University of Zurich^UZH

From Text to Coordinates: Machine-Coding the Location of Historical Battles to Create a New Spatial Conflict Dataset

GEO 511 Master's Thesis

Author
Benjamin Füglister
15-708-183

Supervised by
Prof. Dr. Ross Purves

Faculty representative
Prof. Dr. Ross Purves

24.04.2020
Department of Geography, University of Zurich

UNIVERSITY OF ZURICH

MASTER'S THESIS

From Text to Coordinates: Machine-Coding the Location of Historical Battles to Create a New Spatial Conflict Dataset

Author:
Benjamin FÜGLISTER

Supervisor:
Prof. Dr. Ross PURVES

A thesis submitted in fulfillment of the requirements
for the degree of Master of Geography

in the

Geocomputation Unit
Department of Geography

April 24, 2020

Declaration of Authorship

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.

Signed: B. Fuglister

Date: 25.04.2020

UNIVERSITY OF ZURICH

Abstract

Faculty of Science
Department of Geography

Master of Geography

From Text to Coordinates: Machine-Coding the Location of Historical Battles to Create a New Spatial Conflict Dataset

by Benjamin FÜGLISTER

The spatial language we use is full of ambiguities. The *Battle of Kappel* in 1531 is just one example of a battle with a location that is full of geographic ambiguity. There are over 20 possible location candidates with the name *Kappel*. Therefore, geocoding, i.e. the assignment of geographic coordinates to data, must take this into account. This is of central importance in this thesis, as a new spatial dataset of historical battles is presented. The time from the 12th century onwards was considered with a strong focus on the European bellicosities. A system is introduced which contains a newly trained named entity model as its core. Together with a map-based approach, the system machine-codes thousands of battles from books and assigns coordinates to them. How important the inclusion of context information is, can be shown. For a majority of the battles, the spatial error achieved turns out to be the same as in existing conflict event datasets. This thesis illustrates that existing battle datasets are incomplete and that the creation of historical conflict data for the exploration of new research questions cannot be considered complete.

Acknowledgements

I would like to take this opportunity to thank all the people who have accompanied and supported me during the very intensive time in which this thesis was carried out. I would like to express special thanks to my supervisor, Prof. Dr. Ross Purves, who gave me excellent orientation and was always available for advice. Our meetings were always very pleasant. I would also like to thank Prof. Dr. Lars-Erik Cedermann, who gave me the opportunity to carry this project out. Through him I got such a deep insight into the creation of research data in a field that has always fascinated me. I would also like to thank the members of the group of International Conflict Research at ETH Zurich. In particular Luc Girardin for his technical expertise in the field of computer science and Dr. Carl Müller-Crepon for his extremely useful inputs.

To my family, especially to my parents, and my beloved partner I would like to express my infinite gratitude. Without you I would not be the open-minded personality I am today. I want my son to have a home as protected as I was privileged to enjoy. In addition, I would like to sincerely apologize to you, Lina, for the endless hours during which I nagged you about programming problems. May our future be more peaceful than the geocoded battles of this thesis!

Off to new adventures!

Contents

List of Figures

List of Tables

List of Abbreviations

GIS	**G**ographic **I**nformation **S**ystem
GIR	**G**eographic **I**nformation **R**etrieval
ML	**M**achine **L**earning
NER	**N**ame **E**ntity **R**ecognition
NLP	**N**atural **L**anguage **P**rocessing

Chapter 1

Introduction

In 1541, Süleyman I., Sultan of the Ottoman Empire, could successfully expand his empire and gained control over large parts of modern Hungary. The capturing of Buda (better known as modern Budapest), is one example where warfare preceded territorial conquest. For the next nearly 150 years, the city of Buda was under Turkish rule. A renewed siege by the Habsburgs started in June 1686 and allowed them to recapture the city of Buda, which marked also an end to the Turkish advance into central Europe. A year later, the Ottomans were defeated in today's Montenegro by the Venetians and, in 1688, the often embattled city of Belgrade returned to the Holy Roman Empire. However, this did not prevent the Ottomans from recapturing the city in October 1690. These events are only a few examples of the fact that the European continent is rich in historical battles and wars. The "Military Revolution", which is supposed to have taken place after the first takeover of Buda by the Ottomans, was first conceptualized by Michael Roberts in 1956. He argued that tactical transformations together with the increased resource demand of permanent armies resulted in an "increase in the authority of the state" (Roberts 1995, p. 26). This was then further developed by Charles Tilly and led to his much quoted analogy of *"state makes war and war makes states"*, presented in the publication of *The Formation of National States in Western Europe* (1985). Thus, the idea was spread that warfare is inherently linked to state formation. Nevertheless, there has been little empirical evidence for almost 40 years.
This master thesis does not presume to answer if Tilly's sentence and its implications are right. Rather, it uses sophisticated methods from the field of Geographic Information Retrieval (GIR) to create new spatial conflict data for the time reaching back to the 12th century. By assigning geographic coordinates to the mentioned battles as well as to thousands of others, which where fought spatially spread over the entire continent, this master thesis created entries for a spatial point dataset (see table 1.1 for a minimum example for the battles fought in Buda), which can be used together with other data for spatial analyses. This dataset can provide the basis for future investigations on the connection between state formation and warfare.

The wide use of spatial data and the inevitable respect of the geographical dimension for old as well as for new research questions has found its way into the field of Political Science, especially into Conflict Research. In the past years, huge efforts have been undertaken to create datasets which list conflict events together with attached geographical coordinates (e.g. Raleigh

war_name	battle_name	battle_location	latitude	longitude	battle_date_start	actors
Turkish-Habsburg War	*Siege of Buda*	*Buda*	*47.5*	*19.03*	*4.05.1541*	*Hapsburg, Ottomans*
Later Turkish-Habsburg War	*Siege of Buda*	*Buda*	*47.5*	*19.03*	*17.06.1686*	*Hapsburg, Ottomans*

TABLE 1.1: Example entries of a spatial conflict event dataset

et al. 2010; Sundberg and Melander 2013). However, those datasets only meet the demand of research projects concerned with questions about war and peace regarding this and the previous century. For important questions connected to warfare going beyond this time frame, the available data is not yet in a satisfying state.
Concise historic information about the described battles in modern Hungary as well as, for example, about the younger Bonapartist bellicosities can be found in dictionaries which list thousands of battles. The content of those books served as raw textual data for this master thesis to undertake the attempt to create a comprehensive dataset containing spatial information about historical battles. Claiming that they have created the most comprehensive dataset on historical European battles (1401 to 1900, 2477 battles), including geographical references, the deviation in the number of battles in the dataset of Iyigun, Nunn, and Qian (2017) from the probably first such dataset of Dincecco and Onorato (2016) (800 to 1799, 856 battles) are nevertheless conspicuous. The exact study areas, including the one of this thesis, as well as the used definition of Europe can be found in chapter 3 - Data, Study Area and Preprocessing. While Dincecco and Onorato (2016) have also compiled their dataset by hand, Iyigun, Nunn, and Qian (2017) mention a working time of over six years of manual coding. Only Dincecco and Onorato's 2016 seemingly less extensive dataset has been published and is available to date. As the creation of a spatial conflict dataset obviously leads to different datasets, although mostly based on the same source data[1], the presentation of a new approach, which is provided by this work, is certainly beneficial. Regarding the extensive human workload mentioned by Iyigun, Nunn, and Qian (2017) by manually compiling such a geospatial dataset, the present project relied on a computational approach to present a new spatial dataset. The main idea on how this is achieved is presented after the reader gets more information about the background of the project.

1.1 Context and Background of the Project

How are wars related to state formation? In which context do states emerge? and *How does nationalism change state properties?* These questions lie at the core of the Nationalist State Transformation and Conflict (NASTAC) project, which is funded by the European Research Council (ERC Advanced Grant 2017). The aim of the project is to gain new results to estimate the prospects of state territorial division or power sharing proposals as initiatives to make and preserve peace, especially in multi-ethnic states. The project is hosted by Prof. Lars-Erik Cederman and his team, building the International Conflict Research group at ETH Zurich. This master thesis is affiliated to the NASTAC project and fulfils its need for a spatial dataset containing historical

[1] Both Dincecco and Onorato (2016) and Iyigun, Nunn, and Qian (2017) used Clodfelter (2017)'s extensive dictionary about historical battles to read it by hand and write out each event manually.

European battles from the 15th century onwards as a first data piece to measure the connection between state formation and warfare.

