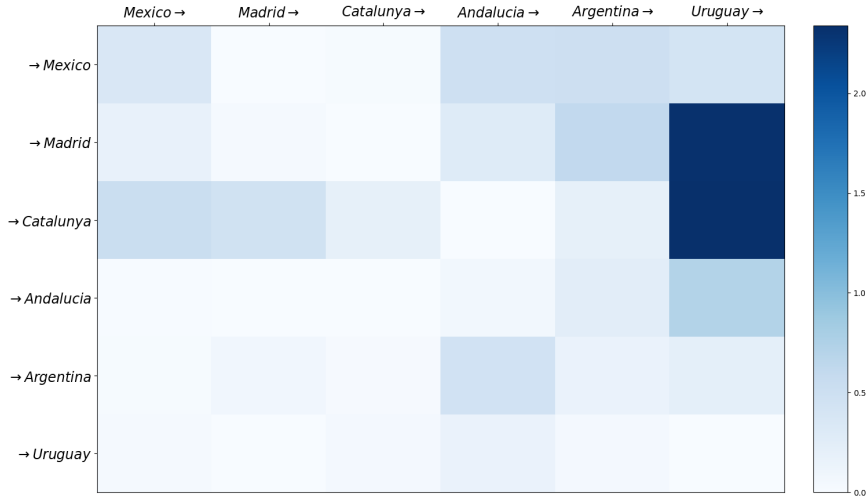


Figure 5.16.: Adjacency matrix of first model run including all regions

The darker the cell, the higher is the influence from the region specified in the column to the region specified in the row. So, the small arrows behind the labels in the columns and rows indicate how the matrix should be interpreted. Note, appendix D lists the matrix in numerical form, showing the values of the adjacency matrix between the processes. One recognizes directly that Uruguay in particular influences Madrid and Catalunya. In general, Catalunya is the region that is most influenced by others. While Madrid and Catalunya tend to be influenced by others, Andalucia is also influencing other regions. Interestingly, Andalucia is the region that influences Argentina the most. But in general, Argentina seems also to be a fairly influential country, as is Mexico. Appendix C plots the timestamps of events per region for the eleven innovations used in the first model run. Although it is not easy to draw the same conclusion at first glance, a closer look shows that the plots support the results of the Hawkes model. In Argentina, Uruguay and Mexico, the events often occur shortly before those in Madrid and Catalunya.

Table 5.3.: Inferred intensity baseline of first model run including all regions

Mexico	Madrid	Catalunya	Andalucia	Argentina	Uruguay
0.00225	0.00323	0.00247	0.00201	0.00219	0.00224

Table 5.3 lists the inferred baseline intensities which describe the likelihood of innovations occurring randomly in these regions. Andalucia has the smallest baseline intensity.

5. Results

It is worth noting, however, that the differences in the baseline intensities between the regions are small. Only Madrid has a slightly higher baseline intensity, suggesting that in Madrid the adoption of innovations is somewhat less triggered by previous adoptions than in the other regions.

Due to the low number of adoptions in Uruguay, I inferred a second network that excludes Uruguay. As mentioned earlier, by focusing only on the remaining five regions, the model can account for all 18 innovations that occur in all of these five regions.

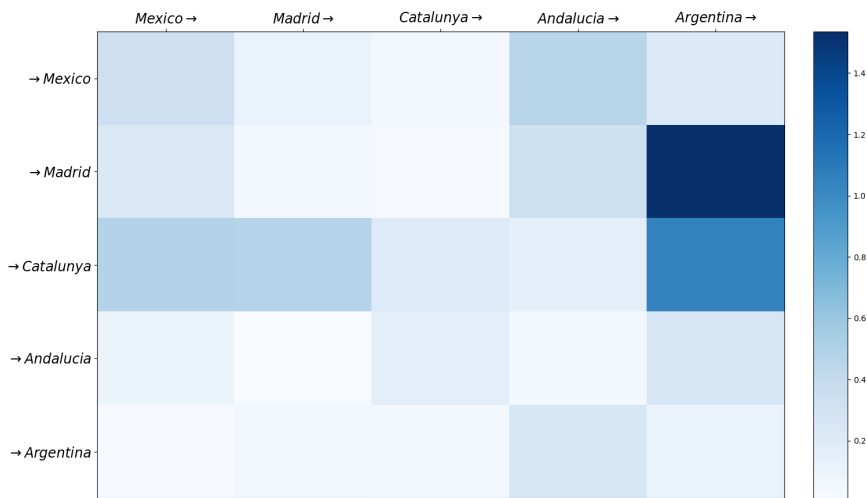


Figure 5.17.: Adjacency matrix of second model run including all regions except Uruguay

Although more data is considered, the output is still similar to the previous model. Mexico, Andalusia and Argentina tend to influence others, while Madrid and especially Catalunya seem to be influenced by many others. Mexico is the region that influences itself the most. Note, the scale shows slightly lower values than before. Thus, the triggering coefficients are smaller, indicating less influence between the processes.

Table 5.4.: Inferred intensity baseline of second model run including all regions except Uruguay

Mexico	Madrid	Catalunya	Andalusia	Argentina
0.00238	0.00354	0.00247	0.00186	0.00185

Table 5.4 lists again the inferred baseline intensities per region. Argentina and Andalusia have almost identical values and have the smallest baseline intensities. They

5. Results

are also smaller than in the previous model run, implying that the two regions in this model appear to be more triggered by past events than in the previous model run. Interestingly, Catalunya has exactly the same baseline intensity as in the previous model. Madrid again has the highest baseline intensity.

In summary, I get similar results when removing Uruguay from the data. Moreover, the output from the Hawkes model can be compared with the results of the spatial impact measure. Both models conclude that Uruguay and Argentina tend to be the driving forces in this network and the European regions seem to be the followers. Mexico can also be classified as an influencing region, but Mexico also strongly influences itself.

6. Discussion

The first few sections in this chapter discuss the findings of the previous chapter and the research questions introduced in chapter 1.2. Specifically, section 6.2 focuses on the research question RQ1, while section 6.3 contains a deeper discussion of the research question RQ2. Finally, section 6.4 points out limitations of the data as well as of the chosen approaches.

6.1. Spread of innovations

Some innovations spread successfully and are used by a relatively high number of users, but a few of them do not spread wide in space and time. Regarding the temporal diffusion most of the innovations considered do not show an s-shaped curve described by Maybaum (2013) and Rogers (2003). Thus, not all innovations do succeed in becoming language norms. Many innovations tend to be short-term popular forms of language. This study examines specific textual innovations, many of which originated on Twitter and therefore tend to become obsolete quickly.

The analysis of the spatial diffusion of linguistic innovations has clearly shown that there are interactions between regions. Although not all innovations are geographically widespread, the majority of innovations did not only occur in one region, but globally in different regions.

The analysis of the spatial properties focus, entropy and spread have shown that whereas some innovations spread only over small geographical areas others diffuse widely in space. Hence, not all phenomena are global, there are also innovations which only occur locally or regionally. Can the classification into regional and global phenomena be explained by the type of linguistic innovation? All morphosyntactic innovations show a global spread, however these are only two. The innovation nr. 2 "ojala" and innovation nr. 44 "-errimo" have both occurred early in the observation period. According to the analysis, both innovations seem to have their origin in Mexico. But whereas Mexico is also the focus region of innovation nr. 44, Andalucia is the focus region of innovation nr. 2. In the latter case this means that the majority of users who have adopted an

innovation are located in Andalusia. The innovations with non-standard orthography are mostly rather local phenomena having their origin and focus region either in Andalusia or Argentina. Most innovations are phraseological innovations that show no specific pattern in terms of the geographical areas over which they spread. The type of the innovation is therefore not an indication of whether the innovation will spread only regionally or globally.

