MSc Thesis (GEO 511)

A Spatial Analysis of German Lemma Variation with Twitter

André Rodrigues

s08-717-522
andre.rodrigues@uzh.ch

Supervisors:

Dr. Hanna Ruch

UFSP Sprache und Raum
Plattenstrasse 54, 8032 Zürich
hanna.ruch@uzh.ch
University of Zurich

Prof. Dr. Ross S. Purves

ross.purves@geo.uzh.ch

Faculty Member:

Prof. Dr. Ross S. Purves

GIScience: Geocomputation Unit
Department of Geography
University of Zurich

Handed in on the 30th of June, 2015

Acknowledgement

This has been a long journey and I would probably not have come this far without the help of some people. They were there to help me when I hit some bump along the road or to point me again in the right direction so I could continue with my journey. Therefore, at this point, I would like to express my deepest gratitude to the following persons:

- Prof. Dr. Ross Purves, my supervisor, for the conversations, the valuable comments and helpful hints, which helped me to stay focussed and not to go over the top.

- Hanna Ruch, my co-supervisor, for her disposition to help me with her insights, and for her countless advices in the field of linguistics.

- Oliver Zihler, for his valuable help with the coding aspects of my thesis and for taking his time to support me when some code I wrote did not work as intended, or at all.

- Prof. Dr. Hans Bickel and Christoph Landolt for allowing me to use their data from the *Schweizerhochdeutsch: Wörterbuch der Standardsprache in der deutschen Schweiz*.

- My parents for making it possible for me to pursue my desired study and for their unlimited support and kindness.

- My sister, for her efforts trying to motivate me and cheer me up during this thesis.

- And finally my friends, who always gave me a good reason to get away from the computer and this thesis to get some fresh air.

“Pluricentric languages are both unifiers and dividers of peoples.”

(Clyne 1992: 1)

Abstract

Over the last few years we witnessed how the number of mobile devices rapidly increased. Smartphones like the iPhone have changed our everyday life and are now an integral part of our daily routine. The majority of such devices holds a GPS receptor, which can provide local based information like availability of restaurants or specific directions. While using those devices we generate an enormous amount of georeferenced data that could lead to new and interesting insights about our surrounding world.

The dispersion of such devices and the emergence of services like Facebook and Twitter have also changed how we communicate. Interaction is no longer restricted to geographical boundaries and can occur independent of time and space. The English language established itself as the main language on the internet and other web services. However, the relationship between place and language still remains relatively unexplored. The majority of the research conducted on this subject focused on the English language, neglecting languages less common on the World Wide Web.

To help overcoming this research gap this thesis uses the micro-blogging service Twitter to analyse lemma variation for the pluricentric German language. Almost half a million of German tweets were collected for this purpose. Since German lemma variation has never been studied on Twitter, a new approach was presented in this study. The noisy characteristic of Twitter data and the dynamics of language have been confirmed to be complex aspects that cannot be fully considered when following an automated approach. Based on a geographical classification and the statistical χ^2 test we could demonstrate that cross-country variation of certain lemmas can also be observed on Twitter and it not restricted to schoolbooks and literature. Based on the insights gained in this study we are able to give some recommendations on how to improve further studies in this subject and hope that this study motivates further research in this field.

Zusammenfassung

Die Anzahl mobiler Geräte ist in den letzen Jahren rasch angestiegen. Smartphones wie das iPhone haben unseren Alltag verändert und sind mittlerweile ein fester Bestandteil dessen. Die Mehrheit dieser Geräte enthält ein GPS-Empfänger. Dieser kann lokal-gebundene Informationen geben, wie der Ort eines Restaurants oder einem helfen eine Adresse zu finden. Während wir diese Geräte benutzen, erzeugen wir eine enorme Menge an georeferenzierten Daten, die zu neuen und interessanten Erkenntnissen über unsere Umgebung beitragen können.

Des Weiteren hat die Verbreitung solcher Geräte, und das Auftauchen von Diensten wie Facebook und Twitter, die Art verändert wie wir kommunizieren. Die Interaktion ist nicht mehr an geographische Grenzen gebunden und kann unabhängig von Zeit und Raum geschehen. Dabei hat sich Englisch als wichtigste Sprache im Internet und anderen Web-Diensten etabliert. Dennoch bleibt die Beziehung zwischen Raum und Sprache weiterhin relativ unerforscht. Die Mehrheit der Studien, die in diesem Feld geführt wurden, konzentrierte sich auf die englische Sprache, während weniger häufige Sprachen im Word Wide Web vernachlässigt wurden.

Um diese Forschungslücke zu mindern, untersucht diese Masterarbeit Lemma Variation in der deutschen Sprache mittels des Micro-Blogging Dienstes Twitter. Beinahe eine halbe Million deutsche Tweets wurden für diesen Zweck gesammelt. Fehlerhafte Twitter Daten und die Dynamik der Sprachen haben sich als zwei Aspekte bestätigt, die nicht vollständig berücksichtigt werden können, wenn ein voll-automatisiertes Vorgehen gewählt wird. Basierend auf einer geographischen Klassierung und dem statistischen χ^2-Test konnten wir zeigen, dass eine internationale Variation von gewissen Lemmas auch auf Twitter ersichtlich ist und sich dieses Phänomen nicht nur auf Literatur und Schulbüchern beschränkt. Die gewonnenen Erkenntnisse dieser Masterarbeit sollen für weitere Studien in diesem Gebiet motivieren und nützliche Vorschläge bei der Anwendung geben.

Contents

List of Figures

List of Tables

List of Abbreviations

API	Application Programming Interface
BFS	Bundesamt für Statistik – (Swiss) Federal Statistical Office
CAT	Communication Accommodation Theory
CMC	Computer-Mediated Communication
DB	Database
DeReKo	Deutsches Referenzkorpus – German Reference Corpus
DoF	Degree of Freedom
ERM	Entity-Relationship-Model
EU	European Union
GIS	Geographic Information System
HTTP	Hypertext Transfer Protocol
IDE	Integrated Development Environment
IDS	Institut für Deutsche Sprache – Institute for German Language
IM	Instant Messaging
REST	Representational State Transfer
SMS	Short Messaging Service
SQL	Structured Query Language
UGC	User-Generated Content
URL	Uniform Resource Locator
WGS 84	World Geodetic System 1984

1. Introduction

The Internet has undergone a tremendous change in the last few years and is now in the era of the Web 2.0. This term, which was coined by in 2005, summarizes various aspects that have changed since the Web 1.0 and constitute nowadays an integral part of the Internet (O'Reilly, 2005). Probably the most important aspect was the shift from a rigid source of information to an interactive medium where everyone could participate (Walsh et al., 2008). Platforms like Facebook[1] and Twitter[2] have emerged where users can post pictures, videos and notifications that can reach from mundane chitchat to breaking news (Java et al., 2007; Poblete et al., 2011). Consequently, every day an enormous amount of so called user-generated content (UGC) is created. These new kinds of data sources provide new possibilities and have therefore called the attention of the scientific community. The interests of the researchers vary considerably and can reach from general analysis (Archambault & Grudin, 2012; Krishnamurthy & Arlitt, 2008), over gender studies (Bamman & Schnoebelen, 2014; Burger et al., 2011), demographic analysis (Baboolall et al., 2013; Mislove et al., 2011; Volkova et al., 2013) to sentiment analysis and opinion mining (Pak & Paroubek, 2010).

1.1 Geography and Language - Research Context and Review

Most of the established social media platforms have enabled their users to share their geographical position. Users of Flickr[3] or Twitter can use the incorporated GPS module of their devices to post their location or to tag their location textually (Graham et al., 2014; Ikawa et al., 2012). This feature allows users to enhance the content of their posts with geographical information. On the provider side this allows to offer location based services like advertisement or local information. Besides, this development has generated a great quantity of geographical referenced data that is accessible through Application Programming Interfaces (APIs) that can be used to gain new insights in various fields of research. Two of these areas are also of relevance for this work, geography and language. And despite them being discussed separately in

[1] www.facebook.com
[2] www.Twitter.com
[3] www.flickr.com

the following two sub-chapters they are highly interconnected and a boundary is hard to define, if not even impossible.

1.1.1 Geography in Social Media

There is a myriad of papers and research that deal with the analysis of the geographical data from UGC on social media. The approaches as well as the purposes can vary considerably, like the purpose which can range from disaster management (Graham et al., 2014; Ikawa et al., 2012) to improved location based services (Cheng et al, 2010).

For instance, Bouillot et al. (2012) argue that all the information inherent to a tweet can be of great importance for decision makers, despite the difficulty to analyse its contents due to their characteristics. They propose an automated process to extract the geographical information contained in tweets. Thereby, they also address the issue of homonyms, i.e. words with the same spelling but with a different meaning. For example, New York can stand for New York City, the state New York or for some places in Great Britain. They propose a geographical hierarchy, i.e. from a course regional division – e.g. state level – to a finer division – e.g. county level, for the end-user to improve the analysis and to reduce the impact on homonyms on the desired results (Bouillot et al., 2012).

Cheng et al. (2010) used a content-based approach to near the location of a Twitter user. The idea behind their approach was to detect words in tweets with a strong geographical affinity that could be used to identify the location of the user. Further, a neighbourhood smoothing model helped to improve the performance of their approach. A similar, but not identical approach was followed by Eisenstein and his colleagues (2010). They believe that the topic of the message and the geographic location is highly correlated in social media posts. By analysing the variations in topics and the lexical variations, they try to predict the position of the user only based on text (Eisenstein et al., 2010).

Rout and his colleagues (2013) on the other hand follow a complete different approach. Instead of using the content of a post to estimate the user location they look at the social ties. Identical to the first law of the Geography by Waldo Tobler: *"Everything is related to everything else, but near things are more related than distant things"* (1970: 236), they argue that a person tends to interact on a more regular basis with other users that are closer to himself. Based on this observation, they create the

hypothesis that the social ties on networks like Facebook and Twitter reflect this behaviour (Rout et al., 2013).

1.1.2 Language Use in Social Media

Language and geography are strongly connected and one can try to infer geographical information based on the used language and vice versa. However, as section 1.1.1 demonstrates one can also focus its attention on one of these aspects and ignoring the other, e.g. by focusing only on language.

A general overview about Internet linguistics is given by Crystal (2011). In his book called *Internet Linguistics – A Student Guide,* Crystal explains in a comprehensive way to what extent the Internet has changed the way of how we express ourselves. Since the ways of how we communicate have changed, e.g. through the introduction of the emoticons (Crystal, 2011: 23), we cannot use methods commonly implemented in the classic literature and simply use them for linguistic analysis on the Internet. A case study about Twitter highlights the problems faced with the language on this medium. New and adapted approaches are required if we want to gain new insight about language use on the Internet (Crystal, 2011).

Bergsma et al. (2013) on the other hand were mainly interested in less common languages like Nepali or Urdu. To see if automated language identifiers perform well on such languages they rely on the vast amount of data that is accessible on Twitter. With new highly trained language identification systems they try to classify those languages with non-Latin scripts. They also describe how the length of a tweet can influence the accuracy of such an identification system. The enormous amount of data that is available on Twitter is also one of the reasons why this thesis works with tweets. However, this and the other reasons will be discussed later in this study.

A more particular approach is followed by Filippova (2012). She used language to predict the gender of users on the video sharing platform YouTube[4]. Due to stylistic choices and lexical differences she tries to predict the gender of the person commenting a video. According to her results, this method achieved an accuracy of about 90% (Filippova, 2012). Bamman et al. (2014) conducted a research in the same subject area. To analyse how gender and lexical variation interact in social media they

[4] www.youtube.com

analysed a corpus based on tweets. According to them this mechanical approach can offer new insights of how gender is constructed (Bamman et al., 2014).

Another interesting topic that has motivated further studies within social media is the communication accommodation theory (CAT). This theory, which has its origins in the sociolinguistics, argues that during a conversation the participants unconsciously tend to converge in their behaviour. This means that one does not only accommodate in the formality of the speech, but also in the linguistic style as well as with non-verbal signals like smile and body movements (Giles & Baker, 2008; Giles et al., 1991). The first that addressed this theory in the environment of social media was the working group around Danescu-Niculescu-Mizil (2011). Based on Twitter conversations, they analysed if this theory also would be confirmed in a medium that has a length restriction, no real-time nature of conversation and that was originally not conceived with the purpose of conversations (Danescu-Niculescu-Mizil et al., 2011).

1.2 Motivation and Goals

The previous chapters showed a small number of the myriad of research that have already been conducted in the field of social media with particular interest in language and geography. Nevertheless, the majority of all these previous research have one aspect in common they all are focused on the English language.

According to Hong et al. (2011) around 50% of tweets are written in English. Making it the most prevalent language on this social media platform. A result that has been confirmed by other studies according to Scheffler (2014). Further, Hong et al. (2011) noted that after English comes Japanese language with a share of 19.1% and Portuguese with 9.6%. Hence, it is obvious why English is the predominant language on which studies on social media have focused. It is more abundant and therefore also easier to get. Due to the higher number of users posting in English, studies based on it have also a higher degree of representativeness and their results are more far-reaching (Scheffler, 2014). Nevertheless, it would be interesting to analyse other languages in social media. That is why this thesis focuses on German Twitter messages, a language that is only used in 1% of the tweets according to Hong et al. (2011). The German language on Twitter has scarcely been studied and this study could help to gain a broadened understanding on the geography and language of the German users of Twitter. However, like the English language with its American, British, Canadian or

Australian varieties, German also consists of three different varieties that are spoken in Germany, Austria and Switzerland.

There are multiple reasons for choosing Twitter as a research medium. As I will discuss later on, Twitter offers various access methods to its data, which allows to decide which data to collect and which parameters have to be fulfilled. Further, Twitter is an interesting communication channel containing posts from users from different origins and backgrounds which helps to raise the representativeness of possible findings (Scheffler, 2014). As mentioned by Hong et al. (2011) multiple languages are represented and can be analysed, without having to change the access. At this point we will not discuss the advantages and disadvantages of Twitter since they will be discussed more detailed later in this work.

1.3 Research

Only little is known about German tweets and the implications that are inherent concerning language and geography. Therefore, this study follows an explorative approach and its main object is to analyse German tweets and to determine what are the possibilities and the limitations of such analysis. Nevertheless, in this work we will focus on some particular aspects and try to answer them:

- Are German Tweets suitable for linguistic and geographical analysis?
- Is the use of the helvetisms, austriazisms and teutonisms equally evident on Twitter as it is on Newspapers and Literature?
- Germany, Austria and Switzerland have different rules for the use of the character <ß>. Is the difference also visible on Twitter?
- Is there an accommodation on the level of lemmas visible between users on Twitter?

1.4 Structure of this Work

To facilitate the understanding of the work presented here, this thesis is structured as following: Chapter two provides the necessary theoretical background in the areas of linguistics and social media. It will illustrate other works that have been conducted on the same or similar subject and depict eventual research gaps. In chapter three the study area and Twitter are introduced. Chapter four focuses on the methodological process.

The individual steps from the data collection to the data evaluation are explained and peculiarities are highlighted. Chapter five presents the obtained results. Problematic aspects are highlighted and issues of representativeness are illustrated. In the following chapter six the results are discussed around the research questions that were defined in chapter 1.3. Finally, chapter seven summarizes the outcome of this work, eventual open points are exposed and recommendations for future studies are given.

2. Theoretical Background

Since this thesis is an interdisciplinary work within the fields of Geography and Linguistics we need some basic knowledge in these areas to enable a more comprehensive understanding. This is of particular importance for the aspects of Linguistics. Therefore, this section will start with a more encyclopaedic approach to introduce some necessary terminology. Afterwards, a brief introduction to the German as a pluricentric language will be provided. This is necessary to understand why despite the geographical proximity of the areas where it is spoken this language has developed three varieties and what the implications of such a development are.

2.1 A Linguistic Overview – Introduction of Terminology

A national language can, according to Schmidlin (2011: 289), enable democratic processes since it enables citizens that can speak it to participate in public discussions and to be part of the decision making process. On the other hand, it can lead to the exclusion of minorities from the political participation as well as despise regional languages and dialects (Schmidlin, 2011: 289).

For analysing and discussing about language, a certain terminology and a definition thereof is needed, since its absence can lead to confusion and misunderstanding. Further, it will help to describe in what sense language varieties can differ. Therefore, to conduct the analysis of the German language and its varieties we will now introduce a terminology that is essential to discuss a pluricentric language. Further, it will also help readers that are less familiarized with the subject of linguistics to understand the thematic discussed here.

- **Grammatical**

  Despite the term grammar being familiar to everyone, a more fine-grained definition is needed. According to Bussmann (1996: 482) there are four areas that can belong to the grammar of a language. First, it stands for the knowledge and study of morphological and syntactic regularities of the natural language. Second, it stands for a system of structural rules that is inherent to every process of linguistic production and comprehension. Third, referring to language theory and

forth, as a description of formal regularities of a natural language (Bussmann, 1996: 482).

- **Lemma**

  According to Bussmann (1996: 667) a lemma is an entry in a lexicon or a dictionary. It can also be referred to as catchword.

- **Lexical**

  Lexical stands for everything related to the vocabulary and the words of a language. It is often the main aspect where language varieties show their differences (Schmidlin, 2011: 72). For instance, an example for the German language would be the term bicycle. While in Germany it is referred to as *Fahrrad*, or short *Rad*, in Switzerland it is called *Velo* and in Austria *Radl*.

- **Phonology**

  It is a sub-discipline of linguistics that deals with the sound of speech that has an impact on the semantics of a word or an utterance, and at the same time the system and relations between these sounds within a language (Bussmann, 1996: 898).

- **Pluricentric**

  A pluricentric language refers to a language that has multiple interacting centres. Each centre, i.e. a nation or a region where the language has gained an official status, has developed a national variety with some distinct codification that distinguishes it from the other varieties (Clyne, 1992). German has only created its own distinguishable codified norms in three nations, which have become centres, Germany, Switzerland and Austria (Ammon et al., 2004: XXXI). Besides German, other examples of pluricentric languages are, among others, English with its centres Great Britain, the United States, and Canada, and Portuguese with its centres Portugal and Brazil (Schmidlin, 2011: 71).

- **Pragmatics**

  The focus of pragmatics, which is a sub-discipline of linguistics, lies on the relationship between natural language expressions and the use of them in specific

situations (Bussmann, 1996: 926). According to Schmidlin (2011: 72) it is probably the most difficult aspect to describe since it only manifests itself in use frequency, routine communicative formulas or use difference in age or gender.

- **Semantics**

  Semantics is concerned with the 'literal' meaning of an expression. Its main focus lies on the description and analysis of the meaning of words (Bussmann, 1996).

- **Syntax**

  Syntax can refer to two different sub disciplines. One lies in the area of semiotics and the other in the area of grammar. The first subcategory deals with for the order and the relationship between the sign and the corresponding reality. Latter subcategory, which is usually meant by linguists, is a system of rules which focus on the implications that basic elements like words or morphemes, can have on sentences of language (Bussmann, 1996: 1169).

- **Variety**

  A linguistic variety is a coherent form of a language which is distinguishable through distinct forms and expressions. Depending on its geographical dispersion it can be considered either as a dialect or a national variety (Bussmann, 1996). An example of a language with multiple varieties is the English language. American, British or Australian English are all varieties of the Standard English. On a national scale, each of these nations has also multiple dialects or regional varieties. A national variety of a standard language can have various functions, an important one is creating identification for the people using it (Clyne, 1995).

This listing above is elementary and is only meant to be a short introduction to a subject within the linguistics that is of interest for this work. As we will see later, all these aspects can vary among and between the varieties, but they do not have to. However, with data that come from the social media, it is not possible to analyse all these elements. For instance, phonological aspects cannot be analysed since most of the actual social media like Twitter or Facebook are based on written language. Therefore, we need to focus on those aspects that are possible to be addressed on these mediums.

2.2 A Short Guide to the German Language

The German language, which belongs to the Western Germanic languages alongside English, Dutch and Frisian, among others, is one of the most important languages in Western and Central Europe (Durrell, 2006). With almost 100 million speakers it is the 10th most frequent language in the world and the most frequent in Europe when considering only native-speakers (Durrell, 2006). German has the status of a nationwide official language in Germany, Switzerland, Austria, Liechtenstein and Luxembourg. These nations also account for the majority of the number of speakers. This makes German the most frequent official language in Europe. Furthermore, German is also recognized as an official language on a regional level in other nations, such as East-Belgium, South Tyrol in Italy, Poland, Slovakia, Czech Republic and France (Ammon et al., 2004; Durrell, 2006). According to Durrell (2006) there are also German-speaking communities outside Europe. The largest located in the United States with over one million members. Further, a considerable number of German speakers live in Argentina, Australia, Brazil, Canada, South Africa and Namibia (Durrell, 2006).

2.2.1 Definition of a Pluricentric Language

However, the aspects mentioned above alone would not be sufficient to make German an interesting subject of study in this work. One of the most interesting characteristics is its pluricentricity. The definition given by Clyne (1992) has already been mentioned in the previous chapter. Another helpful and concise definition of a pluricentric language, was given by Schmidlin (2011: 71):

> *„Von plurizentrischen Standardsprachen spricht man dann, wenn sie in mehr als einem Land als nationale oder regionale offizielle Amtssprache verwendet werden und über eigene, kodifizierte Normen verfügen.“*

This means that for a standard language to be considered as pluricentric it has to be an official language either on a national or on a regional level in more than one country. Further, it has to have its own codified norms that characterise it and make it

distinguishable (Schmidlin, 2011: 71). The use of a variety is further solidified through the media – e.g. newspaper – and schoolbooks (Schmidlin, 2011: 71). In the next chapter we will have a closer look on German and we will illustrate the peculiarities of its national varieties and highlight the differences between them.

2.2.2 German – a Pluricentric Language

As mentioned in the short introduction to chapter 2.2, there are several regions or nations where German has a status of an official language. However, not all of these regions or nations have developed an own variety of German and beacme a centre to the German language. These three nations are Germany, Switzerland and Austria. Besides these so called full centres there are also half-centres like Luxembourg or East-Belgium (Ammon et al., 2004: XXXI). However, the impact that these half-centres have on the standard language is less significant as the one originated by the full centres. That is why in this study, only the varieties from Germany, Switzerland and Austria will be considered.

The different varieties of a standard language can differ on many levels. According to Schmidlin (2011: 72) these differences can occur either on a phonological, grammatical, lexical, semantical or pragmatical level. An example of a phonological difference which is given by Schmidlin (2011: 72) is the pronunciation of monosyllable words that end with a /b/, /d/ or /g/. The terminal devoicing on those words in Southern-Germany, Switzerland and Austria is missing in comparison to the pronunciation in Northern-Germany. An example of a grammatical difference is the word *Salami*, which in standard German is a feminine word and used with the article *die*, while in Switzerland it is a male noun and therefore used with the article *der* (Schmidlin, 2011: 72). An example of a lexical difference is the German term for bicycle: In Germany it is referred to as *Fahrrad* or short *Rad*, while in Switzerland it is called *Velo* and in Austria *Radl*. These lexical differences between these three nations can be very distinct. To clarify the origin of the lexical variants of each variety we need to introduce some further terminology. A lemma variant which is used in Germany is called teutonism, those from Switzerland are called helvetism and those from Austria are called austriazism. There are also lemmas that have the same form, but a different meaning in the different varieties of German. One good example of such a semantic difference between the varieties is the term *Estrich*, which in Austria

and Germany denominates certain type of floor in the house while in Switzerland it stands for an attic (Schmidlin, 2011: 72). Pragmatical differences are probably the hardest to detect in a text. They often occur in typical phrases, various kinds of greetings, as well as in the occurrence of certain words (Schmidlin, 2011: 72).

From all these possible aspects where varieties of a standard language can differ, the most common occur on the lexical and the phonological level which is also the reason why this areas of possible difference are the most studied (Schmidlin, 2011: 107). Nevertheless, despite all the possible differences that have been mentioned in the previous paragraphs the similarities between the varieties of a standard language outweigh by far the differences that might exist (Schmidlin, 2011: 73). For this reason, and contrary to Clyne (1992) who argued that a pluricentric language can simultaneously divide and unite people, Schmidlin (2011: 73) argues that pluricentric languages have rather a unifying than a dividing force for the countries where they are spoken.

2.3 Language on the Internet

2.3.1 Concerns about Language on the Internet

Like other technologies before, e.g. the arriving of the printing in the 15th century or the telegraph in the 19th century, also the emergence of the Internet originated various fears among people. Besides the anxieties concerning issues like privacy, propaganda or crime, there were also concerns about language. These fears reached from the omnipresence of the English language which would supersede other languages, to the fear of language misuse by ignoring formalities and rules (Crystal, 2001: 1). However, similar as it was with the short messaging service (S*MS)* around the year 2000, most of these doomsday prophecies have not become true (Crystal, 2011: 4). Crystal (2011: 4ff.) notes that the availability of a new communication form, such as SMS, rather improved the literacy amongst children, because the use of abbreviations is only possible with a high awareness of the written and the spoken language.

The same can be said about Twitter. Borau et al. (2009) and Grosseck & Holotescu (2008) have both analysed appropriateness of Twitter for educational activities. While Borau et al. (2009) conclude that Twitter is a valuable tool for profession development and to interact with students, Grosseck and Holotescu (2008)

accredit Twitter with capabilities to train cultural and communicative competences without being bound temporally or geographically.

Therefore, according to Crystal (2011; 6), the Internet does not deteriorate language awareness but can rather help to improve it. For instance, by incorporating it in the syllabus students can learn more about different linguistic styles and their appropriateness according to the situation. Further, the Internet has remarkably increased the possibilities to express ourselves. Abbreviations, emoticons or letter omissions have changed the way we write something down. Email, chat and social media, such as Twitter, have on the other hand further diversified the linguistic styles that we use (Crystal, 2011: 7).

2.3.2 Internet Linguistics

The previous chapter gave a short overview of the fears that accompanied the expansion of the Internet in the beginning of the 21st century. Meanwhile, the Internet has become an integral part of our everyday life and is often present when we do our shopping, organize our daily life, and especially when we communicate. Therefore, it is not surprising that many linguists and scientists have addressed the subject of communication on the Internet. This lead to a new field of study called computer-mediated communication (CMC) which became largely known in the 90s (Crystal, 2011: 1). Nevertheless, according to Crystal (2011: 1) the term CMC is a bit vague when the main focus lies on the linguistic aspect of this kind of communications. The web 2.0 enabled us to communicate in many ways. We can send photographs and videos, listen to and play music online, or have a conversation on platforms like Skype[5]. Therefore, Crystal (2011: 1) introduced the term Internet linguistics to emphasize the focus on language and to distinguish it from the broader term CMC. Like CMC it focuses also on chatrooms, blogging services or emails, however, the main interest lies in the language used by users. Since this term best describes the research field in which this thesis is to be situated, we will employ this term from now on in. It also helps to specify other research that has been conducted on the same subject.

[5] www.skype.com

2.3.3 Internet Linguistics as a Field of Research

Until now we have determined the terminology which is essential for the discussion of a pluricentric language and we discussed the fears that accompanied the dissemination of the Internet. To facilitate the understanding of Internet linguistics and to explain why this work can contribute to this relatively young field of science, we will now have a more detailed look into the characteristics of Internet linguistics and what distinguishes it from the regular linguistics. Further, we will have also a look on previous studies[6] conducted in this area. All this theoretical aspects will help us to position Twitter in the context of linguistics as well as this thesis.

Internet linguistics can be seen as a sub-discipline of the study of CMC, as mentioned before in this thesis. While latter comprises any kind of communication form over the Internet, i.e. music, video or text, the first has its focus on the language of this form of communication. However, language use on the Internet cannot be analysed in exactly the same way as language used in a book or in a newspaper. There are multiple channels that can be used to communicate with each other like chatrooms, instant messaging (IM), social networks, email and microblogs. In all these channels language is used differently. Dürscheid (2003) analysed language in the continuum of spoken and written language on different channels. She analysed communication on fax, chatrooms, IM, email and SMS. However, since her work dates from 2003 social networks and microblogging services like Twitter are missing in the discussion. Nevertheless, also the new channels can be integrated in her model.

Dürscheid (2003) argues that the situation and the communication medium that we chose influences the manner of how we communicate, an aspect often emphasised in media research. First of all, there is the choice if we want to transmit our message orally or graphically. Second, the synchronicity of the communications is also an important issue that needs to be considered. For instance, a telephone call is a synchronous form of communication, since one can react even while the other participant is talking. A SMS or a tweet are two forms of asynchronous communication since any reaction is only possible after the message has been sent or posted. There are also communication forms in between referred to as quasi-synchronous like IM, where both participants can write simultaneously but can only see what the other person has

[6] For a comprehensive overview of internet linguistics I invite you to read "Internet Linguistics: A Student Guide" by Crystal (2011), which has been frequently cited in the previous paragraphs. It is simply written and gives you a good first impression of what internet linguistics stands for.

written after the message has been sent (Dürscheid, 2003). Further, there is also an issue relating to the conceptual form of the communication. An email, for example, can represent a very formal way of communication if the addressee is an agency or a bureau and, therefore, it will follow all grammatical rules and formalities. On the other hand it can also be very informal as it is often the case with e.g. birthday greetings. In latter situations, the language used is closer to the spoken language (Dürscheid, 2003; Koch & Oesterreicher, 1985).