1.2 Research Aim and Objectives

The objective of this master thesis is to find a machine-based method to connect the entries of the battle dictionaries, respectively the places where the battles took place, with their physical location on the ground - their coordinates. Even for us humans who have contextual knowledge at our disposal, this is not an easy task because of the ambiguity of geographical place names (see Smith and Crane (2001)). Figure 1.1 provides the reader with a quick overview of information that should be extracted from the dictionary entries. Those variables are requirements of the NASTAC project and are discussed later in this thesis. The general idea is to create a system which automatically extracts this information and distinguishes between the ambiguous place candidates. The general idea is illustrated by the *Battle of Kappel*, which was fought in 1531. There exist over 20 different places exist which are potential candidates regarding the location where the battle could have taken place. In which modern country was the battle fought? Did the battle take place in the German municipality *Kappel* in Rhineland-Palatinate? Or the *Kappel* in Baden-Württemberg? Or was it in Austria or France (for a list of all potential candidates of *Kappel*, see Appendix)?

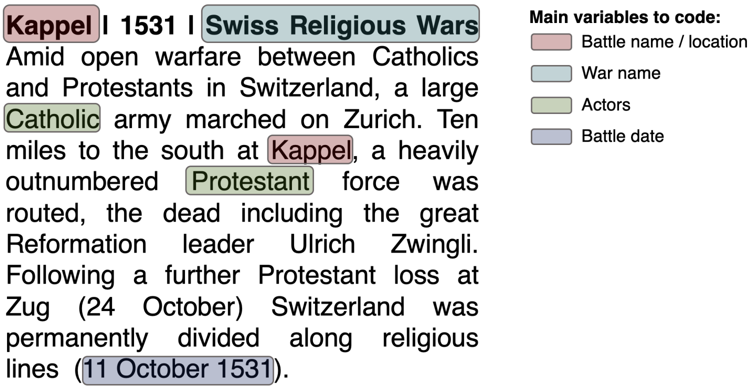

Kappel I 1531 I Swiss Religious Wars
Amid open warfare between Catholics and Protestants in Switzerland, a large Catholic army marched on Zurich. Ten miles to the south at Kappel, a heavily outnumbered Protestant force was routed, the dead including the great Reformation leader Ulrich Zwingli. Following a further Protestant loss at Zug (24 October) Switzerland was permanently divided along religious lines (11 October 1531).

FIGURE 1.1: Example dictionary entry for the *Battle of Kappel* from Jacques (2007) to show which information should be extracted.

Thanks to obtaining further information, in addition to the place name, we were able to figure out that the battle took place in Switzerland. But which of the the following *Kappel* is the right one: *Kappel* in the canton of St.Gallen, Zurich or Solothurn? Since the Protestant forces had to defend Zurich from the Catholic army, it becomes clear that it must be *Kappel* near Zurich. Thus, we can assign the coordinates *47.22811* latitude and *8.52727* longitude, which we derived from a geographical lexicon - a key component of most GIR systems. This is also called gazetteer and is a database containing place names with the corresponding coordinates (see Gazetteers). The question remains, how can a machine decide which *Kappel* the right one is? This especially with the findings of Gritta et al. (2018), who consider the georeferencing of short text passages, such as those to be processed in this project, to be much more difficult. Like Leetaru (2012), they regard the chance of a text having a larger context, a higher chance of getting the right coordinates. Since these coordinates are taken from a gazetteer, as mentioned above, the role of this key component is also critically questioned. These two questions will lead me through my master thesis, alongside my guiding research question, which will be clearly formulated in this chapter.

To give the reader a first insight into what results to expect from processing the battle entries and into how an entry in the resulting battle dataset could look like, a minimum output example of the *Battle of Kappel* can be found in table 1.2.

war_name	battle_name	battle_location	latitude	longitude	battle_date_start	actors
Swiss Religious War	*Battle of Kappel*	*Kappel*	*47.22811*	*8.52727*	*11.10.1531*	*Catholics, Protestants*

TABLE 1.2: Exemplary dataset entry for the *Battle of Kappel*

1.2.1 Research Question

The guiding research question of my master's thesis is:

How is it possible to preserve the information contained in battle dictionaries in order to create a machine-coded geospatial dataset with the highest possible spatial accuracy?

When building a system that makes it possible to locate the battles in space, the aim is to find out which the most important components are. Although the position of a gazetteer is regarded by the literature as being of such central importance, as we shall see later, there is no worldwide historical gazetteer. Therefore, it is to be found out whether one can concentrate at all on a worldwide coverage when working with texts with historical content. From this, I derive the following sub-question:

How exactly can historical events be located in space?

While the answer to this question can be measured, I am simultaneously interested in the battles that cannot be successfully assigned with coordinates. Therefore, reasons why battles cannot be successfully georeferenced, should be elaborated. Since this work is about georeferencing battles, which are described in rather short entries, this makes things more challenging, as Gritta et al. (2018) state. Without being deterred by this, I want to find out how texts with a rather limited context can be georeferenced. This leads to the following secondary sub-question:

How much context information must be available to resolve existing geographical ambiguity?

1.3 Thesis Structure

The structure of the thesis is close to the work stages that have been undertaken. Chapter 2 begins with presenting the most relevant literature from the field of Conflict Research in the context of conflict event data. Further, it discusses the challenges of assigning geographic coordinates to text while presenting the most important steps of this task and the related work from the field of GIR. The data used in this master thesis and the necessary preprocessing steps are presented in chapter 3. How I used NER together with a map-based approach to geocode battles by machine is explained in chapter 4, where the methodological strategy is introduced. In chapter 5, the results of this master thesis are presented, together with the results of the several validation deliverables. Chapter 6 discusses the limitations and potential weaknesses of the dataset and the way it was produced. Additionally, it critically questions the use, relevance and impact of the new dataset. In the final chapter 7 a conclusion is drawn and further work is announced.

Chapter 2

State of Research

Since the presented master thesis is highly interdisciplinary, this chapter serves to reflect on the most important work from the relevant fields and to include their results for the success of this project. For a graphic overview of the interdisciplinary overlap see Figure 2.1.

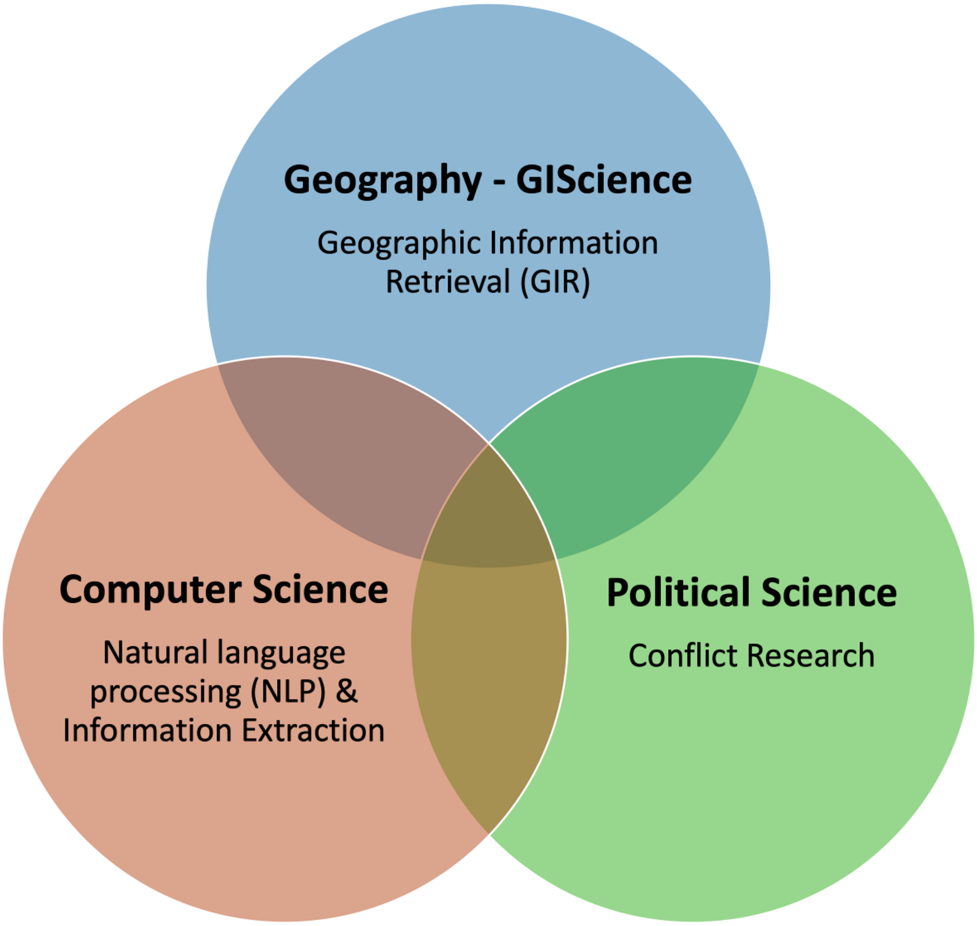

Geography - GIScience
Geographic Information Retrieval (GIR)
Computer Science
Natural language processing (NLP) & Information Extraction
Political Science
Conflict Research

FIGURE 2.1: Venn diagram which shows the different research areas involved in this thesis.