6.2. Factors driving the diffusion of innovations

The first research questions focused on how geographical distance influences the diffusion patterns on online platforms such as Twitter. Research has shown that the diffusion of linguistic innovations in online channels depend on three major factors: geographical distance, population size or density and culture. In this study, the role of geography was explicitly analyzed by examining the relationship between regions considering physical distance and spatial characteristics of linguistic innovation diffusion.

Two regions that are geographically close tend to adopt the same innovations, but not necessarily at the same time. This finding differs from the results of the study by Kamath et al. (2013). They found a positive correlation between distance and the adoption lag of hashtags. The closer two regions are, the more likely they are to adopt hashtags at the same time. Kamath et al. (2013) study the diffusion of innovations on a global grid. In contrast, the regions in this study are not in a geographically continuous space. Only some specific regions on the globe were selected, some of which have very large spatial distances between each other and no direct contact. Therefore, these results are maybe not comparable.

As mentioned in chapter 2.3.2 distance probably plays a role in diffusion of linguistic innovations, because the probability of interaction with people that are spatially close is higher than the probability of interaction with people that are spatially far. However, this relationship does not have to be linear. People from Spain may have more contacts in other regions in Spain, but although Mexico is spatially closer to Spain than Argentina, there is no obvious reason why Europeans should have more contacts to Mexicans than to Argentinians - a large ocean separates both regions from Spain. A critique on Trudgill's gravity model from Boberg (2000) is that when proposing this language diffusion model, the diffusion of innovations was only examined within small regional dialect areas. Boberg (2000) argues that this is a limitation, because for example English-, French- and Spanish speaking communities are spread over many different nations and continents. The space between nations and continents does not have to be planar. Bai-

6. Discussion

ley et al. (1993) argue that some topological features may act as barriers and others as promoters to the diffusion of innovations. For example, rivers can promote the spread of new language forms and mountains can inhibit it. So the accessibility of regions in space is not equal, but it was not taken into account in the wave model, nor in the hierarchical models. The assumption of planar space may work for regional analysis, but as soon as one examines global language diffusion, the models seem to reach their limits.

The influence of the population size and culture on the diffusion of innovations was not of importance in this study. Nevertheless, the selection of the six regions may allow further conclusions about the role of culture. Based on the selection of regions, one can ask whether the factors of culture and physical distance are similar in this work? The Spanish regions Andalusia, Catalunya, and Madrid have small physical distances between them and probably also share a more similar culture than with regions on the other side of the Atlantic ocean. Similarly, Argentina and Uruguay are geographically close and share rather similar cultures. The South American region Mexico is a large country, where again the cultural differences to Latin America and Europe are likely to be greater. Despite the broad concept of what culture actually means, Boberg (2000) concluded in his study about the diffusion of linguistic innovations on the U.S.-Canadian border that culture must be a driving factor of language change because national borders slowed the diffusion of innovations. Hence, he has associated nations with cultures and assumes that people in a country share similar cultures.

The population size in the analyzed regions varies, mainly because of the different spatial area covered. I studied the spatial diffusion on a macro-level rather than on a micro-level. Some regions are very large and cover entire countries, although most of the tweets probably come from big cities, or the capital. No conclusions about the linguistic innovation diffusion between rural and urban areas can be made.

The network of linguistic influence inferred by the Hawkes process shows no clear relation between influence and distance. However, regions close to each other show similar behavior. Argentina and Uruguay are both influencing regions, whereas the European regions are influenced regions. So there is a tendency that spatially close regions behave similar. The biggest influence was estimated from the Uruguayan and Argentinian cluster to the European cluster, although it is also where the biggest physical distance is measured. Hence, in the influence network, there are no denser connections within near regions. In contrast, there are more cross-continental connections.

In summary, the factors driving the diffusion of innovations are complex. Geographic distance alone cannot explain the relations and influences between intercontinental regions. Distance may play a role in diffusion of linguistic innovations, but is certainly

not the only factor. Many innovations in this study originated on Twitter, which may be why the role of geographic distance in the diffusion process is less relevant than in previous studies (Eisenstein et al., 2014; Horvath and Horvath, 1997). Future studies may focus on separately inferring the network of linguistic influence from innovations that were created within Twitter and outside Twitter. Categorizing innovations into two subgroups would allow a more nuanced analysis in terms of their source. I assume that innovations created outside of Twitter are more dependent on space and are spatially clustered, especially in the early innovation phase. Whereas innovations emerging in Twitter are likely to be more “global phenomena” and the role of physical distance is less relevant. There is a lack of research that considers the source of innovations as an aspect that affects diffusion. Future studies on this topic could lead to a better understanding of how the source of innovations affect the spatial characteristic of the diffusion process.

6.3. Network of linguistic influence - Leaders and followers

The second research questions focused on the role that specific Spanish speaking regions play in the diffusion of linguistic innovations. Results have shown that the drivers of Spanish linguistic innovations are mainly Argentinian and Mexican Twitter users. Kulshrestha et al. (2012) address the question whether offline geography still matter in online social networks. They try to answer questions about the importance of transnational links on Twitter or if users preferentially receive followers or follow others from their own country. To conduct this analysis, Kulshrestha et al. (2012) ranked for each country all other countries based on how closely their users followed (followings) or were followed by (followers) users in other countries. They found that more than a third of all social links on Twitter are transnational meaning that they connect two users from different countries. Not surprisingly, the analysis of the closest five follower and following countries for a country show that not only geography but also language appears to be a reason for connections between people in different countries. The closest five follower countries of Spain are Argentina, Bolivia, Ecuador, El Salvador, and Uruguay, whereas the closest five following countries are Argentina, Bolivia, Ecuador, Mexico and Uruguay (Kulshrestha et al., 2012). All countries are geographically distant from Spain but share the same language.

Another observation is that although the follower and following countries of Spain are similar, they are not identical. Mexico is a following country, but not a follower country. Hence, according to these findings, a larger proportion of users in Spain are following

6. Discussion

users in Mexico than vice versa. But these findings from Kulshrestha et al. (2012) are several years ago and recent research lacks new data on this topic.

In addition, Kulshrestha et al. (2012) examine how countries produce and consume transnational tweets, i.e. what proportion of tweets in a country are exported or imported. They found that overall 37.54% of tweets are shared internationally. But a deeper analysis of some countries show that there are significant differences. Some countries are far more internationally integrated than others and some countries import considerably more tweets than they export. Unfortunately, the study offers no data about Spain, Argentina, Uruguay, or Mexico in terms of percentage of imported and exported tweets. Perhaps Mexico is less connected internationally, because results have shown that it is very self-influenced and does not play a significant role in the inferred influence network. Another explanation for Mexico's behavior would be their geographical proximity to the United States, where many Twitter users are located. I assume that despite the different language, Mexicans are strongly connected to the Americans and may be influenced by them. However, this is only an assumption that would be interesting to analyze in future studies.