All these aspects mentioned above can be described in a model – see table 2.1 – that is based on the work by Koch and Oesterreicher (1985) and was further developed by Dürscheid (2003). On the medium side, a communication can be either phonetic or graphic, when considering only language as it is the case in this work. The conceptual level has to be rather seen as a continuum than as distinct classes. Depending on the situation, the formality of the language used is closer to the spoken or to the written language (Koch & Oesterreicher, 1985). Inside this classification we can further differentiate between synchronous and asynchronous communications, as is argued by Dürscheid (2003).

	conceptual oral ⟷ conceptual written
medial oral	Synchronous asynchronous
medial written	quasi-synchronous asynchronous

Table 2.1: Extension of the conceptual model of Koch & Oesterreicher by Dürscheid (2003)

All aspects discussed above, which can be summarised in the model in table 2.1, help to situate the tweets in the context of linguistics. As one might already expect, the general tweet, like the SMS, is written, i.e. medially written, conceptually oral and asynchronous. These expectations are confirmed as a short glimpse of the collected data revealed. Such messages tend to be less formal and to follow grammatical rules less strictly. Nevertheless, we have to keep in mind that we are speaking of a continuum on the conceptual level, and therefore it is quite difficult to classify all tweets in one class.

It is essential to know where to locate tweets in the linguistic context. As explained before, due the nature of the language used on Twitter different approaches might be needed to analyse langue or to implement a part of a language system to analyse the results.

2.4 Research Gap

As we have seen in the preceding chapters Twitter has become an interesting medium to conduct research in various disciplines. Twitter provides easy access, as we will see later, and the real-time nature of the data are just a few reasons why it is interesting to work with Twitter (Han & Baldwin, 2011). Further, we have also seen that a wide range of issues have already been studied based on tweets, ranging from sentiment to geographical analysis. Some studies have also addressed language in social media, as we have mentioned earlier in this work. However, as we have seen, most of the studies concerning language have focused on English. Danet and Herring (2007: 3) note that in 2000 over 90% of the links on secure servers were in English. Further, a survey conducted in 2002 revealed that 56% of all the existing webpages were also written in English. A similar distribution was found also on Twitter. In their study Hong et al. (2011) concluded that more than half of their collected tweets were English. These two aspects confirm the position of the English language as the *lingua franca* on the Internet and on Twitter. Therefore, only few studies have been conducted for the German language which were based on Twitter, e.g. Scheffler (2014) who built a language corpus based on German tweets. There have also been studies focused on code-switching and accommodation for the German language. Androutsopoulos (2006, 2013) has focused on code-switching and accommodation in CMC. One of his main interests lies on the multilingual Internet and how bi- or multilingual communities communicate in German-based forums (Androutsopoulos, 2006). However, we noticed that almost all the research that has been conducted for English or German has focused on a language as an entity. To our knowledge almost none research has been conducted that focused primarily on the level of lemmas, i.e. simple words.

The idea behind this thesis is to make a contribution to this less explored subject. By focusing on lemmas, the basic element of a language, this thesis aims at contributing to a better understanding of regional language variation and

accommodation between regional varieties on the lexical level. Further, it also promotes diversity of linguistic research on Twitter by focusing on German instead of English, despite the availability of English tweets being remarkably higher. By combining all these aspects of geography, the German language and social media this interdisciplinary study tries to contribute to an interesting subject based on a language that mighty have been neglected in the past.

3. Study Area and Data

3.1 The Study Area

The study area is well defined trough the nature of this thesis. Since we are looking at the pluricentric German language the extent of the study area is limited by the dimensions of this three national centres, i.e. Switzerland, Germany and Austria, as displayed in figure 3.1. This means that tweets that have been sent from outside these three nations will not be taken into consideration for the further analysis, even if the user itself is a resident of one of the three centres. Furthermore, Tweets from the other German speaking regions, e.g. Luxembourg or Liechtenstein, will also be ignored, since they can only be considered as half-centres as explained in chapter 2.3.3. Nevertheless, for the purposes of the analysis it makes sense to focus the attention on these three full-centres.

Figure 3.1: Study area with the highlighted centres of the German language: Germany, Switzerland and Austria.

The decision to ignore tweets that have been sent outside the study area has also some implications for the analysis of the results. Germans, Austrians and Swiss can also post

messages from outside their country of residence. As well as a tourist can send tweets within the boundaries of the study area. However, this aspect will be seized again later on.

3.1.2 Germany

Germany is in terms of population and area the largest of the three centres. It has about 80.7 million inhabitants of which 8.7 % have a different nationality than German, which makes it the smallest percentage of the three countries. Further, it is also the most densely populated nation of the three considered in here.

Population (in million)	80.767
Density (Inhabitants/km^2)	226
Area (in km^2)	357‘340
Foreign population (in %)	8.7
Official Language(s)	1

Table 3.1: Key figures of Germany for 2013. Source: DESTATIS (2015)

According to the *Variantenwörterbuch des Deutschen* by Ammon et al. (2004) while the majority has a comprehensive knowledge of German only a small proportion of the population has rudimentary or none knowledge about the language. There are also some minority languages in certain regions, like *Sorbian* in Lausitz – in the federal state of Saxony – that are accepted by the state. Nonetheless, German is the only official language, even if it is not expressly mentioned in the constitution (Ammon et al., 2004).

3.1.3 Switzerland

Switzerland is the smallest of the three full-centres. According to the Bundesamt für Statistik (BFS) (2013a), The Swiss Federal Statistical Office, its population had grown to 8.1 million by 2013. With 23.8% of foreigners by 2013 it is also the centre with the highest percentage of immigrants (Swiss Federal Statistical Office, 2013a).

A unique characteristic of Switzerland compared to Germany and Austria is that it has nationwide four official languages and not only one. The four languages are German, French, Italian and Romansch. As displayed in table 3.2 German is the most

widespread language with 64.9%, followed by French with 22.6% and Italian with 8.3%. Interestingly, due to the high percentage of immigrants, some languages like English or Portuguese are more frequent then the forth official language, the Romansch (Ammon et al., 2004; Swiss Federal Statistical Office, 2012). It is spoken in only a few municipalities and contrary to the other three official languages the municipalities do not form an interconnected geographical region (figure 3.2).

Population (in million)	8.139
Density (Inhabitants/km2)	203.5
Area (in km2)	41’285
Foreign population (in %)	23
Official Language(s) German French Italian Romansch	4 64.9% 22.6% 8.3% 0.5%

Table 3.2: Key figures of Switzerland for 2013.
Source: (Swiss Federal Statistical Office, 2013a)

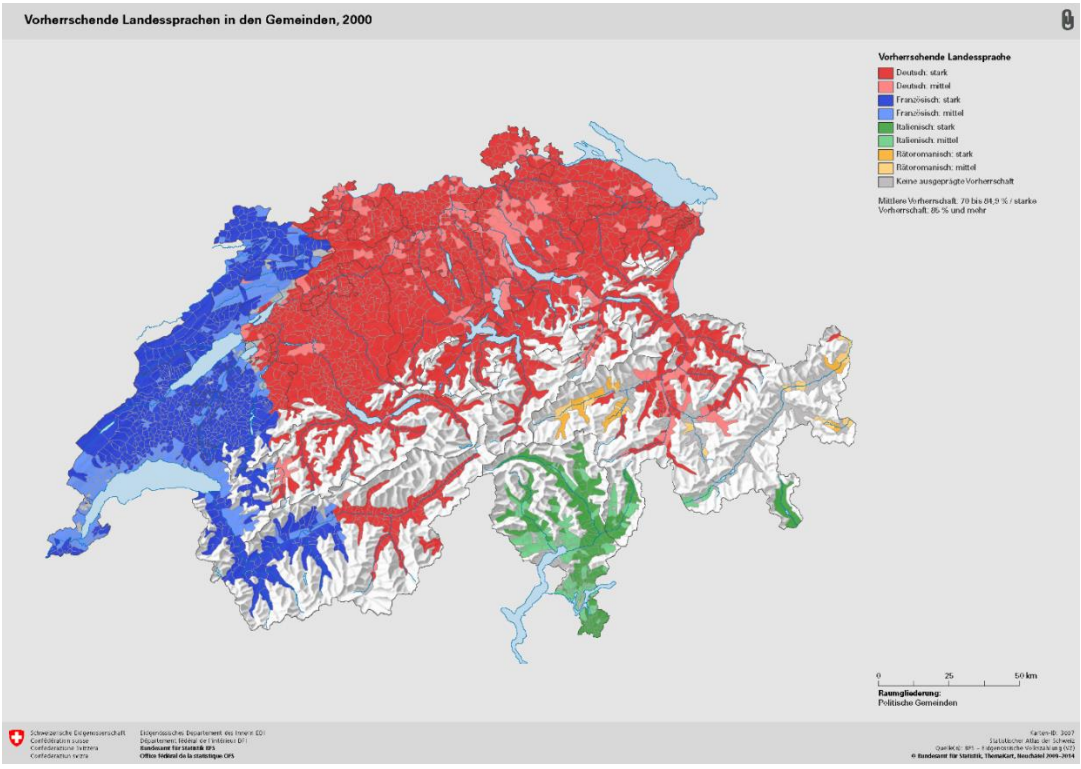

Figure 3.2: Distribution of the four official languages in Switzerland in 2000 on the level of the municipalities. Red for German, blue for French, green for Italian, and yellow for Romansh. Source: (Swiss Federal Statistical Office, 2013b).

3.1.4 Austria

Austria is the last of the three centres discussed in this thesis. It has a population in the same range as Switzerland but dispersed over an area twice as big. The immigrants in Austria account for 12.5% of the population, which positions Austria between Switzerland and Germany.

Population (in million)	8.477
Density (Inhabitants/km^2)	101
Area (in km^2)	83’878
Foreign population (in %)	12.5
Official Language(s)	1

Table 3.3: Key figures of Austria for 2013. Source: (Statistics Austria, 2014).

German is the official language for Austria and contrary to Germany it is specified in the Austrian constitution. Art. 8, § 1 says that despite the rights that the minority languages have, German is the only official language for the entire republic (Bundeskanzleramt, 2015).

3.2 About Twitter

In the early 2006 Twitter Inc., a company based in San Francisco, California, founded the microblogging platform called Twitter. On March 21st, Jack Dorsey, a co-founder of Twitter sent the very first message with the famous 140 characters restriction – the first “Tweet” (Twitter, 2014). Since this day, Twitter has experienced fast growth. According to the latest numbers published by the company itself in 2014 there are about 284 million monthly active users, 500 million tweets are sent every day, and more than 35 languages are supported by now (Twitter, 2014). Weerkamp et al. (2011) note that Twitter has become one of the most important platforms for real-time information sharing and it is widely used for event detection, mining opinions and media analysis. According to the Webpage ALEXA (2014) Twitter is the 9th most visited webpage and the second, after Facebook, if we only take social networks into consideration.

Eleta & Goldbeck (2014) argue that Twitter exhibits simultaneously characteristics that are typical for information sharing platforms and social networks. However, contrary to other social networks like Facebook, relationships in Twitter do not need to be based on reciprocity (Eleta & Golbeck, 2014). The relationships on Twitter consist of users following or being followed by other users. As mentioned before this kind of relationships, as it is general for microblogging services, can be asymmetrically. For instance, on Facebook both participants have to agree on their relationship before being able to see the posts of the other (Zappavigna, 2012: 27). On Twitter there is no need for this kind of reciprocity, since a user can follow someone without being followed by the same, which is often the case with politicians, celebrities or corporations (Kwak, et al., 2010).

According to Eleta & Goldbeck (2014) there are three kind of Twitter posts, or tweets as they are known. First, it can be a public tweet with a content ranging from trivial information to political statements, and containing images and links to other sites on the Internet (Poblete et al., 2011). Second, it can be a direct reply to another user, characterized by a "@" proceeding the username, e.g. *@BarackObama*. And finally, it can be a reposting called *Retweet* which helps to disseminate the information even further (Kwak et al., 2010). Another characteristic of Twitter is the symbol "#", referred to as hashtag, followed by a word, indicating that the tweets belongs to a particular subject, e.g. *#WorldCup* (Poblete et al., 2011). In 2014 the Swiss radio station SRF 3 even selected hashtag as word of the year (Schweizer Radio und Fernsehen, 2014).

3.2.1 Demographics of Twitter

Data concerning the demographics is of particular importance since it helps us to understand who or what we are dealing with. Nevertheless, Twitter does not publish any demographics about its users, neither global, nor on the level of individual nations. Meaning that we have an enormous data source without knowing much about the population behind Twitter (Mislove et al., 2011). To overcome this gap there are two established approaches an academic and a non-academic approach.

There have been many academic efforts trying to extract demographic information from the Tweets – or other Internet sources in general. Mislove et al. (2011) have dedicated a study to this subject trying to extract information about

gender, location and ethnicity from Twitter users in the United States. For this purpose they examined locations, first and last names, and compared them to information from the national census or from Google Maps[7] (Mislove et al., 2011). Filippova (2012) analysed the social environment and language of YouTube[8] users to predict gender. A different approach in this field was conducted by O'Connor and his colleagues (2010) who analysed how the language on social networks is influenced by demographic factors. However, since all this approaches depend on which tweets you are analysing, the comparability of the numbers require some caution.

Opposing the academic approaches are the non-academic ones. Such analyses are often conducted by companies dedicated to marketing and online analysis such as PewResearchCenter[9] or comScore[10]. In a report published by the PewResearchCenter Duggan et al. (2015) conclude that in 2014, 23% of all adult Americans on the Internet used Twitter. The same report notes that it is more popular amongst men than women, and that the majority of the users are between 18-29 years old and come from an urban or suburban environment (Duggan et al., 2015).

There are also numbers concerning the study area of this thesis. For example, according to Steiner (2012) in June 2012 over 58% of the users in Switzerland were older than 35 years, and 76% were men. But as mentioned by the author the numbers are estimated by the Google AdPlanner (now called Google Display Planner) Platform. This can lead to some less reasonable results, like the group of users between 0-17 years (Steiner, 2012).

There are three obvious problems when dealing with demographics on Twitter. First, the academic approaches would allow a reasonable comparison but demand more resources to execute, and are often restricted to a specific study area. Second, the data from analytic companies need to be handled with great caution. Often, the methods and calculus behind the numbers are not clear which makes it almost impossible to compare the statistics from different platforms. And third, yet there has not been much research in the demographics of Twitter users in our study area, i.e. Switzerland, Germany and Austria. Therefore, it is not possible to make a satisfying comparison of the users of these nations.

[7] maps.google.com
[8] www.youtube.com
[9] www.pewresearch.org
[10] www.comscore.com

3.2.2 Twitter API

Twitter offers its developers many ways to incorporate its service on their applications. And one of the most effective ways to access tweets with a low delay is through an API. Twitter offers various kinds of APIs according to the specific needs. The following three are probably the most used. First, there is the Streaming API that enables you to determine the parameters of the data that shall be returned. The second method is a special variation of the Streaming API called Firehose. The Firehose allows you to access all the public data that is available, however, it is subjected to special requirements (Twitter Developers, 2015c). Some of the presented research in this paper have been based on this access method, e.g. the work by Hu et al. (2013). The last method is the *Representational State Transfer* (REST) API. This method enables a more complex implementation of Twitter into any application, allowing you also to write tweets or to access the user profiles, and to change the requests without having to shut down the connection (Twitter Developers, 2015a). However, since only the streaming API was used in this thesis the other two will not be discussed further.

The concept behind the conventional streaming API is easy to understand. You create a Hypertext Transfer Protocol (HTTP) request that can hold, theoretically, for an indefinitely long period of time if there are no external interferences. This connection can then be used simultaneously for sending requests and parsing results (Twitter Developers, 2015c). On the other hand, this is one of the drawbacks of this access since it requires a permanent connection to the Internet.

Further, the Streaming API offers you the possibility to implement request parameters if needed. Some of them were also used in this work. The settings used in this work will be discussed later in the methods section. At this point we will only have a look at the possibilities provided by Twitter. For instance, you can define the language of the tweets you want to collect. By doing so only the posts that were identified as being written in the specified language will be accessed by the API. Further, you can set the request that the posts need a geolocation if you are interested in geographical analysis. Also useful is the possibility to define a tracking list. It is a comma-separated list containing all words or phrases that determines which tweets should be streamed (Twitter Developers, 2015c).

There are more parameters provided by Twitter that can be defined according the personal needs. We invite you to visit the Twitter developers site on the Internet for a comprehensive overview.

3.2.3 Geocoding

There are three ways of adding geographical information to a tweet. The most common is through the settings of your Twitter profile where you can specify your location. Though, in many cases these locations might differ from the actual position from where the user is tweeting. The second problem with this is that there are no restrictions on what you can add as location. This can lead to locations like *Mordor*, a place from the novel *Lord of the Rings* (Graham et al., 2013). The second method is by mentioning the location in the tweet itself as displayed in figure 3.4. Nevertheless, we face here the same problems as with the location in the user profile since every user can write anything as a location. Further, there are some issues concerning geographical ambiguity and geographical perception by the user (Graham et al., 2014). Finally, users can allow Twitter to access the GPS-tracker of the mobile device and to share its coordinates. This exhibits the most accurate geographical information of the three methods mentioned in this chapter.

Figure 3.3: Location added in the tweet to announce where to find the user.

3.2.4 Automated and Semi-Automated Programs – Bots and Cyborgs

Due to its popularity and openness, Twitter has, like other social media platforms, become a target for automated programs, also known as bots (Chu et al., 2012). According to Zhang and Paxson (2011) 16% of the active exhibit automated characteristics. While early bots were mainly used to propagate automatically generated content, contemporary bots have become much more complex. These so called social bots are capable of sending requests, looking for information on the Internet and also of interacting with other users by sending requests or by participating in discussion. These properties have made the distinction between human-like and bot-like behaviour much more difficult (Ferrara et al., 2014). Furthermore, in between the categories human and bot two new categories have emerged according to Chu et al. (2012). These bot-assisted humans or human-assisted bots are often referred to as cyborgs and are now very common on Twitter. Cyborgs exhibit characteristics common to humans and to automated programs and, therefore, highly complex algorithms are needed to deal with them (Chu et al., 2012).

The impact that these programs have on social media, including Twitter, are twofold according to Ferrara et al. (2014) as well as Chu et al. (2012). On one side, most of these bots are inoffensive and can help the user to stay updated as it is the case with a news feed. On the other side, there are also bots with the purpose to distribute malware or spam. They can also artificially influence political decisions or manipulate stock exchanges (Ferrara et al., 2014). For this reason many efforts have been made in the area of automated bot and spam recognition. For instance, Yardi et al. (2009) and Grier et al. (2010) have both addressed the issue of spam on Twitter. There has also been research conducted with other social media like Facebook, Flickr[11] or Internet chatrooms that can be related to Twitter. Gianvecchio et al. (2008) addressed the issue of bot recognition in Internet chatrooms.

However, all these works implemented complex algorithms to detect and analyse tweets posted by this automated and semi-automated programs. Since such programs can also influence the occurrence of a certain lemma in our DB we also need to address this subject. Nevertheless, since such complex algorithms can be a work on their own, we will only implement a rudimentary approach that will help us to eliminate the tweets created by simple bots.

[11] www.flickr.com

3.2.5 Twitter as an Object of Studies for Linguistic

Twitter is a very interesting medium that has originated many studies including this thesis. Ranging from geographical analysis to gender detection, various research fields are represented. However, this work, like all the others, is also affected by some issues that are inherent to Twitter as a research medium. Some aspects, like automated and semi-automated programs have already been discussed in a previous chapter. Nevertheless, there are further aspects that need some attention, especially when analysing language, as it happens in this study.

- Both, Zappavigna (2012) and Crystal (2011), note that Twitter is a temporal strongly bound medium. Catch-Phrases and current events can influence the properties of the collected data (Crystal, 2011; Zappavigna, 2012: 19). Even the time of day can have an influence on the occurrence of certain topics or words appearing on Twitter (Zappavigna, 2012: 19).

- Even if the language recognition of Twitter can help us to extract only the tweets that we want, it is not absolutely error proofed. The same holds true for collecting data with a keyword list. German tweets can contain English terms and vice versa (Crystal, 2011). This means that even if the code was set to only download German tweets it is possible that also posts in other languages can be collected and stored in the DB.

- Another point that was mentioned by Crystal (2011) and Zappavigna (2012: 19) concerns the retweet. The possibility to repost something already said is one of the main characteristics of Twitter, which obviously can influence the number of occurrences of a particular lemma in the DB. However, in the spoken language this is a very seldom phenomenon. Furthermore, someone from Switzerland can retweet a post from Germany, which will also affect the analysis of the collected data.

- An issue that might need some consideration as well are the posts that are longer than the 140 allowed characters. While Crystal (2011) asks if such tweets are linguistically relevant, Zappavigna (2012: 21) argues that such tweets may be

removed if one can define a procedure to do so. This would be of particular importance if a part of speech analysis should be incorporated, since the missing part can be essential for the analysis.

- There are also issues concerning the orthography used in Twitter. Laboreiro et al. (2010), which focused on tokenizing micro-blogging messages, highlighted some particular aspects. First, due to length restrictions, characters and white spaces are skipped, and new kinds of abbreviations are created, e.g. “gr8” for great. Second, because of the conversational nature of Twitter, oral markers like emoticons and missing or unusual punctuation are common. And finally, as we have discussed previously, the majority of the tweets are sent from mobile devices like smartphones or tablets. Writing on such devices is prone to spelling errors and users tend to not correct their misspellings (Laboreiro et al., 2010).

4. Methods

In this section we will have a closer look at the methods implemented in this thesis. The implementation of the database (DB) and the data acquisition took place on a computer running on Linux Ubuntu[12] inside the network of the University of Zurich. During the entire time there were no technological issues or errors that caused an interruption in the streaming. After the data collection the entire DB and software was transferred to a second computer for the analysis. For an overview of all the software used in this work please consult the Appendix B.

4.1 Coding and Database Implementation

4.1.1 Coding

Most of the code used in this work is based on existing scripts from the thesis of Oliver Zihler (2013) and is written in the object oriented programing language of Java. The adapted code with the streaming API was run on the Eclipse IDE for Java Developers[13], an open source program, available on the Internet. Eclipse was also used to generate the required output files since it allows an automated and more efficient processing of the data. The output files, always written in the plain text format .txt, could then be used in Excel or ArcGIS 10.2. Since the majority of the coding follows the state of the art, only the segments that facilitate the understanding are added at the specific point or in the appendix. All the other codes can be found on the appended CD.

The first step was to create an account on Twitter and register as a developer. With the registration you receive the necessary tokens which enabled us to establish a connection to Twitter through the streaming API. However, before starting collecting data one can set some parameters which help to reduce the amount of undesired data in the dataset. The methods that can be used were already mentioned in section 3.2.2. Now, we will have a closer look at the language and geolocation settings. Since the list with the keywords requires more attention it will be discussed separately and more detailed in chapter 4.2.

[12] www.ubuntu.com/desktop
[13] www.eclipse.org

```
public static void main(String[] args) throws TwitterException {

// Parameter configurations
String file = "…/resources/inputfile/keywords.txt";
String host = "localhost";
String port = "5432";
String database = "postgres";
String user = "postgres";
String password = "postgres";
String[] keywordsToFilterFor = getKeywords(file);
for (String s : keywordsToFilterFor) {
      System.out.println(s);
}
String[] languages = { "de" };

boolean mustHaveGeolocation = true;
boolean printIncomingTweets = true;

// System initialisation
AbstractDBConnector postgresConnector = new PGDBConnector(host, port,
                    database, user, password);
```

Figure 4.1: The settings concerning the language and the geolocation embedded in the code.

Figure 4.1 above illustrates how the parameters for the language and the geolocation were implemented in this work. First, the language was set to `"de"` for German. Since we are only interested in the German language it makes sense to look only for German Tweets and to remove other languages since they would increase the number of undesired data in the DB and increase the possibility to hit the rate limit. An approach also followed by Scheffler (2014). The other aspect concerns the geolocation. Given that we want also to conduct a geographical analysis and investigate the distribution of the varieties of the language the availability of geolocation is indispensable. This is taken into account with the expression "`boolean mustHaveGeolocation = true`".

These two parameters already help to precise which data should be collected and to reduce the amount of undesired data that in a further step would need to be removed.

4.1.2 Database

The DB is based on the open source program PostgreSQL[14]. One of the many advantages of this DB system is that it incorporates native programming interfaces for Java, which facilitates the interaction of the DB with the existing code that is written in Java too (PostgreSQL, 2014).

[14] www.postgresql.org

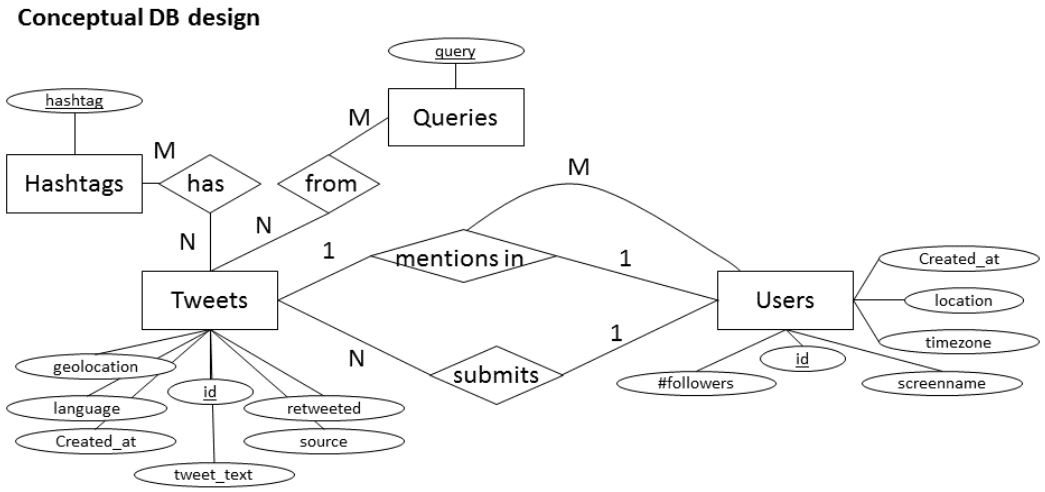

Figure 4.2: Conceptual design of the DB.

Figure 4.2 shows de conceptual design of the used DB, often referred to as the Entity-Relationship Model (ERM), as developed by Chen (1976). As the figure depicts the DB consists of 7 tables. The two tables *tweets* and *users* are the most important since they contain all the information that is analysed in this work. Nevertheless, the other tables have also vital function since they contain all the necessary additional information. Each tweet and user has an ID which fulfils the function of a Primary Key (PK). These PKs are used to reference additional information in other tables to the respective tweet or user. The reference occurs through a *Foreign Key* (FK). To facilitate the understanding of this issue here is an example: The ID in the *users* table has the function of a PK. Every user has an ID which is unique and that helps to identify him. On the other hand, we might want to know which user tweeted to which user. This relationship is stored in the table *submits*. However, in this table the user ID has only the function of a FK that helps to reference it to the user in the table *users*. The PK in the table *submits* consists of three FKs that only in combination form the PK.

The other tables, or entities, are mainly used to store additional information like mentioned hashtags or who tweeted to whom. Further, there is another table called

spatial_ref_sys that was implemented in the DB with the installation of the Add-On PostGIS. This Add-On and the respective table contain all the necessary information to extend the created DB in PostgreSQL enabling it to support geographical queries (PostGIS, 2015).

It is clear that the discussion of the DB might fall a bit short in this work. Since, the entire process of how to design and implement a DB could be a work at its own, we decided to give only a short overview in here. For a more comprehensive discussion of the ERM we invite you to consult the work of Chen (1976). Additionally, in every library and book store further literature about DB implementation can be found.

4.2 Keyword List

As mentioned before Twitter allows to implement a keyword list called track to specify keywords. If a tweet contains a word or a word-compound included in the list it will trigger a download of this tweet (Twitter Developers, 2015c). Since the objective of this thesis is to analyse the dispersion and the variation in the use of different German lemmas it was very clear that such an input list would be essential for this purpose.

4.2.1 Keyword list – First Stage

To create the keyword file for Twitter for this thesis we resorted to the work of the Institute for German Language (IDS). Based on their German Reference Corpus (Deutsches Referenzkorpus - DeReKo) the IDS created the German Word Frequency List called DeReWo (Institut für Deutsche Sprache Programmbereich Korpuslinguistik, 2013). Based on the reference corpus, DeReWo lists the most common lemmas according to their occurrence frequency. However, since a simple count of the occurrence of the lemmas would not make sense, they used frequency classes (Institut für Deutsche Sprache Programmbereich Korpuslinguistik, 2013). According to the IDS (2013) these classes represent a relative frequency compared to the most frequent German words *der*, *die*, *das*, which has been proven to be more accurate and useful.

Due to the restriction imposed by the Twitter API concerning the length of the keyword list, it was decided to select the 100 most frequent lemmas that exhibited at least one different variety in Germany, Switzerland or Austria. For this intent, starting with the most frequent class, each lemma in the DeReWo was manually compared to

the respective entry in the *Variantenwörterbuch des Deutschen* by Ammon et al. (2004). If a lemma consisted of different forms for Germany, Austria and Switzerland all the forms were added to the track, otherwise it was ignored and the next one was analysed. For the case of Switzerland it was also possible to compare the lemmas with the *Schweizerhochdeutsch: Wörterbuch der Standardsprache in der deutschen Schweiz* by Bickel and Landolt (2012), especially if some uncertainty prevailed.
After having extracted the 100 most frequent lemmas and their varieties the data was revised and a comma delimited file was created as an input for Eclipse and the streaming API.

This kind of approach is not entirely new. Scheffler (2014) used a similar approach to create a corpus based entirely on German tweets. But since she did not want to compare specific lemmas, she used very-high frequency terms in the track. By doing so she was able to stream almost every single German tweet with a relative short input list (Scheffler, 2014).

4.2.2 Rethinking the Keyword list

After having collected data from July 23rd to October 23rd, i.e. for 93 straight days, the stream was interrupted for an interim analysis to see if the data acquisition was going as planned and if the data would fit the purpose of this thesis.

The analysis showed that some of the lemmas used in the track did not generate any entry in the DB during this period, e.g. the word *Zollstab* that stands for a folding ruler. Therefore, it was decided to adapt the tracking list to increase the number of entries in the DB. Furthermore, since due to the size of Germany it can be assumed that there are more Germans and therefore more German terms on Twitter, it was also decided to address this problem.