On the one hand there is the field of Conflict Research, which deals with causes and emergence, effects and cessation of conflicts. It focuses on all kind of wars and their battles as special cases of conflicts. Conflicts are described by Bonacker and Imbusch (1999) as social facts in which at least two parties are involved, and which are based on differences in the constellation of interests. Carl Clausewitz (1883) regarded war as a continuation of politics only by other means - where other means can be seen as organised violence by which people are killed. The definition of Clausewitz refers above all to wars in the period post 30-year war 1618-1648) until the time when the Weberanian[1] principle of state monopoly of power was over. For many of today's violent conflicts, a criminalization, commercialization and denationalisation of wars is visible (see Singer (2001) and Münkler (2002)), which undermines this definition. The definition also does not apply to the time before the Military Revolution as wars were waged with less of a systematic planning than afterwards, when war was a centralized state affair, with permanent armed forces and large armies - which at the same time led to a certain control of violence (Roberts 1995).

Geography as the discipline in which this master thesis is written, has also traditionally been concerned with wars and their effects for a long time. However, since this master thesis is completely focused on the quantitative research of warfare, the work of Political Geography will not be discussed, as it is mainly concerned with qualitative research. Nevertheless, the task which can be considered the main challenge of this work - adding coordinates to text - can be solved by a sub-discipline of geography, namely Geographic Information Retrieval (GIR). GIR has a strong overlap with the field of Computer Sciences and the diciplines of Information Extraction and Natural Language Processing (NLP). In the following sections, the mentioned disciplines and their most important works in connection with the collection of quantitative conflict data will be discussed. The main focus is on how to methodically generate quantitative conflict data from historical texts.

2.1 Conflict Research - From War Lists to Event Datasets

Quantitative research on warfare started essentially with lists of wars. With their seminal works "Statistics of Deadly Quarrels" by Lewis F. Richardson (1960) and "A study of war" by Quincy Wright (1942), these two researchers were the first who presented comprehensive lists of wars for statistical analysis. While Richardson reflected on how his data collection could be better represented, he made a suggestion with implications for his discipline which he probably could not yet estimate. He suggested to map "deadly quarrels" through their locations by means of their coordinates. Without access to Geographic Information System (GIS) technology, Richardson did not find his idea as purposeful. Incorporating Richardson's and Wright's preliminary work, lists of wars continued to be compiled. To systematically collect research data on wars, the Correlates of War Project (COW)[2] was initiated by Singer and Small (1982) in 1963. Although massive efforts have been made to improve the data about wars, Brecke (1999) noted in 1999, when publishing his own list of wars, that there is still too little systematic data available on wars. But what is more unsatisfactory about all these first data collections on wars is, from the point of view of a geographer, the sparse consideration and integration of the geographical aspects. For

[1] See Weber (1921)

[2] http://www.correlatesofwar.org/

example, Brecke's data was assigned to only one of 12 world regions without any further consideration of the spatial distribution of wars and in his most recent data publication one would have to add geographic variables oneself. That the geographic component was undoubtedly omitted has something to do with the fact that it is not entirely trivial to locate a phenomenon like a war in space. Where does the spatial extend of a war begin and where does it end? Such an attempt to record armed conflicts and their spatial extent was undertaken by Hallberg (2012). He codes center points and a corresponding radius to denote the spatial extent of conflict zones. I am not aware of any similar data on historical conflicts. However, individual conflict events are easier to locate in space. Contrary to earlier analyses in which wars were examined at the level of states, state of the art analyses are often described as disaggregated and work with individual observations of conflicts. *Disaggregation* simply means that the resolution of statistical data is increased. This leads to wars being split into their subunits of individual events or battles. Since such events are both temporally and spatially more limited than the entirety of the associated war, they can also be better represented as point datasets. The first steps to collect such event data were probably taken by Azar (1980) and were published as the Conflict and Peace Data Bank (COPDAB). However, this data was also not suitable for spatial analysis due to the lack of coordinates. To analyse spatial patterns in the subnational level of conflicts, Weidmann (2015) says that datasets with individual observations of conflicts are necessary. Nowadays, researchers can download this data from various platforms with the goal of compiling individual conflict events. These now include information on the longitude and latitude, which provide the approximate location of the events. The use of such conflict event data is widely spread in state of the art conflict research (e.g. Hulaman, Kathman, and Shannon 2014; Themnér and Wallensteen 2014; Urdal and Hoelscher 2012). Branch (2016) is critical of these approaches. For him the question is whether the accurate representation of political events and entities through GIS data is possible at all. He uses the term "measurement validity", which questions the representation by typical GIS representations of points, lines and polygons. For example, he criticizes the representation of pre-modern political entities by clearly defined polygon representations. Similarly, in this thesis one has to question whether the creation of a point dataset about historical battles can represent them in a suitable way. Clear is that a battle did not take place at a single point, but rather is a phenomenon with a certain spatial extension. But if a battle is represented by a polygon, the question would also arise as to where it should be limited. As Purves et al. (2018) recalls: "all measurements of location in geographical information systems are subject to error and hence to uncertainty", the question of the appropriate representation must be adapted to the intended use. Since this is the case for the NASTAC project, Branch's (2016) concerns can be put aside for the time being.

In the same style as contemporary[3] conflict datasets, attempts have also been made to create similar datasets for historical time periods (before 1900). These can be used to study the connection between warfare and state formation and nationalism, and appease the appetite for more data regarding analyses undertaken by studies such as the NASTAC project. It is not yet clear which methods will be used in the NASTAC project. Contemporary conflict data is often used

[3] By "contemporary" I mean datasets that try to cover the recent past. These represent in a way dynamic datasets, because they are updated at regular intervals. This is of course not the case for historical datasets, which can therefore be described as static.

by conflict researchers in spatial analysis, where the data points are aggregated to grid cells (Backer, Bhavnani, and Huth 2016). The underlying principle of those methods are regression analysis. For these methods, the exact georeferencing of a battle therefore results to which cell a battle is aggregated. The implications of the accuracy of georeferenced battle data on those methods will be briefly discussed at a later stage.

Before the historical battle datasets are explored in more detail, the two different ways in which contemporary conflict event datasets are constructed are examined. This is necessary because much more is known about the strengths and weaknesses of contemporary datasets than it is the case for the newly emerged historical datasets that have not yet been used too often. This also has to do with the fact that Iyigun, Nunn, and Qian (2017) have not yet published their dataset. In the following, it comes to the distinction between datasets created by hand and those, which tried to extract data by machine. An initial overview can be found in Table 2.1.

	ACLED	**UCDP GED**	**GDELT**	Dataset by **Dincecco and Onorato**	Dataset by **Iyigun, Nunn, and Qian**
Data type	Point dataset	Point dataset	Point dataset	Point dataset	Point dataset
Time coverage	1997 - 2019	1975 - 2018	1979 - today	800 - 1799	1401 - 1900
Spatial coverage	World	World	World	Europe	Europe
Creation method	Hand-coded	Hand-coded	Machine-coded	Hand-coded	Hand-coded
Number of entries	92,963	142,901	69+ M	856	2477
Source data	News	News	News	Battle dictionary	Battle dictionary
Created by	Raleigh et al.	Sundberg and Melander	Leetaru and Schrodt	Dincecco and Onorato	Iyigun, Nunn, and Qian
Published	2010	2013	2013	2016	Unpublished

TABLE 2.1: Comparison of the datasets, which are described in more detail in the following subchapters.

2.1.1 Conflict Event Data - Hand-Coded

The conflict event datasets that probably receive the most attention from the research community are the following: *Armed Conflict Location & Event Data Project (ACLED)* by Raleigh et al. (2010), Militarized Interstate Dispute Location (MIDLOC) by Braithwaite (2010) or *UCDP Georeferenced Event Dataset (GED)* by Sundberg and Melander (2013). These datasets represent the "new wave of disaggregated conflict data" explained by Gleditsch, Metternich, and Ruggeri (2014) in their article on the increase in data-oriented research projects in the *Journal of Peace Research*. Part of this "wave of disaggregation" is also the listing of spatial coordinates. While most of these datasets have a global spatial coverage, their temporal coverage is different. MIDLOC's data points come from events from the period 1816 to 2010, those from ACLED from 1997 to 2019 and UCDP GED from 1975 to 2018.

How are they created?

All datasets mentioned are based on media texts and were compiled each coded by hand. Such coding tasks are often performed by undergraduate students (see Gleditsch, Metternich, and Ruggeri (2014)), but the data from UCDP GED was only coded by experienced experts (see Stina (2019)). Hand coding means that the coder reads the media text containing the information about a conflict event, and writes out the necessary variables such as actors, number of fatalities, date etc., as well as information about the location of the event, which is then recorded as geographic coordinates. According to Stina (2019), the quality of the UCDP data, set by the coders, is monitored by numerous algorithms. In addition, coding is done with visualizations on maps to keep the quality of the geoinformation high.