Toole et al. (2012, p. 8) concluded in their study about the spreading process of the innovation Twitter that "early adopting cities tend to be those with large, young, and tech-savvy populations". Linguistic research has shown that young people tend to use more innovative language forms (Androustopoulos, 2005; Merchant, 2001). Unfortunately, the data contains no information about the demographic characteristics of Twitter users. Thus, there is no information about the age of a Twitter user. Some studies tried to estimate the age of Twitter users based on their names (Longley and Adnan, 2016; Pavalanathan and Eisenstein, 2015), but this is outside the scope of this work. It was mentioned earlier that Twitter users are not a representative sample of the population, but nevertheless I include some numbers about the median population age of the regions. According to United Nations population statistics the median age of the Argentinian population in 2020 is 31.5 years (United Nations, 2019). Uruguay has a median age of 35.8 years and Mexico has a median age of only 29.2 years in 2020 (United Nations, 2019). Spain, on contrary, has a median age of 44.9 years in 2020, which is significantly higher than in the other regions (United Nations, 2019). The numbers show that the population in Latin America and South America is younger than the population in Spain. So even though the median age of Twitter users does not have to be equally distributed, it is still likely that Argentine (and Uruguayan) Twitter users are younger than Spanish Twitter users and thereby probably also more innovative. Note that this approach alone does not explain the results, as Mexico has the youngest population but

6. Discussion

is not the most innovative. How actively Twitter users are probably also plays a role in how innovative regions are. As table 3.2 shows, Twitter users in Argentina and Uruguay are particularly active.

Moreover, Wolfram and Schilling-Estes (2003) argue that in denser populated areas more interpersonal contact can be found, which in turn promotes the diffusion of innovations. Perhaps in Argentina, a larger percentage of Twitter users can be located in cities, probably especially in Buenos Aires. Whereas Spanish speaking European Twitter users may be diffused more equally in space.

Interestingly, Argentina and Mexico do not seem to influence each other despite being in similar time zones. Thus, the time zones do not seem to have an influence on the connections between regions.

A Hawkes process model is used to infer the structure of the unknown underlying network over which innovations propagate (Zhao et al., 2015). But it is not only used to model an unknown network, but also to simulate future events and predict the time and volume of future events (Rizoiu et al., 2017). Hence, the inferred network of a Hawkes process concurrently visualizes the most likely diffusion of a future innovation. New Spanish language forms are therefore more likely to transfer from Latin America to Europe than vice versa.

6.4. Materials and methods

6.4.1. Data-related limitations

The Twitter dataset brings some limitations. Due to the immense data volume of tweets in the online archive, only tweets from the first week of each month are considered in the analysis. The gaps bring some uncertainties since one does not know how the data behave in the missing time periods.

Even though the way how innovations were defined in this study allows an interesting selection of innovations, the approach also has some weaknesses. The infrequent occurrence of some innovations shows that just because people perceive some language forms as innovative, there is no guarantee that these phenomena actually show an increasing popularity over time. Other studies define innovations by their increasing popularity. They determine innovations from the data itself by looking for words that occur more and more often (Eisenstein et al., 2014). In addition, due to the time-consuming pre-processing, I consider only a small set of innovations, many of which do not occur often. For these two reasons, the dataset of innovations is rather small for the methods used in this study.

6. Discussion

As mentioned earlier, the innovations may be biased because of the selection method. Since linguists speaking European Spanish have selected the innovations, new language forms from Spain may be over-represented. Hence, the set of innovations is not a representative sample of language forms from all analyzed regions.

Furthermore, considering more popular innovations would allow the exclusion of retweets. The use of an innovation in a retweet is though an example of a stimulation - the user has perceived the new language form. But would the retweeting user use the linguistic innovation her- or himself? In other words, there is an input but not necessarily also an output. Just because people have seen a new language form, they do not necessarily have to use it in their own daily life. In research, both approaches can be found. Some studies take retweets into account and some do not. For example Eisenstein (2014) and Eisenstein et al. (2012) exclude retweets in their analysis, but Jones (2015) and Kamath et al. (2013) likely include retweets because they did not comment on their specific handling.

A final limiting factor of the dataset is the preprocessing approach used to get tweets with some geoinformation. Only a few users geotag their tweets. In order to extend the set of tweets with some geoinformation, the location information provided in the user profile was also used to locate tweets. So, I made the simplifying assumption that all tweets from a user are posted within the region defined in the user profile. Due to the increasing number of spatial studies using Twitter, there is a lot of research on techniques to obtain geospatial information from Twitter data. Several techniques have been proposed to infer locations on Twitter, e.g., by using the message context, the social network of a Twitter user, the information in the user profile, or the information about time zones (Ajao et al., 2015). For example Mahmud et al. (2012, 2014) present a new algorithm to identify the home location of Twitter users by using the content of their tweet and their tweeting behavior. By analyzing movement variations, they are even able to predict whether a user was travelling and use that information to improve the accuracy of location detection. So the inference of Twitter user's home location is a complex field of research on its own, which is not explored in more depth in this study. Most tweets could be localized by the information in the user profile and were most likely posted there, but note that a small uncertainty remains in the data.

6.4.2. Critical evaluation of methodologies

I used two different methods to estimate the influence or impact regions have on others in terms of language change. Thereby, the results can be compared and evaluated. This is particularly interesting due to the lack of ground truth data.

6. Discussion

In both used methodologies one has to deal with some uncertainties. The fundamental property of the spatial impact measure is time, just as in the Hawkes process. But while the Hawkes model considers random occurrences of events, the measure of spatial impact does not. The latter measure is less robust, because of the fixed comparison of points in time. For instance, if one innovation occurred in Uruguay and then 10 innovations occurred in Madrid, the spatial impact measure assumes that Uruguay fully influenced Madrid. Uruguay counts the fewest adopters, with only 426 Twitter users adopting an innovation. Even though Uruguay is not one of the regions where the analyzed innovations are likely to have emerged, new linguistic forms tend to be adopted early there. Therefore, the influence of Uruguay probably tends to be overestimated in the spatial impact model. Additionally, the spatial impact measure does not model the extent to which regions influence themselves. Self-influence is possible and is modeled in the Hawkes process, which is an advantage of this method.

The results of the two modeling approaches are similar - the Latin American regions influence the European regions. Uruguay seems to have again a role as an influencing region. In the Hawkes process it is unclear how the different amount of data per region affects the modelled processes. In the point plots in appendix C, one can also see how different the number of events per region is. In addition, the graphs also visualize that sometimes only very few events occur over the entire observation period. Usually Hawkes processes are used to model large networks with a few hundred or thousand nodes and even a larger number of events (can easily be in the millions) (Nickel and Le, 2020). For example, the ADM4-model proposed by Zhou et al. (2013) is originally designed to model the influences between people and thereby study social networks. Inferring a network of users of a platform can easily involve thousands of nodes in the resulting adjacency matrix. So the approach is designed to model large networks with millions of events and possibly a sparse structure of the network (Nickel and Le, 2020). The dimensions of the model in this study are very small with a network of only six nodes inferred from a few thousand events. The model is small scaled and the limited data is at the lower bound for this methodology. More regions and innovations would make the model more robust.

Another weakness of the Hawkes model that is worth to mention is the strong dependence on the input parameter "decay". The decay is beside the timestamps of events the only input to the model. The decay is estimated using maximum likelihood. However, Rizoiu et al. (2017) list some practical concerns of using maximum likelihood estimation for parameter estimation. First, they argue that the shape of the log of the likelihood function sometimes is complex with multiple local maxima. Thus, it may be difficult

6. Discussion

to identify the global maximum of the function. RizoIU et al. (2017) suggest to use several sets of initial values for the estimation, although they also note that it does not mitigate the concern entirely. Second, RizoIU et al. (2017) discuss concerns about edge effects. They argue that there may be unobserved events before the modeling period, which could have had an impact on the intensity function. One way to address this issue is the base intensity parameter that represents this edge effect. Finally, RizoIU et al. (2017) mention concerns about the computational costs for evaluating the maximum log-likelihood.

In addition, RizoIU et al. (2017) point out that power-law kernels provide higher prediction performances for social media compared to exponential kernels. The two kernels have a similar function, which is why it probably would not have influenced the model because of the few events inferred in this study. Nonetheless, in future studies it would be interesting to model the Hawkes process with a kernel well optimized for the data.