4.2.3 Keyword list – Second Stage

To increase the presence of austriazisms and helvetisms relatively to the number of teutonisms in the database a second keyword list needed to be created. For the austriazisms we resorted to the Protocol-Nr.10 from the European Commission (Europäische Kommission, 2010). When Austria joined the *European Union* (EU) in 1995, after an intense debate, they succeeded in the recognition of 23 words that are considered of importance for Austria by the EU (Wanzeck, 2010). The 23 austriazisms

and their common German counterparts, as displayed in table 4.1, were added to the keyword list.

Austriazism	**Standard German**	***English***
Beiried	Roastbeef	*Roastbeef*
Eierschwammerl	Pfifferlinge	*chanterelle*
Erdäpfel	Kartoffeln	*potato*
Faschiertes	Hackfleisch	*minced meat*
Fisolen	Grüne Bohnen	*green bean*
Grammeln	Grieben	*greaves*
Hüferl	Hüfte	*haunch*
Karifol	Blumenkohl	*cauliflower*
Kohlsprossen	Rosenkohl	*Brussels sprouts*
Kren	Meerrettich	*horseradish*
Lungenbraten	Filet	*filet*
Marillen	Aprikosen	*apricot*
Melanzani	Aubergine	*aubergine*
Nuss	Kugel	*ball*
Obers	Sahne	*cream*
Paradeiser	Tomaten	*tomato*
Powidl	Pflaumenmus	*plum butter*
Ribisel	Johannisbeere	*red currant*
Rostbraten	Hochrippe	*rib eye*
Schlögel	Keule	*joint*
Topfen	Quark	*curd*
Vogerlsalat	Feldsalat	*corn salad*
Weichseln	Sauerkirschen	*morello cherry*

Table 4.1: The 23 Austriazisms from Protocol-Nr. 10 (Europäische Kommission, 2010).

Since there is no such list for helvetisms, we resorted to the listing of helvetisms that can be found on Wikipedia (2014). All the helvetisms and their German counterparts where extracted from the web site and subsequent compared to DeReWo to extract the frequency. Only the 100 lemmas belonging to the most frequent classes where selected and incorporated in the keyword list, together with their common German counterparts. Another difference between Germany, Austria and Switzerland lies in the use of the <ß> that stands for <ss>. While in Switzerland the <ß>, in Germany and Austria the use is compulsory according to the grammatical rules (Ammon et al., 2004; Duden, 2013b). For taking this also into account all the helvetisms containing a <ss> were also written down with a <ß> and incorporated in the list.

However, even after having deleted all the duplicates the new and augmented list hit the limit for the keyword list imposed by Twitter. Despite there being no such information on the official Twitter channels our case showed that the limit is at 400 keywords per list. To reduce the number of lemmas in the list two approaches were followed. First, due to the nature of the streaming API (2015b) we do not need to add for all words with a <ss> the same word with an <ß> to it. It is enough to add simply a <ß> to the list to collect all words with a <ß> in it. And second, all the lemmas from the first keyword list that had generated none or only few entries in the DB were removed. All the new added lemmas, i.e. the austriazisms, helvetisms, and their German counterparts were kept in the new list to enable a comparison.

With this new adjusted keyword list we collected further data over a period of time of 70 days. These two stages combined resulted in a period of 163 days were data was collected. However, in those 4 days where the code was started or interrupted we did not collect data over the entire day.

4.3 Handling of automated and semi-automated Programs

4.3.1 Identification of Bots and Cyborgs

As mentioned in section 3.2.3, bots have become highly complex in the last few years and are therefore difficult to detect, especially the cyborgs. Usually, one would require complex algorithms to detect with confidence the majority of users that could be considered to be bots and cyborgs and to extract their tweets. However, the creation and implementation of such an algorithm could be a work on its own as describes in section 3.2.3. Nevertheless, for this thesis there is one possible aspect that based on simple structures could help to detect the rudimentary automated programs.

According to Chu et al. (2012) it is currently expensive and non-practical to run bots on mobile devices implying that most of these automated programs are run on stationary devices and, therefore, should have the same coordinate for all the tweets. However, one of the requirements during the data acquisition was the necessity to be georeferenced. This reduces already the number of bots since it would not make much sense to post the location of a device sending spam or malware over Twitter. Further, for the collected data we can use the coordinates of each post and analyse the geographical distribution. The idea behind this concept is rather simple. If a user has posted an extremely high number of tweets from the same coordinate the possibility

of being a bot or cyborg is relatively high. Users exhibiting such a behaviour can then be manually further examined to assess if it is indeed a bot or not.

Nevertheless, it is obvious that this method is not perfect and that it exhibits some flaws. First, users posting from their desktop PC will also exhibit only one coordinate for all their posts. The question is where to set the threshold for the number of tweets. Second, since we use a keyword list to filter the data streaming it is possible that only one tweet from a bot exhibited a lemma from the input list. Such a post would remain undetected in the data set used in this thesis. And third, due to the simplicity of this approach more complex programs will not be detected. Nevertheless, this approach can further help to reduce the amount of undesired data from our DB.

```
SELECT userid, count(t.tweet_text) AS tweet_Count, count
(DISTINCT geolocation) AS dif_coord FROM tweets t GROUP BY
t.userid;
```

Figure 4.3: SQL command to compare the number of tweets and coordinates.

Figure 4.2 above depicts the SQL code used to search the collected data for bots. For each user this code counts the number of different coordinates from which all tweets were sent.

4.3.1 Bot and Cyborg Removal from the Database

The section above introduced a possible approach to deal with simple bots and cyborgs in the data set. The next step would be to remove them. Based on the resulting list one can access a method to extract the bots from the DB. We decided to go for the simple coefficient:

$$\lambda_i = \frac{Number\ of\ Tweets_i}{Number\ of\ Coordinates_i}, \quad i \in Users$$

The coefficient λ tends towards 0 if one user posts an enormous number of Tweets from one single location, and equals 1 if every single Tweet was posted from a different coordinate. The subsequent is where to set the threshold defining which users will

be deleted and which retained. After a manual analysis of the respective λ it was decided to set the threshold at λ = 0.05, i.e. 0.05 posts per coordinate. This value might seem arbitrary, however, the manual analysis revealed that from the 47 users with a λ under 0.05, 36 could be considered as bots or semi-automated programs, 5 could not be classified properly, and 6 were human. The number of human users started to increase significantly after a λ of 0.05.

Nevertheless, as mentioned before this method is not perfect and there will still be tweets from bots and cyborgs in the DB. And of course one could also argue about the threshold that might seem arbitrary. One has to ponder between practicality and accuracy.

The removal of these bots from the DB consisted of various steps. Due to the structures of the DB (figure 4.1) the users and their tweets could not simply be deleted. Prior to deleting the tweets from the table tweets, one had to delete all the foreign keys in the other tables that consisted partially by the ID of the tweets. This implied to delete the users and their tweets in the tables tweet_to_hashtags, submits, tweet_to_queries before being able to delete the posts from the table tweets. The entire code used for the removal of the tweets can be found in the appendix C.

4.4 Data Cleaning

The removal of bots and cyborgs with the method explained above reduced the number of tweets from 490'596 to 415'152 as can be seen on table 5.3 later in this work. Nevertheless, there is no guarantee that the remaining tweets contain the exact lemmas defined by the input keywords used with the API. The reason for this lies within the API itself. For instance, by defining the word *Amt* (agency, department) the API will also consider tweets containing the word *gesamt* (total, entire), which have no connection at all except the string sequence *AMT* (Twitter Developers, 2015b). Therefore, to minimize the amount of undesired tweets in our DB, the data has to be revised with respect to the lemmas in the keyword list.

4.4.1 Pre-processing of the Data

One possibility to extract all the tweets that contain a certain keyword is by using a simple SQL query like:

```
SELECT * FROM tweets WHERE tweet_text LIKE '%keyword%';
```

Figure 4.4: SQL command to extract tweets by keyword.

However, such a query induces the same problem as the Twitter API, which is described in the paragraphs above. To overcome this problem, a different code needed to be written in Java. The code used in this step is depicted in the appendix at the end of this work. The idea of this code is to extract all words and the respective compounds, based on the keywords that were used for filtering. Furthermore, the code counts the number of occurrences in tweets and shows the respective tweet IDs.

4.4.2 Extracting the Keywords from the DB

The code used in the pre-processing generated a list with 90'892 keywords and after removing the duplicates with Microsoft Excel a total of 31'373 remained. However, since our interest lies also in the spatial distribution of these lemmas all the words that only occurred once in the DB were also removed. This means that in the end 17'766 different keywords remained that were equal to the lemmas defined previously or at least a compound based on such a keyword.

Nevertheless, the number of remaining keywords was still too large and contained also lemmas that were of no use for this work. To eliminate those words a process of various steps was followed. During each step the resulting keywords, their frequency and the respective tweets where extracted and stored. The steps were organized as following:

1. All the keywords that matched perfectly the lemmas defined in the beginning of this work were extracted. Since there can be differences in the use of upper or lower case, all the keywords and lemmas were converted to lower case. However, not all lemmas are represented here since some only have triggered one download that was removed in a previous step or did not generate an entry in the DB at all, like the old Swiss word for father *Ätti*.

2. To incorporate also the plural forms of the words and their compounds we used the Snowball Stemmer[15]. This stemmer is free to use and supports various languages, including German. Another advantage is its capability to be implemented in Java (Snowball, 2014). Both, the list with the input lemmas and the keywords extracted with Java were run through the stemmer. Matching results

[15] snowball.tartarus.org

were then also extracted.

3. Because of the uncertainty induced by the stemmer the two most common plural endings in German, i.e. –s and –en, were verified by hand. For this purpose Microsoft Excel was used and all the lemmas were extracted from the keyword and replaced by a comma. The endings of –en and –s where then analysed. Most of the keywords had been correctly identified by the stemmer in the previous step, but there were still a few remaining with the ending -s, including *Velos* (bicycles) or *Handys* (mobilephones). The endings on –en were all identified by the stemmer.

   Based on a random sample some of the classified lemmas were analysed manually to conclude if this step did function as planned.

4. As explained before a further difference in the German varieties of Germany, Austria and Switzerland lies in the use of <ß>. While in Germany and Austria the <ß> is compulsory according to grammatical rules, in Switzerland the use of <ss> prevails. To be capable of analysing its distribution all the lemmas on the keyword list containing a <ss> or a <ß> were once written the other way round. These words were then compared with all the keywords from the Java code and if matching, extracted and added to the list with the other lemmas.

At this point some caution in the handling of the data is advised. Since <ß> and <ss> stand for the same, some programs, like Microsoft Excel, tend to handle this lemmas as duplicates. For instance, *Fuss* and *Fuß* (foot) will be seen as identical and one will be removed as a duplicate. Therefore, with focus on <ss> and <ß>, the keywords and the lemmas were compared with the dataset before the removing of any duplicates.

The result and the discussion of this cleaning is depicted in chapter 5. A complete list of all the keywords that were extracted during this step can be found in the Appendix D.

4.5 Geographical Analysis

4.5.1 Geographical and Projected Coordinate System

The coordinates of the collected tweets are stored in the format EPSG: 4326, also known as *World Geodetic System 1984* (WGS 84). One of the problems of this system

is, that is a global approximation and is not as accurate on regional level. However, as it can be expected, that the tweets will be distributed globally, WGS 84 is the best choice. Further, since the geolocation of the tweets is also stored in this format (section 3.2.2), it has the advantage, that there is no necessity to conduct any geographic transformation at the beginning.

For the geographical analysis we use the software ArcGIS 10.2 by ESRI which allows us to manage data, create maps or to change the geographical reference system, just to name a few. The latter aspect is of particular importance for the analysis. As mentioned in the paragraph above, WGS 84 has some inherent advantages and the coordinates are represented in the well-known longitude and latitude format. However, it is not suitable for any areal analysis since the unit used by this system is degree. To overcome this significant problem the maps of Germany, Switzerland and Austria used in this work will all have to be transformed to the projected coordinate system *Europe Albers Equal Area Conic*. This projection enables us to represent the areas in Europe proportionally to the real area on earth (Esri, 2014a). And second, the ground unit are meters which facilitates any areal analysis. The geographical origin of the tweets that will be depicted as points on a map can also be transformed to *Europe Albers Equal Area Conic* projection to be further analysed.

4.5.2 Density Maps

It would be possible to simply represent the origin of the Tweets as points on a map. However, such a representation can be confusing and none or only few insights could be gained from it. One approach to facilitate the interpretation of such a map is by using a Kernel Density Map. ArcGIS has already a kernel function incorporated which allows us to determine the occurrence of a certain event over a unit area, which in the case of the *Europe Albers Equal Area Conic* projection used here would be meters (Esri, 2014c).

The idea behind the kernel function is to calculate the density based on a neighbourhood. The neighbourhood can be defined according to the respective needs or we can resort to a spatial variant of *Silverman's Rule of Thumb* (Gaussian approximation) which is calculated by the software based on the input dataset (Esri, 2014c). To facilitate the comparison over Germany, Switzerland and Austria with respect to the differences in area we opted to use the rule of thumb by Silverman, since

it would define a neighbourhood best suited for the entire study area. Further, instead of using a classification we opted for a stretched visualization with a histogram equalization. According to Esri (2014b) this stretch method is best suited when a high number of pixels values "are closely grouped together". This can be expected since according to our expectations the German tweets have its origins also mainly in urban and suburban areas as it is in the US (Duggan et al., 2015). With the stretched visualization the results are stretched along a colour ramp for the representation. The reason for this choice is the difference in numbers between the three centres. A classification of the entire region in various classes could be problematic since the majority of the tweets were sent from Germany. This could lead to a shift, independent of using *natural breaks* or *quantile* classification.

4.5.3 National Subdivisions

In some cases it might be interesting to analyse the data in respect to just one nation. There would be two possibilities for this purpose. The first option is to use Java itself to conduct a point-in-polygon test that would extract all tweets that have its origins in the polygon of the respective country. The second option, which probably is the simplest approach and the one used in this thesis, is to use ArcGIS for this task. This approach is divided in two sub-task. First, one has to generate a text or a CSV file that can be used as an input for the tables in ArcGIS. But, since this files often contain an enormous number of rows, i.e. tweets, it is not possible to simply open the file in ArcGIS. The answer is to create a dBASE in ArcGIS that enables the handling of such files. This is done with the function *Table to Table* that can be found under *Conversion Tools*. And second, to extract the information of a certain region we can use either the function *Clip* under *Analysis Tools* if you are working with vector datasets. Or you can use the function *Extract by Mask* under *Spatial Analyst Tools* when dealing with raster data sets.

In this work both functions, *Clip* as well as *Extract by Mask* where used according to the data format used. Each tweet can be represented on the map as a vector, i.e. as a point. On the other hand the results of the kernel density function, as explained in the previous section, are raster dataset implying the use of the *Extract by Mask* function.

4.6 χ^2-Test

All the methods explained above are necessary to handle or analyse the data collected. Nevertheless, they do not show if there is any difference in the use of a certain lemma over the three centres of the German language. For this purpose we need a statistical test that allows us to determine if there are any significant divergences in the use of a certain lemma. The test that will be used in this work is the χ^2- *test* (chi-Squared test). To understand how this test can help to determine if the use of teutonisms, helvetisms and austriazisms is regionally bound, we will first give a short explanation of this test and then discuss how the χ^2- test can be implemented with the collected tweets.

The chi-squared test enables us to compare an observed frequency with one that theoretically would be expected (Storrer, 2009: 260). Further, it has the advantage that it allows to include more classes than the z-test, which only allows two values to be compared (Freedman et al., 2007: 523). Both, Storrer (2009: 260) and Elpelt & Hartung (2004: 150), work with the following two hypothesis:

- H_0: The sample derives from a population with a determined probability distribution.
- H_1: The sample does not derive from the same population.

However, to implement this test some requisites have to be met. First, the observed values have to be absolute frequencies and not just relative. Second, the observed values *n* need to be divided into *k* classes with x_i elements. Further, t_i is the number of the expected number of elements in each class. Finally, the expected frequencies t_i have to be ≥ 5 (Storrer, 2009: 260). Elpelt and Hartung (2004: 150) argue that the test is precise enough if less than 20% of the used classes have an absolute frequency that is lower than 5 and none has less than 1 count. On the other hand, Devore et al. (2014: 396) argue that the expected counts have to be at least 5, like mentioned by Storrer (2009: 260). Further, since we only have three classes once one class has less than 5 entries we already would have fallen below the limit mentioned by Elpelt and Hartung (240: 150). For these reasons we decided to set the minimum requisite at 5.

To conduct the test we first need to determine the desired level of significance α and the *degrees of freedom* (DoF) ν. The DoF is given by the number of classes minus the linear relationship between the observed x_i and minus the number of the estimated

parameters. This elements help to determine the critical value $\chi^2{}_{\alpha,\nu}$ to whom our calculated value χ^2 will be compared (Storrer, 2009: 260). A list with the needed $\chi^2{}_{\alpha,\nu}$ can be found in Storrer (2009: 260) or in any other major statistics book. The formula used to calculate χ^2 is:

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - t_i)^2}{t_i} = \sum_{i=1}^{k} \frac{(observed - expected)^2}{expected}$$

Based on the relationship between $\chi^2{}_{\alpha,\nu}$ and χ^2 we can accept or refute our working hypothesis. If $\chi^2{}_{\alpha,\nu} \leq \chi^2$ we can refute H_0, i.e. our sample does not have its origins in a population with a certain distribution. If $\chi^2{}_{\alpha,\nu} > \chi^2$ there is no reason to refute H_0 (Storrer, 2009: 260).

After explaining how the chi-squared test works we will now demonstrate how this test can be implemented in this work. The idea is simple, if there is no difference in the use of a certain lemma the geographical distribution has to be similar or identical to the one of all the collected tweets. However, with the helvetisms, teutonisms and with the austriazisms this should be different. As explained earlier in this work these lemmas should show a different distribution. The helvetisms should be overrepresented in Switzerland and underrepresented in the other three centres. Similar happens with the austriazisms and Austria, and the teutonisms and Germany.

The population that we will use is given by the distribution of all the collected tweets. Its percentage can be used to define the expected frequencies of the lemma distributions that we will use to calculate χ^2. The observed occurrences will be aggregated nationally to create our three classes: Germany, Switzerland and Austria. We have a DoF of two, since we have three classes and there is no linear relationship between the parameters and we only want to estimate one parameter. Further, I will opt for the common level of significance of 0.05 i.e. that the probable error is 5%. Therefore, our critical value $\chi^2{}_{\alpha,\nu}$ will be 5.991 (Storrer, 2009: 230, 365). The working hypothesis that will be used are the same as given above. Furthermore, as explained above, the expected absolute frequency has to be at least 5. Due to this limitation and the percentage distribution of Switzerland, i.e. 5.599%, each lemma has theoretically

to appear at least 90 times over the three centres. This reduces the amount of lemmas in our list that are suitable for our test to 209.

5. Results

5.1 Preliminary Results

5.1.1 User Activity

Data acquisition took place over a period of 163 days, i.e. between June 23rd and January 22nd with an interruption from October 23rd to November 14th. During this period a total of 490'596 tweets were streamed by the API and stored in the DB. A first analysis shows that 70'692 users were responsible for all the data collected. Furthermore, as figure 5.1 depicts, the activity of the various users is not as uniform as one might expect. While most users have generated only one or two entries in the DB, the ten most active users have generated 72'626 Tweets, i.e. almost 15% of the Tweets were posted by 0.01% of the users.

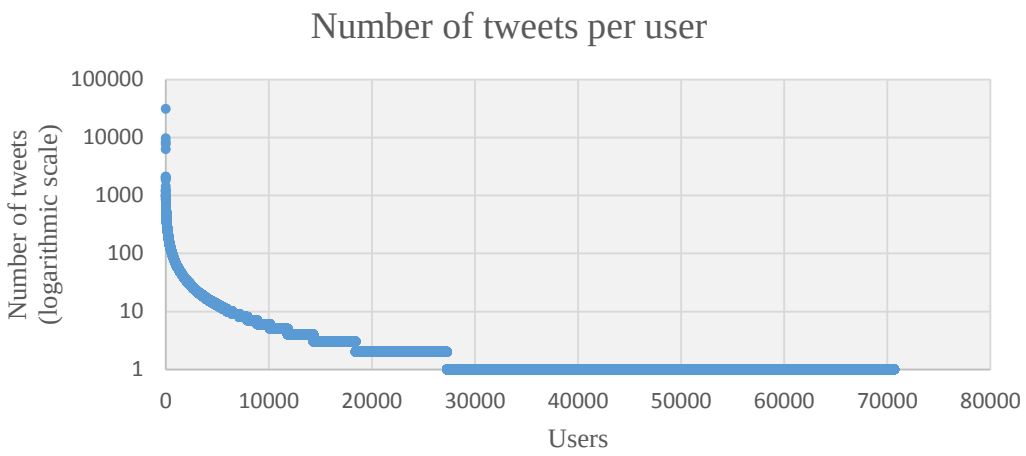

Figure 5.1: Number of tweets per user.

5.1.2 Temporal Distribution

The first stage of data acquisition took place between July 23rd and October 23rd. During these 93 days, based on the first keyword list, 289'705 tweets where streamed. During the 70 days of the second stage which took place during November 14th to January 22nd, 200'891 tweets were collected, based on the keywords on the second list. Nevertheless, with an average of 3009.8 tweets per day for both phases combined, one

can argue that the temporal distribution is relatively homogenous. This statement is further endorsed by the figure below.

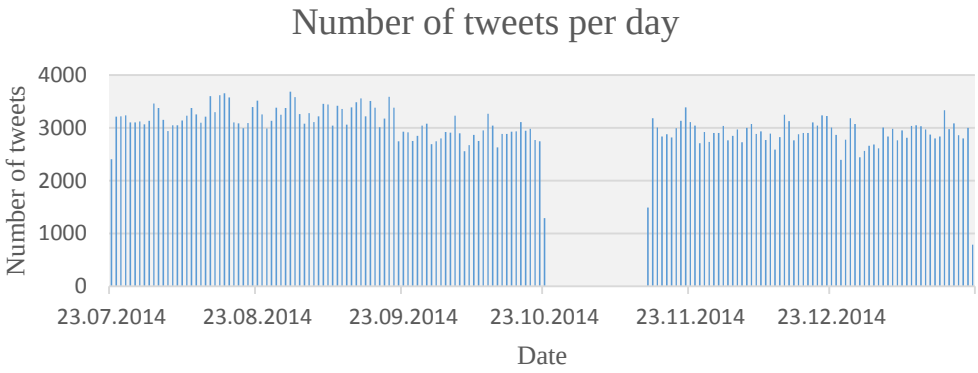

Figure 5.2: Daily entries of tweets in the DB.

5.1.3 Geographical Distribution

According to our expectations, the data collected has its origins all over the world. Figure 5.3 depicts these circumstances. Created with ArcGIS, it depicts the global distribution of the collected tweets on a map with the WGS84 system. Despite the global distribution of the tweets we can see that the majority lies in Europe. There is also a considerable amount of tweets in the United States and in South America.

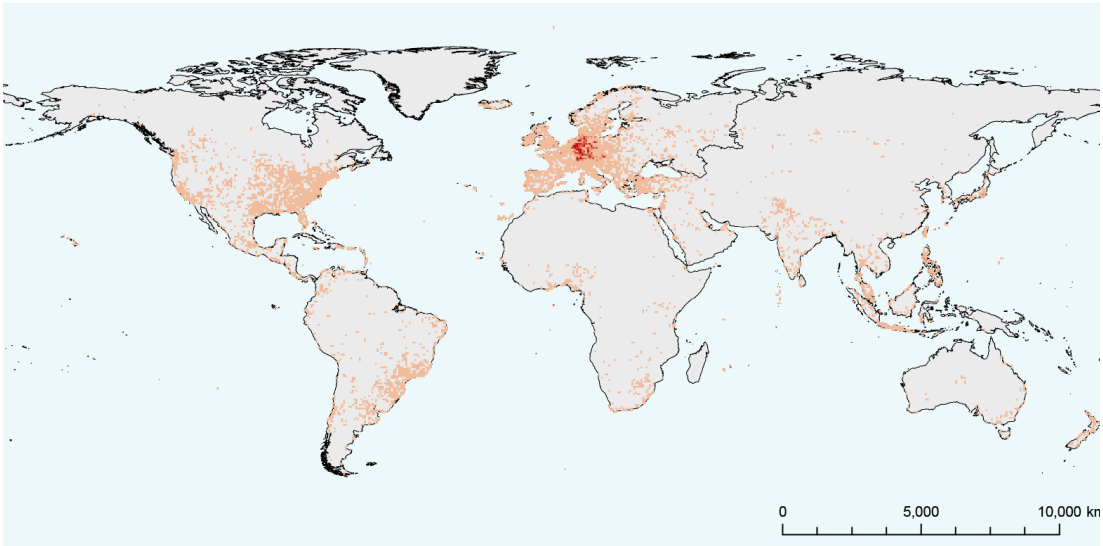

Figure 5.3: Geographical distribution of the collected tweets.

Nevertheless, it would be premature to draw some conclusions from the image above, since it is very vague and suggests a rather vast distribution. A more profound analysis with ArcGIS presents a more conclusive overview. First, from the 490’596 collected tweets 427’084 are situated in the study area defined by the centres (figure 3.1), i.e. 87.05% of the data derived from Germany, Switzerland or Austria. This implies that most of the tweets recognized by Twitter as being German, and containing at least one of the lemmas in the input list, originate indeed from the area where German is spoken. And second, it was clear, that in our DB there would also be data stored from people outside our study area. A German speaking person on a holiday abroad is not counted and therefore removed from further analysis. On the other side, users that are situated inside the study area but living in a different nation than Germany, Switzerland or Austria, will be taken into consideration, since the Tweet was sent from one of these three countries.

This point was predictable since the conceptual outline presented earlier in this work. Nevertheless, it would require rather complex algorithms to take all these aspects into consideration. However, this issue has to be taken into account when drawing any conclusions.

Total number of tweets	490'596
Tweets inside the study area	427'084
in Germany	379’925 (88.96)*
in Switzerland	21'677 (5.08)*
in Austria	25'482 (5.97)*
* in % of the tweets inside the study area	

Table 5.1: Key numbers of the geographical distribution.

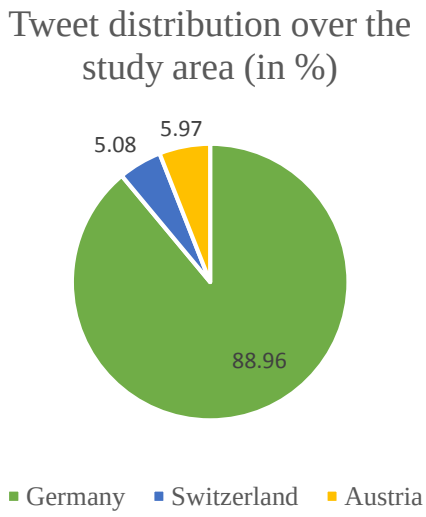

Figure 5.4: Graphical representation of the distribution inside the study area.

More interesting than the overall distribution of the Tweets inside the study area is the distribution over the three national centres discussed in this thesis. From the 427’084 Tweets that were posted from inside our study area, 379’925, i.e. almost 89%, came

from Germany. The other 11% are more or less evenly divided over Switzerland with 5.08%, and Austria with 5.97%. Figure 5.4 depicts this situation graphically. The distribution itself is not entirely surprising. Due to the size and the population it could be expected that most of the data collected has its origins in Germany. Nevertheless, it would premature to draw any overall conclusions at this point because of a few issues. First, we do not know the geographical distribution of the users. Second, it has still to be determined if some of the most active users (see section 5.1.1) can be considered as humans or as automated or semi-automated programs, an aspect that will be discussed in the following chapter.

5.1.4 Bot Identification and Extraction

Table 5.2 lists the 25 most active users and the number of coordinates where they are tweeting from. As we can see the three most frequents users are tweeting from one single coordinate. However, as mentioned in the methods, since the applied method has its flaws, it would be precipitated to draw any conclusions before analysing the data manually. The column on the right in table 5.2 shows the results of the manual test and if the user is a human or in some way computer-assisted or fully automated.

The outcome corresponds the anticipated results. It is obvious that all the users that only posted one tweet will also have only one coordinate. However, these users are not the main target of the approach presented here, despite the possibility that one of these messages was indeed posted by a bot or a cyborg. The target of this approach were the most active users since they account for a big part of the data – see chapter 5.1.1 – and have therefore a bigger impact on the results.

User Nr.	**# Tweets**	**# of different Coordinates**	**Bot/Cyborg**
1	30883	1	Yes
2	9578	1	Yes
3	8282	1	Yes
4	7688	1518	Yes
5	6203	1152	Yes
6	2097	1597	Not
7	2094	863	Not
8	2019	342	Not
9	1938	4	Yes

10	1844	1395	Not
11	1423	1223	Not
12	1237	776	Not
13	1217	2	Yes
14	1189	925	Not
15	1168	632	Not
16	1037	600	Not
17	1032	636	Not
18	982	344	Not
19	969	812	Not
20	963	720	Not
21	948	698	Not
22	858	1	Yes
23	848	1	Yes
24	824	185	Not
25	815	470	Not

Table 5.2: The 25 most active users and their number of different coordinates.

The three images that follow give examples of the messages posted by the three most active users listed in table 5.2. The first two users are automated programs that tweet about the playlist of the respective radio stations. User number three posts about trends in Germany. From the other users that were listed as bots or cyborgs in the list above some are classical news feeds as mentioned by Ferrara et al. (2014), e.g. user number twenty-two. Further, there are also some meteorological services posting about weather conditions, which is the case with the users thirteen and twenty-three.

Figure 5.5: The most active user.