Definition of an Event

While searching for raw data from global newswire reporting, the question arises which events to include in the dataset and which to leave out. How the stored events were conceptualized is different for all datasets. UCDP GED defines a conflict event as *"The incidence of the use of armed force by an organized actor against another organized actor, or against civilians, resulting in at least 1 direct death in either the best, low or high estimate categories at a specific location and for a specific temporal duration"* (Sundberg and Melander 2013, p. 524). To be included in the latest version of the UCDP GED dataset, an event had to be part of an armed conflict that claimed more than 25 lives each year. Therefore, in UCDP only events are contained, which also led to fatalities, in ACLED all events (or example also events with only injured persons) are recorded. In her article *In data we trust? A comparison of UCDP GED and ACLED conflict events datasets* Eck (2012) takes up the question of whether it makes sense to use a somewhat broader definition of a conflict event (like ACLED) in order to obtain more data points. She particularly points out that the more inclusive approach raises the problem that it is difficult to compare events within a dataset. She gives the example that the massacre of Srebrenica with more than 8000 victims receives the same weight as a deadly ambush of a sniper. This could be bypassed if the possibility exists to extract the number of fatalities from the text source. Then, the user can at least subdivide the dataset into their own selection and apply their own definition of event to a certain extent.

Selection and Description Bias

The fact that the datasets are almost entirely based on media texts, is certainly not an unproblematic aspect of these data collections. While Weidmann (2015) focuses on civil wars, he points out that international media rarely report completely objectively on conflicts. Thus, it cannot be assumed that all violent conflict events receive media coverage. Hence, Weidmann further points out that this "selection bias" can be based on two different aspects: On the one hand there are probably conflict events that occur in remote areas, far away from the world public, and on the other hand not every event is reported, even if there are fatalities, simply because it is apparently not "sensational enough". Regarding to Earl et al. (2004) the problem of the "description bias" is that if the event is taken up by the media, it must objectively reproduce all the variables sought and these are often not reproduced in their entirety. Certainly, information on conflict events can also be obtained from other raw data, for example from governmental or non-governmental organisations, but this information is often not available for each individual event.

Data Quality

In this subsection, special attention is paid to the quality of the spatial component, as this is central to the research question presented. When investigating the spatial accuracy of the UCDP

GED dataset, Weidmann (2015) found out that the further an event took place from the next settlement, the greater the error is concerning the coordinates set within the dataset. He found this out by comparing a subset of UCDP GED with a US Army dataset (SIGACTS) published via WikiLeaks. In addition, Weidman reports that when SIGACTS data is taken as ground truth and compared to UCDP GED, 80% of geocoded events have a spatial error of less than 50 kilometers. Eck (2012) also made statements on data quality while comparing the two datasets UCDP GED and ACLED. Regarding the handling of the coding for the spatial dimension of the ACLED dataset, Eck names two main problems: The first is simply the incorrect assignment of coordinates. Coders are apparently not sufficiently aware of the fact that when assigning coordinates, confusion can easily occur and wrong coordinates of locations with the same name are assigned. As Eck illustrates with examples, this leads to entries that are sometimes wrongly located by over 100 kilometres. UCDP GED seems to minimize this problem by trying to keep the data quality high through a triple checking process. It uses *SpatialKey*[4], a software to visualize geographical data, to detect possible errors graphically. In order for coders to achieve a high level of precision and to be able to assign the correct coordinates to ambiguous place names, Eck argues that high quality gazetteers are necessary on the one hand, and good map material on the other hand, so that the coders can locate the assigned places on maps. The second main problem is the incorrect assignment of geoprecision codes. Geoprecision codes are used to communicate uncertainties in the assignment of coordinates to the user. Integers are assigned, the higher the number the more inaccurate the data. Eck has studied the geoprecision codes of ACLED and found that they often present themselves better than the precision really is. Sometimes an exact coordinate is implied, although the source text does not actually say anything closer than "near town Y". A reliable geopresicion code would be helpful for many users to assess whether the data is useful for their research or not. For example, if one is interested in whether historical battles were fought in urban areas or in remote fields, it depends on how accurate the coordinates are.

2.1.2 Conflict Event Data - Historical Battle Datasets

In the same way that disaggregated spatial datasets can be used to address interesting research questions for the recent past, the creation of such datasets for the study of even older processes for which modern data is still scarce is expected to yield new research results. The already mentioned datasets of Iyigun, Nunn, and Qian (2017) and Dincecco and Onorato (2016) are the only ones known to me which have tried to record historical battles in the same style as the contemporary hand coded datasets presented above. Since the data situation for the creation of such datasets is uneven - the latest mass media is in no way comparable to the limited information about historical events - the compilation of such datasets is more difficult. Fortunately, historians have already taken over much of this work and produced far-reaching dictionaries on historical battles. Among the most important works are the books of the following authors: Jacques (2007), Clodfelter (2017), Sweetman (2004), Laffin (1986), and Harbottle (1904). These are all written in English. Both datasets, that of Iyigun, Nunn, and Qian as well as that of Dincecco and Onorato, were largely based on the battle dictionary of Clodfelter (2017). At the same time, as this work

[4] https://www.spatialkey.com/

appears to be extremely comprehensive, Clodfelter (2017) makes it clear that in a project such as his, which attempts to depict all the conflict events of the past since 1492, a selection must inevitably be made. Although Iyigun, Nunn, and Qian also used the Brecke[5] data collection mentioned above, it contains only war events and no battle data. While no data collection can be all-encompassing, it should be comprehensible that if one uses an existing selection, one is using its pre-selection. Even though not both datasets are published, the following sections will discuss how Iyigun, Nunn, and Qian and Dincecco and Onorato created their datasets by hand. It is clear that encoding historical battles involves similar problems as encoding contemporary conflict events, although there are additional hurdles.

Definition of a Battle

Just as one has to define which contemporary conflict events one want to take into account, one should think about which historical battles to code. But one should reflect beforehand on what a battle actually is. Iyigun, Nunn, and Qian define a battle as: "a location with conflict" (Iyigun, Nunn, and Qian 2017, p. 7). While Dinecco do not clearly define what they mean by a battle, they do state that they regard a battle as a disaggregated unit of a war. They assume that Clodfelter does not list all battles, but still contains the most important ones. At this point it must be mentioned, however, that it is extremely difficult to judge the importance of a battle. The historian Sir Edward Creasy is quoted here: "I need hardly remark that it is not the number of killed and wounded in a battle that determines its general historical importance" (Creasy 1851, p. 3). In the *Battle of Valmy*, for example, which took place about 100 kilometres east of Paris on September 20, 1792, only about 300 soldiers of the conflict's respective parties died (which is small compared to the more than 50,000 lives lost in the Battle of Waterloo 23 years later). Nevertheless, the battle won by the French Revolutionary Army continued to have an impact well beyond September 20, without the approximately 70,000 soldiers of the two armies actually meeting, since the only thing that happened was that the cannons fired for several hours. Battles may have gone down in history for many different reasons, be it because a large number of people lost their lives or because a significant army commander was killed. Not only the spatial component of a battle or the place where it was fought, as recorded by Iyigun, Nunn, and Qian's definition of a battle, but also the temporal component is an essential characteristic of a battle. For example, some battles can last for several days or even weeks and others were fought on a single day. Their data collection seems to be a reflection of the listing of battles of the mentioned historians, since they only brought their books by hand into a computer readable format to use them for analysis. By doing so, they took a broader approach and made a less restrictive selection of their battles. Thus, massacres and sieges were partly included in the data collection.

How Are They Created?

Both existing battle datasets were compiled by hand. Both Iyigun, Nunn, and Qian (2017) and Dincecco and Onorato (2016) state that they worked their way through Clodfelder's work and wrote out all the battles. Dincecco states that: "Historical accounts cannot pinpoint the exact

[5] Brecke (1999)

geographical locations of military conflict" (Dincecco and Onorato 2016, p. 9). With this consideration they justify that they approximate the locality by assigning the coordinates of the next known settlement to the battle. This is a feasible approach, since many historical battles are named after a nearby town anyway. Iyigun, Nunn, and Qian proceeded in the same way. They claim to have worked their way through Clodfelter over a period of six years, while georeferencing the respective battles. They state that a major barrier to this process was that given place names of battles had several possible locations and were therefore ambiguous. At the same time as Iyigun, Nunn, and Qian read out the localities, they also tried to gather information about the number of victims. However, they state that they only found reliable data for a good third of their battles, while they also consulted other sources.

Data Quality

As described in the introduction, both datasets contain a different number of entries, although they neither cover the same time periods, nor do they use the same stakeout of Europe. However, a closer comparison of the two datasets has not been possible up to now, because only one of the two datasets has been published. Clodfelter assumes that the quality of the historically handed down facts about battles decreases the further they took place in time and away from the "Western industrial world" they took place (Clodfelter 2017). For historical events, some of which date back several centuries, "selection bias" as well as "description bias" seem to be of immense significance. The observation has also been made that the aforementioned battle dictionaries have a tendency to describe a strongly Europe-oriented view. For example, in the case of war events outside of Europe, often only the commanders of the European armies are mentioned, and their activities are not reflected. Therefore, it can be assumed that these dictionaries primarily depict European/Western history and that this focus is also transferred to the datasets derived from them.