I modeled the process at the region level by aggregating all events occurring in a region into a group. Another option would have been to model the influence at the user level and only later group users into communities and infer spatial diffusion. Kersgaw et al. (2017) model user influence on language adoption, but not through a Hawkes process. Nevertheless, they distinguish between an analysis on the user level and on the community level. By grouping users into communities only later, they were able to analyze inter and intra community effects. Thus, they examine the influence that exists internal and external to communities and conclude that the influence of communities is greater than that of individuals.

Despite all the recent research on Hawkes process models described in chapter 2.4.3, the models lack geography. Addressing spatial issues in network modeling with a Hawkes process has rarely been done so far. Recently, Nickel and Le (2020) analyze the influence that states in the U.S. have on each other regarding the adoption of the same policy. They found that states tend to be more influential on their neighbors or on other regional states. However, New York and California share a similar behaviour regarding the adoption of policies, even though located on different coasts. In addition, Cai et al. (2021) proposed a topological Hawkes process which is based on the assumption that an event is excited not only by its history but also by its topological neighbors. They used the example of a mobile network, where the network stations build a topological structure. Alarms spreading across the network are not independent of their network station, but are excited or inhibited by events that occurred in the same network station.

6. Discussion

Although the topological Hawkes process proposed by Cai et al. (2021) was originally not designed for spatial analysis, it would probably fit well for it. In the future, hopefully, more analyses will analyze spatial data in a Hawkes process. The modeling and results in this work have shown that Hawkes processes can be used to study the spatial diffusion of linguistic innovations.

7. Conclusion

7.1. Summary and major findings

In this work, I analyzed the spatio-temporal dynamics of the diffusion of linguistic innovations through the social media channel Twitter. With the increasing availability of user-generated online media content, a new interdisciplinary research field combining geography and linguistics emerges. To analyse the spatial diffusion, over 57 million tweets with geoinformation between 2012 and 2019 are considered.

The analysis consisted of two key parts: First, the study examined the spatial properties of the diffusion process through descriptive statistics with the measures focus, entropy, spread and impact. Second, the study focused on inferring the network of linguistic influence by a Hawkes process.

The results of the first part of the analysis have shown that even though not all innovations diffuse wide in space, many innovations spread globally, suggesting that there are some interactions between Twitter users of the analyzed regions. The probability of a global diffusion of a phenomenon increases as the popularity of an innovation increases. Spatial distance does not seem to fully explain patterns of the innovation diffusion. Hence, continental borders do not seem to act as a linguistic barrier in this study.

The findings of the second stage of the analysis have shown that mainly Argentinian and Uruguayan Twitter users influence European Twitter users. By using the Hawkes process, I could distinguish between leaders and followers of language change. This approach is based on the idea that past adoptions of innovations trigger the occurrence of future adoptions of innovations. Finally, it is worth noting that the study demonstrates the successful application of a Hawkes model to infer the unknown network of linguistic influence.

7.2. Future work

For future studies, there are several interesting directions. The Hawkes process model not only allows to infer the structure of the unknown underlying network over which

7. Conclusion

innovations propagate, but also to simulate future innovations and predict the time and volume of future innovations. Further work may therefore focus on popularity predictions of different linguistic innovations on social platforms based on a Hawkes process model.

Furthermore, upcoming studies could not only model linguistic influence, but also try to infer the network of other cultural aspects. For example this could be done by analyzing the relationship between regions through observations on other platforms, such as information diffusion or rumor propagation on other social networks. Another example that could be interesting to explore further is the analysis of the diffusion of music tracks and videos. Music is a cultural aspect and the investigation in its distribution can lead to a better understanding of social relations between Spanish speaking regions. All in all, there are many exciting opportunities to analyze linguistic influence or even cultural influence between regions using a Hawkes process.

References

- Ajao, Oluwaseun, Jun Hong, and Weiru Liu (2015). “A survey of location inference techniques on Twitter.” In: *Journal of information science*. ISSN: 1741-6485. DOI: 10.1177/0165551510000000.
- Alvari, Hamidreza and Paulo Shakarian (2019). “Hawkes Process for Understanding the influence of Pathogenic Social Media Accounts.” In:
- Androutsopoulos, Jannis K. (2005). “Research on Youth-Language.” In: *Sociolinguistics/Soziolinguistik*, pp. 1496–1505.
- Bacry, E., M. Bompain, S. Gaïffas, and S. Poulsen (2017a). *tick*. URL: [2https://x-datainitiative.github.io/tick/index.html](https://x-datainitiative.github.io/tick/index.html) (visited on 03/10/2021).
- Bacry, Emmanuel, Martin Bompain, Stéphane Gaïffas, and Soren V. Poulsen (2017b). “tick: A Python library for statistical learning, with a particular emphasis on time-dependent modeling.” In: *arXiv*, pp. 1–8. ISSN: 23318422.
- Bacry, Emmanuel, Martin Bompain, Stéphane Gaïffas, and Jean-Francois Muzy (2020). “Sparse and low-rank multivariate Hawkes processes.” In: *Journal of Machine Learning Research* 21.1-32.
- Bacry, Emmanuel, Iacopo Mastromatteo, and Jean-François Muzy (2015). “Hawkes Processes in Finance.” In: *Market Microstructure and Liquidity* 01.01, p. 1550005. ISSN: 2382-6266. DOI: 10.1142/s2382626615500057.
- Bahamou, Achraf, Maud Doumergue, and Philippe Donnat (2019). “Hawkes processes for credit indices time series analysis: How random are trades arrival times?” In: *arXiv*. ISSN: 23318422.
- Bailey, Charles-James N. (1973). “Variation and Linguistic Theorie.” In: pp. 1–169. DOI: 10.1117/12.717847.
- Bailey, Guy, Tom Wikle, Jan Tillery, and Lori Sand (1993). “Some patterns of linguistic diffusion.” In: *Language Variation and Change* 5.
- Balusu, Murali Raghu Babu, Taha Merghani, and Jacob Eisenstein (2018). “Stylistic Variation in Social Media Part-of-Speech Tagging.” In: pp. 11–19. DOI: 10.18653/v1/w18-1602.
- Baptista, Rui (2001). “Geographical Clusters and Innovation Diffusion.” In: *Technological Forecasting and Social Change* 66.1, pp. 31–46. ISSN: 00401625. DOI: 10.1016/S0040-1625(99)00057-8.