Figure 5.6: The second most active user.

Figure 5.7: The third most active user.

Nevertheless, the distinction between bot and not-bot is even after this classification approach, not that simple. User number four and five can be seen as examples. The tweets from these users have been sent from more than 1000 coordinates which could lead to the assumption that these users are not automated or semi-automated programs. However, a quick inspection revealed that these two users can indeed be considered as bots or cyborgs. The inspection reveal that user four is also a news feed that uses the coordinates to locate the news. User number five posts about job vacancies and the coordinates are used to inform in which region this vacancy is available.

Anyhow, the method used in here proved to fit its purpose and helped to identify a considerable amount of bots or cyborgs. However, as mentioned in section 4.3, this method only works with simple bots and cyborgs and is not capable to identify all of these programs as some examples have demonstrated. Nevertheless, despite the flaws mentioned earlier, we will use this method to revise our dataset. For the explanation of how and which threshold is used please consult section 4.3.

5.2 Final Results

5.2.1 An Overview

Previous in this work (see chapter 4.3 and 4.4) it was explained on how the collected data could be cleaned. The cleaning process consisted of two essential steps. First, a very elementary method was implemented to remove users that were suspected to be

bots or cyborgs. And second, all tweets that did not contain any of the desired lemmas were also removed from the dataset. To understand the implications that this steps had on the collected data they will now be briefly discussed in the order of their implementation.

5.2.1.1 After the Removal of Bots and Cyborgs

The implementation of the method discussed in section 4.3.1 resulted in a removal of 48 users. With these users a total of 75’444 tweets were deleted, which represented a reduction of 15.38%. The average number of tweets that were collected each day during the two phases dropped from 3009.8 to 2546.9 posts per day.

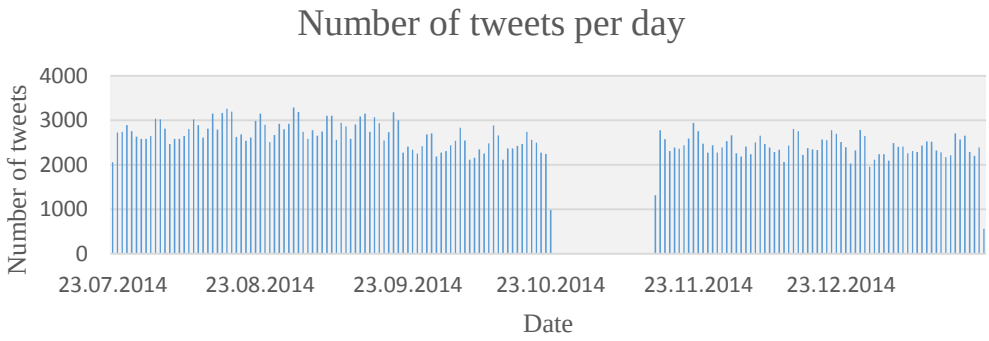

Figure 5.8: Daily entries of tweets in the DB after cleaning.

Table 5.3 below depicts some key figures after the reduction of the dataset. To facilitate the comparison it contains also the numbers previously depicted in table 5.1 and the differences that have resulted.

Number of tweets:	**After removal**	**Before removal**	**Absolute Difference**	**Difference in %**
Total	415’152	490'596	75’444	15.4
Tweets inside the study area	352’835	427'084	74’249	17.4
in Germany	308’126 (87.33)*	379’925 (88.96)*	71’799	18.9
in Switzerland	19’881 (5.63)*	21'677 (5.08)*	1’796	8.3
in Austria	24’828 (7.04)*	25'482 (5.97)*	654	2.6
* in % of the tweets inside the study area				

Table 5.3: Key numbers after and before bot and cyborg removal.

As the numbers above depict almost the entire data which was removed was situated inside our study area. From the 75'444 posts removed only 1195 had its origins outside Germany, Switzerland, or Austria. With over 70'000 tweets, Germany in particular was the most affected by the implementation of the removal method. Nevertheless, also Switzerland and Austria suffered a decrease in the number of tweets but on a smaller scale. The geographical distribution of the tweets over the three centres retained almost identical proportions with changes in the order of about 1%.

5.2.1.2 After Lemma Removal

To facilitate the understanding and interpretation of the results at this point we need to have a look at all the numbers involved in this step and the methods implemented beneath (see chapter 4.4). The keyword list that was used for Twitter to specify which tweets to download consisted of 567 distinct lemmas. These 567 lemmas spawned a total of 51'373 different lemma variations. However, as explained previously, all the variations that only appeared once in the database were removed. This left us with a remaining of 17'767 different variations. In this amount of variations only 411 from the 567 input lemmas appeared unaltered in the database. Nevertheless, based on the methods explained in chapter 4.4 a total of 882 keywords were selected and the IDs of the tweets in which they occurred were extracted. Since a post can contain multiple lemmas all the duplicate IDs were removed, which resulted in 407'009 unique tweet IDs. At this point 8'143 tweets were removed from the dataset, since they did not contain any of the specified lemmas.

At this point one might be surprised that only 8'143 tweets were removed during this step despite taking only 882 keywords into consideration instead of the 17'767. However, this is not that surprising. Some of the lemmas, specially the prepositions like *mit* (with) or *auf* (on, at, to or in), are very common and account for a big number of entries in the database.

5.2.2 Geographical Distribution for the three Centres

As mentioned before our primary focus lies in the three centres Germany, Switzerland and Austria. Nevertheless, at this point it might be interesting to see from which nations the tweets were posted. This helps to see from where all the German tweets are sent. However, once again, in this point some attention is required since the

numbers do not show if the person posting is a native, a German-speaking tourist or a German-speaking person living there.

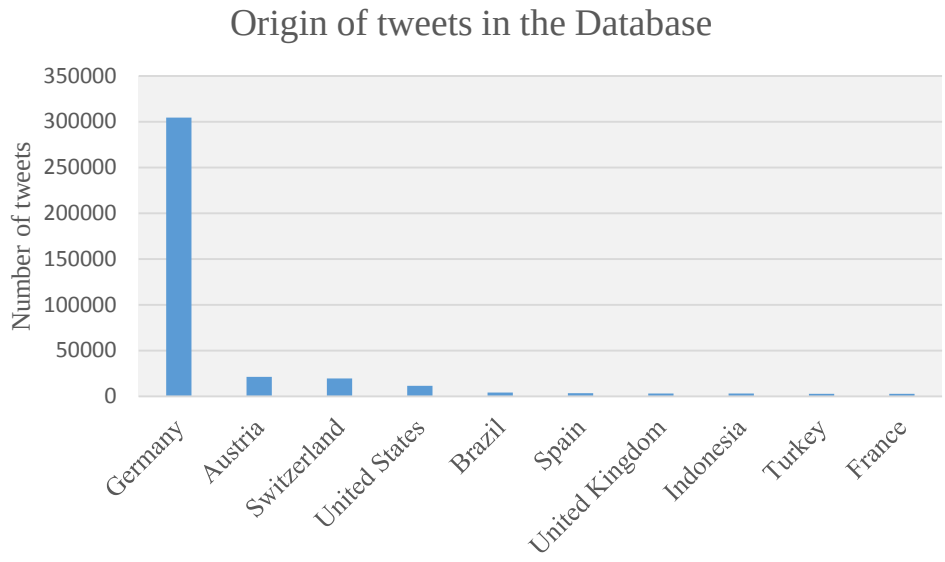

Figure 5.9: The ten most frequent origin nations of the collected tweets.

Figure 5.9 depicts the ten most common nations from where the tweets in the DB were sent. Unsurprisingly, most of the tweets were posted from Germany. With over 300’000 tweets Germany is by far the most active nation concerning German tweets. With 21’441 respectively 19’604 tweets follow Austria and Switzerland. The proportions before and after removal remained almost identical, meaning that all the three centres discussed in here were affected equally by the methods implemented. Interestingly is the fact that 11’441 from the tweets in the DB were sent from the United States. However this can be justified by the high number of German speakers in the United States as mentioned by Durrell (2006). The complete list of the national distribution of the collected tweets can be found on the Appendix D.2.

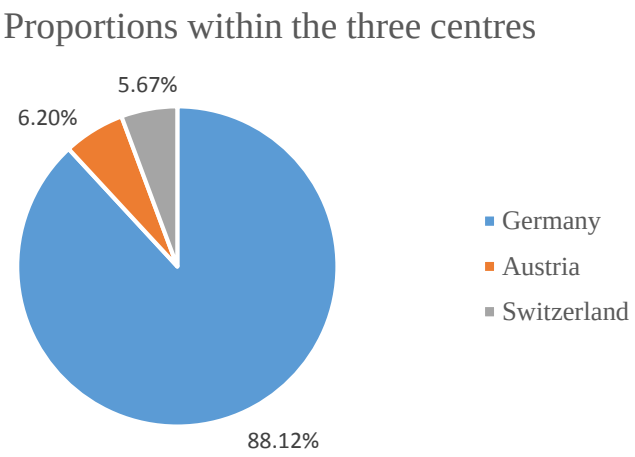

Figure 5.10: Relative distribution of the tweets over the three centres.

However, the main interest of this work lies on the three centres Germany, Switzerland and Austria. Therefore, we will not discuss the global distribution again like in chapter 5.1.3, which helped to get a first insight, but rather focus on our study area.

Figure 5.10 above depicts the relative proportion of the posts inside the study area, i.e. the three centres. As we can see 88.1% of the tweets posted within the study area were posted from Germany, 6.2% from Austria, and 5.7 % from Switzerland. This distribution can and will be used to determine if there is any significant deviation of the tweets from a certain lemma variation, e.g. *Velo* and *Fahrrad*.

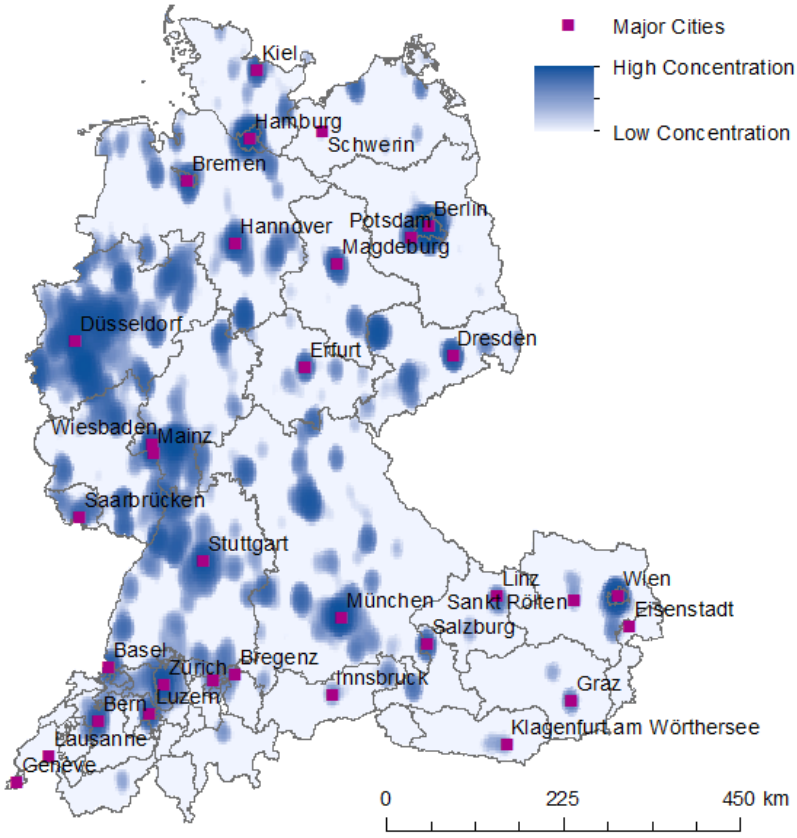

Figure 5.11: Origin of the tweets in the DB.

Based on the clean adjusted dataset, figure 5.11 above depicts a density map that was created according to the steps explained in chapter 4.4.2. The colour ramp was chosen so that the darker colour represents a higher density and vice versa, i.e. the darker the colour the more tweets where sent from this region.

The results are not surprising. As could be expected the density of tweets is the highest in urban and suburban centres. In Switzerland the highest density levels are reached in the greater area of Bern, Basel, Zurich and Luzern. In the region of St. Gallen the density is slightly lower. In Austria the highest densities are in the greater areas of Wien, Salzburg, Linz, Graz and Kizbühel. The cities Klagenfurt and Innsbruck show a little lower activity level. In Germany we can observe the same pattern. A high number of tweets have been sent from the metropolitan areas of Frankfurt, Berlin, Munich, and Hamburg. The Ruhr region and the area around Cologne show also high levels. In summary, one can conclude that the tweeting activity is higher in the urban areas of the study area. Rural regions tend to show none or only a low level of activity.

5.2.3 Results of the χ^2-Test

Since it would take too long to present the obtained results for all the 209 lemmas, we will present a few cases that are representative and will allow an understanding of the entire list. In section D.3 you can find a list with all the calculated values, i.e. the observed and expected values for each centres, as well as the calculated χ^2 value. The abbreviation DE stands for Germany, AT for Austria and CH for Switzerland. Further, the endings _obs and _exp define if it is an observed or an expected value and the χ^2 column lists the results of the statistical test. At this point it is also important to bear in mind that the critical value for this test is 5.991. All values that are over will cause the rejection of the working hypothesis H_0, which states that the sample derives from a population with a determined probability distribution (Storrer, 2009).

From the 209 original lemmas 10 did not fulfil the requirement concerning the number of expected values per class. For 87 lemmas the calculated χ^2 value was less than 5.991 meaning that there is no reason to reject our working hypothesis. This means that with respect to the geographical distribution of all the tweets in the DB there is no difference in the use of the particular lemma over the three centres. As an example, the word *schauen* (to look) has a χ^2 of about 0.11, as can be seen on table 5.4. If we look at the observed and the expected values we see that the numbers are very similar, meaning that this word is used with the same relative frequency in all three centres.

Lemma	DE_obs	AT_obs	CH_obs	Total	DE_exp	AT_exp	CH_exp	χ^2
schauen	3659	260	236	4155	3665.81	256.55	232.64	0.11

Table 5.4: χ^2 calculations for the lemma schauen.

Figure 5.12 demonstrates this behaviour graphically. The term *schauen* appeared in a total of 4155 tweets. As we can see the origins of those posts lie all over the study area. At first it might seem that Germany is overrepresented but one has to keep in mind that the calculation is based on the relative distribution of the collected tweets and Germany is responsible for about 88% of the tweets.

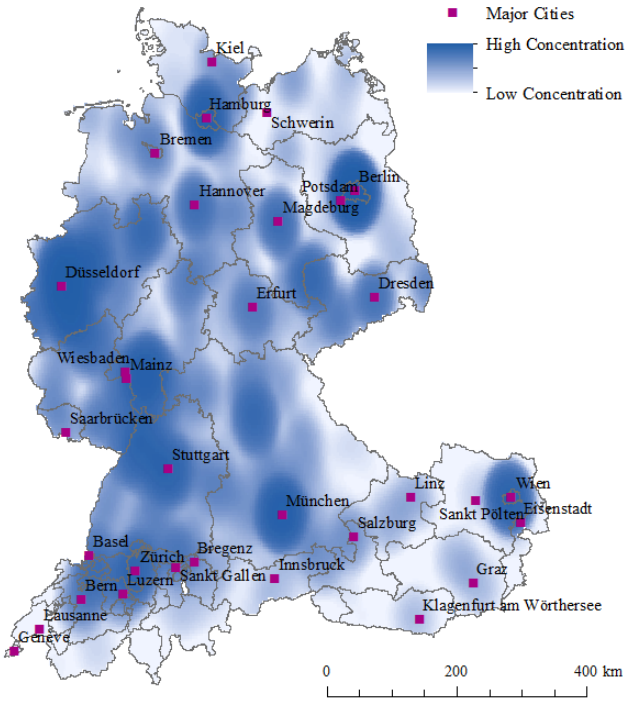

Figure 5.12: Origin of the lemma schauen.

The analysis for the remaining 112 lemmas yielded a value equal or bigger than 5.991 causing the rejection of the working hypothesis, meaning that the probability behind the distribution of the lemma is different from the one of the entirety of the tweets in the DB. Table 5.5 below presents an example to clarify the situation. The word *Urlaub* (holidays) has a calculated χ^2 of 40.02 meaning that there is a significant difference between the observed and expected distribution. As the values in the table below depict *Urlaub* is slightly overrepresented in Germany by about 3.5% and by 1.9% in Austria. However, in Switzerland it is clearly underrepresented by about 58.05%. The explanation for this behaviour is simple. In the German language the words *Ferien* and *Urlaub* are synonyms. However, in Germany these two words have slightly different meanings and people tend to distinguish between them, in Switzerland they do not. And finally, *Urlaub* is more common in Germany and Austria while in Switzerland the word *Ferien* is preferred.

Lemma	**DE_obs**	**AT_obs**	**CH_obs**	**Total**	**DE_exp**	**AT_exp**	**CH_exp**	**χ^2**
Urlaub	1828	126	47	2001	1765.41	123.55	112.04	40.02

Table 5.5: χ^2 calculations for the lemma Urlaub.

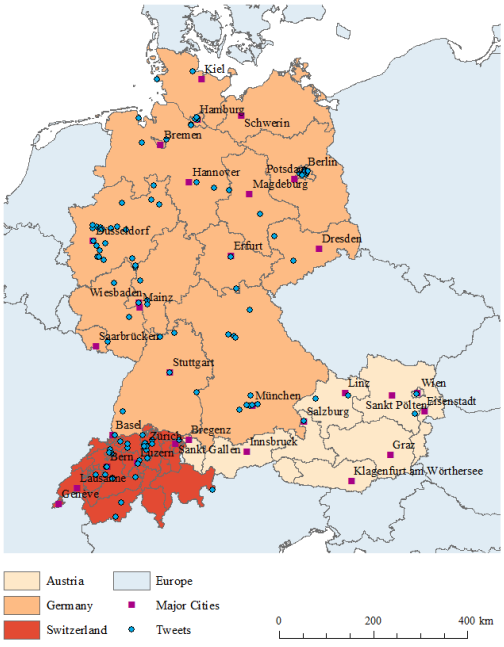

Figure 5.13: Origin of posts containing the word Fuss.

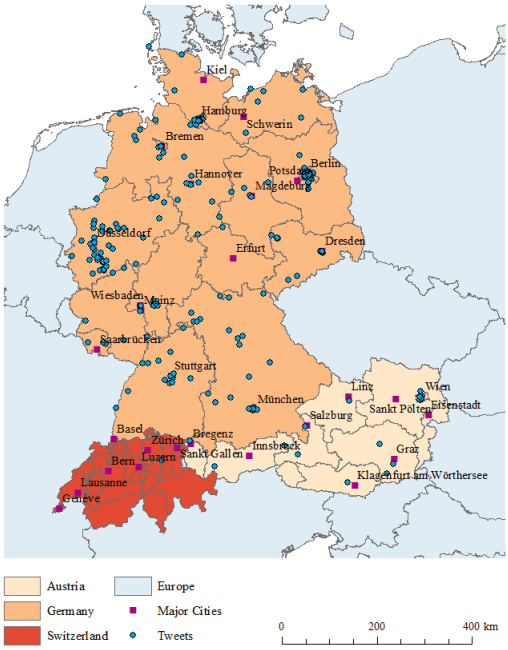

Figure 5.14: Origin of posts containing the word Fuß.

Another interesting case concerns the word *Fuss* (foot). As mentioned before another difference of the German variety of Germany, Austria and Switzerland lies in the use of the letter <ß>, which stands for <ss>. This implies that in Austria and Germany people tend to write *Fuß* while those from Switzerland write *Fuss* (Ammon et al., 2004; LXI). This behaviour is also evident from the results, as it is depicted by table 5.6 below. As we can see for *Fuss* Switzerland is clearly overrepresented while for Germany and Austria the observed numbers are smaller than the expected ones. In the case where <ß> is used instead of <ss>, the behaviour is the opposite. Switzerland is underrepresented while the other two centres are slightly overrepresented. However, interestingly, for the plural form of *Fuß* there is no reason to reject our assumption, i.e. our working hypothesis, that there is no difference in the use of this word since the χ^2 is smaller than 5.991 as it is depicted in the appendix D.3.

Lemma	DE_obs	AT_obs	CH_obs	Total	DE_exp	AT_exp	CH_exp	χ^2
Fuss	90	5	36	131	115.58	8.09	7.33	118.87
Fuß	212	17	1	230	202.92	14.20	12.88	11.91

Table 5.6: χ^2 calculations for the lemma Fuss and Fuß.

The figures 5.13 and 5.14 above illustrate the described behaviour graphically. We can see that the lemma *Fuss* appears very often in Switzerland while, in comparison to the size of the nation, is appears less frequent in Germany and Austria. The opposite happens when we look at the lemma *Fuß*. The number of occurrences for Switzerland decreased remarkably to one while increasing in the other two centres. There is now also a visible peak of tweets with its origins in the metropolitan area of Vienna. Inside Germany the number of occurrences has also increased as the comparison between figures 5.13 and 5.14 illustrates. All this changes arise from the circumstance that <ß> is compulsory in Germany and Austria but not in Switzerland (Ammon et al., 2004).

These examples above only depict a small part of the conducted analysis and are mostly meant to give an overview of the results that were found. As mentioned previously a list with all the calculated χ^2 values is depicted in the Appendices. In the next chapter we will have a comprehensive discussion of the results and to determine its expressiveness.

However, one aspect that the χ^2-test fails to answer is the issue about an eventual language accommodation between users. The results obtained by this test do not allow us any statement on this issue. The reasons that caused these results will be discussed in the following chapter.

6. Discussion

In this chapter we will discuss the results that were presented in the previous chapters. Some of the results were preliminary and helped to get a better understanding of the data and are worth a short discussion. Nevertheless, the main focus lies on the final results, especially on the results of the statistical test. To facilitate the understanding of this chapter we will first discuss the results in a more comprehensive way. We will have a general look on the results, on the geographical distribution and the results of the χ^2-test. Then, some very important issues concerning Twitter as a medium of research as well as when working with lemmas and tokens are discussed. These aspects were also described by other authors and were confirmed by this work. To highlight their importance, they will be discussed separately. During this discussion we will try to answer the research questions that were mentioned in chapter 1.3.

6.1 Insights gained during the Process

As mentioned above the main focus lies on the results obtained with the adjusted dataset. Nevertheless, some of the insights gained during the process are also worthy to be mentioned:

- A small number of users is responsible for the majority of the collected tweets. This behaviour is not new and was already described by other researchers. In this paper we showed that the same behaviour applies also for the German Twitters. Most of the users on Twitter are rather passive and only send a tweet once in a while, confirming that in most cases Twitter is used as a channel to get news or to see what other people are doing. However, as Huberman et al. (2008) argue, the number of posts is influenced by the number of followers. An interesting aspect would be to analyse if the number of social ties has an influence on the pace of the accommodation process. But to be capable of analysing long term accommodation one would require longer time periods and the issue if accommodation can still be analysed after two users have first met is still open (Danescu-Niculescu-Mizil et al., 2011).

- During the data collection period a constant amount of tweets could be streamed through the API. This demonstrated that Twitter, besides its easy accessibility, is a reliable data source that can be implemented in various research areas.

- Based on the assumption that tweets from automated and semi-automated programs are posted from stationary systems, we were capable to implement a simple geographical method to reduce the number of undesired data (Chu et al., 2012). By comparing the activity of each user and the number of different coordinates exhibited by his tweets, we could identify possible bots and cyborgs and remove them and the respective tweets from the dataset. Nevertheless, this method is very rudimentary and could be improved by incorporating other aspects like the time of posting, as done by Zhang and Paxson (2011). Further, as mentioned by Ferrara et al. (2014), bots and cyborgs have become much more sophisticated and are now a research field of its own.

6.2 Final Results

After discussing the insights gained through the preliminary results we will now discuss the results obtained with the adjusted dataset. While discussing the geographical distribution and the χ^2-test results we will try to answer the research questions defined at the beginning.

6.2.1 Geographical Distribution

The geographical analysis conducted in this thesis has not brought many new insights about the geographical distribution of German posts on Twitter. As figure 5.9 in chapter 5.2.2 depicts there is a clear peak in Germany with over 300’000 tweets. This is no surprise since due to the size of the population and the higher degree of activity of the Twitter users one could expect that Germany would account for the majority of the collected tweets. On the other hand Switzerland and Austria exhibit a similar behaviour with both nations accounting for around 20’000 tweets each. Despite this similarities in the number of tweets and the size of the population, there are some interesting differences that require some attention. First, Switzerland has four official languages, around 60% of the population speaks German as first language, while Austria has only one official language (Ammon et al., 2004: XXXVIII). Further,

according to the respective statistical offices, 23% of the Swiss population are foreigners. In Austria they only account for 12.5% of the population (Statistics Austria, 2014; Swiss Federal Statistical Office, 2013a). The interpretation of these results could lead to the assumption that the Swiss people tend to tweet more than the Austrians. However, the interpretation of these national differences is not that simple and it would require an identical demographic analysis for all the three centres to be capable to allow a viable comparison. At this point we cannot conduct such a comparison since most of the demographics available about Twitter are non-academic. Studies like those from Steiner (2012) or Smith and Brenner (2012) enable a better understanding of this microblogging service but due to different calculus and classifications a comparison cannot be recommended. There are other reasons why a comparison based merely on these numbers would not be advisable. First, as Smith and Brenner (2012) have noted for the United States, the cultural background and the educational level are both important factors influencing the activity on Twitter. Second, Grossenbacher (2014: 104) found out that for Switzerland the French-speaking Twitter users are by far more numerous and active than their German-speaking counterpart, despite the French-speaking population being smaller than the German-speaking. And third, as my own example shows, the percentage of foreign population does not mean that those people do not use German as first language. Even being a Portuguese and not a Swiss citizen, I would consider myself a native speaker of Portuguese and German and use both languages equally. Similar to Grossenbacher (2014: 104), Androutsopoulos (2015) argues that the language practices in social networks are very individualised. Aspects like gender, individuality and conversation topic can also influence in which language one expresses itself.

Furthermore, English is still the *lingua franca* as was showed by (Danet & Herring, 2007: 3; Hong et al., 2011) on Twitter and the chosen words on the keyword list used in here could have also had their impact on the number of collected tweets. For instance, the list of austriazisms based on protocol nr. 10 (Europäische Kommission, 2010) did only generate a few entries in the DB and almost none fulfilled the requirements imposed by the χ^2-test. Therefore, in a future study one could try to recreate the same approach used in this thesis with other words.

Besides Germany, Austria and Switzerland, there are further nations from where tweets have been posted. The United States are by far the most active nation with more

than 14’000 tweets, followed by Brazil with about 4’000. As mentioned by Durrell (2006) there are significant numbers of German speakers outside Europe, especially in the United States, Canada, Namibia and some South American countries. The findings of this work have shown, that this behaviour is partially visible on Twitter. Other nations like Argentina or Canada rank amongst the 22 most active nations for German tweets. Other regions where German is being spoken like Belgium and the Netherlands are also amongst the most active regions. However, we have to keep in mind that the interpretation of these results is somehow arguable since the user does not have to be a native but can also be a tourist or a German speaker living there. One possibility to discern between these various groups could be the location information on the user profiles. But, as Hecht et al. (2011) noted, 34% of all users do not provide real geographical information and those who do only provide information on a city-level.

Figure 6.1 below depicts the geographical concentration of the collected tweets within the study area. Further, major cities are also depicted by red squares. As was explained chapter 3.2.1 most of the activity was concentrated in the urban and sub-urban areas. The distribution of the collected German tweets reflects the same behaviour that Smith and Brenner (2012) found for the US population. It is not entirely surprising that a similar behaviour was found for the German tweets. According to numbers published by the World Bank (2015) and the national statistical offices, e.g. Swiss Federal Statistical Office (2015), the majority of the population of Germany, Austria and Switzerland lives and works in cities and metropolitan areas. This makes it statistically very probable that the number of messages being posted from such areas is significantly higher.

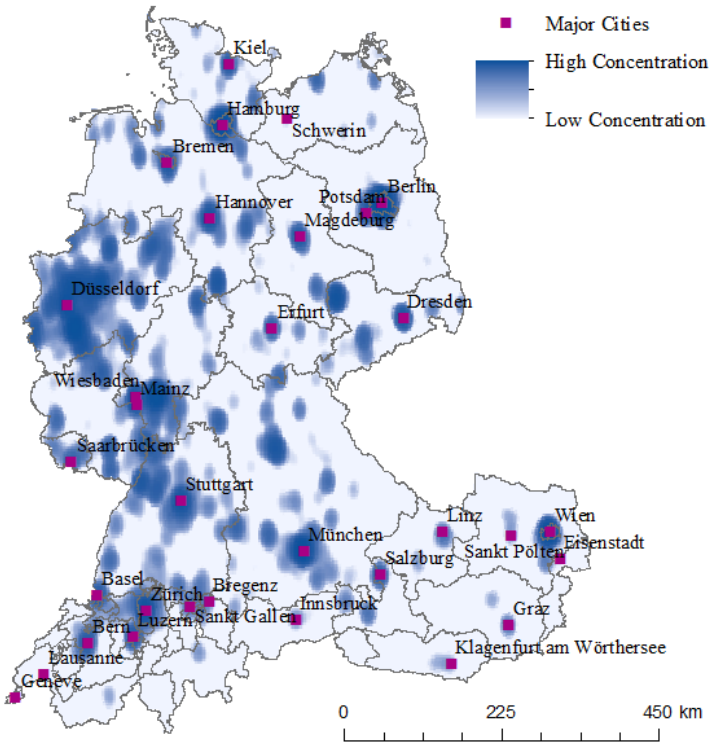

Figure 6.1: Concentration level of tweets in comparison to major urban areas.