2.2 Conflict Data and Geographic Information Retrieval

As mentioned above, the creation of a dataset by hand and in particular the georeferencing of conflict data is very labor-intensive. For this reason, there were early efforts to leave this work to machines. An example, which fully relies on the machine coding of events and presents itself as a challenging one compared to the more established hand coded datasets like UCDP GED or ACLED, is GDELT (Global Data on Events, Location and Tone) by Leetaru and Schrodt (2013). Algorithms read the text source, in this case newspaper articles, and recognize the most important actors as well as the time and location of the mentioned event and classify it before it becomes an entry in the dataset. The fact that this kind of data processing is possible at all is due to the enormous progress in the field of computational linguistics, also known as natural language processing (NLP). Hirschberg and Manning (2015) name four key factors that have made this development possible. These are: (1) a tremendous increase in computing power, (2) the availability of large amounts of textual data, (3) a rapid development in the field of machine learning (ML), and (4) a broader understanding of the human language. Since space plays such a central role in our human language, this development was accompanied by the establishment of another branch of research, namely Geographic Information Retrievals (GIR). Larson defines GIR as: "an applied research area that combines aspects of DBMS research, User Interface Research, GIS research, and Information Retrieval research, ... concerned with indexing, searching, retrieving and browsing of geo-referenced information sources, and the design of systems to accomplish these tasks effectively and efficient" (Larson 1996, p. 81). Purves et al. (2018) adopt this definition to emphasize the aspect of processed textual data being "unstructured". That it makes sense to further intensify this field of research is shown by the example of Leetaru (2011), which implies that the use of geographic information is pervasive in our language. To take newspapers as an example, according to him, a location is mentioned every 200 to 300 words. Additionally, if one reads through the battle dictionaries one will notice that they are very rich in geographical references. Those references can also be named *toponyms*. A toponym is *the general name for any place or geographical entity* (United Nations. Dept. of Economic and Social Affairs 1974, p. 68). These range from names of villages, cities and states to geographical features such as mountains, rivers, lakes, moors or bays. How these toponyms are used to assign coordinates to the described battles will be explained in chapter 4 about the method.

While various approaches exist to extract these toponyms, to my knowledge they have not yet been used to geocode historical conflict events. This may have to do with the fact that we are still waiting for the quality of the automatic geocoding of conflict events to reach the desired accuracy. This statement is derived from the research of Hammond and Weidmann (2014), who investigated the quality of the geocoded data of GDELT and compared it to ACLED and UCDP GED. They merely give the quality of the gocoded data the rating "mediocre" and consider the dataset unsuitable for spatial analysis at the subnational level. However, one should not shy away from creating computer-based datasets, since not all research is done on a "micro-level". This is the level of research Hammond and Weidmann (2014) would like to use the new development of computer-based datasets for, but if one takes a closer look at current algorithms, it becomes clear that this is not yet possible. This is mainly due to the fact that the use of Gazetteers is an integral part of current GIR systems. Machine-coded coordinates are expressions of entries stored in gazetteers. If an event took place outside of a town, in a place that has no special name, these coordinates are not stored with an entry anywhere. Therefore, only coordinates of localities or other known places which have an entry in a gazetteer can be assigned. Otherwise the event must be assigned to the next settlement or similar. In this procedure, however, inaccuracy is inevitably included. Specific literature on gazetteers will be presented in an upcoming separate subchapter.

Georeferencing, which is part of GIR, is divided by McCurley into the following two main steps, namely (1.) "Geoparsing" as "the process of recognizing geographic context" and (2.) "Geocoding" which refers to "the process of assigning geographic coordinates" (McCurley 2001, p.2). The former is also known as "toponym recognition" whereas the latter is also known as "toponym resolution". Before I go into detail about the different approaches mentioned in the literature and the difficulties which need to be overcome, I will first discuss the diversity of spatial language and point out that spatial language does not simply end by defining a place name.

2.2.1 Not Only Toponyms - the Diversity of Spatial Language

Telling someone where something happens or has happened, describing where you can find something or where you want to go, is a fundamental part of our communication. The basic concept that often underlies this communication is the relationship between "located object" and "reference object" (Coventry and Garrod 2004). The former describes the phenomenon one wants to localize. In the case of this thesis it is the site of a historical battle, which is described by the explanations in the battle dictionaries. The location of a battle is often described by so-called "reference objects", which are often some kind of reference to geographical locations or toponyms. In the literature, other expressions like "figure" or "primary object" respectively "ground" or "secondary object" are also used (see Talmy (1983) and Langacker (1986)). The simplest such sentence with a spatial meaning consists of these two objects (nouns), a verb and the connecting spatial preposition (Coventry and Garrod 2004). Purves et al. (2018) see the weaknesses of many GIR applications in the fact that these refer primarily to the localization of the "reference object" and not beyond to the deeper lying relationship basis spanned by the connecting preposition.

According to Levinson (2003), there are three frames of reference into which spatial language can be divided. An example from each of the reference frames will follow. As one will see, the examples are rich in spatial relations, so only one frame per example is marked in color. The first one is called *intrinsic* and consists of putting the "located object" in relation to another object:

> "Taira Komemori marched north from Kyoto against the Minamoto rebel Yoshinaka to secure the fortress at Hiuchi, then met the full Minamoto army in the mountains at Kurikara, near Tsubata in Toyama, below the ridge at Tonamiyama." (Jacques 2007, p. 553)

The second is called *relative* and takes the perspective of the viewer:

> "On the right at Jena were 48,000 under the Prince of Hohenlohe, on the left the Duke of Brunswick had 63,000 at Auerstadt, eleven miles further north, and the remainder of the Prussian army was in the rear between Jena and Weimar." (Sweetman 2004, p. 87)

The third frame of reference is of type *absolute* and uses a coordinate system with predefined spatial directions:

> "During the Russian siege of Plevna, south of the Danube, Prince Alexander Konstantinovich Immeritinski was sent south against the powerful position at Loftche." (Jacques 2007, p. 595)

The same observation that Purves et al. (2018) considers insufficiently implemented for many GIR applications, namely that not enough attention is paid to this frames of reference, has also been denounced by Eck (2012) as a weakness of manually coded conflict datasets. She complains, for example, that the meaning of the preposition "near" and the resulting vagueness of spatial language is not sufficiently taken into account. This is particularly difficult, as different perceptions exist for the quantification of the term "near" and may also differ from context to context. This was shown by Derungs and Purves (2016) who investigated the different use of "near" in the context of different US cities. With their work they were also able to demonstrate that when using the construct *PLACE 1* "near" *PLACE 2*, the former has less population in the majority of cases and "near" thus establishes an asymmetrical relationship between places. For a deeper discussion of vague concepts see the work of Bennett (2010). In order to take account of the vagueness of the spatial language, geographic precision codes, which have already been mentioned, can at least call the user for attention regarding the use of affected data points. Since I am not aware of any machine coding systems that can accurately implement all the concepts mentioned, especially that of spatial vagueness, I continue to discuss the two main steps in georeferencing text data.

2.2.2 Geoparsing

Geoparsing describes the first step of georeferencing text. It is about identifying possible toponym candidates in a text. These can consist of simple place names or other geographical features such as mountains, lakes, canyons and so on. Leidner and Lieberman (2011) distinguish between three different approaches to recognize these geographical references in a text:

1. The simplest, though the most rudimentary, approach is to match the words of the text to be processed word by word with predefined lists. The geoparsing algorithms of GDELT are in principle based on such an approach[6].

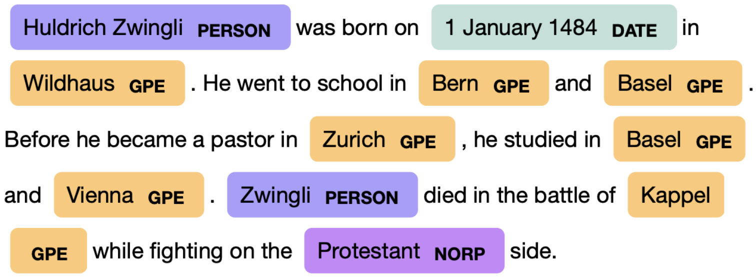
2. The second way to identify toponyms in text is to try to establish rules. This can be achieved by using regular expressions. Since the suffix "-kon" is a frequently used suffix of Swiss-German place names, one could, for example, set up the rule: [A-Z].+kon and would treat all words beginning with a capital letter and ending with "kon" as place names.

3. The third approach uses ML to calculate the probability whether a word with the given context represents a place name or not. Named Entity Recognition (NER), a sub-task of NLP which involves assigning text snippets to predefined categories such as places, persons or organisations, can be performed under the use of machine learning models. For this purpose, training data is used to find out in which context a place name typically occurs.