References

- Blank, Grant (2017). “The Digital Divide Among Twitter Users and Its Implications for Social Research.” In: *Social Science Computer Review* 35.6, pp. 679–697. ISSN: 15528286. DOI: 10.1177/0894439316671698.
- Boberg, Charles (2000). “Geolinguistic diffusion and the U . S . – Canada border.” In: *Language Variation and Change* 12.2000, pp. 1–24.
- Britain, David (2018). “Space, Diffusion and Mobility.” In: *The Handbook of Language Variation and Change*. Ed. by J.K. Chambers and Natalie Schilling. Second. Blackwell Publishers, pp. 471–500.
- Brodersen, Andres, Salvatore Scellato, and Mirjam Wattenhofer (2012). “YouTube Around the World: Geographic Popularity of Videos.” In: *International World Wide Web Conference Committee (IW3C2) WWW 2012*.
- Cai, Ruichu, Siyu Wu, Jie Qiao, Zhifeng Hao, Keli Zhang, and Xi Zhang (2021). “THP: Topological Hawkes Processes for Learning Granger Causality on Event Sequences.” In: 14.8, pp. 1–12. URL: <http://arxiv.org/abs/2105.10884>.
- Chu, Zi, Steven Gianvecchio, Haining Wang, and Sushil Jajodia (2010). “Who is tweeting on twitter: Human, bot, or cyborg?” In: *Proceedings - Annual Computer Security Applications Conference, ACSAC*, pp. 21–30. ISSN: 10639527. DOI: 10.1145/1920261.1920265.
- Crystal, David (2011). *Internet Linguistics*. ISBN: 9780415602686. DOI: 10.4324/9780203830901.
- Danescu-Niculescu-Mizil, Cristian, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts (2013). “No country for old members: User lifecycle and linguistic change in online communities.” In: *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web*, pp. 307–317.
- Ding, Jiinvan, Luis Gravano, and Narayanan Shivakumar (2000). “Computing geographical scopes of web resources.” In: *Proceedings of the 26th International Conference on Very Large Data Bases, VLDB’00*, pp. 545–556.
- Doyle, Gabriel (2014). “Mapping dialectal variation by querying social media.” In: *14th Conference of the European Chapter of the Association for Computational Linguistics 2014, EACL 2014*, pp. 98–106. DOI: 10.3115/v1/e14-1011.
- Dwyer, Barry (2016). “Chapter 2 - Mathematical background.” In: *Systems Analysis and Synthesis*. Ed. by Barry Dwyer. Boston: Morgan Kaufmann, pp. 23–78. ISBN: 978-0-12-805304-1. DOI: <https://doi.org/10.1016/B978-0-12-805304-1.00011-4>. URL: <https://www.sciencedirect.com/science/article/pii/B9780128053041000114>.
- Eisenstein, Jacob (2013). “What to do about bad language on the internet.” In: *NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference* June, pp. 359–369.

References

- Eisenstein, Jacob (2014). “Identifying regional dialects in on-line social media.” In: *The Handbook of Dialectology*. DOI: 10.1002/9781118827628.ch21.
- Eisenstein, Jacob, Brendan O’Connor, Noah A. Smith, and Eric P. Xing (2012). “Mapping the geographical diffusion of new words.” In: *arXiv.org*.
- (2014). “Diffusion of lexical change in social media.” In: *PLoS ONE* 9.11, pp. 1–13. ISSN: 19326203. DOI: 10.1371/journal.pone.0113114.
- Esri (2015). *ArcGIS Hub - Countries WGS84*. URL: https://hub.arcgis.com/datasets/a21fdb46d23e4ef896f31475217cbb08_1/explore?location=-0.001057%2C4.230974%2C0.02 (visited on 02/09/2021).
- Eurostat (2020). *Geographical information and maps - Administrative Units / Statistical Units*. URL: <https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/administrative-units-statistical-units/nuts> (visited on 02/09/2021).
- Farajtabar, Mehrdad, Jiachen Yang, Xiaojing Ye, Huan Xu, Rakshit Trivedi, Elias Khalil, Shuang Li, Le Song, and Hongyuan Zha (2017). “Fake News Mitigation via Point Process Based Intervention.” In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1097–1106. URL: <http://proceedings.mlr.press/v70/farajtabar17a.html>.
- Finin, Tim, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze (2010). “Annotating Named Entities in Twitter Data with Crowdsourcing.” In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk 2010*. January, pp. 80–88.
- GeoMidpoint.com (2021). *Geographic Midpoint Calculator*. URL: <http://www.geomidpoint.com> (visited on 03/15/2021).
- Gonçalves, Bruno and David Sanchez (2014). “Crowdsourcing dialect characterization through twitter.” In: *PLoS ONE* 9.11, pp. 1–6. ISSN: 19326203. DOI: 10.1371/journal.pone.0112074.
- Graham, Mark, Scott A. Hale, and Devin Gaffney (2013). “Where in the world are you ? Geolocation and language identification in Twitter.” In: *Professional Geographer* 2013, pp. 1–17.
- Grier, Chris, Kurt Thomas, Vern Paxson, and Michael Zhang (2010). “@spam: The Underground on 140 Characters or Less Categories and Subject Descriptors.” In: *Proceedings of the 17th ACM conference on Computer and communications security*, pp. 27–37. ISSN: 15437221. URL: <http://portal.acm.org/citation.cfm?id=1866307.1866311>.
- Grieve, Jack, Andrea Nini, and Diansheng Guo (2018). “Mapping Lexical Innovation on American Social Media.” In: *Journal of English Linguistics* 46.4, pp. 293–319. ISSN: 15525457. DOI: 10.1177/0075424218793191.

References

- Hawkes, Alan G. (1971). “Spectra of some self-exciting and mutually exciting point processes.” In:
- Helmstetter, Agnès and Didier Sornette (2003). “Predictability in the Epidemic-Type Aftershock Sequence model of interacting triggered seismicity.” In: *Journal of Geophysical Research: Solid Earth* 108.B10. ISSN: 0148-0227. DOI: 10.1029/2003jb002485.
- Horvath, Barbara M. and Ronald J. Horvath (1997). “The geolinguistics of a sound change in progress: /l/ vocalization in Australia.” In: *University of Pennsylvania Working Papers in Linguistics* 4, pp. 109–124. ISSN: 0075-4242. DOI: 10.1177/007542429702500207.
- Huang, Yuan, Diansheng Guo, Alice Kasakoff, and Jack Grieve (2016). “Understanding U.S. regional linguistic variation with Twitter data analysis.” In: *Computers, Environment and Urban Systems* 59, pp. 244–255. ISSN: 01989715. DOI: 10.1016/j.compenvurbsys.2015.12.003. URL: <http://dx.doi.org/10.1016/j.compenvurbsys.2015.12.003>.
- Internet Archive (2020). *Twitter Stream*. URL: <https://archive.org/details/twitterstream?and%5B%5D=year%3A%222014%22> (visited on 09/28/2020).
- Jones, Lucy (2010). *The chnaging face of spelling on the internet*. Tech. rep.
- Jones, Taylor (2015). *Toward a description of African American vernacular english dialect regions using "black twitter"*. Vol. 90. 4, pp. 403–440. ISBN: 0003128334. DOI: 10.1215/00031283-3442117.
- Kamath, Krishna Y., James Caverlee, Kyumin Lee, and Zhiyuan Cheng (2013). “Spatio-temporal dynamics of online memes: A study of geo-tagged tweets.” In: *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web*, pp. 667–677.
- Kersgaw, Daniel, Matthew Rowe, Anastasios Noulas, and Patrick Stacey (2017). “Birds of a Feather Talk Together: User Influence on Language Adoption.” In: *Proceedings of the 50th Hawaii International Conference on System Sciences (2017)*, pp. 1851–1860. DOI: 10.24251/hicss.2017.225.
- Kobayashi, Ryota and Renaud Lambiotte (2016). “TiDeH: Time-dependent Hawkes process for predicting retweet dynamics.” In: *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016* ICWSM, pp. 191–200.
- Kulshrestha, Juhi, Farshad Kooti, Ashkan Nikraves, and Krishna P. Gummadi (2012). “Geographic dissection of the Twitter network.” In: *ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, pp. 202–209.
- Labov, William (2001). *Principles of Linguistic Change, vol. 2: Social Factors*. Oxford: Blackwell Publishers.
- (2003). “Principles of Linguistic Change . Volume 2 : Social Factors by William Labov.” In: 45.4.