The discussion of the insights gained during the process and the geographical distribution have enabled us to address the first of the specified research questions at this point:

> *1. Are German Tweets suitable for linguistic and geographical analysis?*

The simple answer to this question is: Yes, they are. Despite their occurrence being by far lower than with tweets written in English, as was showed by Hong et al. (2011), we were capable of collecting almost half a million tweets over a time period of 163 days. Further, we have to consider that all the collected tweets were georeferenced. This reduces the amount of available posts that can be streamed from Twitter, however, on the other hand it preserves a high level of geographical details according to Mocanu et al. (2013). Furthermore, we discovered that even by using a keyword list and defining the language of the tweets we end up with very noisy data. According to Han and Baldwin (2011) this issue is inherent to Twitter and can hamper a fully- or semi-automated approach like the one followed in this study. Nevertheless, despite all

the issues this study has demonstrated that German tweets can still be used for geographical and linguistic analysis. More sophisticated algorithms, as well as the incorporation of a part of speech analysis could further improve the approach presented here and the obtained results.

6.2.2 χ^2-Test

Some of the calculated χ^2 values were presented in chapter 5.2.3 and a complete list with the results for the 209 lemmas can be found in the appendix. The interpretation of the calculated values is not that simple since the results are very ambivalent. From the 209 lemmas that were analysed, 10 did in the end not fulfil all the requirements imposed by the χ^2-test and can therefore not be taken into consideration. From the remaining 199 lemmas, 87 did generate a χ^2 value smaller than 5.99, i.e. there is no reason to reject our working hypothesis. The remaining 112 caused a rejection of the hypothesis H_0 implying that the observed distribution of these 112 lemmas differed from the one we expected.

The first thing that we notice is that the prepositions *mit*, *auf*, *zum* and *mehr* which were the most frequent lemmas, caused the rejection of H_0, meaning that they were not used equally over the three nations. Nevertheless, this is not very expressive. The differences in the respective varieties lies mainly in the use of them, i.e. a syntactical difference, and would require a part of speech analysis. It is one possibility that the differences between the observed and expected numbers derive from the differences in the use of them. However, in this study it is not possible to draw any conclusion in the differences between the three national centres discussed here. But as Jurafsky et al. (2002) demonstrated for the English language by using the function words *to*, *that*, *of*, and *you*, such words can help to better understand the complex relationship between syntactically and semantically definition of other lemmas.

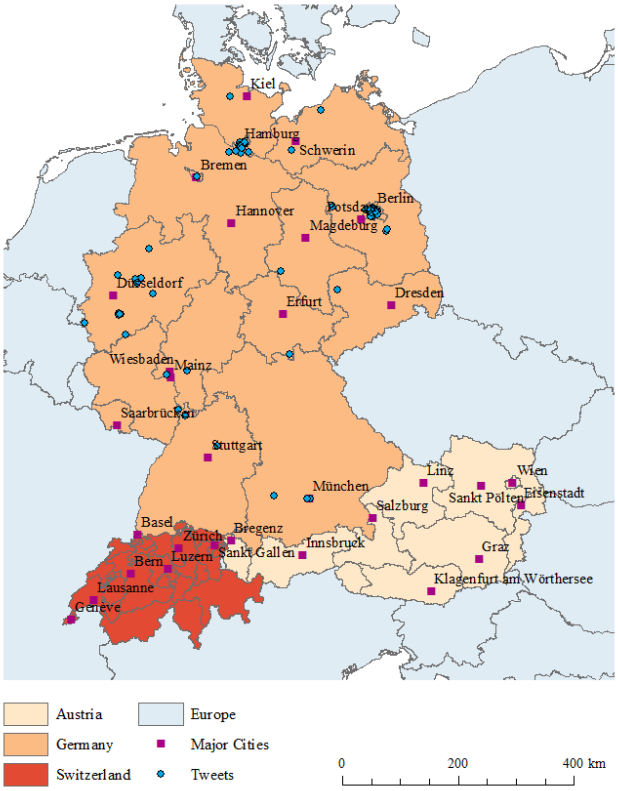

Figure 6.2: Density map of the geographical occurrences of the lemma Kiez.

Nevertheless, not all the results that lead to a rejection of H_0 are so intriguing. Some have met our expectations and proved that in social media the dispersion of a lemma variety can be observed. The example of the German word for foot has already been given in chapter 5. But there are more such lemmas. For instance, *Kiez* is used in the north-east of Germany as synonym for *Quartier*, *Viertel* or *Stadtteil* (neighbourhood, district) while in the remaining regions of Germany, as well as Austria and Switzerland the latter terms are used. This behaviour is endorsed by our analysis. Neither from Switzerland, nor from Austria did anyone post a message containing the word *Kiez*. All the tweets were posted from inside Germany (table 6.1). Further, as the figure above depicts, despite the origins being distributed all over Germany, the areas with most activity lie in the north. And with the hot-spots Hamburg and Berlin the observations met the expectations.

Lemma	DE_obs	AT_obs	CH_obs	Total	DE_exp	AT_exp	CH_exp	χ2
Kiez	176	0	0	176	155.28	10.87	9.85	23.49

Table 6.1: χ2 calculations for the lemma Kiez.

Another good example that demonstrates exemplary the variation of lemmas are the German words for holidays *Urlaub* and *Ferien*. As shortly mentioned in chapter 5, both terms mean the same, however, they are used in different circumstances. While in Germany and Austria *Urlaub* is used as the common word for holidays, *Ferien* is only used in the sense of school holidays. In Switzerland *Ferien* is the common word for holidays and *Urlaub* is used rarer and only in the sense of leave of absence (Duden, 2013a, 2013c). This behaviour can be depicted by the collected tweets that contain these lemmas.

Lemma	DE_obs	AT_obs	CH_obs	Total	DE_exp	AT_exp	CH_exp	χ2
Urlaub	1828	126	47	2001	1765.41	123.55	112.04	40.02
Ferien	545	36	67	648	571.70	40.01	36.28	27.66

Table 6.2: χ2 calculations for the lemma Urlaub and Ferien.

As we can see in table 6.1 above and graphically in figures 6.3 and 6.4 below, the term *Urlaub* is clearly underrepresented in Switzerland. While the observed and the expected numbers for Austria and Germany coincide the same cannot be said about Switzerland. Less than half of the expected tweets were indeed sent from Switzerland. However, when we look at the term *Ferien* the exact opposite is the case. This term occurs less frequent in Austria and Germany and is therefore slightly underrepresented. In Switzerland the expected values is exceeded by almost 100%.

Like *Kiez*, or *Ferien* and *Urlaub*, there are other lemmas where the results met the expectations. For example *parken* (to park) is also clearly underrepresented in Switzerland, since they use the word *parkieren*. Interesting is, however, that Austria is also underrepresented despite using the same lemma as Germany. Other examples are *Pfannkuchen* (pancake), *Flur* (corridor) or *Bürgermeister* (mayor). Such a clear result is also often exhibited by lemmas related to politics or sports, like the previous *Bürgermeister*, but also *Bundesregierung* (federal government), *Endspiel* (final) or *Spieltag* (matchday).

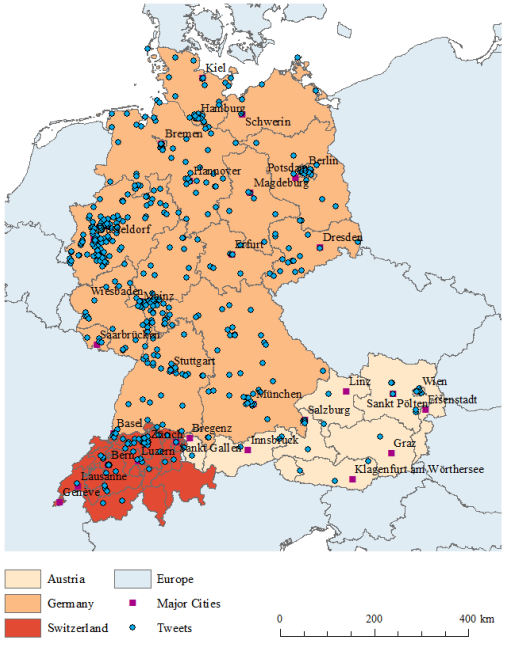

Figure 6.3: Origin of posts containing Ferien.

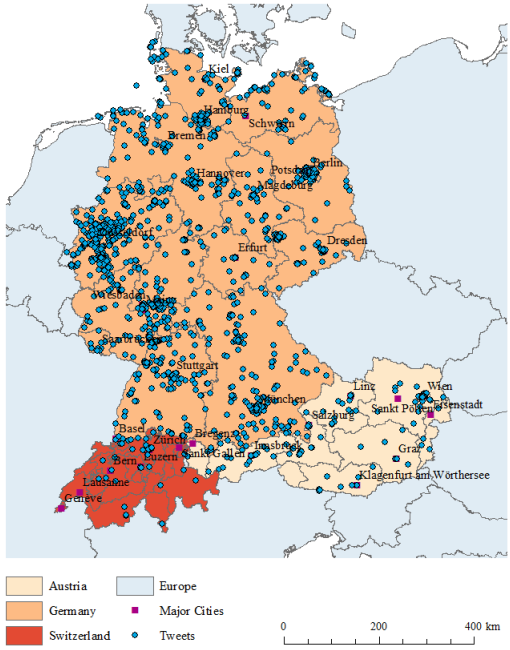

Figure 6.4: Origin of posts containing Urlaub.

Nevertheless, not all rejected lemmas are that obvious. For instance, *Bericht* (report), *Bildung* (education) or *drücken* (push) led to a rejection of H_0 despite being used similarly in the literature in all the three centres.

Until now we had a look at those lemmas whose analysis caused the working hypothesis to be rejected. Now we will discuss those 87 that gave us no reason to reject H_0. Like *laufen* (to walk) in chapter 5, there are other words that are used equally over the study area, e.g. *Gebäude* (building), *schlagen* (to hit) or *Tomaten* (tomatoes). However, the latter one is a bit curious, in Austria there is the synonym *Paradeiser*, nevertheless, there is no reason to refute the assumption that tomato is used with a different frequency in Austria than it is in Switzerland and Germany. Unfortunately, *Paradeiser* did not appear often enough in the DB to enable a comparison.

Nevertheless, some of the lemmas which gave us no reason to reject H_0 were expected to be otherwise. For example, all the lemmas that contained an <ß> should be underrepresented in Switzerland. This has proven to be true with the singular form foot, i.e. *Fuss* respective *Fuß*. However, this is not the case with the plural form *Füße* and *Füßen*. In Switzerland their occurrence is slightly lower as expected but since the calculated χ^2 value is smaller than 5.991 there is no statistical evidence to reject the hypothesis H_0 with an α of 0.05, i.e. with an error probability of 5%. However, this

single example does not enable us to draw any conclusions. Only the examples given above contained an <ß> and fulfilled all the necessary parameters. We would need more lemmas with these two forms to make a comparison and to draw a more definite conclusion.

The insights gained through the results of the χ^2-test and the discussion above allow us at this point to give a comprehensive answer to the following research question:

> *2. Is the use of helvetisms, austriazisms and teutonisms also evident in Twitter as it is in Literature?*

Despite the answer to this question being rather affirmative, some caution is advised. The examples discussed with the lemmas *Kiez*, *Urlab* and *Ferien* have demonstrated that Twitter can indeed be used to analyse regional lemma variation. However, *Ferien* and *Urlaub* are one of the few pairs that fulfilled all the requirements imposed by the statistical test and that could be compared to each other. The majority of the pairs that were defined in our keyword list did not generate enough entries to be compared to each other and, therefore, only single lemmas could be analysed. The occurrence of certain lemmas depends strongly on the nature of the word, e.g. if it is a word like *Faschiertes* (minced meat) the probability to find a message containing such a lemma is very small. On the other side, if you use for example words that have its origins in sports or politics like *Endspiel* (final), the probability to find such a message is relative high. On the other hand some lemmas are very ambiguous, as we will discuss more detailed later, and cannot be compared to its pair without any further analysis, as it was the case with the lemma *Viertel* that could not be compared to *Kiez*. These issues could be addressed by incorporating a part of speech analysis and by collecting data over a longer period of time than it was done for this thesis. Despite these issues, the obtained results indicate that Twitter is an appropriate medium to analyse regional lemma variation.

At this point we can also answer the next research question that was defined in the beginning of this paper:

> *3. Is the difference of the use of <ß> and <ss> visible on Twitter?*

Compared to the other rather general research questions analysed by this study this question is very precise and might surprise to have been included. However, since the difference lies in the use of a distinct character, this issue could also be addressed in

this study. Unfortunately, we have to assert that this question cannot be fully answered. As discussed in section 6.2.2, only the variations of foot, i.e. *Fuss* and *Fuß* – and its plural form – fulfilled all the requirements imposed by the χ^2-test. While for the singular the difference is clearly visible with *Fuss* being overrepresented in Switzerland while being underrepresented in Austria and Germany, and the exact opposite with *Fuß*, the same cannot be said about the plural. The collected data gives us no reason to assume that there is any difference of its use between all the three centres. We would require more examples of this kind of variation to be capable of answering this question.

6.3 Issues with Twitter as an Object of Studies

Earlier in this thesis (chapter 3.2.5) we looked at issues that need some attention when conducting a research based on Twitter, especially when focusing on linguistic. Since the described aspects are inherent to this research medium, they also affected this thesis. To explain the implications we will discuss two aspects that were important for this study as well as how we could approach them in an oncoming study.

- As mentioned by Crystal (2011: 41), a tweet can contain terms in other languages, i.e. German tweets can also contain English words and otherwise. For this purpose, we defined a keyword list with German words and specified in the algorithm to only look for posts that were recognized as German by the language detection system of Twitter (Twitter Developers, 2015b). However, both this methods are not faultless. For instance, the word *Mais* can be understand as maize and noise, but it is also the Portuguese word for more and the French word for but. This lead to the curious situation that the vast majority of the tweets containing this lemma were written in Portuguese and not in German as specified by our API parameters. Interesting was also the fact that only a few French tweets could be found in the DB, despite *mais* being a very common lemma.

- As discussed earlier orthography is an important aspect on Twitter. Emoticons, abbreviations and character omissions can difficult an analysis of the message content (Laboreiro et al., 2010). Some examples could also be found in the collected data. For instance, one user wrote *"…auffe couch…"* instead of *"…auf die*

Couch…". Another wrote *daß*, which in Germany must also be written as *dass*. There were also cases, where people used numbers instead of words, e.g. *2beiner* (biped) or *2monate* (two months). Emoticons could also be found in the collected data. However, to facilitate the processing of the data, all the special characters like [,], (,) or /, were deleted at the beginning. Furthermore, the distinction between deliberateness and mistake is not always obvious. Two users used the word *Seeen*, which probably was a mistake, since the German plural for sea is *Seen*, written with only two <e>. Five users used the word *Spaaaaß* instead of *Spaß* (fun). However, this time the misspelling was probably used intentionally to emphasize the amount of fun, as can be confirmed by the tweet depicted by figure 6.5. Nevertheless, to facilitate the processing of the collected data such posts were removed, including those from the examples given above.

Figure 6.5: Use of misspelled words on Twitter.

These two aspects were of particular importance for this study and needed to be addressed during this thesis. The implications of the other aspects mentioned in chapter 3.2.5 on this work can be seen as reduced. The data was collected over a period of 163 days. The program used to collect our data ran 24 hours a day, except the 4 days where it was launched or stopped. With this approach the influence of catch-phrases and current events on the dataset, as mentioned by Crystal (2011) and Zappavigna (2012: 19), was minimized. Further, since we only looked at lemmas, it was irrelevant for the analysis if our dataset contained truncated tweets. However, if in a later study one would like to incorporate a part of speech analysis, it would be recommended to follow the idea of Zappavigna (2012: 21) and to remove them from the data. Finally, we did not analyse the issue of retweets in this study. Based on the established mark-up RT, that stands for retweet, a simple SQL query returned that only 247 tweets in the DB could be considered retweets (Kwak et al., 2010). The small number of occurrences and the missing information of the origin of those posts, did not justify a

comprehensive analysis and we decided to keep them in the dataset. However, for an extensive linguistic analysis the retweets would require more attention. As Crystal (2011) and Zappavigna (2012: 19) note this phenomenon is very popular on Twitter, but very seldom in the spoken language. Further, since a German can retweet a post from Switzerland or Austria, and vice versa, this will influence the distribution of the collected lemmas. In his previously unreleased thesis, Mändli (2015) focused on methods to analyse Twitter data. Thereby he also addressed the issue of the retweets. The insights gained in his study help in the future to address this subject.

6.4 Working with Tokens on Twitter

The previous section discussed some issues that are inherent to Twitter as a research medium and that we faced during this study. Now we will discuss some of the problems faced on the lemma level, i.e. on the tokens used in this work. Some of these issues might have been already mentioned shortly. For the sake of completeness they are mentioned again in this section.

As Nerius (2007: 18) mentions in his work, language itself is not static but rather dynamic. This is especially true for language on a medium like the Internet and in particular on Twitter. Space restriction and actual events can induce new trends that often are accompanied with new catchphrases, Internet memes or even new and modified words (Zappavigna, 2012: 100ff.). These are not the only aspects that make tokenization and automatic processing of Twitter data problematic. Laboreiro and his colleagues (2010) discussed in their work the most common issues when running some automatic processing over microblogging services such as Twitter. Some of the issues they discussed, can also be found in the collected dataset in this work. Due to the processing steps implemented in here mainly two aspects from Laboreiro et al. (2010) are of concern. First, as a part of modern communication emoticons can be found in the data. However, since we focused on the extraction of the defined lemmas, all the special characters, except # and @, were removed to facilitate an automatic processing. This affected mainly all emoticons and Uniform Resource Locators (URLs). And second, spelling errors are very abundant in Twitter messages as it was also demonstrated by Han and Baldwin (2011). In the previous chapter we have presented two examples of users posting messages containing *Seeen* and *Spaaaaß*. While the first can be considered as an error and could therefore be normalised with an approach like

Han and Baldwin (2011) presented, the same cannot be said about the latter. It can be assumed the user wanted thereby to emphasize the meaning of the word implying that it was a deliberated misspelling that cannot be simply removed. These are just two examples that can be found in the dataset amongst others. There are also mixtures between letters and numbers, e.g. *10117mittagspause* or *3meter5*. To facilitate an automated processing of the dataset such words were removed, similar to what Zhao et al. (2011) did. But we have to keep in mind that for a linguistic analysis such stylistic variations could be interesting but would require a more differentiated treatment.

The previous paragraphs overlap strongly with the last point mentioned in 6.3. Nevertheless, this aspect cannot be emphasized enough since it can affect the collected data as well as the final results. But there are more issues that need attention when working with lemmas. The following three subchapters will address the main points that were also of importance for this work. The awareness of them could help to avoid some issues or to improve the way of collection and handling the data.

6.4.1 Ambiguity

The points mentioned above are more of semantical or grammatical nature. Despite making an automatic processing of the collected tweets rather difficult and problematic, there is another very important issue when working with lemmas: ambiguity. As mentioned by MacDonald et al. (1994) language can be ambiguous at any given point in a sentence and this on many levels, i.e. lexical, grammatical or phonological amongst others. For instance, in English the word *watch* can either mean an object that indicates time, or a verb which is synonymous with *to observe*. This word shows not only an ambiguity on lexical level but also on a grammatical by being either a noun or a verb, depending on the context (MacDonald et al., 1994). Despite this example being for the English language the same holds also true for German. For instance, the word *Leiter* can either stand for a ladder or a manager. *Wirtschaft* can stand for the economy or for a pub or an inn. The word *Tor* can either mean a gate, a goal in football or a fool. *Unternehmen* can either be used as a noun, meaning a corporation, or as the verb *to undertake*. All these examples could be found in the collected data. These ambiguity issues make it also difficult for us to interpret the calculated χ^2 values since only one of the meaning was intended and no part of speech analysis was carried out.

Nevertheless, this issue arose already in an early stage of this work, when creating the track used for Twitter (see chapter 4.2). DeReWo only lists single words which can be problematic in case of ambiguity. The word *Leiter* can once more be used as an example. While it is used similar in all three centres as a ladder is shows a variation when used as manager. The same issue persists when analysing the occurrences after the data collection. Since both varieties are downloaded a simple algorithm is not enough to extract and divide them correctly. To be capable of extracting the correct meaning of the word one would require a more complex algorithm with an incorporated part of speech recognition tool.

6.4.2 Twitter API and Tokens

Another issue lies in the Twitter API. As mentioned in chapter 3.2.2 Twitter allows you to implement a keyword list (Twitter Developers, 2015b). One of the advantages lies in the flexibility of the used algorithm by Twitter. For instance, if you define the word *Glück* (luck), the algorithm used by Twitter will also download tweets that contain the word *Glücksgefühl* (feeling of happiness). The word *Arbeit* (work) will also trigger the download of messages containing *Arbeitslosigkeit* (unemployment). This allows us to analyse word compounds without having to explicitly include them on the keyword list. This can be very useful since the size of the this list is limited.

Nevertheless, this very same aspect has also a serious disadvantage since it causes a lot of noise in the dataset. A good example that can be found in the collected data is the word *Amt* (department). Besides all the desired word compounds like *Standesamt* (register office) or *amtlich* (official) it has also generated entries based on words like *Amtrak* (railroad service in the US), *zugespamt* (spammed) or *gesamt* (whole, total). The process to detect and eliminate such words can hamper an automated process as it was followed by this thesis and can increase the resources to handle it.

6.4.3 Occurrence of Lemmas

Another problem that one has to consider when working with tokens on Twitter is their frequency. First of all, Germany has more inhabitants and the German population tends to use Twitter more frequent than the Swiss or the Austrians. This aspect can be incorporated in the used statistical test as it was done in here with the χ^2-test. The relative geographical distribution is used to calculate the estimated frequencies of the

lemmas. However, the occurrence of the tokens themselves cannot be influenced. As mentioned in chapter 5.2.1.2 only 411 from the original 567 input lemmas did generate any entry in the databased. 156 tokens did not appear in any of the tweets collected during the observation period, like *Ätti*, a rather old Swiss German word for father. Further, some lemmas and their compounds did only generate one entry, which cannot be used to analyse any geographical distribution, e.g. the Austrian word *Faschiertes* (minced meat). Another issue that has an impact on the occurrences is the requirement imposed by the χ^2-test, which requires an expected frequency of at least 5. This reduces the number of tokens that can be used even further.

All this aspects mentioned in the paragraph above have an implication on the comparability of the used lemmas. Even if similar word compounds are merged to one single class they often do not fulfil the requirements, especially the minimum number of observations, imposed by the test that is used here. For instance, the term *Velo* and *Fahrrad* mean both bicycle but the first is used in Switzerland and latter is used in Germany. While *Fahrrad* appears in enough tweets allowing us to conduct an analysis on its distribution, the same cannot be done with *Velo*, which only appeared 33 times. This makes it difficult to conduct a comparison and profound analysis, since not every lemma and its counterpart appeared in enough tweets during the period of time when data was being collected.

6.5 Lemma Accommodation

Based on the Accommodation theory by Giles et al. (1991) we also wanted to access the appropriateness of Twitter to study language and lemma accommodation for the German language on Twitter in this study. Like Tamburrini et al. (2015) who analysed how users on Twitter changed word usage according to their conversation-partner we wanted to analyse if users changed their lemma usage according to the geographical affiliation of their followers. However, the approach followed in this study did not promote such an analysis as was already mentioned at the end of chapter 6.2.2. Our data collection focused primarily on lemmas and, therefore, on the message itself and not on the users.

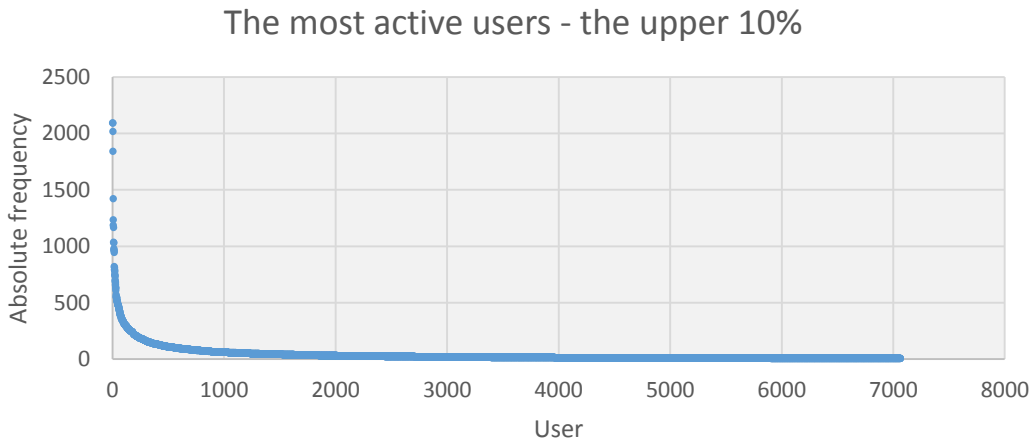

Figure 6.6: The 10% most active users and the number of posted tweets.

As it is depicted by figure 6.5 only a small proportion of the users has posted more than one message. The 10% most active users from the 70'644 users in the DB are responsible for 72.9% of the collected tweets. This implies that the vast majority of the users has only posted one tweet which is not enough to analyse accommodation on social media. Further, for this study we only collected data over a period of 163 days, which probably is not enough to conclude if there has been an accommodation by the users. Finally, the terms of the Twitter APIs are more restrictive concerning requests on the user level, which was another reason why in this thesis we focused mainly on lemmas.

> *4. Is there an accommodation on the level of lemmas visible between users on Twitter?*

Therefore, unfortunately, the research question above remains unanswered. This study followed rather the same approach as Gonçalves and Sánchez (2014) in their paper but instead of focusing on dialects we focused on lemma variations. Nevertheless, the insights gained by this study could help oncoming studies by identifying interesting cases, i.e. lemmas or users, on which to base further research. By adapting the keyword list it will possible to ascertain if there is also a communication accommodation for the German language on the micro-blogging service Twitter. Further, the time taken to collect data would need to be increased. For instance, Gonçalves and Sánchez (2014) collected georeferenced tweets over the course of two years in comparison to

the 163 days here. We are confident that by changing these parameters, we can also study accommodation for the German language on Twitter.

6.6 The Research Questions – A Synopsis

In the beginning of this study, i.e. chapter 1.3, we defined some research questions that would be addressed by this thesis. These questions have been discussed in this chapter 6. At this point we want to provide a short synopsis concerning the research questions.

Unfortunately, not all the questions imposed at the beginning could be fully answered. The first two could be answered with a certain level of reliability. The issue of <ß> and <ss> could not be completely answered, but the results indicate that its use can also be analysed on Twitter if more data was available. The question focusing on lemma accommodation could not be answered. Nevertheless, we think that this study has made its contribution to the state of the art of how analysing lemma variation on Twitter. It addressed an issue, i.e. geographical language variation, which according to Gonçalves and Sánchez (2014) is still poorly understood and, to our knowledge, it is the first paper analysing lemma variation of German tweets. We proposed a semi-automated approach that allows us to collect data suitable for an analysis of lemma variation. Due to its explorative nature the approach followed here identified aspects were problems might occur, helping further research to avoid the same problems. And through the gained insights recommendations can be given that would improve the results of upcoming research on this subject.

7. Conclusion

As mentioned in the beginning of this thesis, the idea was to conduct an explorative study on lemma variation on Twitter for the German language. Only few previous research has in some way dealt with this issue, and mainly for the English language. English remains the lingua franca on the Internet and it is no surprise that most of the conducted linguistic analysis focused on the English language. Nevertheless, studies based on other languages are equally important to broaden our understanding of how we communicate with each other across borders. The work presented here tried to contribute to this wider understanding by exploring the properties, possibilities and limitations of Twitter and access its appropriateness for studies concerning lemma variation and accommodation for the German Language.

7.1 Achievements

During the work process underlying this thesis valuable achievements were obtained that are worth to be mentioned. The following list notes the most important:

- Through a comprehensive literature review this work was set in the context of other studies that were dedicated to geography and linguistics. It was demonstrated that language is often strongly connected to space and should not be analysed separately on social media platforms, and vice versa.

- As an interdisciplinary thesis this study gives non-linguists as well as non-geographers a simple but comprehensive overview of the unfamiliar research area and how these fields can be combined to study lemma variation on a micro-blogging service.

- We presented an approach of how to collect data which could be used to conduct a study about German lemma variation on Twitter.

- We confirmed that Twitter is a very constant source of real time data. A simple access method to the available public data is a further aspect which makes it a very appealing medium to conduct linguistic or geographical research.

- The implementation of a simple geographical concept helped to remove a considerable part of tweets that were generated by automated and semi-automated programs from our data set.

- The raw dataset was adjusted to our needs and data that did not fit the purpose of this study was removed.

- The anticipated geographical distribution, with most of the tweets originating from urban areas, could also be confirmed for the German language in this thesis.

- This study demonstrates that working alone with lemmas, i.e. tokens, can be problematic on multiple levels. First, the words can be ambiguous. Second, the search algorithm of Twitter will also consider compounds based on the input token. Despite this being a useful feature it will also generate a lot of undesired data.

- The implementation of the χ^2-test enabled us to analyse lemma variation over the three nations of the study area. The comparison between the observed and expected values helped to determine where the lemmas were over- or under-represented.

- The obtained results and gained insights during the entire process were discussed to allow a general interpretation. Thereby, aspects were mentioned and highlighted where this study could be improved and other studies could be based on.