All these approaches have their advantages and disadvantages. The first two approaches have the advantage that they are rather simple and easy to implement and yet can deliver a respectable result. However, the creation of suitable lists can be very labor-intensive. With such a simple approach, as matching words with reference lists, it is also not possible to detect geo/non-geo ambiguities (McCurley 2001). These ambiguities are present when toponyms are used in the context of other things as well, such as the names of people or other frequently used terms (Amitay et al. 2004). This can be seen in the following example: not only does the first name of

[6] The approach chosen by GDELT to determine the location of the extracted events is described in the article by Leetaru (2012)

legendary Michael Jordan names several places around the globe, but his last name is also used as toponym for various populated places (i.a. Jordan [United States, Washington], Jordan [United States, Missouri]) and even for the middle east country of the Hashemite Kingdom of Jordan. Geoparsing approaches, which are based on ML, cope better with these ambiguities, although, with the disadvantage that training data must be available. However, Lample et al. (2016) was able to show that with the help of neural networks a significantly lesser amount of training data must be available. This bypasses the bottleneck mentioned above and makes NER an extremely competitive approach. Teitler et al. (2008), for example, has shown that the identification of geographical references is well feasible with NER. Named Entity Recognition describes the task of recognising "real-world objects" such as a person's name, places, organisations, dates or products in text. An example for labeled named entities can be found in Figure 2.2.



Huldrich Zwingli PERSON was born on 1 January 1484 DATE in Wildhaus GPE . He went to school in Bern GPE and Basel GPE . Before he became a pastor in Zurich GPE , he studied in Basel GPE and Vienna GPE . Zwingli PERSON died in the battle of Kappel GPE while fighting on the Protestant NORP side.

FIGURE 2.2: Visualization of named entities for example sentences. The *spaCy* function *displacy* was used for this. The labels mean the following: PERSON - Person names; NORP - Nationalities or religious or political groups; GPE - Countries, cities, states; DATE - Absolute or relative dates.

Gritta et al. (2018) compare and report the most important measurements of selected NER parsers[7]. For the validation of a NER system, the concepts of Precision, Recall and F-score are often used:

$$precision = \frac{Total\ number\ of\ entities\ recognised\ that\ are\ correct}{Total\ number\ of\ entities\ recognised}$$

$$recall = \frac{Total\ number\ of\ entities\ recognised\ that\ are\ correct}{Total\ number\ of\ entities\ present\ in\ the\ text}$$

$$F = 2 * \frac{precision\ * recall}{precision + recall}$$

The measure of *Precision* tells us how many of the recognized entities were actually recognized correctly. So, if there are seven locations (labeled with *"GPE"* by the *spaCy* library) mentioned in a battle entry and a NER model notices five of them, of which three are actually correct, then the *Precision* would be 3/5 and the *Recall* 3/7. The *F-score* combines these two measurements. The NER systems examined by Gritta et al. (2018) achieve *Recall* values between 68.6% and

[7] The following NER parsers were used for the comparison: NCRF++ by Yang and Zhang (2018), Spacy NLP/NER by Honnibal and Montani (2017) and Google Cloud NLP.

87.2%, *Precision* values between 79.9% and 91% and *F1* values between 74.9% and 88.6%. Nevertheless, NER also has its weaknesses, so Gritta et al. (2018) complained about the lack of understanding of metonymy of NER systems. Metonymy is defined by Lakoff and Johnson (1980) as "one entity to refer to another that is related to it" (p.35). Leveling and Hartrumpf (2008) show that 17.05% of location names regarding German textual data is used in this sense. In the sentence "Königsberg decided to invade Austria", the author is referring to the leadership of the late medieval land of Prussia and not to its physical lands on the Baltic Sea. When processing textual data of historical conflicts, this difficulty should be kept in mind, as the frequent occurrence of such uses can be expected. Another important point that Gritta et al. (2018) mention is the increased difficulty in accurately georeferencing shorter text passages over longer ones. This is because there is less context information available and therefore less geographic information to geocode the text. This step will be discussed below.

2.2.3 Geocoding

Once the corresponding toponyms in the text are correctly recognized, whether by means of a list comparison, with established rules or by means of ML, correct coordinates can be assigned. This process is not as trivial as it sounds, since the proportion of geo/geo ambiguities is very high. This type of ambiguity occurs when there are several possible place candidates with the same name in different locations. Smith and Crane (2001) for example found, while processing the Perseus digital library, that 92% of the discovered toponyms refer potentially to more than one place. In the example of the *Battle of Kappel* mentioned in the introduction, there are 26 different locations with the name *Kappel*, 8 different locations with the name *Zurich* and five different *Zugs* [8]. Which three of all these candidates did the author of the text had in mind? This has to be analyzed by means of geocoding. An additional complexity, which makes geocoding more difficult, is the fact that there are often several names for a location or that they are spelled slightly differently. These differences may be due to the fact that the place names are borrowed from other languages or a historical name is used (Smith and Crane 2001). Smith and Mann (2003) found out that whether one processes current news or historical texts, has a decisive influence on the success level of geocoding. Historical texts therefore do not only contain more geographical references, they also have a higher density of ambiguities. Smith and Crane (2001) point out that the proportion of ambiguity in place names varies also greatly from region to region. Europe has the lowest percentage of similar place names for different locations and thus coordinates (16.6%). With 57.1%, over half of all place names in North and Central America refer to more than one location. In Europe, however, the proportion of places with more than one name is comparatively low (18.2%), whereas places in Africa (27.0%) and Asia (32.7%) more often have multiple names.

To clarify the ambiguities, different approaches are utilized. In principle, however, all of them are concerned with finding the right candidate from a number of possible ones and assigning the corresponding coordinates. Buscaldi (2011) and Buscaldi and Rosso (2008b) divide these into three different categories:

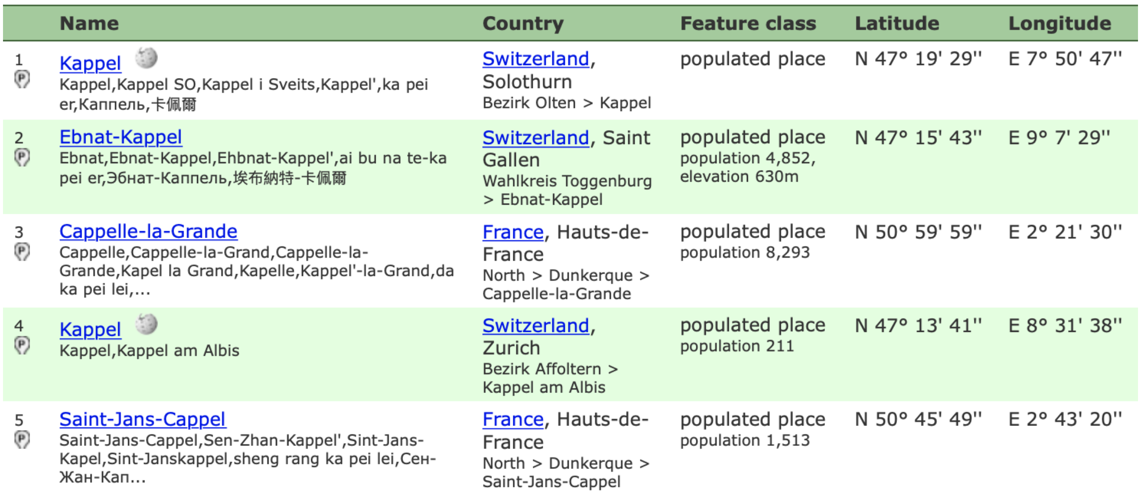
[8] All these locations, some of which have also multiple spellings, can be found either in GeoNames by Wick (2012) or The Getty Research Institute (2017) or in both. See Appendix

1. Map-based: This approach makes use of the spatial relationship of the toponyms candidates and calculates for example their spatial distances to identify the right locations.
2. Knowledge-based: This approach uses external knowledge such as population statistics, gazetteers or online encyclopedias.
3. Data-driven: Those approaches that use ML.

Smith and Crane (2001) were able to show that approaches, which are based on calculating the spatial distance concerning other toponyms mentioned in the text and selecting those candidates whose overall distance is the smallest, can achieve desired results. As Purves et al. (2018) explain, the principle of spatial autocorrelation underlies these approaches. Brunner and Purves (2008) showed that ambiguous toponyms often have an autocorrelated distribution. This should be reflected, particularly when working with a map-based method. The disadvantage of working with distances is that there must always be enough geographical references to be able to calculate those distances. Smith and Crane (2001), however, see the advantage of their approach in the fact that spatial distances are constant over time and therefore suitable for the geocoding of historical texts. Other approaches that implement additional knowledge (knowledge-based), on the other hand, would have to take temporal variation into account, since neither population statistics nor political boundaries are constant over time. Purves et al. (2018) point out that many successful approaches use additional context information to geocode successfully. Buscaldi and Magnini (2010) for example have shown that the place where newspaper publishers are located, which can be called "source", is an important contextual information to resolve ambiguities. How the use of additional context information can lead to better geocoding will be shown in the course of this work. Roller et al. (2012) show a data-driven approach and use all information of a text, including non-spatial context. An assumption that is often used in geocoding text is that the name of a geographical entity within a text is always the same. This is based on Gale, Church, and Yarowsky's (1992) principle of "one sense per discourse".