References

- Leetaru, Kalev H., Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook (2013). "Mapping the global Twitter heartbeat: The geography of Twitter." In: *First Monday* 18.5, pp. 18–20.
- Li, Linna, Michael F. Goodchild, and Bo Xu (2013). "Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr." In: *Cartography and Geographic Information Science* 40.2, pp. 61–77. ISSN: 15230406. DOI: 10.1080/15230406.2013.777139.
- Longley, Paul A. and Muhammad Adnan (2016). "Geo-temporal Twitter demographics." In: *International Journal of Geographical Information Science* 30.2, pp. 369–389. ISSN: 13623087. DOI: 10.1080/13658816.2015.1089441. URL: <http://dx.doi.org/10.1080/13658816.2015.1089441>.
- Mahmud, Jalal, Jeffrey Nichols, and Clemens Drews (2012). "Where is this tweet from?: Inferring home locations of Twitter users." In: *ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, pp. 511–514.
- (2014). "Home location identification of twitter users." In: *ACM Transactions on Intelligent Systems and Technology* 5.3. ISSN: 21576912. DOI: 10.1145/2528548.
- Maybaum, Rebecca (2013). "Language Change as a Social Process: Diffusion Patterns of Lexical Innovations in Twitter." In: *Berkeley Linguistics Society*, pp. 152–166. URL: <http://dx.doi.org/10.3765/bls.v39i1.3877>.
- Merchant, Guy (2001). "Teenagers in cyberspace: an investigation of language use and language change in internet chatrooms." In: *Journal of Research in Reading* 24.3, pp. 293–306.
- Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist (2011). "Understanding the Demographics of Twitter Users." In: *Int'l AAAI Conference on Weblogs and Social Media (ICWSM)* January.
- Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley (2013). "Is the sample good enough? Comparing data from twitter's streaming API with Twitter's firehose." In: *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, pp. 400–408.
- Nickel, Maximilian and Matthew Le (2020). "Learning multivariate hawkes processes at scale." In: *arXiv*. ISSN: 23318422.
- Nie, H. Ruda, Xiuzhen Zhang, Minyi Li, Anil Dolgun, and James Baglin (2020). *Modelling User Influence and Rumor Propagation on Twitter using Hawkes Processes*.
- Niwattanakul, Suphakit, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu (2013). "Using of jaccard coefficient for keywords similarity." In: *Lecture Notes in Engineering and Computer Science* 2202, pp. 380–384. ISSN: 20780958.

References

- Ogata, Yoshihiko (1988). “Statistical models for earthquake occurrences and residual analysis for point processes.” In: *Journal of the American Statistical Association* 83.401, pp. 9–27. ISSN: 1537274X. DOI: 10.1080/01621459.1988.10478560.
- (1999). “Seismicity analysis through point-process modeling: A review.” In: *Pure and Applied Geophysics* 155.2-4, pp. 471–507. ISSN: 00334553. DOI: 10.1007/s000240050275.
- Pavalanathan, Umashanthi and Jacob Eisenstein (2015). “Confounds and consequences in geo-tagged twitter data.” In: *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing* September, pp. 2138–2148. DOI: 10.18653/v1/d15-1256.
- Reinhart, Alex and Joel Greenhouse (2018). “Self-exciting point processes with spatial covariates: modelling the dynamics of crime.” In: *Journal of the Royal Statistical Society. Series C: Applied Statistics* 67.5, pp. 1305–1329. ISSN: 14679876. DOI: 10.1111/rssc.12277.
- Rizoiu, Marian Andrei, Young Lee, Swapnil Mishra, and Lexing Xie (2017). “A tutorial on Hawkes processes for events in social media.” In: *arXiv*, pp. 1–26. ISSN: 23318422.
- Rogers, Everett M. (2003). *Diffusion of Innovations - Fifth Edition*. New York: Simon and Schuster.
- Sadilek, Adam, Henry Kautz, and Jeffrey P. Bigham (2012). “Finding your friends and following them to where you are.” In: *WSDM 2012 - Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, pp. 723–732. DOI: 10.1145/2124295.2124380.
- Sloan, Luke and Jeffrey Morgan (2015). “Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on twitter.” In: *PLoS ONE* 10.11, pp. 1–15. ISSN: 19326203. DOI: 10.1371/journal.pone.0142209.
- Sloan, Luke, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, and Omer Rana (2013). “Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter.” In: *Sociological Research Online* 18.7.
- Sornig, Karl (1981). *Lexical innovation: A study of slang, colloquialisms and casual speech*. John Benjamins Publishing.
- Stomakhin, Alexey, Martin B. Short, and Andrea L. Bertozzi (2011). “Reconstruction of missing data in social networks based on temporal patterns of interactions.” In: *Inverse Problems* 27.11, pp. 1–15. ISSN: 02665611. DOI: 10.1088/0266-5611/27/11/115013.
- Takahashi, Tetsuro, Shuya Abe, and Nobuyuki Igata (2011). “Can Twitter be an alternative of real-world sensors?” In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6763 LNCS.PART 3, pp. 240–249. ISSN: 03029743. DOI: 10.1007/978-3-642-21616-9_{_}27.

References

- Toole, Jameson L., Meeyoung Cha, and Marta C. González (2012). “Modeling the adoption of innovations in the presence of geographic and media influences.” In: *PLoS ONE* 7.1. ISSN: 19326203. DOI: 10.1371/journal.pone.0029528.
- Tredici, Marco Del and Raquel Fernandez (2018). “The Road to success: Assessing the fate of linguistic innovations in online communities.” In: *arXiv*. ISSN: 23318422.
- Trudgill, Peter (1974). “Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography.” In: *Language in Society* 3.2, pp. 215–246. ISSN: 14698013. DOI: 10.1017/S0047404500004358.
- Twitter (2021). *Counting characters*. URL: <https://developer.twitter.com/en/docs/counting-characters> (visited on 06/18/2021).
- United Nations (2019). *World Population Prospects 2019*. URL: <https://population.un.org/wpp/DataQuery/> (visited on 06/06/2021).
- Wolfram, Walt and Natalie Schilling-Estes (2003). “Dialectology and Linguistic Diffusion.” In: *The Handbook of Historical Linguistics*. Ed. by Brian D. Joseph and Richard D. Janda. Blackwell Publishers. Chap. 24, pp. 713–735. URL: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470756393#page=727>.
- Wood-Doughty, Zach, Michael Smith, David Broniatowski, and Mark Dredze (2017). “How Does Twitter User Behavior Vary Across Demographic Groups?” In: pp. 83–89. DOI: 10.18653/v1/w17-2912.
- Yang, Yi and Jacob Eisenstein (2017). “Overcoming Language Variation in Sentiment Analysis with Social Attention.” In: *Transactions of the Association for Computational Linguistics* 5, pp. 295–307.
- Zhang, Xin, Ding Ding Han, Ruiqi Yang, and Ziqiao Zhang (2017). “Users’ participation and social influence during information spreading on Twitter.” In: *PLoS ONE* 12.9, pp. 1–17. ISSN: 19326203. DOI: 10.1371/journal.pone.0183290.
- Zhao, Qingyuan, Murat A. Erdogdu, Hera Y. He, Anand Rajaraman, and Jure Leskovec (2015). “SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity.” In:
- Zhou, Ke, Hongyuan Zha, and Le Song (2013). “Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes.” In: *Journal of Machine Learning Research* 31, pp. 641–649. ISSN: 15337928.
- Zhu, Shixiang and Yao Xie (2019). “Spatial-Temporal-Textual Point Processes with Applications in Crime Linkage Detection.” In: *arXiv*, pp. 1–35. ISSN: 23318422.

Appendix

A. List of linguistic innovations

1. **En mi libro XXX explico cómo... - ‘In my book XXX I explain...’**

The title of the (invented) book is an ironic comment on whatever its content is said to be, which is normally something that the user typically does. For instance, “In my book ‘Go to hell, you slow driver’ I explain how to drive slowly and carefully”.