7.2 Insights

Many studies have been undertaken that were based on Twitter. However, most of the conducted research focused on the English language since it is the lingua franca on the Internet and social media platforms, and only few were dedicated to the subject of language variation over space. This work contributed to fill this research gap by analysing lemma variation over the three national centres of the German language. The analysed data demonstrated that lemma variation can be also studied on Twitter and is not only restricted to literature, news media or schoolbooks. As we have seen language variety can manifest itself on various linguistic aspects, with the lexical and

phonological being the most common. We had to limit ourselves to the lexical aspects due to the medial nature of Twitter, ignoring therefore all the other aspects where language varieties might differ. The selected keywords, ambiguity, tokenization and other issues related to the access and processing methods implemented in this work have proven to be complex aspects. Without the implementation of a system that enables us to analyse the content of a tweet some uncertainties remain. Nevertheless, this thesis confirmed that the variations of the German language can also be analysed on Twitter, which could motivate further research in this subject.

Twitter has proven to be an interesting data source for data retrieval due to its accessibility and the real-time character of the collected data. However, the data is remarkably noisy which complicates every automated approach. This is partially due to the Twitter API, which offers a certain amount of flexibility by considering not only the exact lemmas in the keyword list but also other words that contain the same character order. This enabled us to consider word compounds but increased remarkably the amount of undesired data in the dataset. Therefore, data cleaning becomes a vital aspect of any analysis that is based on Twitter. More sophisticated methods as well as an implementation of a part of speech system could help to reduce the amount of data in the collected dataset and are essential to obtain valid results.

We pursued an automated approach that would allow us to process a great amount of data. But like other studies that followed a similar approach, we had to conclude that it is a complicated process. Language is a very dynamic aspect of our lives and the use of emoticons and unintentional or intentional misspellings make it difficult to follow an automated approach. An automating is further hampered by the special characters of the German language, i.e. <ä>, <ö>, <ü> and <ß>. Existing algorithms had to be adjusted to enable an analysis of the German language during this study. We demonstrated that a full automated approach is not the most adequate. Automated processing steps can help to handle a great amount of data as we have demonstrated in this thesis but a manual analysis is indispensable to study the multifaceted subject of language. Even with complex, self-learning algorithms language remains a very social phenomenon, making it impossible to look at it only with the help of computers.

7.3 Future Work

We saw the potential of Twitter to study lemma variation and accommodation for the German language, but there is no doubt that further research is necessary. Furthermore, this work could not answer all the questions posed by itself, and some issues are still unanswered and would require further efforts. We did not apply a part of speech analysis to our data. This could improve the data cleaning and therefore our results. Further, we did not look at the accommodation aspect at all. Data was only collected over 163 days which is a short period to analyse the accommodation phenomenon. However, based on the data gathered in this work one could try to determine interesting candidates, i.e. users, and analyse them further. This could help us to broaden our understanding of how the German language varies on the Internet. This thesis tried to make its contribution to this interesting research field and the insights gained could help other research to avoid certain issues. There are three particular aspects that could be of importance for oncoming research.

First, the lemmas used in the track need to be chosen differently. Some of the words did not generate any entry in the DB at all, making it difficult to compare it to its counterpart. Another problem is imposed by prepositions. Prepositions are some of the most frequent words in any language and by including them in the track they will generate an enormous number of entries in the DB. However, most of the prepositions used in the track are valid in the three centres analysed in here. The differences are only apparent through the context, making an automatic processing more complex and more time consuming. One possibility to overcome this problem is by using only substantives.

Second, another problem that has to be overcome is the number of entries for a certain lemma. The statistical test used in here requires an absolute frequency of at least 5 per class, inducing that each lemma has to appear at least 100 times in the study area. This is often a problem with specific words like the Austrian *Faschiertes* (minced meat). Further, there is the problem of the possibility that someone is posting about minced meat is rather low. This problem could partly be solved by collecting data over a longer period of time. It is obvious that the 163 days used for this work are not enough.

Third, since the focus lies on the German Language it is impossible to try to reproduce the results presented in here with data from other nations, where Twitter is

used more frequently. In an oncoming study one could try to conduct the same kind of research based on a different platform like Stähli et al. (2011) did with the SMS for Switzerland. For a German wide analysis one could consider platforms like Facebook or WhatsApp[16]. Nevertheless, other issues like privacy concerns and data accessibility could arise. Twitter on the other hand, offers numerous ways of accessing data of users that publicly post their tweets, which makes it very inviting as a research medium.

All these points mentioned above are recommendations that could improve results or help other researchers to avoid some issues that were faced here. However, as mentioned by Zappavigna (2012: 24) it is difficult to give any recommendation on how to collect data on Twitter since the access policies can always change and any recommendation could be outdated by the time. Further, the Internet is an ever-changing medium and like Facebook in 2004 and Twitter in 2006 other platforms could emerge and change the way of how we communicate with each other (Facebook, 2015; Twitter, 2014).

This study has shown that Twitter is a very interesting medium that facilitates the access to a huge amount of real time data from various regions, cultures and is written in different languages. The main intention of this work was to explore the possibilities offered by Twitter to study lemma variation over the national centre of the German language. Based on gained data and insights, future studies could be designed. For instance, like Dürscheid and Frick (2014) who compared SMS and WhatsApp, one could also include Twitter in this comparison. Like SMS, also Twitter has a length restriction and it would be interesting to see if both mediums have something in common or if they differ completely.

All the aspects that were mentioned in this thesis, the possibilities and the limitations, as well as the real-time nature of the tweets and the challenges imposed by the analysis of tweets and users, make this field of research so interesting and worthy for further investigations.

[16] www.whatsapp.com

Bibliography

ALEXA. (2014). The top 500 sites on the web. Retrieved February 17, 2015, from http://www.alexa.com/topsites

Ammon, U., Bickel, H., Ebner, J., Esterhammer, R., Gasser, M., Hofer, L., … others. (2004). *Variantenwörterbuch des Deutschen. Die deutsche Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. Berlin/New York: De Gruyter.

Androutsopoulos, J. (2006). Language Choice and Code Switching in German-Based Diasporic Web Forums, 340–361.

Androutsopoulos, J. (2013). Code-switching in computer-mediated communication. In S. Herring, D. Stein, & T. Virtanen (Eds.), *Pragmatics of computer-mediated communication* (pp. 667–694). Berlin: De Gruyter.

Androutsopoulos, J. (2015). Networked multilingualism: Some language practices on Facebook and their implications. *International Journal of Bilingualism*, *19*(2), 185–205.

Archambault, A., & Grudin, J. (2012). A longitudinal study of facebook, linkedin, & twitter use. *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems - CHI '12*, 2741. doi:10.1145/2207676.2208671

Baboolall, D., Hammond, R., Joseph, K., & Todhunter, I. (2013). Mapping Twittsburgh : Visualizing Twitter Data & Neighborhood Demographics.

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation. *Journal of Sociolinguistics*, *18*(2), 135–160.

Bergsma, S., Mcnamee, P., Bagdouri, M., Fink, C., & Wilson, T. (2012). Language Identification for Creating Language-Specific Twitter Collections. In *Proceedings of the second workshop on language in social media* (pp. 65–74). Association for Computational Linguistics.

Bickel, H., & Landolt, C. (2012). *Schweizerhochdeutsch: Wörterbuch der Standardsprache in der deutschen Schweiz*. Mannheim/Zürich: Dudenverlag.

Borau, K., Ullrich, C., Feng, J., & Shen, R. (2009). Microblogging for language learning: Using twitter to train communicative and cultural competence. In *Lecture Notes in Computer Science (including*

subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). doi:10.1007/978-3-642-03426-8_10

Bouillot, F., Poncelet, P., & Roche, M. (2012). How and why exploit tweet ' s location information ? In J. Gensel, D. Josselin, & D. Vandenbroucke (Eds.), *AGILE'2012: 15th International Conference on Geographic Information Science* (pp. 24–27). Avignon.

Bundeskanzleramt. (2015). Bundes-Verfassungsgesetz. Retrieved February 19, 2015, from https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10000138

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating Gender on Twitter, 1301–1309.

Bussmann, H. (1996). *Routledge Dictionary of Language and Linguistics*. New York: Routledge.

Chen, P. P.-S. (1976). The Entity-Relationship Model - Toward a Unified View of Data. *ACM Transactions on Database Systems*, *1*(1), 9–36. doi:10.1145/320434.320440

Cheng, Z., Caverlee, J., & Lee, K. (2010). You Are Where You Tweet : A Content-Based Approach to Geo-locating Twitter Users. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 759–768. doi:10.1145/1871437.1871535

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 811–824. doi:10.1109/TDSC.2012.75

Clyne, M. G. (1992). *Pluricentric Languages: Differing Norms in Different Nations*. (M. G. Clyne, Ed.) (Contributi.). Berlin: De Gruyter.

Clyne, M. G. (1995). *The German language in a changing Europe*. Cambridge: Cambridge University Press.

Crystal, D. (2001). *Language and the Internet*. Cambridge: Cambridge University Press.

Crystal, D. (2011). *Internet Linguistics: A Student Guide*. Abingdon: Routledge.

Danescu-Niculescu-Mizil, C., Gamon, M., & Dumais, S. (2011). Mark My Words ! Linguistic Style Accommodation in Social Media. In *Proceedings of the 20th international conference on World wide web* (pp. 745–754). Hyderabad, India: ACM.

Danet, B., & Herring, S. C. (2007). Introduction: Welcome to the Multilingual Internet. In B. Danet & S. C. Herring (Eds.), *The multilingual Internet: Language, culture, and communication online* (pp. 3–39). New York: Oxford University Press.

Devore, J., Farnum, N., & Doi, J. (2014). *Applied Statistics for Engineers and Scientists* (Third Edit.). Stamford, US: Cengage Learning.

Duden. (2013a). Ferien, die. Retrieved June 20, 2015, from http://www.duden.de/rechtschreibung/Ferien

Duden. (2013b). ss und ß. Retrieved June 28, 2015, from http://www.duden.de/sprachwissen/rechtschreibregeln/doppel-s-und-scharfes-s

Duden. (2013c). Urlaub, der. Retrieved June 20, 2015, from http://www.duden.de/rechtschreibung/Urlaub

Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., & Madden, M. (2015). Demographics of Key Social Networking Platforms. Retrieved February 17, 2015, from http://www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms-2/

Durrell, M. (2006). German. In K. Brown (Ed.), *Encyclopedia of Language & Linguistics (Second Edition)* (Second Edi., pp. 42–45). Oxford: Elsevier. doi:http://dx.doi.org/10.1016/B0-08-044854-2/02197-0

Dürscheid, C. (2003). Medienkommunikation im Kontinuum von Mündlichkeit und Schriftlichkeit. Theoretische und empirische Probleme. *Zeitschrift Für Angewandte Linguistik.*, *38*, 37–56.

Dürscheid, C., & Frick, K. (2014). Keyboard-to-Screen-Kommunikation gestern und heute: SMS und WhatsApp im Vergleich. In A. Mathias, J. Runkehl, & T. Siever (Eds.), *Sprachen? Vielfalt! Sprache und Kommunikation in der Gesellschaft und den Medien. Eine Online-Festschrift zum Jubiläum für Peter Schlobinski (= Networx 64)* (pp. 149–181). Hannover. Retrieved from www.mediensprache.net/networx/

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 1277–1287). Association for Computational Linguistics.

Eleta, I., & Golbeck, J. (2014). Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior*, *41*, 424–432. doi:10.1016/j.chb.2014.05.005

Elpelt, B., & Hartung, J. (2004). *Grundkurs Statistik : Lehr- und Übungsbuch der angewandten Statistik* (Third Edit.). München: Oldenbourg.

Esri. (2014a). Albers Equal Area Conic. Retrieved March 8, 2015, from http://resources.arcgis.com/en/help/main/10.2/index.html#//003r0000001n000000

Esri. (2014b). Drawing a continuous raster dataset such as an orthophoto. Retrieved March 13, 2015, from http://resources.arcgis.com/en/help/main/10.2/index.html#//009t00000077000000

Esri. (2014c). Kernel Density (Spatial Analyst). Retrieved March 13, 2015, from http://resources.arcgis.com/en/help/main/10.2/index.html#//009z0000000s000000

Europäische Kommission. (2010). Beachtung Österreichischer Ausdrücke. Retrieved February 9, 2015, from http://ec.europa.eu/geninfo/query/index.do?filterNum=10&queryText=österreichische+ausdrücke&summary=summary&more_options_source=global&more_options_date=*&more_options_date_from=&more_options_date_to=&more_options_language=de&more_options_f_form

Facebook. (2015). Über Facebook. Retrieved June 1, 2015, from https://www.facebook.com/facebook/info?tab=page_info

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2014). The Rise of Social Bots. *arXiv Preprint arXiv:1407.5225*, 1–11. Retrieved from http://arxiv.org/abs/1407.5225

Filippova, K. (2012). User Demographics and Language in an Implicit Social Network. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1478–1488). Jeju Island, Korea: Association for Computational Linguistics.

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics* (Fourth Edi.). New York, US: W.W. Norton & Company.

Gianvecchio, S., Xie, M., Wu, Z., & Wang, H. (2008). Measurement and Classification of Humans and Bots in Internet Chat. In *USENIX security symposium* (pp. 155–170).

Giles, H., & Baker, S. C. (2008). Communication Accomodation Theory.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequences. In H. Giles, N. Coupland, & J. Coupland (Eds.), *Contexts of accomodation: Developments in applied sociolinguistics* (pp. 1–68). Cambridge University Press.

Gonçalves, B., & Sánchez, D. (2014). Crowdsourcing Dialect Characterization through Twitter. *arXiv Preprint arXiv:1407.7094.*, 1–10.

Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the world are you ? Geolocation and language identification in Twitter. *The Professional Geographer*, *66*(4), 568–578.

Grier, C., Thomas, K., Paxson, V., & Zhang, M. (2010). @ spam : The Underground on 140 Characters or Less ∗ Categories and Subject Descriptors. *Proceedings of the 17th ACM Conference on Computer and Communications Security*, 27–37. doi:10.1145/1866307.1866311

Grosseck, G., & Holotescu, C. (2008). Can we use Twitter for educational activities? *The 4th International Scientific Conference of eLearning and Software for Education*, (June), 1–8. Retrieved from http://www.cblt.soton.ac.uk/multimedia/PDFsMM09/Can we use twitter for educational activities.pdf

Grossenbacher, T. (2014). *Studying Human Mobility Through Geotagged Social Media Content*. University of Zurich.

Han, B., & Baldwin, T. (2011). Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 368–378). Association for Computational Linguistics.

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber ' s Heart : The Dynamics of the " Location " Field in User Profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 237–246). ACM. doi:10.1145/1978942.1978976

Hong, L., Convertino, G., & Chi, E. H. (2011). Language Matters in Twitter : A Large Scale Study. In *IV International AAAI Conference on Weblogs and Social Media* (pp. 518–521).

Hu, Y., Talamadupula, K., & Kambhampati, S. (2013). Dude , srsly ? : The Surprisingly Formal Nature of Twitter ' s Language.

Huberman, B. A., Romero, D. M., & Wu, F. (2008). Social networks that matter: Twitter under the microscope. Available at SSRN 1313405. doi:10.2139/ssrn.1313405

Ikawa, Y., Enoki, M., & Tatsubori, M. (2012). Location Inference using Microblog Messages (pp. 687–690). Lyon, France.

Institut für Deutsche Sprache Programmbereich Korpuslinguistik. (2013). Korpusbasierte Wortgrundformenliste DeReWo, v-ww-bll-320000g-2012-12-31-1.0, mit Benutzerdokumentation. Mannheim, Deutschland. Retrieved from http://www.ids-mannheim.de/derewo

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why We Twitter : Understanding Microblogging. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (pp. 56–65). doi:10.1145/1348549.1348556

Jurafsky, D., Bell, A., & Girand, C. (2002). The Role of the Lemma in Form Variation. *Laboratory Phonology*, *7*, 3–34.

Koch, P., & Oesterreicher, W. (1985). Sprache der Nähe — Sprache der Distanz. *Romanistisches Jahrbuch*, *36*, 15–43.

Krishnamurthy, B., & Arlitt, M. (2008). A Few Chirps About Twitter. *Proceedings of the First Workshop on Online Social Networks (WOSP '08)*, 19–24. doi:10.1145/1397735.1397741

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a Social Network or a News Media? Categories and Subject Descriptors. In *Proceedings of the 19th international conference on World Wide Web* (pp. 591–600). Raleigh, North Carolina: ACM.

Laboreiro, G., Sarmento, L., Teixeira, J., & Oliveira, E. (2010). Tokenizing micro-blogging messages using a text classification approach. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data* (pp. 81–88). Toronto: ACM. doi:10.1145/1871840.1871853

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). Lexical Nature of Syntactic Ambiguity Resolution. *Psychological Review*, *101*(4), 676–703.

Mändli, U. (2015). *Qualitative und Quantitative Methoden zur Analyse von (Geo)Twitter-Daten*. University of Zurich.

Mislove, A., Lehmann, S., Ahn, Y., Onnela, J., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media* (pp. 554–557). AAAI Press.

Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., & Vespignani, A. (2013). The Twitter of Babel: mapping world languages through microblogging platforms. *PloS One*, *8*(4), e61981. doi:10.1371/journal.pone.0061981

Nerius, D. (2007). *Deutsche Orthographie* (4th ed.). Hildesheim, Germany: OLMS.

O'Connor, B., Smith, N. A., & Xing, E. P. (2010). A Latent Variable Model for Geographic Lexical Variation, (October), 1277–1287.

O'Reilly, T. (2005). What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. Retrieved March 22, 2015, from http://www.oreilly.com/pub/a//web2/archive/what-is-web-20.html

Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREC*, *10*, 1320–1326.

Poblete, B., Garcia, R., Mendoza, M., & Jaimes, A. (2011). Do All Birds Tweet the Same ?: Characterizing Twitter Around the World Categories and Subject Descriptors. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1025–1030). ACM.

PostGIS. (2015). About PostGIS. Retrieved March 16, 2015, from http://postgis.net/

PostgreSQL. (2014). PostgreSQL. Retrieved from http://www.postgresql.org/

Rout, D., Preotiuc-Pietro, D., Bontcheva, K., & Cohn, T. (2013). Where ' s @ wally ? A Classification Approach to Geolocating Users Based on their Social Ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media* (pp. 11–20). ACM.

Scheffler, T. (2014). A German Twitter Snapshot. In N. Calzolari (Conference Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 2284–2289). Reykjavik, Iceland: European Language Resources Association (ELRA).

Schmidlin, R. (2011). *Die Vielfalt des Deutschen: Standard und Variation : Gebrauch, Einschätzung und Kodifizierung einer plurizentrischen Sprache*. Berlin: De Gruyter.

Schweizer Radio und Fernsehen. (2014). # ist das Wort des Jahres 2014. Retrieved June 21, 2015, from http://www.srf.ch/radio-srf-3/highlights/ist-das-wort-des-jahres-2014

Smith, A., & Brenner, J. (2012). *Twitter Use 2012*.

Snowball. (2014). Snowball-Download. Retrieved April 24, 2015, from http://snowball.tartarus.org/index.php

Stähli, A., Dürscheid, C., & Béguelin, M.-J. (2011). sms4science: Korpusdaten, Literaturüberblick und Forschungsfragen. *Themenheft Linguistik Online*, *48*(4), 3–18. Retrieved from http://www.linguistik-online.de/48_11/staehliDuerscheidBeguelin.html

Statistics Austria. (2014). Statistisches Jahrbuch Österreichs. Retrieved February 19, 2015, from http://www.statistik.at/web_en/publications_services/statistisches_jahrbuch/index.html

Steiner, S. (2012). Twitter-Nutzung in der Schweiz – Juni 2012. Retrieved February 17, 2015, from http://alike.ch/twitter-nutzung-in-der-schweiz-juni-2012/

Storrer, H. H. (2009). *Der X^2- Test. Einführung in die mathematische Behandlung der Naturwissenschaften II*. Basel: Birkhäuser.

Swiss Federal Statistical Office. (2012). Languages and religions – Data, indicators. Retrieved February 19, 2015, from http://www.bfs.admin.ch/bfs/portal/en/index/themen/01/05/blank/key/sprachen.html

Swiss Federal Statistical Office. (2013a). Population - Key figures. Retrieved February 19, 2015, from http://www.bfs.admin.ch/bfs/portal/en/index/themen/01/01/key.html

Swiss Federal Statistical Office. (2013b). Vorherrschende Landessprachen in den Gemeinden, 2000. Retrieved February 19, 2015, from http://www.atlas.bfs.admin.ch/maps/13/map/mapIdOnly/0_de.html

Swiss Federal Statistical Office. (2015). Permanent resident population in urban and rural areas. Retrieved June 20, 2015, from http://www.bfs.admin.ch/bfs/portal/en/index/themen/01/02/blank/key/raeumliche_verteilung/agglomerationen.html

Tamburrini, N., Cinnirella, M., Jansen, V. a. a., & Bryden, J. (2015). Twitter users change word usage according to conversation-partner social identity. *Social Networks*, *40*, 84–89. doi:10.1016/j.socnet.2014.07.004

Tobler, W. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*, 234–240. doi:10.1126/science.11.277.620

Twitter. (2014). Über Twitter. Retrieved from https://about.twitter.com/de/company

Twitter Developers. (2015a). REST APIs. Retrieved February 15, 2015, from https://dev.twitter.com/rest/public

Twitter Developers. (2015b). Streaming API request parameters. Retrieved April 24, 2015, from https://dev.twitter.com/streaming/overview/request-parameters

Twitter Developers. (2015c). The Streaming APIs. Retrieved February 15, 2015, from https://dev.twitter.com/streaming/overview

Volkova, S., Wilson, T., & Yarowsky, D. (2013). Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media. *EMNLP*, (October), 1815–1827.

Walsh, G., Kilian, T., & Hass, B. H. (2008). Grundlagen des Web 2.0. In *Web 2.0 Neue Perspektiven für Marketing und Medien* (pp. 3–21). Berlin: Springer. Retrieved from http://www.springerlink.com/index/10.1007/978-3-540-73701-8

Wanzeck, C. (2010). Lexik nationaler Varietäten. In *Lexikologie: Beschreibung von Wort und Wortschatz im Deutschen* (UTB, Vol., pp. 114–124). Göttingen: Vandenhoeck & Ruprecht.

Weerkamp, W., Carter, S., & Tsagkias, M. (2011). How People use Twitter in Different Languages. In *Proceedings of the ACM WebSci'11* (p. 1). Koblenz, Germany.

Wikipedia. (2014). Liste von Helvetismen. Retrieved November 7, 2014, from http://de.wikipedia.org/wiki/Liste_von_Helvetismen

World Bank. (2015). Rural population (% of total population). Retrieved June 20, 2015, from http://data.worldbank.org/indicator/SP.RUR.TOTL.ZS

Yardi, S., Romero, D., Schoenebeck, G., & Boyd, D. (2009). Detecting spam in a Twitter network. *First Monday*, *15*(1).

Zappavigna, M. (2012). *Discourse of Twitter and Social Media: How We Use Language to Create Affiliation on the Web* (1st ed.). UK: Bloomsbury Academic.

Zhang, C. M., & Paxson, V. (2011). Detecting and analyzing automated activity on twitter. *Passive and Active Measurement, 6579*, 102–111. doi:10.1007/978-3-642-19260-9_11

Zhao, S., Zhong, L., Wickramasuriya, J., & Vasudevan, V. (2011). *Human as Real-Time Sensors of Social and Physical Events: A Case Study of Twitter and Sports Games*. *arXiv preprint arXiv:1106.4300*. doi:10.1145/2309996.2310053

Zihler, O. (2013). *Implementation of a Spatially-Aware Image Search Engine and Its Evaluation using Crowdsourced Relevance Judgements*. University of Zurich.

Appendices

A. Lemmas

A.1 Austriazisms

Abgeordnete	Faschingskrapfen	Landkreis	schauen
absperren	Fernsprecher	laufen	schlagen
abstimmen	Fest	Leiter	Schlögel
AHS	Festl	letschert	schwach
Aktion	Finale	Lungenbraten	See
Amt	Fisolen	Mädchen	sohin
Amt der Landesregierung	Foto	Mädel	somit
Amtsraum	Frankfurter	Madl	Spezi
Anbot	Freund	Magistrat	Spital
Angeklagte	Freunderl	Mandatar	staatlich
ankündigen	Fuss	Marillen	Staatsbürgerkunde
anlangen	Gang	Markt	Staatsliga
Ausschuss	Gebäude	Marktgemeinde	Stadtteil
Autobahnknoten	Gemeinde	Masen	steigen
Beihilfe	Geschäftsführer	Massel	sterben
Beilage	Goal	Maßel	Team
Beisel	Grammeln	Melanzani	Telefon
Beiz	Gymnasium	Minister	tief
Bericht	Haberer	Ministerin	Tod
Bezirk	Haxen	mit	Topfen
Bildung	hinsitzen	Musi	Treffer
Bub	hocken	na	Umgebung
Bummerl	Hüferl	Nationalrat	Unterstützung
Bundesland	insbesondere	nieder	Vater
Bundesliga	Jänner	niedersetzen	Verhandlungsgegenstand
Bundesminister	Karifol	niedersitzen	verkosten
Bundesregierung	Karte	Nuss	versperren
Bürger	kehren	nützen	Viertel
Bürgermeister	Kilbi	Obers	Vogerlsalat
Bürgermeisterin	Kirchtag	Obfrau	Vorstand
Dachboden	Kirchweih	Obmann	Vorständin
Dati	Kirtag	obwohl	Weichseln
Dienstzimmer	klären	Offert	Werks
Direktor	Klasse	Paradeiser	Wienerle
Dirndl	Klaße	Pastor	Wirtschaft
drücken	Klub	Pastorin	Wirtshaus
Dult	Knabe	per	zeigen
einheben	Knoten	Pfarrer	Zollstab

einlassen	Kohlsprossen	Powidl	zum
einlaßen	kosten	Preisausschreiben	Zuschauer
Entscheidung	Krankenhaus	Professor	Zuschuss
erweisen	Krapfen	Profeßor	Zuschuß
exekutieren	Kren	Ribisel	Zuseher
Fan	Land	Rostbraten	Zuseherin
Faschiertes	Landeshauptmann	Runde	zusperren

A.2 Helvetisms

abklären	Fan	Lohn	speditiv
Ableger	fehlbar	Mädchen	Spital
abschliessen	Ferien	Mais	Spunten
abserbeln	Fest	Marktflecken	Staat
absitzen	Final	Match	Staatskunde
abstimmen	Föteli	Matur	Staatsrat
Advokat	Foto	Matura	Stadtammann
ahnden	Fraktion	Mehr	Stadtpräsident
Aktion	Frankfurterli	Meitli	Stadtteil
allenfalls	Freund	Meitschi	Stand
Ammann	Führerausweis	Messe	Standeskomißion
Amt	fuhrwerken	Meter	Standeskommission
Ämter	Fürsprech	Mietzins	Stapi
Amtsbezirk	Fürsprecher	Morgenessen	Steueramt
Amtsstelle	Gang	Morgeneßen	stimmen
anbetreffen	GAV	Musik	stossen
Angeklagte	Gebäude	Mutation	stoßen
Angeschuldigte	Gemeinde	mutieren	Subsidium
ankünden	Gemeindeammann	Natel	Supporter
ankündigen	Gemeindehauptmann	Nati	Team
Anlegen	Gemeindepräsident	Nationalliga	Teilstaat
Ätti	Gesamtarbeitsvertrag	Nationalrat	Telefon
auf	Geschäft	nein	tief
aufgestellt	Geschäftsführer	nieder	Tod
Ausgang	Geschäftsleiter	Nomination	Töff
Automobilist	Ghüder	nützen	tönen
Baute	Gliedermeter	Obligatorium	Traktandenliste
bedingt	Glück	Obmann	Traktandum
Beilage	Goal	obschon	Trassee
Bein	Goalie	Offerte	Traßee
Beitrag	Güsel	Ort	Treffer
Beiz	Gymi	Ortsbürger	treten
Bericht	Gymnasium	Ortsbürgerin	Umgebung
Berliner	Hag	Ortsgemeinde	Umgelände
betreffen	harzig	Ortsvorsteher	unbedingt

betreiben	Hektare	parkieren	Unterdach
Bezirk	Hektaren	Parlamentsmitglied	Unternehmung
Billett	Hinschied	Parteipräsident	Unterstützung
bis anhin	hinsitzen	Parterre	Vater
blochen	hocken	per	Velo
blutt	innert	Pfarrer	verdanken
bohnern	insbesondere	plagieren	Verwaltungsrat
Bub	Januar	Poulet	vorab
Büez	Jupe	Pult	vorführen
Bundesrat	Kader	Quartier	vorgängig
Burger	Kanti	Rapport	Vorsteher
Bürger	Kanton	rassig	Vorsteherin
Burgerin	Kantonsschule	raßig	Vorsteher
Büro	Kantonßchule	Ratsmitglied	Wegleitung
Chauffeur	Kategorie	Redaktor	weisen
Check	Kehricht	Regierungspräsident	Werk
Chilbi	Kilbi	Regierungsrat	Wettbewerb
degustieren	klären	Rektorin	Wienerli
Departement	Knabe	Rektor	Wirtschaft
Departmentschef	Kollege	Runde	wischen
Departmentsvorsteher	Komission	Sanität	Wissenschafter
Detailhandel	Komposition	scharf	Wißenschafter
Direktion	Korridor	schauen	würzig
Direktor	Kreuz	schlagen	zentral
Dopplemeter	Landamman	schlapp	Zmittag
eidgenössisch	Landammann	Schulzimmer	Zmorge
eindrücklich	Landkreis	schwach	Zmorgen
Einwohnergemeinde	laufen	See	zügeln
Entscheid	Lehrer	serbeln	Zuschauer
erheben	Lehrerin	sitzen	Zustupf
Estrich	Lehrtochter	somit	
Fahrausweis	Leiter	Sonderangebot	