2.2.4 Gazetteers

This subsection discusses *gazetteers* as one of the most important components of most GIR systems. Hill (2000) defines a *gazetteers* as *"geospatial dictionaries of geographic names"* with the three main components *name, location* and *type*. The *name* simply refers to the name of the location, where a place can of course be known by several names. The coordinates, most often longitude and latitude values of a place, are usually stored under *location*. Under *type* is often noted what kind of geographic feature the entry describes. Examples are "populated place" or "river". For an exemplary gazetteer entry, Figure 2.3 shows the first five entries for the online query of GeoNames by Wick (2012), an often used world gazetteer, for the search word "Kappel". There are many different types of gazetteers. While some cover the whole world (e.g. GeoNames and Getty Thesaurus of Geographic Names (TGN)), others are limited to single countries. At this point, some important literature on worldwide gazetteers should be mentioned, as they will be applied in this project. GeoNames is one of the largest and most used gazetteers (Ahlers 2013). Provided under the *Creative Commons* license, users are invited to extend or improve the entries.



	Name	Country	Feature class	Latitude	Longitude
1	Kappel Kappel,Kappel SO,Kappel i Sveits,Kappel',ka pei er,Каппель,卡佩爾	Switzerland, Solothurn Bezirk Olten > Kappel	populated place	N 47° 19' 29''	E 7° 50' 47''
2	Ebnat-Kappel Ebnat,Ebnat-Kappel,Ehbnat-Kappel',ai bu na te-ka pei er,Эбнат-Каппель,埃布納特-卡佩爾	Switzerland, Saint Gallen Wahlkreis Toggenburg > Ebnat-Kappel	populated place population 4,852, elevation 630m	N 47° 15' 43''	E 9° 7' 29''
3	Cappelle-la-Grande Cappelle,Cappelle-la-Grand,Cappelle-la-Grande,Kapel la Grand,Kapelle,Kappel'-la-Grand,da ka pei lei,...	France, Hauts-de-France North > Dunkerque > Cappelle-la-Grande	populated place population 8,293	N 50° 59' 59''	E 2° 21' 30''
4	Kappel Kappel,Kappel am Albis	Switzerland, Zurich Bezirk Affoltern > Kappel am Albis	populated place population 211	N 47° 13' 41''	E 8° 31' 38''
5	Saint-Jans-Cappel Saint-Jans-Cappel,Sen-Zhan-Kappel',Sint-Jans-Kapel,Sint-Janskappel,sheng rang ka pei lei,Сен-Жан-Кап...	France, Hauts-de-France North > Dunkerque > Saint-Jans-Cappel	populated place population 1,513	N 50° 45' 49''	E 2° 43' 20''

FIGURE 2.3: The first five results for the online query of GeoNames with the search text "Kappel". The fourth entry indicates the place where the battle of Kappel took place in 1531.

While some historical place names can be found in the "alternative names" section, GeoNames does not actually maintain historical entries. To search for historical names, Getty Thesaurus of Geographic Names (TGN)[9] is better suited, although not as extensive as GeoNames. Up so far, the consideration of the time component has received only sparse attention from gazetteers. The development of place names is therefore not often reflected in worldwide gazetteers (Southall, Mostern, and Berman 2011). Even if Grossner, Janowicz, and Kessler (2016) consider the benefits of a historical world gazetteer to be high, there is no such gazetteer apart from the Peripleo initiative, which has so far only been of a limited extend.

Acheson, De Sabbata, and Purves (2017) have found that global gazetteers vary greatly in the quality of the data depending on the region. In their study, the investigation of the spatial distribution of gazetteer entries of TGN and GeoNames has revealed patterns which are unlikely to reflect reality. In North America and Europ, they also found a higher density of place names than on other continents. In some cases, major differences have already been found between neighbouring countries, which probably has more to do with the active contribution of users than with local realities. Ahlers (2013) was able to show similar results with his study, which revealed large differences in accuracy between countries. Manguinhas, Martins, and Borbinha (2008) mention the lack of historical information and the simple representation of places by means of centroid coordinates as the biggest limitation in using GeoNames to process texts with historical context. This is also the reason why machine coded event datasets are not expected to produce sufficient datasets for the use in micro-level studies, as Hammond and Weidmann (2014) hope. This is true at least as long as the assigned coordinates, read from a gazetteer, only represent the centroid of a location and not the exact position of an event. Micro-level analyses, which, for example, want to investigate whether conflict events tend to take place in urban areas or in the countryside, therefore do not find a suitable data basis.

[9] See The Getty Research Institute (2017)

Chapter 3

Data, Study Area and Preprocessing

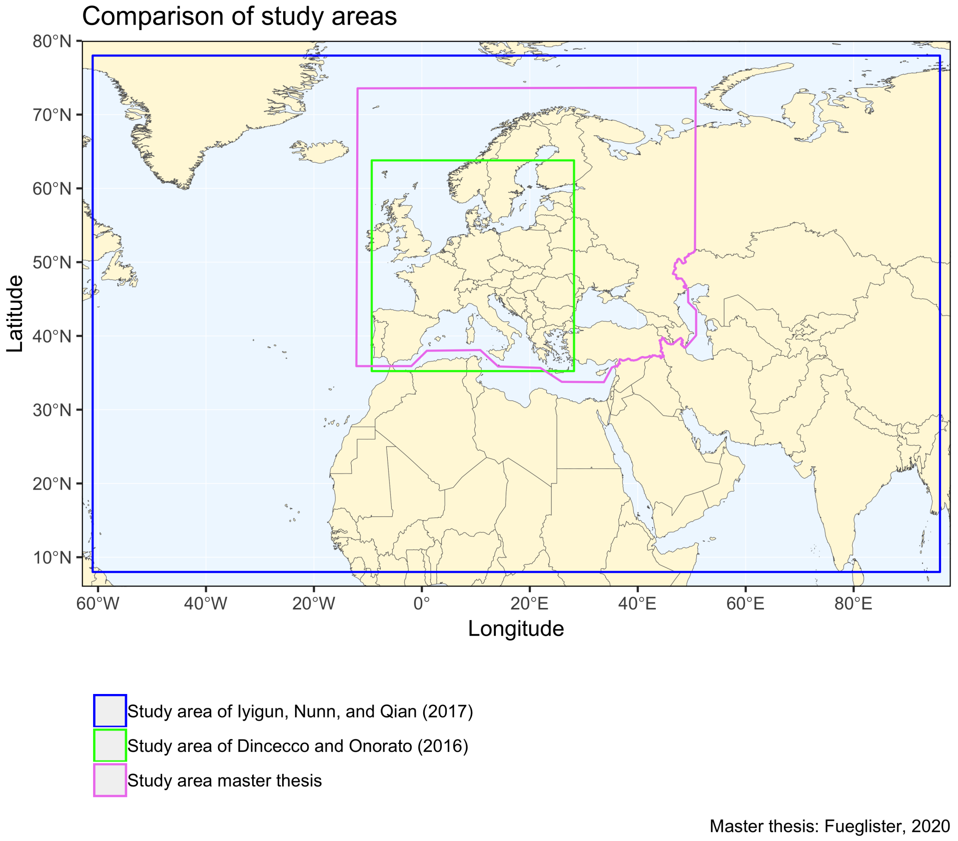
The requirements for a historical battle dataset from the NASTAC project in terms of temporal coverage are that at least the period between the Peace of Westphalia (1648) and the beginning of the World Wars should be covered. Spatially speaking, as many European historical battles as possible should be georeferenced. However, most battle dictionaries cover a longer period of time and are not only limited to Europe. For this reason, a more inclusive approach was initially adopted and an attempt was made to create a global dataset. The possibility to restrict the data set spatially at a later point in time remained in any case. In terms of temporal aspects, a similar approach was followed, with the attempt to code all battles contained in a dictionary. But this was also done with the idea that the dataset can be limited at any time. However, Europe should always be the main focus in order to meet the requirements of the NASTAC project. It was also assumed, that it would be easier to obtain a higher quality of georeferenced data for Europe than for other world regions. This hypothesis is based on the research of Smith and Crane (2001), who found a deeper ambiguity in European place names.

In the following, the dictionary used as the main source for the dataset will be introduced to the reader and explained why it was chosen. Subsequently, the preprocessing steps that have been carried out will be discussed.