[Productive phraseology]

2. **Ojalá - ‘I wish’**

Ojalá is a rather special adverb in Spanish, because it introduces sentences. These sentences have a verb in the subjunctive in Standard Spanish, but in Twitter Spanish infinitives and gerunds are allowed.

[Morphosyntactic innovation]

3. **Acompáñenme/acompañadme en / a ver esta triste historia - ‘Come and see this sad story’**

Used to comment on incoherences said by other users (such tweets typically show pictures of or links to someone else’s tweets, which say contradictory things).

[Productive multimedia phraseology]

4. **Emosido engañados. - ‘We’ve been lied to’**

Its meaning is quite literal, although it’s often used ironically. The catch is its non-standard orthography (in Standard Spanish it should read Hemos sido engañados) and its origin is a graphitti that went viral. There’s a song from 2020 and in 2020 someone found the place where the picture had been taken, but we don’t have data from 2020!

[Fixed phraseology. Non-standard orthography]

5. **En fin/Ay, la hipotenusa/hipocondría/hipotalámica - ‘You know, the hypotenuse’** Wordplay, from “hypocresy”, just about any Word that starts with hypo-

Appendix

can be used. It's used to comment on contradictions and incoherences (you know, people complain a lot on Twitter!).

[Productive phraseology]

6. Gracias por venir a mi TedTalk. - 'Thanks for coming to my Ted Talk'

Used as a way to end a thread or some kind of long and hence potentially boring explanation.

[Fixed phraseology]

7. Inyustisia. - 'injustice'

Literal meaning, but orthography is used to imitate Portuguese pronunciation: it's an imitation of Cristiano Ronaldo's complaints. Cristiano Ronaldo's video is from 2012.

[Non-standard orthography]

8. jaja para k kieres saber eso - 'hahaha why do you want to know that'

Ironically used to signal that some question is quite on point but has no easy answer (or an answer people don't want to hear). It's origin is an answer in Yahoo answers from 2009, but for some reason it went viral on 2016, unclear why. One of the articles says it started in Mexico, confirmed by the data

[Fixed phraseology. Non-standard orthography]

9. ke me da iwá - 'I really don't care'

Literal meaning.

[Fixed phraseology. Non-standard orthography]

10. Madre mía los haters - 'OMG the haters'

Literal meaning.

[Fixed phraseology]

12. Salta la sorpresa en Las Gaunas - 'Surprise at Las Gaunas'

Las Gaunas is the football stadium of a not very good team, so this is a sentence that can be heard when radio or tv commentators are retransmitting a match there and the local team scores. In Twitter is used to comment on any unexpected event.

[Fixed phraseology, not originated in Twitter, but it's probably fair to assume that its meaning got expanded mostly there.]

Appendix

13. Suélteme del brazo, señor/a - ‘Stop grabbing my arm, sir/madam’

As a way to signal that someone is insisting on discussing a topic after the point where you (or everyone) are sick of it already.

[Fixed phraseology]

15. Bro, sos famoso/a - ‘Bro, you’re famous’

What you say after someone’s tweet got viral. It’s quite clearly Argentinian/Uruguayan Spanish.

[Fixed phraseology]

16. Cuando lo pides por AliExpress, Cuando te llega - ‘When you ask for it at AliExpress / When you receive it’

Used to add pictures or tweets of something that looked better than it ended up being.

[Productive multimedia phraseology]

17. Lo que pasó a continuación te sorprenderá... - ‘What happened next will surprise you’

Used to comment on contradictions by some other user. It imitates clickbait headlines.

[Productive multimedia phraseology]

18. Me lo creo. Y me cuadra - ‘I believe so. And it checks out’

Literal meaning.

[Fixed phraseology]

19. Ni un tuit sin su errata/etarra/rata - ‘Not a single tweet without a typo’

Literal meaning, but with wordplay based on the word errata ‘typo’.

[Fixed phraseology]

20. No lo sé, Rick, parece falso. - ‘I don’t know, Rick, it looks fake’

Used to comment on something you believe it’s a lie. The name (Rick) comes from an American TV show, although it seems that they never actually said this sentence.

[Fixed phraseology]

21. No tengo pruebas, pero tampoco dudas. - ‘I have no evidence, but no doubts either’

Literal meaning.

[Fixed phraseology]

22. Que me pise la cara - ‘Please, let s/he step on my face’

To indicate that you have such a big crush on someone that you wouldn’t mind even if they stepped on your face.

[Fixed phraseology]

23. Se tenía que decir y se dijo - ‘It had to be said and it was said’

Quite literal meaning, although often used ironically.

[Fixed phraseology]

24. Un saludo a Pablo Sobrado - ‘Regards to Pablo Sobrado’

Pablo Sobrado is a guy who collected and retweeted funny tweets. Someone once commented on a funny thing saying Regards to Pablo Sobrado, implying that they already knew that Pablo Sobrado would see that funny thing and comment on it and it ended up becoming viral.

[Fixed phraseology]

25. yen2 - ‘going’

Playing with the pronunciation of numbers to incorporate them in as shortened orthography (as in h8 ‘hate’..)

[Non-standard orthography]

26. Noso3 - ‘we’

Playing with the pronunciation of numbers to incorporate them in as shortened orthography (as in h8 ‘hate’..). The cool thing here is that 3 implies a non-existing ending in Standard Spanish (-es), but which is the ending that is more spread right now to refer to non-binary people or as gender-inclusive form in general.

[Non-standard orthography + Morphosyntactic innovation]

27. abro melón - ‘I’ll open the melon’

It means ‘Let’s start talking about this topic’. It most likely did not originate on Twitter, but became very frequent there due to the fact that people are always looking for new topics of discussion.

[Fixed phraseology]

29. La ciencia no se ace sola, ay k acerla - ‘Science is not done on its own, it has to be done’

Used to comment after someone gives actual data on a topic, but also to indicate that you need to leave Twitter because work is calling. Both the literal sentence and the non-standard orthography come from a Twitter account which is a parody of a standard scientist and writes with impossible spelling and a rather childish style.

[Fixed phraseology. Non-standard orthography]

30. ¿Sabén quién era ese niño? / Ese niño era el mismísimo XX - ‘Do you know who was that kid? That kid was the very XX’

Used to comment on either incredible or completely boring stories about children. The child is often said to be either Albert Einstein or Adolf Hitler (depending on whether you’re commenting on clever or mean things).

[Fixed phraseology]

31. No te lo perdonaré jamás [X]. Jamás. - ‘I will never forgive you, X. Ever.’

Literal meaning, but typically used ironically. It started with a tweet by a politician on January 5th 2016 due to something she disliked from the Three wise men parade)

[Semi-productive phraseology]

32. Que Twitter haga su magia - ‘Let Twitter do its magic’

Used when you’re asking the community for something (adopting a kitty, finding a job, just anything really...)

[Fixed phraseology]

33. Me pide me perdona - ‘S/he should ask for my forgiveness’

Literal meaning. The syntax is completely wrong and the sentence actually has not real meaning in Spanish, but it was coined by a second-language speaker (of Chinese origin) and went viral after one of his employees told the story where he said it.

[Fixed phraseology]

34. ¿Pregunta? Pregunta. - ‘¿Question? Question’

The structure of these tweets is very abstract: you ask something and reply (affirma-

Appendix

tively) with the exact same words. For instance: Am I trying to open a beer with my bare hands? I am trying to open a beer with my bare hands. (In Spanish this works better, because you can use the exact same sentence, you don't need the subject-verb inversion that you have in English or German questions: I am > am I, mostly because you don't need to say the subject.)

[Productive phraseology]

Coloquialismos

39. que emocionó a Steven Spielberg - 'that made Spielberg excited'

It comes from the ad of a TV series (the TV series that made Spielberg excited) and is used to exaggerate the characteristics of whatever thing you want to praise.