A.3 Teutonisms

Abfall	entscheidender Bedeutung	Kneipe	Rummel
Abgeordnete	Entscheidung	Knochenarbeit	Sahne
Abitur	Erdgeschoss	Kollektivvertrag	Sanitätsdienst
abklären	Erdgeschoß	Korridor	Sauerkirschen
abnibbeln	Eren	kosten	schauen
abschliessen	erheben	Kraftakt	Scheck
abschließen	Erste Bürgermeister	Krankenhaus	schieben
absperren	erweisen	Krapfen	schlagen
abstimmen	eventuell	Kreis	schlapp

Abteilung	Fahrer	Kreuz	Schreibtisch
Aktion	Fahrrad	Kugel	Schufterei
akündigen	Fakultät	Land	schuldig
Amt	Fan	Landesminister	schwach
amtliche Entscheidung	Faschingskrapfen	Landkreis	seine Stimme abgeben
Amtsraum	fegen	langsam	Senator
Amtszimmer	Feldsalat	Lärm	Senatsverwaltung
anbetreffen	Fernsprecher	laufen	Söller
Änderung	Fest	Lehrer	somit
angeben	Festl	Lehrerin	Speicher
Angebot	Fete	Leiter	Spezi
Angeklagte	Filet	Leitfaden	Spieltag
ankünden	Filiale	Lenker	spontan
Anlange	Finale	Mädchen	staatlich
anlangen	Finanzamt	Mädel	staatliche Einrichtung des Kantons
Anwalt	Flur	Madl	Staatsbürgerkunde
Aprikosen	Formation	Mädle	Stadtoberhaupt
Arbeit	Foto	Maloche	Stadtteil
Ärger	Fraktion	Marktflecken	steigen
Aubergine	Frankfurter	Massel	sterben
Aufsichtsrat	Freund	Maßel	Strapaze
Ausschuss	fröhlich	Meer	Studienrat
Ausschuß	Frühstück	Meerrettich	Tagesordnung
Auszubildende	Führerschein	Mehrheit	Tarifvertrag
Autofahrer	Fuss	Messe	Team
Bahnkörper	Gang	Meße	Telefon
Behörde	Gebäude	Miete	Themenpunkt
Beihilfe	Gehalt	Minister	tief
Bein	Gemeinde	Ministerin	Tod
Beisel	Gemeinschaftskunde	Ministerium	toll
Beisl	Geschäftsführer	Ministerpräsident	Tomaten
beitreiben	Geschäftsleiter	mit Bewährungsfrist	Tor
Beiz	Glück	Mittagessen	Torhüter
Bekannter	Grieben	Mittageßen	Trasse
Bericht	Grüne Bohnen	Mobiltelefon	Traße
Berliner	gucken	Motorrad	Treffer
Berufsfahrer	gustieren	Musi	treten
Beschluss	gut drauf	Musik	umgänglich
Beschluß	gut gelaunt	na	Umgebung
Beschreibung	Gymnasium	nackt	umziehen
bestrafen	Hackfleisch	Nationalmannschaft	Unternehmen
betreffen	Handy	nee	unternehmen
Bildung	hantieren	nein	Untersstützung
binnen	Hax	nieder	Urlaub

bisher	Haxen	niedersitzen	Vater
blank	Hektar	Nominierung	Verhandlungsgegenstand
blocken	herummachen	nutzen	verkosten
bluffen	Hinscheiden	nützen	Verpflichtung
Blumenkohl	Hochrippe	Oberbürgermeister	Viertel
Bockwurst	höchstens	obwohl	von zentraler
Boden	hocken	ohne Bewährungsfrist	vorab
bohnern	Hüfte	parken	Vorsitzende
Bub	Huhn	Pastor	Vorstand
Bühne	innerhalb	Pastorin	Vorständin
Bundesland	insbesondere	Pfannkuchen	wachsen
Bundesliga	Januar	Pfarrer	Werks
Bundesminister	Johannibeere	Pflaumenmus	Wiener
Bundesregierung	Joppe	Pflicht	Wirtschaft
Bundesstaat	Junge	Pflichtfach	Wirtshaus
Bundeßtaat	Kader	Pinte	Wissenschaftler
Bundestag	kahl	Plackerei	Wißenschaftler
Bürger	Karte	prahlen	zäh
Bürgermeister	kehren	Präsident des Senats	Zaun
Bürgermeisterin	Keule	Preisausschreiben	zeigen
Büro	kiecken	Quark	Zollstock
Dachboden	Kiez	rasch	zügig
den Wohnsitz wechseln	Kirbe	Rechtsanwalt	Zugzusammenstellung
Dienstzimmer	Kirchtag	Redakteur	zum
Direktor	Kirchweih	regierende Bürgermeister	Zuschauer
drücken	Kirmes	Regierung	zuschliessen
Dult	Kirta	Rektor	Zuschuss
Dusel	klären	Rektorin	Zuschuß
eindrucksvoll	Klasse	Rettung	zusperren
Einzelhandel	Klaße	Rock	zuvor
Endspiel	klingen	Rosenkohl	

B. Software

ArcGIS 10.2

A mapping software created by ESRI which combines data and Geography. With incorporated functions that can be adapted to the respective needs, the possibility to write own functions, geographic and projected systems transformations ArcGIS facilitates the analysis and planning of geographical data.
More information on: www.arcgis.com

Eclipse IDE

The Eclipse Java *Integrated Development Environment* (IDE) is an open source desktop IDE which allows you to program software and other automated programs based on the language Java. The default packages can be customized and extended to accommodate the respective needs.
More information on: eclipse.org

Microsoft Excel

A table calculation program created by Microsoft. Incorporated functions and the possibility to create own macros facilitate the analysis and representation of data.
More information on: products.office.com/de-ch/excel

PostgreSQL

It is an open source object-relational database system. It works on multiple platforms and its programming interface supports multiple programing languages including Java, which was used for this work. The pgAdmin graphical tool facilitates the management and development of the entire dataset. Queries on PostgreSQL are conducted on the *Structured Query Language* (SQL). Add-Ons like PostGIS allow you to expand the capabilities of the database system, enabling also support for geographic objects and geographical queries.
More information on: www.postgresql.org

C. Code

In this appendix only the code that was relevant for the analysis is listed. The implementation of the DB and the connection to Twitter by the API followed the state of the art. Therefore, there is no reason to list them at this point. However, they can be found in the CD annexed to this thesis.

C. 1 SQL

```
-- Tweet count
select count (*) from tweets;

-- Daily count
select t.created_at, count(*) from tweets t GROUP BY t.created_at;

-- Count of unique user IDs
select count(distinct t.userid) from tweets t;

-- Get unique user IDs
select distinct (t.userid) from tweets t;

-- Count aggregated by unique user ID
select distinct t.userid, count(*) from tweets t group by t.userid;

-- Select particular user from DB
select t.tweet_text, created_at, geolocation from tweets t where userid = 2834847965;

-- Count direct replies
select count(*) from tweets t where t.tweet_text like '@%';

-- Number of different coordinates for each user (Bots/Cyborgs)
SELECT userid, count(t.tweet_text) AS tweet_Count, count (DISTINCT geolocation) AS
dif_coord FROM tweets t GROUP BY t.userid;

-- Removal of bots and cyborgs from the DB according to the implemented method.
delete from tweet_to_hashtags where tweet_id IN (

    select id from tweets where userid IN
    (
      161262801,160874621,1336218432,115089595,203007243,2315299130,345423903,1868998
60,28348
      47965,2834844017,117756259,2834830029,36744887,282492573,40224678,139789733,109
2190045,
      566539016,243764296,542768859,2348965490,2798503111,1380163830,278035810,610134
154,4030
      1402,125298958,271776179,89898739,40232085,40241473,40217418,2147525082,4021728
6,715283
      70,29430823,33509894,2838722758,24881646,19202150,30421313,329938758,33513508,2
94190333      ,124131780,2887896783,72859719,33630376
    )
);

delete from submits where tweetid IN (

    select id from tweets where userid IN
    (
      161262801,160874621,1336218432,115089595,203007243,2315299130,345423903,1868998
60,28348
      47965,2834844017,117756259,2834830029,36744887,282492573,40224678,139789733,109
2190045,
      566539016,243764296,542768859,2348965490,2798503111,1380163830,278035810,610134
154,4030
      1402,125298958,271776179,89898739,40232085,40241473,40217418,2147525082,4021728
6,715283
      70,29430823,33509894,2838722758,24881646,19202150,30421313,329938758,33513508,2
94190333      ,124131780,2887896783,72859719,33630376
    )
);
```

```
delete from tweet_to_queries where tweet_id IN (

    select id from tweets where userid IN
    (
      161262801,160874621,1336218432,115089595,203007243,2315299130,345423903,1868998
60,28348
      47965,2834844017,117756259,2834830029,36744887,282492573,40224678,139789733,109
2190045,
      566539016,243764296,542768859,2348965490,2798503111,1380163830,278035810,610134
154,4030
      1402,125298958,271776179,89898739,40232085,40241473,40217418,2147525082,4021728
6,715283
      70,29430823,33509894,2838722758,24881646,19202150,30421313,329938758,33513508,2
94190333      ,124131780,2887896783,72859719,33630376
    )
);

Delete from tweets where userid IN (

      161262801,160874621,1336218432,115089595,203007243,2315299130,345423903,1868998
60,28348
      47965,2834844017,117756259,2834830029,36744887,282492573,40224678,139789733,109
2190045,
      566539016,243764296,542768859,2348965490,2798503111,1380163830,278035810,610134
154,4030
      1402,125298958,271776179,89898739,40232085,40241473,40217418,2147525082,4021728
6,715283
      70,29430823,33509894,2838722758,24881646,19202150,30421313,329938758,33513508,2
94190333      ,124131780,2887896783,72859719,33630376
);
```

C.2 Java

All the following codes have in common that they start with the import of necessary java extensions. Some of them are already incorporated in Eclipse, i.e. Java, which is visible at the reference *starting with import java…*. Others are very specific to the processing step and needed to be added as external libraries. Common to the following code snippets are the java imports:

```
import java.io.BufferedReader;
import java.io.BufferedWriter;
import java.io.FileInputStream;
import java.io.FileNotFoundException;
import java.io.FileOutputStream;
import java.io.IOException;
import java.io.InputStreamReader;
import java.io.OutputStreamWriter;
import
java.io.UnsupportedEncodingException;
import java.sql.ResultSet;
import java.sql.SQLException;
import java.sql.Statement;
import java.util.ArrayList;
import java.util.Set;
import java.util.TreeMap;
import java.util.TreeSet;
```

Besides this general imports there are also the specific references. This references were maintained in the code to facilitate the understanding while the common references have been removed and replaced by the place holder (… other imports).

C.2.1 Lemmas in the tweets

```
package main;

(… other imports)

import db.AbstractDBConnector;
import db.PGDBConnector;

public class InvertedIndex_Adaptation {

    public static void main(String[] args) throws Exception {

        String host = "localhost";
        String port = "5432";
        String database = "MA_Clean";
        String user = "postgres";
        String password = "postgres";

        AbstractDBConnector con = new PGDBConnector(host, port, database, user,
password);

        // Get the whole database as inverted index
        TreeMap<String, ArrayList<Long>> invertedIndex = getInvertedIndex(con);

        /*
         * This is new: search for keywords that contain searchKeyword (fragments)
May be used to retrieve composites like Bauamt, Standesamt, amtierend, etc. Will
also retrieve things like insgesAMT, etc.
         */
        // Enter your keywords (here, "Amt" is an example keyword. Multiple
keywords possible)
        ArrayList<String> searchKeywords = new ArrayList<String>();

        BufferedReader reader = new BufferedReader(new InputStreamReader(new
FileInputStream("C:/Users/André/Documents/UZH/2014_2015_MasterArbeit/Daten/Keywords
/lists_for_eclipse/complete_teutonismen_utf8.txt"), "UTF-8"));
        String line = null;
        while ((line = reader.readLine()) != null) {
            String[] splits = line.toLowerCase().split(",");
            for (String l : splits) {
                searchKeywords.add(l);
            }
        }

        reader.close();


//      searchKeywords.add("amt");

        // Search the index for the keywords.
        // Returns a subset of the index, where the String key is a word that
contained one of the searchKeywords
        TreeMap<String, ArrayList<Long>> subset = getTweetIds(invertedIndex,
searchKeywords);

        // Optionally, do something with the subset, e.g. count how many tweets
were found for all searchKeywords
        // Could be used to e.g. manually remove all the words that are not an
actual subset of searchKeywords (e.g. insgesamt in the case of "Amt")
        System.out.println("Found: " + subset.keySet().size() + " different words
for given search terms");
        System.out.println("Search Results: ");

        int tweetCounter = 0; // used to count the overall number of tweets found
        for (String k : subset.keySet()) { // iterate through all the keywords k
that contained a search keyword

            ArrayList<Long> keywordTweetIds = subset.get(k); // get all the
tweetids for a keyword k
//          System.out.println(k + "; Found: " + keywordTweetIds.size() + " tweets
for search terms: " + keywordTweetIds);
```

```
            System.out.print(k + ";" + keywordTweetIds.size() + ";"); // print the
keyword k
            tweetCounter += keywordTweetIds.size(); // update the overall number of
tweet ids found
            for (Long l : keywordTweetIds) { // print all the tweet ids found for a
keyword k that contained a search keyword
                System.out.print(l + ", ");
            }
            System.out.println();

//          System.out.println(k + " Found: " + keywordTweetIds.size() + " tweets
for search terms: " + l);
        }
        // print the overall tweet count for all search keywords
        System.out.println("Found " + tweetCounter + " tweets for the keywords
containing the search terms");

        // look up the tweet id's in the database and get the corresponding full
tweet text with getTweetForIds
        TreeMap<Long, String> tweetsForId = new TreeMap<>(); // this will contain
the tweet id and the whole tweet text
        for (String k : subset.keySet()) { // for all keywords k that contained a
search keyword
            ArrayList<Long> tweetIds = subset.get(k); // extract all the ids for a
certain keyword that contained a search keyword
            getTweetForIds(con, tweetsForId, tweetIds); // get all tweets from the
database using the tweetids and con, and store the results in tweetsForId<TweetId,
TweetText>
        }


        con.closeConnection();

    }

    /**
     * Write a whole index to a file
     *
     * @param invertedIndex
     * @param kw
     * @param invertedIndexFile
     * @throws UnsupportedEncodingException
     * @throws FileNotFoundException
     * @throws IOException
     */
    public static void writeInvertedIndexToFile(TreeMap<String, ArrayList<Long>>
invertedIndex, TreeSet<String> kw, String invertedIndexFile) throws
UnsupportedEncodingException, FileNotFoundException, IOException {
        BufferedWriter writer = new BufferedWriter(new OutputStreamWriter(new
FileOutputStream(invertedIndexFile), "UTF-8"));
        for (String k : kw) {
            ArrayList<Long> list = invertedIndex.get(k);
            int cnt = 0;
            if (list != null) {
                cnt = list.size();
            }
            writer.write(k + ": " + cnt + "\n");

        }
        writer.close();
    }


    /**
     * Retrieving the tweets from the database for a number of tweetIds
     *
     * @param con
     * @param tweetsForId
     * @param tweetIds
     */
    public static void getTweetForIds(AbstractDBConnector con, TreeMap<Long,
String> tweetsForId, ArrayList<Long> tweetIds) {
        if (tweetsForId == null) {
            return;
        }
        if (tweetIds == null || tweetIds.size() == 0) {
```

```
            return;
        }
        try {
            Statement st = con.getConnection().createStatement();
            String sql = "Select id, LOWER (tweet_text) from tweets where ";

            for (int i = 0; i < tweetIds.size() - 1; ++i) {
                sql += "id = " + tweetIds.get(i) + " OR ";
            }
            sql += "id = " + tweetIds.get(tweetIds.size() - 1) + ";";
//          System.out.println(sql);
            ResultSet res = st.executeQuery(sql);
            while (res.next()) {
                Long tweetId = res.getLong(1);
                String tweetText = res.getString(2);
                tweetsForId.put(tweetId, tweetText);
            }
            st.close();
        } catch (SQLException s) {
            s.printStackTrace();
        }
    }

    /**
     * Retrieving the whole database of tweets as an inverted index.
     *
     * @param con
     * @return
     * @throws SQLException
     */
    public static TreeMap<String, ArrayList<Long>>
getInvertedIndex(AbstractDBConnector con) throws SQLException {
        TreeMap<String, ArrayList<Long>> invertedIndex = new TreeMap<>();

        Statement st = con.getConnection().createStatement();

        String sql = "Select id, tweet_text from tweets;";

        ResultSet res = st.executeQuery(sql);

        while (res.next()) {

            Long tweetId = res.getLong(1);
            String text = res.getString(2);
            String[] split = text.toLowerCase().replaceAll("[^A-ZÄÖÜa-zäöü0-9ß]", "
").split(" ");;
            for (String s : split) {
                ArrayList<Long> list = invertedIndex.get(s);
                if (list == null) {
                    list = new ArrayList<>();
                    invertedIndex.put(s, list);
                }
                list.add(tweetId);
            }
        }
        st.close();
        return invertedIndex;
    }

    public static TreeMap<String, ArrayList<Long>> getTweetIds(TreeMap<String,
ArrayList<Long>> inputIndex, ArrayList<String> searchKeywords) {

        Set<String> keys = inputIndex.keySet();
        TreeMap<String, ArrayList<Long>> keysSubset = new TreeMap<>();

        for (String searchKeyword : searchKeywords) {
            String tmp = searchKeyword.toLowerCase().replaceAll("[^A-ZÄÖÜa-zäöü0-
9ß]", " ");

            for (String key : keys) {
                if (key.contains(tmp)) {
                    keysSubset.put(key, inputIndex.get(key));
                }
            }
        }
```

```
        return keysSubset;
    }
}
```

C.2.2 Stemmer

```
package stemmer;

(… other imports)

import org.tartarus.snowball.ext.germanStemmer;


public class Stemmer {

    public static void main(String[] args) throws Exception {

        String txt = "Input;Resulting Stem;\n" + getStemms();

        BufferedWriter writer = new BufferedWriter(new OutputStreamWriter(new
FileOutputStream(new
File("C:/Users/André/Documents/UZH/2014_2015_MasterArbeit/Daten/Keywords/lists_for_
eclipse/Analyse/Analyse/lemma_liste_after_Stemmer.txt")),"UTF-8"));
        writer.write(txt);
        writer.close();

    }

    private static String getStemms() throws Exception {

        String resultStemms = "";

//      Create an ArrayList where all the Lemmas are stored that need to be
analyzed with the Stemmer in a next step
        ArrayList<String> keywords = new ArrayList<String>();

        BufferedReader reader = new BufferedReader(new InputStreamReader(new
FileInputStream("C:/Users/André/Documents/UZH/2014_2015_MasterArbeit/Daten/Keywords
/lists_for_eclipse/complete_lemma_liste.txt"),"UTF-8"));
        String line = null;
        while ((line = reader.readLine()) != null) {
            String[] splits = line.toLowerCase().split(",");
            for (String l : splits) {
                keywords.add(l);
            }
        }
        reader.close();


//      For-Loop to read each Lemma in the Stemmer and then to give it out

        for (int i=0; i<keywords.size(); i++){

            germanStemmer stemmer = new germanStemmer();
            String lemma = keywords.get(i);
            stemmer.setCurrent(lemma);
            if (stemmer.stem()){
//              System.out.println("Input: " + lemma + "\tStem: "
+stemmer.getCurrent());
                resultStemms += lemma + ";" + stemmer.getCurrent() + ";\n";
            }

        }

        return resultStemms;
    }
}
```

C.2.3 .txt-File for ArcGIS

```
package main;

(… other imports)

import org.postgis.PGgeometry;

import db.AbstractDBConnector;
import db.PGDBConnector;

public class Create_txt_ArcGIS {

    private static final String host = "localhost";
    private static final String port = "5432";
    private static final String database = "MA_Clean";
    private static final String user = "postgres";
    private static final String password = "postgres";
    private static String keyword;
    private int ObjectID;
    private static final String filePath
="C:/Users/André/Documents/UZH/2014_2015_MasterArbeit/Daten/Keywords/lists_for_ecli
pse/Analyse/Analyse/ids/ArcGIS_complete_clean.txt";
    private static AbstractDBConnector db;


    public static void main(String[] args) throws IOException, SQLException,
InterruptedException {
        db = new PGDBConnector(host, port, database, user, password);
            String txt = "ID, LAT, LON, TEXT, \n" + getLocations();

        BufferedWriter writer = new BufferedWriter(new OutputStreamWriter(new
FileOutputStream(new File(filePath)),"UTF-8"));
        writer.write(txt);
        writer.close();
    }



    private static String getLocations() throws SQLException {
        db = new PGDBConnector(host, port, database, user, password);
        Statement s = db.getConnection().createStatement();
        String resultLocations = "";
        String file =
"C:/Users/André/Documents/UZH/2014_2015_MasterArbeit/Daten/Keywords/lists_for_eclip
se/Analyse/Analyse/ids/complete_ids.txt";
        String[] keywordsToFilterFor = getKeywords(file);

        for (int i = 0; i < keywordsToFilterFor.length; i++) {
            String keyword = keywordsToFilterFor[i].toLowerCase();
//          System.out.println(keyword);
//          String sqlQuery = "SELECT id,tweet_text, geolocation FROM tweets where
tweet_text like '%"+keyword+"%';";
            String sqlQuery = "SELECT id,tweet_text, geolocation FROM tweets where
id = "+keyword+";";
//          System.out.println(sqlQuery);

            ResultSet set = s.executeQuery(sqlQuery);

            while (set.next()) {
            String id = set.getString(1);
            String tweet = set.getString(2).toLowerCase().replaceAll("[^A-ZÄÖÜa-
zäöü0-9ß@#]", " ");
            PGgeometry geom = (PGgeometry) set.getObject(3);
            resultLocations += id + "," + geom.getGeometry().getLastPoint().y + ","
+ geom.getGeometry().getLastPoint().x + "," + tweet + ",\n";
            }

        System.out.println(resultLocations);
        }
        return resultLocations;

    }
```

```
    private static String[] getKeywords(String file) {
        try {
            Reader r = new InputStreamReader(
                    new FileInputStream(new File(file)), "UTF-8");
            BufferedReader reader = new BufferedReader(r);
            String line = null;
            Set<String> words = new HashSet<String>();
            while ((line = reader.readLine()) != null) {
                String[] w = line.toLowerCase().split(",");
                for (String w1 : w) {
                    words.add(w1.trim());
                }
            }
            reader.close();

            String[] wordArray = new String[words.size()];
            return words.toArray(wordArray);
        } catch (IOException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        }

        return null;
    }

}
```

D. Results

D.1 Number of Occurrences of the Selected Lemmas

Keywords	Frequency	Keywords	Frequency
mit	124044	bein	1265
auf	113817	tod	1189
zum	39243	check	1178
mehr	18742	goal	1157
na	12904	krankenhaus	1120
foto	11481	nutzen	1100
nein	11162	angebot	1094
glück	5893	freunde	1085
team	5140	boden	1051
schauen	4449	frankfurter	1038
musik	4437	amt	1002
gang	4278	unbedingt	988
gucken	4263	rock	977
see	4194	bisher	969
arbeit	4145	tolle	953
toll	3687	lehrer	936
klasse	3594	entscheidung	933
freund	3566	telefon	922
berliner	3394	werk	899
fest	3345	aktion	893
laufen	3219	staat	826
land	3121	tief	816
fan	2889	finale	813
bundesliga	2699	fans	792
markt	2667	kreuz	765
per	2609	kosten	762
nee	2530	fotos	746
handy	2477	gymnasium	735
urlaub	2473	ferien	724
sitzen	2374	wirtschaft	711
runde	2199	freunden	703
messe	2194	meter	698
obwohl	2168	stimmen	697
stand	2123	kollegen	694
langsam	1987	januar	689
frühstück	1985	bühne	685
tor	1922	aufs	662
junge	1915	burger	662
vater	1815	beitrag	657
final	1745	bundestag	625
mädchen	1694	schlagen	618
wiener	1658	bericht	615
büro	1610	mais	567
ort	1592	kreis	566
zeigen	1500	mittagessen	560
karte	1442	bildung	542
sterben	1435	tollen	539
meer	1412	fahrrad	508
arbeiten	1309	drücken	495
kollege	494	pflicht	206
viertel	480	blocken	205
zuschauer	475	beine	203

bürger	473	scharf	203
somit	465	arbeite	202
unternehmen	459	wettbewerb	202
spieltag	453	bundesregierung	199
spontan	442	mehrere	198
tolles	437	vorstand	197
unterstützung	424	mehrheit	191
schwach	421	ärger	187
toller	414	senator	186
steigen	375	leiter	184
match	374	füße	183
kneipe	366	kiez	181
gemeinde	351	entscheiden	178
mädels	336	fuss	176
wirtshaus	330	laufe	172
minister	317	tomaten	169
eren	311	zuvor	169
regierung	310	mädel	167
gebäude	309	flur	163
klub	308	beinen	160
wachsen	304	gehalten	149
nackt	298	behörden	145
umgebung	298	höchstens	141
treten	289	teams	141
lauf	286	ausschuss	139
fahrer	279	support	139
karten	275	blank	136
kirmes	269	klären	135
innerhalb	268	schieben	133
jungen	265	kader	132
dirndl	264	führerschein	131
fuß	264	fröhlich	129
sahne	262	pfannkuchen	127
fraktion	250	lehrerin	124
geschäft	250	füßen	123
nieder	241	umgang	123
frühstücken	240	huhn	121
tore	234	parken	118
bürgermeister	229	handys	117
ahs	224	betreiben	113
bezirk	221	messer	113
speicher	218	abteilung	111
schreibtisch	217	berichten	110
kategorie	209	fakultät	109
professor	209	insbesondere	106
eventuell	208	weisen	105
filiale	104	rocken	66
abstimmen	101	bundesrat	64
quartier	101	kehren	64
kugel	99	lärm	64
autofahrer	97	orte	64
schlapp	96	dachboden	63
gehalt	95	nationalmannschaft	61
bub	94	gangster	60
anwalt	93	rettung	60
bundesland	93	aktionen	58
ausgang	91	anlegen	58
hagen	91	geschäfte	57

motorrad	90	knoten	57
klingen	89	standen	57
miete	89	aubergine	56
tode	89	einzelhandel	56
kanton	88	schuldig	56
spital	88	geschäftsführer	55
mehreren	87	lande	54
nuss	87	ministerin	53
treffer	86	musiker	53
hocken	83	änderung	52
pfarrer	82	feste	52
quark	82	stoßen	52
runden	82	todes	52
vorab	82	finanzamt	51
landkreis	81	zentrale	51
umziehen	81	abitur	50
beschreibung	79	ämter	50
angeben	78	behörde	50
landen	78	supporter	50
ministerpräsident	77	bekannter	49
staaten	77	erheben	49
angeboten	76	landes	49
ärgern	75	tollsten	49
ministerium	75	abschließen	48
angebote	73	messen	48
checken	71	nackte	46
junger	71	vorsitzende	46
stadtteil	71	zuschauen	46
binnen	70	metern	45
entscheidungen	69	zentral	45
wischen	69	abgeordnete	44
zaun	69	rosenkohl	44
aufgestellt	68	tiefen	44
berichte	68	tollste	44
lohn	68	direktor	43
toren	68	langsamer	43
bedingt	66	nati	43
spezi	43	vaters	28
entscheid	42	abgeordneten	27
beschluss	41	dult	27
bockwurst	41	hüfte	27
filet	41	junges	27
kirchtag	41	lehrern	27
kren	41	pastorin	27
tiefe	41	rummel	27
beilage	40	speichern	27
finals	40	vorsitzender	27
nackten	40	änderungen	26
tiefer	40	aufsichtsrat	26
torhüter	40	endspiel	26
wissenschaftler	40	nationalrat	26
nominierung	38	zentraler	26
tagesordnung	38	zollstock	26
redakteur	36	bürgermeisterin	25
urlaubs	36	entscheide	25
tollem	35	kreise	25
ammann	34	oberen	25
ärgerlich	34	orten	25

fröhliche	34	goalie	24
hackfleisch	34	oberbürgermeister	24
lohnen	34	pers	24
musi	34	schwachen	24
verdanken	34	zügig	24
freundes	33	bestrafen	23
fröhlichen	33	bürgern	23
mieten	33	eindrucksvoll	23
vorsitzenden	33	kirchweih	23
arbeitest	32	ministerpräsidenten	23
chauffeur	32	rocks	23
einlassen	32	spontanen	23
festen	32	zuschauern	23
spontane	32	fegen	22
burgers	31	lenker	22
krapfen	31	verpflichtung	22
rechtsanwalt	31	allenfalls	21
klassen	30	baute	21
rasch	30	erdgeschoss	21
staatlich	30	filialen	20
arbeiter	29	gange	20
blumenkohl	29	matches	20
feldsalat	29	matura	20
fete	29	spontaner	20
keule	29	ableger	19
velo	29	burgern	19
werke	29	kneipen	19
goals	28	runder	19
schwache	19	burgerlich	14
spezies	19	glücks	14
tiefsten	19	mieter	14
ärgere	18	regierenden	14
beisl	18	rocker	14
entscheidend	18	stossen	14
fröhliches	18	telefone	14
kirtag	18	vorabend	14
langsamen	18	werks	14
madl	18	wirtschaftliche	14
nützen	18	beitragen	13
pult	18	beiz	13
scharfer	18	frühstücks	13
sonderangebot	18	haxen	13
wirtschaftlich	18	kreisen	13
ankündigen	17	meerrettich	13
checke	17	pflichten	13
entscheidende	17	rapport	13
frühstücke	17	zmittag	13
jüngeren	17	arbeits	12
spontanes	17	bezirke	12
verkosten	17	burgermeister	12
bundesminister	16	hag	12
eindrücken	16	knabe	12
fester	16	kreises	12
fraktionen	16	leitfaden	12
hüften	16	scharfen	12
innert	16	schlappe	12
mutieren	16	schwacher	12
scheck	16	tönen	12