[Productive phraseology]

44. -érrimo

Intensification suffix that in Standard Spanish can be used only with 10 adjectives but that in colloquial Spanish can be used with any adjective (and even other words). Did not originate on Twitter, but is quite productive there.

[Morphosyntax]

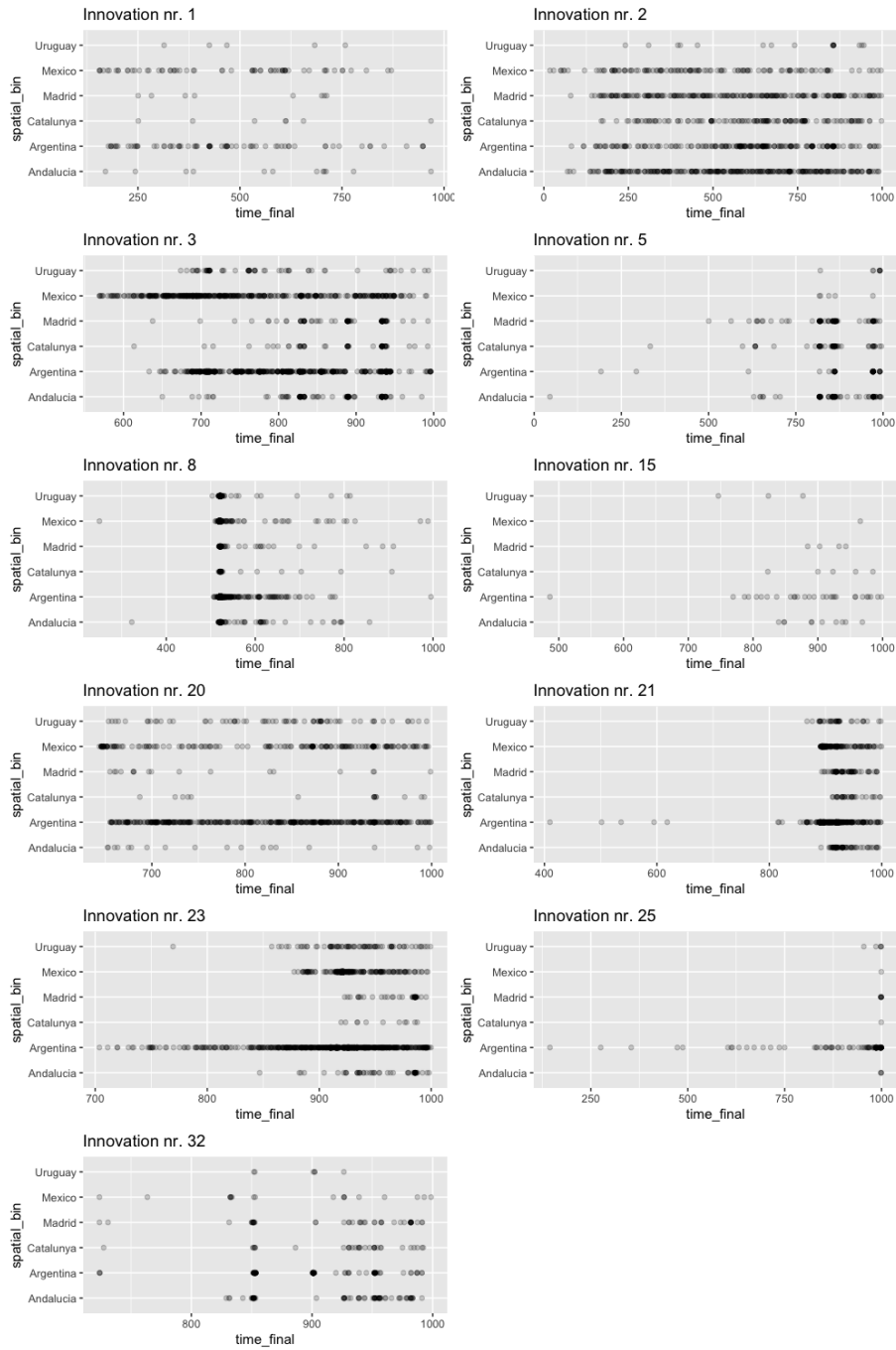
Appendix

B. Documentation Twitter corpus

Year	Month	Day	Notes
2012	01 - 12	01 - 07	No missing data in 2012
2013	01 - 12	01 - 07	No missing data in 2013
2014	01	25.-31. Dec. 2013	month is missing - previous month used
2014	02	06 - 12	
2014	03 - 12	01 - 07	
2015	01	25.-31. Dec 2014	month is missing - previous month used
2015	02	NA	month and previous month is missing - gap could not be filled
2015	03	23 - 29	
2015	04 - 10	01 - 07	
2015	11	02 - 08	
2015	12	01 - 05, 16, 17	
2016	01 - 02	01 - 07	
2016	03	01, 03, 08, 10, 11, 13, 14	
2016	04 - 05	01 - 07	
2016	06	01, 09- 14	
2016	07 - 09	01 - 07	
2016	10	01 - 05, 07, 08	
2016	11 - 12	01 - 07	
2017	01	01 - 07	
2017	02	01, 02, 05 - 09	
2017	03 - 05	01 - 07	
2017	06	01, 02, 04- 08	
2017	07	01 - 07	
2017	08	01, 03 - 08	
2017	09 - 12	01 - 07	
2018	01	02 - 08	
2018	02	01, 02, 05 - 09	
2018	03 - 04	01 - 07	
2018	05	02 - 07, 24	
2018	06	07 - 13	
2018	07	01 - 03, 12 - 15	
2018	08 - 12	01 - 07	
2019	01	01, 02, 10, 13 - 16	
2019	02	02 - 08	
2019	03	01 - 07	
2019	04	04- 10	
2019	05 - 09	01 - 07	
2019	10	01 - 05, 07, 08	
2019	11 - 12	01 - 07	

Appendix

C. Point plots



Appendix

D. Adjacency matrix in numbers

D.1 Adjacency Matrix of Model 1

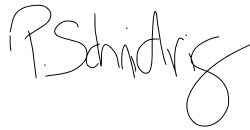
	Mexico	Madrid	Catalunya	Andalucia	Argentina	Uruguay
Mexico	0.35836	0.00003	0.02121	0.52326	0.50878	0.42731
Madrid	0.17637	0.03893	0.00332	0.30559	0.62875	2.30455
Catalunya	0.53860	0.46519	0.20417	0.00000	0.20837	2.32255
Andalucia	0.01482	0.00926	0.00281	0.08138	0.25198	0.71929
Argentina	0.01864	0.08491	0.02793	0.46837	0.15640	0.22632
Uruguay	0.03792	0.00000	0.04789	0.16973	0.04120	0.00000

D.2 Adjacency Matrix of Model 1

	Mexico	Madrid	Catalunya	Andalucia	Argentina
Mexico	0.34250	0.11593	0.04470	0.46065	0.22331
Madrid	0.23234	0.04611	0.03531	0.33849	1.53496
Catalunya	0.47901	0.47286	0.21251	0.15747	1.05071
Andalucia	0.10075	0.01155	0.16488	0.04200	0.26273
Argentina	0.02978	0.05776	0.05262	0.26132	0.11482

Personal Declaration

I hereby declare that the submitted thesis is the result of my own, independent work.
All external sources are explicitly acknowledged in the thesis.

A handwritten signature in black ink, appearing to read 'P. Schmidrig'. The signature is written in a cursive style with a large, looped 'P' and 'g'.

Patricia Schmidrig
Zurich, 25.06.2021