staatliche	16	urlauber	12
stoße	16	betreffen	11
telefonisch	16	billett	11
unternehmens	16	chilbi	11
werken	16	fuße	11
bühnen	15	kantonsschule	11
büros	15	kreuze	11
dusel	15	messern	11
erweisen	15	obmann	11
geschäftlich	15	pastor	11
maloche	15	poulet	11
prahlen	15	regierungen	11
regierender	15	scharfes	11
scharfe	15	schuldigen	11
supporten	15	spieltage	11
zäh	15	tiefste	11
zmorge	15	wirtschaftlichen	11
abklären	14	auszubildende	10
angeklagte	14	departement	10
fahrern	10	entscheidest	8
finales	10	finalen	8
jänner	10	hax	8
kanti	10	kahl	8
klubs	10	kategorien	8
kreuzen	10	mehrheitlich	8
kreuzer	10	mitem	8
malochen	10	mutation	8
mehrheiten	10	regierende	8
mobiltelefon	10	staatlichen	8
nominierungen	10	staatliches	8
obwohls	10	töff	8
pinte	10	zuseher	8
professoren	10	abschliessen	7
rektor	10	bundestags	7
schlappen	10	bürgerlich	7
senatsverwaltung	10	dati	7
staates	10	entscheidenden	7
versperren	10	festem	7
wirtschafts	10	formation	7
zauner	10	haxe	7
aufem	9	jungem	7
autofahrern	9	jupe	7
beilagen	9	keulen	7
betreffend	9	knaben	7
buben	9	landeshauptmann	7
bundesstaat	9	landest	7
bürgerliche	9	langsame	7
checks	9	nomination	7
gebäuden	9	rasche	7
gemeinden	9	redakteure	7
gymi	9	schuldige	7
komposition	9	spieltagen	7
korridor	9	tieferen	7
meere	9	wirtschaften	7
nackter	9	abgeordneter	6
natel	9	amts	6
pflaumenmus	9	angeklagter	6
staatlicher	9	arbeitende	6

staats	9	arbeitenden	6
tarifvertrag	9	ausschusses	6
telefonen	9	bauten	6
vorführen	9	berlinern	6
zentralen	9	bürgerin	6
absitzen	8	bürgerliches	6
absperren	8	fraktions	6
aprikosen	8	fröhlicher	6
aufer	8	geschäften	6
blanke	8	hektar	6
kirta	6	wienerle	5
magistratisches	6	wieners	5
marktgemeinde	6	wirtschaftlicher	5
musikern	6	zuschuss	5
parkieren	6	abteilungen	4
pflichtfach	6	angebotene	4
schlögel	6	ärgerlichen	4
senators	6	beisel	4
spieltages	6	blanker	4
stadtteile	6	bundestages	4
stoß	6	bürgers	4
tiefere	6	checker	4
verpflichten	6	estrich	4
verpflichtungen	6	eventuellen	4
zügeln	6	eventueller	4
absperrung	5	finaler	4
ämtern	5	föteli	4
auberginen	5	fröhlichkeit	4
berichtigung	5	gemeindepräsident	4
bildungs	5	geschäftlichen	4
bürgermeisters	5	geschäfts	4
dienstzimmer	5	geschäftsführern	4
fahrausweis	5	geschäftsführung	4
festes	5	kraftakt	4
flure	5	lands	4
gangen	5	lärms	4
gebäudes	5	leitern	4
kirbe	5	lohner	4
lander	5	mädchens	4
langsames	5	madlene	4
langsamkeit	5	magistrat	4
meitli	5	masen	4
meitschi	5	matchen	4
messers	5	meeres	4
motorräder	5	obschon	4
motorrädern	5	raschen	4
nacktes	5	ratsmitglied	4
schulzimmer	5	redakteuren	4
spieltags	5	regierungsrat	4
stadtpräsident	5	rektorin	4
stande	5	rockst	4
tiefes	5	sanitäter	4
tollster	5	schreibtischen	4
tores	5	schwachem	4
torhütern	5	schwaches	4
treffern	5	staate	4
urlaube	5	standest	4
urlauben	5	supporters	4

vorgang	5	telefonische	4
umgangen	4	kantone	3
unternehmungen	4	kehricht	3
velos	4	kilbi	3
verwaltungsrat	4	lehrers	3
vorstands	4	lohne	3
wettbewerben	4	marillen	3
wettbewerbs	4	marktes	3
wienern	4	meeren	3
würzig	4	mehrerer	3
zusperren	4	meters	3
abschließend	3	ministers	3
abschließende	3	morgenessen	3
angebotenen	3	nacktem	3
angeklagten	3	obfrau	3
aufen	3	offert	3
beihilfe	3	parterre	3
berichts	3	pfarrers	3
beschreibungen	3	professore	3
betreffende	3	rocke	3
betreffenden	3	rockers	3
bezirken	3	rundes	3
bezirks	3	sanitätsdienst	3
bundesstaaten	3	staatsbürgerkunde	3
bundestagung	3	staatsrat	3
bürgerlicher	3	staatsräte	3
checkst	3	stapi	3
direktion	3	supportern	3
duselig	3	supports	3
eindrucksvolle	3	teamen	3
entscheidender	3	telefons	3
entscheider	3	tiefs	3
eventuelle	3	tiefstes	3
exekutieren	3	umgebungen	3
fluren	3	unternehmung	3
frankfurtern	3	urlaubern	3
gangs	3	verpflichtend	3
gehalts	3	viertels	3
gemeinschaftskunde	3	wettbewerbe	3
geschäftliches	3	wettbewerber	3
geschäftsführender	3	zähe	3
glücken	3	zentrales	3
gymnasiums	3	abschließendes	2
haberer	3	aktions	2
hage	3	anbot	2
hager	3	anboten	2
joppe	3	angebots	2
jungens	3	arbeitendes	2
kaders	3	arbeitern	2
kahle	3	aufsichtsräte	2
autofahrers	2	mehren	2
beinern	2	melanzani	2
berliners	2	ministern	2
blanken	2	mite	2
bluffen	2	mittagesse	2
bube	2	mobiltelefonen	2
bundeslandes	2	niederer	2
bundesligaen	2	offerte	2

bürgerlichen	2	orts	2
detailhandel	2	ortsvorsteher	2
direktoren	2	paradeiser	2
eindrücklich	2	parteipräsident	2
eindrücklichsten	2	pere	2
einzelhandels	2	pflichte	2
entscheidig	2	pinter	2
entscheidungs	2	regierungspräsident	2
fahrers	2	scharfem	2
faschingskrapfen	2	schärferen	2
festl	2	schreibtische	2
filete	2	schuldigung	2
fisolen	2	schwächeren	2
führerscheine	2	seeen	2
fuhrwerk	2	senatoren	2
fußen	2	spezie	2
gangstern	2	spitals	2
gehaltes	2	spontanste	2
geschäftes	2	stadtteilen	2
geschäftliche	2	standes	2
geschäftsführenden	2	stands	2
geschäftsleitung	2	stöß	2
goali	2	teamer	2
güsel	2	telefonischen	2
hagener	2	tiefem	2
hantieren	2	tieferer	2
hinschied	2	tiefster	2
hinsitzen	2	tolleren	2
huhns	2	topfen	2
kantonen	2	trasse	2
kiezen	2	urlaubes	2
kneiper	2	vatern	2
krankenhauses	2	vogerlsalat	2
kreuzes	2	werkes	2
kugele	2	wissenschafter	2
landkreise	2	wissenschaftlern	2
langsamste	2	würzige	2
langsamsten	2	würzigem	2
lenkern	2	zähen	2
letschert	2	zügigen	2
malocher	2		

D.2 Geographical Distribution of the Tweets

To simplify the table all the overseas territories were aggregated to the respective nation. For instance, Saint Martin and French Guiana were aggregated to France. This method was implemented for all nations equally.

Ranking	**Country**	**ISO Code**	**Frequency**
1	Germany	DE	304515
2	Austria	AT	21441
3	Switzerland	CH	19604

4	United States	US	11441
5	Brazil	BR	4106
6	Spain	ES	3542
7	United Kingdom	GB	2983
8	Indonesia	ID	2956
9	Turkey	TR	2899
10	France	FR	2785
11	Italy	IT	2670
12	Netherlands	NL	2356
13	Philippines	PH	1752
14	Belgium	BE	1157
15	Greece	GR	1139
16	Portugal	PT	898
17	Thailand	TH	822
18	Malaysia	MY	807
19	Sweden	SE	785
20	Canada	CA	752
21	Mexico	MX	716
22	Argentina	AR	686
23	India	IN	637
24	Russian Federation	RU	623
25	Japan	JP	595
26	Denmark	DK	584
27	China	CN	508
28	Poland	PL	507
29	Croatia	HR	503
30	United Arab Emirates	AE	456
31	Ireland	IE	450
32	New Zealand	NZ	411
33	Australia	AU	399
34	Serbia	RS	395
35	South Africa	ZA	383
36	Israel	IL	378
37	Czech Republic	CZ	357
38	Colombia	CO	331
39	Chile	CL	323
40	Nigeria	NG	311
41	Egypt	EG	310
42	Luxembourg	LU	307
43	Norway	NO	292
44	Saudi Arabia	SA	276
45	Hungary	HU	275
46	Singapore	SG	233
47	Pakistan	PK	231
48	Finland	FI	204
49	Malta	MT	202
50	South Korea	KR	198
51	Venezuela	VE	187
52	Ecuador	EC	182
53	Ukraine	UA	169
54	Iceland	IS	146
55	Peru	PE	127
56	Slovenia	SI	127
57	Ghana	GH	125
58	Guatemala	GT	121
59	Bulgaria	BG	119
60	Vietnam	VN	111
61	Bosnia and Herzegovina	BA	106

62	Nepal	NP	101
63	Dominican Republic	DO	100
64	Tunisia	TN	87
65	Panama	PA	85
66	Costa Rica	CR	84
67	Kenya	KE	82
68	Lebanon	LB	82
69	Romania	RO	79
70	Kuwait	KW	78
71	Morocco	MA	73
72	Paraguay	PY	69
73	Latvia	LV	64
74	Cyprus	CY	60
75	Qatar	QA	57
76	The Former Yugoslav Republic of Macedonia	MK	55
77	Uruguay	UY	54
78	Cambodia	KH	52
79	Belarus	BY	50
80	Georgia	GE	48
81	Estonia	EE	47
82	Sri Lanka	LK	46
83	Jamaica	JM	43
84	Liechtenstein	LI	43
85	Slovakia	SK	43
86	Cameroon	CM	40
87	Senegal	SN	39
88	El Salvador	SV	39
89	Palestinian Territory	PS	35
90	Tanzania	TZ	33
91	Lithuania	LT	32
92	Azerbaijan	AZ	31
93	Bahrain	BH	30
94	Bahamas	BS	29
95	Montenegro	ME	29
96	Barbados	BB	28
97	Jordan	JO	28
98	Iraq	IQ	27
99	Kazakhstan	KZ	26
100	Namibia	NA	26
101	Albania	AL	24
102	Algeria	DZ	22
103	Monaco	MC	22
104	Maldives	MV	22
105	Oman	OM	21
106	Zimbabwe	ZW	21
107	Vatican City	VA	20
108	Cabo Verde	CV	15
109	Curacao	CW	15
110	Honduras	HN	14
111	Trinidad and Tobago	TT	14
112	Laos	LA	13
113	Mongolia	MN	13
114	Nicaragua	NI	13
115	Bangladesh	BD	12
116	Mozambique	MZ	12
117	Moldova	MD	11
118	Angola	AO	10

119	Botswana	BW	10
120	Côte d'Ivoire	CI	10
121	Myanmar	MM	10
122	Uzbekistan	UZ	10
123	Yemen	YE	10
124	Armenia	AM	9
125	Libya	LY	9
126	Uganda	UG	9
127	Bolivia	BO	8
128	Cuba	CU	8
129	Iran	IR	8
130	Suriname	SR	8
131	Andorra	AD	7
132	Afghanistan	AF	7
133	Mauritius	MU	7
134	Congo DRC	CD	6
135	Antigua and Barbuda	AG	5
136	Bermuda	BM	5
137	Haiti	HT	5
138	Brunei Darussalam	BN	4
139	Congo	CG	4
140	Gabon	GA	4
141	Somalia	SO	4
142	Zambia	ZM	4
143	Belize	BZ	3
144	Dominica	DM	3
145	Gambia	GM	3
146	Jersey	JE	3
147	Lesotho	LS	3
148	Mali	ML	3
149	Malawi	MW	3
150	Sudan	SD	3
151	Sint Maarten	SX	3
152	Turks and Caicos Islands	TC	3
153	Saint Vincent and the Grenadines	VC	3
154	Benin	BJ	2
155	Fiji	FJ	2
156	Guernsey	GG	2
157	Guyana	GY	2
158	North Korea	KP	2
159	Saint Lucia	LC	2
160	Mauritania	MR	2
161	Syria	SY	2
162	Aruba	AW	1
163	Burkina Faso	BF	1
164	Djibouti	DJ	1
165	Ethiopia	ET	1
166	Faroe Islands	FO	1
167	Greenland	GL	1
168	Isle of Man	IM	1
169	British Indian Ocean Territory	IO	1
170	Cayman Islands	KY	1
171	Liberia	LR	1
172	Rwanda	RW	1
173	Solomon Islands	SB	1
174	Seychelles	SC	1
175	San Marino	SM	1
176	Swaziland	SZ	1

177	Anguilla	AI	0
178	Antarctica	AQ	0
179	Burundi	BI	0
180	Bhutan	BT	0
181	Bouvet Island	BV	0
182	Cocos Islands	CC	0
183	Central African Republic	CF	0
184	Cook Islands	CK	0
185	Christmas Island	CX	0
186	Eritrea	ER	0
187	Falkland Islands	FK	0
188	Micronesia	FM	0
189	Grenada	GD	0
190	Gibraltar	GI	0
191	Guinea	GN	0
192	Equatorial Guinea	GQ	0
193	South Georgia	GS	0
194	Guinea-Bissau	GW	0
195	Heard Island and McDonald Islands	HM	0
196	Kyrgyzstan	KG	0
197	Kiribati	KI	0
198	Comoros	KM	0
199	Saint Kitts and Nevis	KN	0
200	Madagascar	MG	0
201	Marshall Islands	MH	0
202	Montserrat	MS	0
203	Niger	NE	0
204	Norfolk Island	NF	0
205	Nauru	NR	0
206	New Zealand	NU	0
207	Papua New Guinea	PG	0
208	Pitcairn	PN	0
209	Palau	PW	0
210	Saint Helena	SH	0
211	Sierra Leone	SL	0
212	South Sudan	SS	0
213	Sao Tome and Principe	ST	0
214	Chad	TD	0
215	French Southern Territories	TF	0
216	Togo	TG	0
217	Tajikistan	TJ	0
218	Tokelau	TK	0
219	Timor-Leste	TL	0
220	Turkmenistan	TM	0
221	Tonga	TO	0
222	Tuvalu	TV	0
223	United States Minor Outlying Islands	UM	0
224	British Virgin Islands	VG	0
225	Vanuatu	VU	0
226	Samoa	WS	0

D.3 χ^2-test

Lemma	DE_obs	AT_obs	CH_obs	Total	DE_exp	AT_exp	CH_exp	χ^2
mit	95806	6233	6528	108567	95784.818	6703.496	6078.685	66.239
auf	87570	5286	5189	98045	86501.630	6053.813	5489.557	127.034
zum	29283	1813	2064	33160	29255.893	2047.472	1856.634	50.037
mehr	15041	1003	836	16880	14892.626	1042.260	945.114	15.554
na	7494	560	463	8517	7514.247	525.884	476.868	2.671
foto	6405	873	519	7797	6879.017	481.428	436.555	366.721
nein	7740	636	635	9011	7950.086	556.386	504.527	50.684
glück	4595	212	346	5153	4546.309	318.173	288.517	47.404
team	1911	150	187	2248	1983.331	138.803	125.866	33.234
schauen	3659	260	236	4155	3665.809	256.552	232.639	0.108
musik	3410	214	210	3834	3382.602	236.731	214.666	2.506
gang	644	37	80	761	671.403	46.988	42.609	36.055
gucken	3837	77	71	3985	3515.824	246.055	223.121	249.205
see	1642	830	224	2696	2378.585	166.465	150.950	2908.321
arbeit	3485	163	149	3797	3349.959	234.447	212.595	46.240
toll	3040	200	186	3426	3022.638	211.539	191.822	0.906
klasse	2977	131	93	3201	2824.129	197.647	179.225	72.231
freund	2574	183	99	2856	2519.748	176.344	159.908	24.619
berliner	2623	18	15	2656	2343.295	163.995	148.710	283.581
fest	2108	98	116	2322	2048.618	143.372	130.009	17.590
laufen	2594	160	170	2924	2579.742	180.543	163.715	2.658
land	1828	176	145	2149	1895.987	132.691	120.323	21.635
fan	1827	142	82	2051	1809.525	126.640	114.836	11.421
bundesliga	1237	37	12	1286	1134.592	79.404	72.003	81.892
markt	1904	143	51	2098	1850.991	129.542	117.467	40.526
per	1462	92	129	1683	1484.851	103.917	94.231	14.547
nee	2054	37	39	2130	1879.224	131.517	119.259	138.194
handy	2019	114	121	2254	1988.624	139.174	126.202	5.232
urlaub	1828	126	47	2001	1765.411	123.552	112.036	40.021
sitzen	1952	109	78	2139	1887.164	132.073	119.763	20.822
runde	1833	109	96	2038	1798.055	125.837	114.108	5.805
messe	1926	74	47	2047	1805.996	126.393	114.612	69.578
obwohl	1754	126	100	1980	1746.884	122.256	110.861	1.208
stand	1594	57	96	1747	1541.316	107.869	97.815	25.823
langsam	1660	94	91	1845	1627.778	113.920	103.302	5.586
frühstück	1476	91	49	1616	1425.740	99.780	90.480	21.561
tor	1622	38	49	1709	1507.790	105.523	95.687	74.637
junge	1469	85	105	1659	1463.677	102.435	92.888	4.566
vater	1423	64	74	1561	1377.215	96.384	87.401	14.458
final	228	20	58	306	269.973	18.894	17.133	104.070
mädchen	1265	72	32	1369	1207.820	84.529	76.651	30.574
wiener	209	933	11	1153	1017.251	71.192	64.557	11119.111
büro	1337	47	103	1487	1311.927	91.815	83.257	27.035
ort	1169	72	108	1349	1190.175	83.294	75.531	15.866
zeigen	1203	76	95	1374	1212.232	84.838	76.931	5.235
karte	1039	51	43	1133	999.606	69.957	63.437	13.274
sterben	1116	62	55	1233	1087.832	76.132	69.036	6.206
meer	670	36	44	750	661.698	46.309	41.993	2.495
arbeiten	1076	74	55	1205	1063.129	74.403	67.468	2.462
bein	334	19	28	381	336.143	23.525	21.332	2.968
tod	914	28	54	996	878.736	61.498	55.766	19.718
check	604	49	81	734	647.582	45.321	41.097	41.976
goal	38	3	17	58	51.171	3.581	3.247	61.725
krankenhaus	969	75	3	1047	923.731	64.647	58.622	56.652

nutzen	908	46	65	1019	899.028	62.918	57.054	5.745
angebot	975	29	36	1040	917.555	64.215	58.230	31.395
freunde	809	97	44	950	838.151	58.658	53.191	27.664
boden	825	46	62	933	823.153	57.608	52.239	4.167
frankfurter	932	7	4	943	831.975	58.226	52.799	102.195
amt	271	16	26	313	276.149	19.326	17.525	4.767
unbedingt	837	50	43	930	820.506	57.423	52.071	2.871
rock	476	29	16	521	459.660	32.169	29.171	6.840
bisher	812	36	32	880	776.393	54.336	49.271	13.875
tolle	770	39	50	859	757.865	53.039	48.096	3.986
lehrer	674	47	83	804	709.341	49.643	45.016	33.952
entscheidung	761	59	33	853	752.572	52.669	47.760	5.417
telefon	786	25	47	858	756.983	52.977	48.040	15.910
werk	451	18	24	493	434.956	30.440	27.603	6.146
aktion	741	39	44	824	726.986	50.878	46.136	3.142
staat	524	97	66	687	606.116	42.419	38.465	101.065
tief	662	36	52	750	661.698	46.309	41.993	4.680
finale	607	31	27	665	586.706	41.061	37.233	5.980
fans	634	36	26	696	614.056	42.975	38.969	6.096
kreuz	665	34	19	718	633.466	44.333	40.201	15.159
kosten	574	59	48	681	600.822	42.049	38.129	10.587
fotos	548	44	32	624	550.533	38.529	34.938	1.036
gymnasium	407	17	3	427	376.727	26.365	23.908	24.044
ferien	545	36	67	648	571.707	40.011	36.282	27.658
wirtschaft	554	51	56	661	583.177	40.814	37.010	13.747
freunden	525	56	24	605	533.770	37.356	33.874	12.328
meter	490	39	50	579	510.831	35.750	32.418	10.680
stimmen	536	41	50	627	553.180	38.714	35.106	6.988
kollegen	602	21	14	637	562.003	39.332	35.666	24.552
januar	554	10	69	633	558.473	39.085	35.442	53.454
bühne	581	26	27	634	559.356	39.146	35.498	7.287
aufs	536	33	21	590	520.536	36.430	33.034	5.166
burger	465	19	8	492	434.074	30.379	27.547	20.336
beitrag	511	36	46	593	523.183	36.615	33.202	5.227
bundestag	559	6	2	567	500.244	35.010	31.746	58.811
schlagen	502	36	22	560	494.068	34.577	31.354	2.977
bericht	456	41	43	540	476.423	33.342	30.235	8.024
mais	44	1	9	54	47.642	3.334	3.023	13.727
kreis	432	10	76	518	457.013	31.984	29.003	92.635
mittagessen	374	49	29	452	398.784	27.909	25.308	18.018
bildung	420	44	24	488	430.545	30.132	27.323	7.045
tollen	433	31	25	489	431.427	30.193	27.379	0.234
fahrrad	439	19	7	465	410.253	28.712	26.035	19.217
drücken	417	15	22	454	400.548	28.032	25.420	7.195
kollege	402	9	42	453	399.666	27.971	25.364	23.792
viertel	394	30	8	432	381.138	26.674	24.188	11.682
zuschauer	398	26	23	447	394.372	27.600	25.028	0.290
bürger	371	28	26	425	374.962	26.242	23.796	0.364
somit	329	34	63	426	375.845	26.303	23.852	72.345
unternehmen	292	30	33	355	313.204	21.920	19.877	13.079
spieltag	373	4	7	384	338.790	23.710	21.500	29.619
spontan	379	18	24	421	371.433	25.995	23.572	2.621
tolles	365	20	23	408	359.964	25.192	22.844	1.142
unterstützung	318	29	37	384	338.790	23.710	21.500	13.630
schwach	319	38	21	378	333.496	23.340	21.164	9.840
toller	331	29	23	383	337.907	23.648	21.444	1.465
steigen	284	29	29	342	301.734	21.117	19.149	9.053
match	146	11	23	180	158.808	11.114	10.078	17.602

kneipe	303	5	2	310	273.502	19.141	17.357	27.216
gemeinde	269	30	31	330	291.147	20.376	18.477	14.718
mädels	296	17	1	314	277.031	19.388	17.581	17.231
wirtshaus	261	30	9	300	264.679	18.524	16.797	10.781
minister	156	26	3	185	163.219	11.423	10.358	24.149
eren	154	1	1	156	137.633	9.632	8.734	16.531
regierung	195	45	12	252	222.331	15.560	14.110	59.378
gebäude	230	18	14	262	231.153	16.177	14.669	0.242
klub	94	15	16	125	110.283	7.718	6.999	20.851
wachsen	241	13	19	273	240.858	16.856	15.285	1.785
nackt	249	15	18	282	248.799	17.412	15.789	0.644
umgebung	228	19	17	264	232.918	16.301	14.781	0.884
treten	237	15	11	263	232.036	16.239	14.725	1.143
lauf	222	18	21	261	230.271	16.116	14.613	3.309
fahrer	225	19	12	256	225.860	15.807	14.333	1.028
karten	221	20	7	248	218.802	15.313	13.886	4.871
kirmes	245	0	1	246	217.037	15.189	13.774	30.638
innerhalb	217	22	12	251	221.448	15.498	14.054	3.117
jungen	205	16	9	230	202.921	14.201	12.878	1.417
dirndl	206	25	4	235	207.332	14.510	13.158	13.966
fuß	212	17	1	230	202.921	14.201	12.878	11.913
sahne	149	7	9	165	145.574	10.188	9.238	1.084
fraktion	229	9	5	243	214.390	15.004	13.606	8.841
geschäft	185	25	13	223	196.745	13.769	12.486	9.883
nieder	169	32	18	219	193.216	13.522	12.262	30.970
frühstücken	205	7	4	216	190.569	13.337	12.094	9.521
tore	174	13	11	198	174.688	12.226	11.086	0.052
bürgermeister	189	25	2	216	190.569	13.337	12.094	18.637
ahs	35	16	1	52	45.878	3.211	2.911	54.777
bezirk	91	113	9	213	187.922	13.152	11.926	808.756
speicher	173	5	18	196	172.924	12.102	10.974	8.666
schreibtisch	199	7	4	210	185.276	12.967	11.758	8.881
kategorie	164	8	12	184	162.337	11.361	10.302	1.291
professor	114	11	9	134	118.223	8.274	7.503	1.348
eventuell	164	19	15	198	174.688	12.226	11.086	5.790
pflicht	172	6	10	188	165.866	11.608	10.526	2.963
blocken	171	15	7	193	170.277	11.917	10.806	2.141
beine	167	6	14	187	164.983	11.546	10.470	3.879
scharf	142	16	15	173	152.632	10.682	9.686	6.303
arbeite	162	11	10	183	161.454	11.299	10.246	0.016
wettbewerb	137	9	34	180	158.808	11.114	10.078	60.178
bundes-regierung	168	12	1	181	159.690	11.176	10.134	8.726
mehrere	164	8	9	181	159.690	11.176	10.134	1.146
vorstand	172	9	5	186	164.101	11.485	10.414	3.732
mehrheit	140	14	17	171	150.867	10.558	9.574	7.664
ärger	156	8	8	172	151.750	10.620	9.630	1.041
senator	127	2	7	136	119.988	8.397	7.615	5.333
leiter	150	4	8	162	142.927	10.003	9.070	4.079
füße	155	11	4	170	149.985	10.497	9.518	3.391
kiez	176	0	0	176	155.279	10.867	9.854	23.487
entscheiden	146	8	7	161	142.045	9.941	9.014	0.939
fuss	90	5	36	131	115.577	8.089	7.335	118.868
laufe	132	8	14	154	135.869	9.509	8.622	3.703
tomaten	136	5	7	148	130.575	9.138	8.287	2.299
zuvor	138	11	5	154	135.869	9.509	8.622	1.789
mädel	141	7	3	151	133.222	9.324	8.455	4.552
flur	148	1	2	151	133.222	9.324	8.455	13.998

beinen	130	7	9	146	128.811	9.015	8.175	0.545
gehalten	124	5	8	137	120.870	8.459	7.671	1.510
behörden	136	0	6	142	125.282	8.768	7.951	10.163
höchstens	106	8	10	124	109.401	7.656	6.943	1.467
teams	95	8	6	109	96.167	6.730	6.103	0.255
ausschuss	118	5	3	126	111.165	7.780	7.055	3.744
support	100	4	10	114	100.578	7.039	6.383	3.365
blank	41	1	1	43	37.937	2.655	2.408	2.102
klären	123	3	4	130	114.694	8.027	7.279	5.226
schieben	117	6	5	128	112.930	7.903	7.167	1.260
kader	94	5	3	102	89.991	6.298	5.711	1.733
führerschein	115	8	3	126	111.165	7.780	7.055	2.469
fröhlich	86	7	8	101	89.109	6.236	5.655	1.174
pfannkuchen	112	0	3	115	101.460	7.101	6.439	10.032
lehrerin	93	14	5	112	98.814	6.915	6.271	7.857
füßen	102	6	1	109	96.167	6.730	6.103	4.700
umgang	90	9	12	111	97.931	6.854	6.215	6.699
huhn	83	4	10	97	85.580	5.989	5.431	4.582
parken	106	1	2	109	96.167	6.730	6.103	8.643
handys	97	6	6	109	96.167	6.730	6.103	0.088
betreiben	88	6	11	105	92.638	6.483	5.879	4.729
messer	91	4	7	102	89.991	6.298	5.711	1.141
abteilung	81	10	8	99	87.344	6.113	5.543	4.022
berichten	89	7	6	102	89.991	6.298	5.711	0.104
fakultät	82	26	0	108	95.285	6.668	6.047	63.940
insbesondere	84	3	8	95	83.815	5.866	5.319	2.752
weisen	86	3	6	95	83.815	5.866	5.319	1.544
filiale	91	11	1	103	90.873	6.360	5.767	7.326
abstimmen	75	8	9	92	81.168	5.681	5.151	4.292
quartier	35	8	15	58	51.171	3.581	3.247	53.096
kugel	68	12	11	91	80.286	5.619	5.095	15.970
autofahrer	79	8	6	93	82.051	5.742	5.207	1.122
schlapp	81	3	7	91	80.286	5.619	5.095	1.939
gehalt	82	5	3	90	79.404	5.557	5.039	0.966
bub	63	5	4	72	63.523	4.446	4.031	0.074
anwalt	74	5	2	81	71.463	5.001	4.535	1.507
bundesland	80	5	2	87	76.757	5.372	4.871	1.855
ausgang	75	4	7	86	75.875	5.310	4.815	1.325
hagen	73	2	2	77	67.934	4.754	4.311	3.212
motorrad	64	3	3	70	61.759	4.322	3.919	0.701

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.

Zurich, 30th June 2015

André Rodrigues