3.1 Study Area

The exact study area of this master thesis can be seen in Figure 3.1. It is roughly oriented on a definition of Europe, which extends from the ural mountains to the Atlantic coast, including the area of present-day Turkey, which also provides the demarcation to the Middle East. The study area of Iyigun, Nunn, and Qian (2017) is much broader. It ranges from "approximately eight to 78 degrees latitude and from -61 to 96 degrees longitude" (Iyigun, Nunn, and Qian 2017, p. 14). Dincecco and Onorato (2016) also limited themselves to a narrower definition of Europe. Since their field of study is not specified exactly, the bounding box of the battles they coded was used as a reference area.

The exact definition of the study area is important because in the chapter "Results", I will compare the number of extracted and georeferenced battles of the different datasets.


Comparison of study areas
80°N
70°N
60°N
50°N
40°N
30°N
20°N
10°N
Latitude
60°W
40°W
20°W
0°
20°E
40°E
60°E
80°E
Longitude
Study area of Iyigun, Nunn, and Qian (2017)
Study area of Dincecco and Onorato (2016)
Study area master thesis

Master thesis: Fueglister, 2020

FIGURE 3.1: Comparison of study areas

3.2 Data

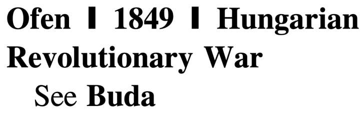
Since there are several dictionaries which list historical battles, it was necessary to select those works in advance, which, at first sight, have the desired spatial and temporal coverage and at the same time appear to be suitable for reading-in by computer. The English battle dictionary of Jacques (2007) with the title: *Dictionary of Battles and Sieges: A Guide to 8,500 Battles from Antiquity through the Twenty-first Century* served as the main source from which the historical battles were extracted. As a further source, the dictionary of Sweetman (2004) was considered suitable. First, the method presented in chapter 4 was applied to the work of Jacques and then its transferability to the work of Sweetman (2004) was tested in order to extract and code even more battles. As mentioned above, Iyigun, Nunn, and Qian, as well as Dincecco and Onorato, concentrated mainly on Clodfelter's dictionary when creating their hand-coded datasets. While Clodfelter and the statistics it contains enjoy wide acceptance and are also used by many studies as a data source (see e.g. Lacina and Gleditsch (2005) and Valentino, Huth, and Balch-Lindsay (2004)), the structure of the dictionary is not suitable for machine processing. Mentioned conflict events are described together in a continuous text. This makes it very difficult to extract individual events and to clearly assign the mentioned place names. The "one sense per discourse" principle of Gale, Church, and Yarowsky (1992) and the associated assumptions to facilitate the georeferencing of place names are therefore not valid. Since the internal structure in Jacques's dictionary, as well as that of Sweetman, is already organized at the level of individual battles, it was not necessary to recognise the events during the data processing. Therefore, full attention could be given to the geographical references contained therein. This does not contradict Purves et al.'s (Purves et al.) emphasis on the "unstructured" nature of GIR data, since the individual battles in these works are described in continuous text. What made the georeferencing of the battles seem easier was the fact that the short sections describing a battle in the dictionary of Jacques are organized by the title of the battle, which often contains the name of the place where the battle took place. But these still had to be disambiguated. How this was done will be explained in the next chapter.

Iyigun, Nunn, and Qian mention that they consider the quality of Jacques and Clodfelter to be equal, but mention that Clodfelter would contain 115 battles more for the period of 1400 to 1900. How this was found out is not clear but it is worth noting that Jacques and Clodfelter did not choose the same criteria for whether or not a battle should be mentioned as such in their work. Jacques states that his aim was to list all historical battles as comprehensively as possible and to not make any selection. Jacques also contains some naval and air battles. To georeference these would already be difficult to do by hand, let alone by machine. He has also listed some massacres, but as he says, their title is politically controversial anyway. What is also disputed among war historians is, according to Jacques, the basis on which a battle is now called a battle. In rare cases, Jacques also lists several war-like actions under one battle entry, which would be listed separately from others. This of course also allows for a certain discrepancy in the counting of battles. One criterion that Jacques has imposed on himself is that the battles considered must all have been submitted in writing and must be referenced by at least two different sources. These sources must also be consistent about date, happening, participants and output. That it makes sense to rely on several sources, is supported by the story of the *Battle of Tannenberg* (1914). As Jacques tells, the battle was actually fought in Frogenau, but the Germans, who were the winner of the battle, used their victory to forget the defeat of 1412 of the teutonic knights, which took place in Tannenberg. These are villages which are several kilometres apart.

As Jacques points out, he placed particular emphasis on researching battles beyond English literature so as not to incorporate this bias into his work.

While Jacques sees geography as the number one source for naming battles and states that: "the great majority of battles are named for their geographic location" (Jacques 2007, p. XV), this does not mean that there is a consensus on the naming of a battle. He gives several examples for this: a French battle that took place in 1914 is called the "*Battle of Guise*", but is known by Germans as the "*Schlacht bei St. Quentin*". Of course, this leads to different results when searching the included geographic references, for example in GeoNames. Also, battles are known by different names but only because there are different names for the same places. If one searches for these place references on GeoNames, one quickly notices that these place names have other ambiguities. Depending on which place name gives a battle its title, georeferencing can also be different and can make it even more difficult in case of ambiguities. What should have made georeferencing potentially easier, however, is the fact that Jacques has partially "translated" historical place names into today's designation. These are either given in brackets or in one of 2500 cross-references, which mainly provide other names for battles.

For the main entry of a battle, however, Jacques uses the place name that was used at the time of



Ofen | 1849 | Hungarian Revolutionary War
See **Buda**

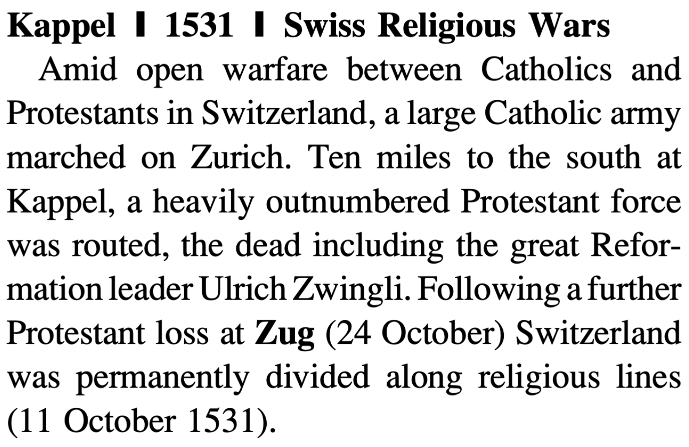
FIGURE 3.2: Example of a cross reference to another battle which took place in *Buda*, respectively *Budapest*, in German also called *Ofen*.

the battle. An example is the battle that was fought in the year 636 near the present city of Hilla, Iraq. The battle, which was the starting point for the takeover of Mesopotania by the Arabs, causing the Persians to retreat, is known by many names. Jacques mentions the following: Qadisiyya, Kadasiya, Kadesiah, Cadesia and Ghadesiyeh. Battle of Qadisiyya is the name used by Jacques for the main battle entry. The other names are listed as cross-references. That the naming is not a simple matter becomes apparent when one considers other languages as well. The place is known in German as Kadesia or under the Arabic spelling: al-Qādisīya or معركة القادسيّة, which would drive the whole thing to the extreme.
The fact that Jacques has often made translations from historical place names into today's English place names relativised the problem that no worldwide historical gazetteer exists. For this reason, an attempt was made to create the dataset using existing worldwide gazetteers. In the next chapter, one can also learn how these were created to fit specifically this project.
The battle dictionary by Sweetman has a very similar structure to that of Jacques. However, it only contains European battles from the end of antiquity to the end of the Second World War. Since the work is limited to Europe, the book was intended to be used purely to cross-check the specific needs of the NASTAC project for an even broader search of battles. As it turned out later, this data linkage is a difficult undertaking to realise.
Another point that justifies the use of Jacques battle dictionary as the main source for the data set, is the high amount of information that is relatively structured in the book and therefore easier to extract. This is not the case in Sweetman's book. While the book by Jacques is available as an e-book (PDF format), only a print version of Sweetman exists. In order to obtain a computer readable version, it was scanned in the DigiCenter[1] of the ETH Library. Since the two PDF files initally contained only image files and the text could not yet be extracted, an optical character recognition software was used first. Adobe Acrobat DC was used for this. The text was then ready for the preprocessing, as explained below. But before the reader can learn more about these steps, the internal structure of the books is described in more detail. This is necessary to understand the chosen method in chapter 4.

3.2.1 Internal Structure

As can be seen in Figure 3.3, using the example of the *Battle of Kappel*, the battle entries for Jaques are structured as follows: First, the name of the battle is mentioned. By random sampling, it has been found that in most cases this is the name of the locality where the battle took place. The name is followed by the year in which the battle took place. Also, part of the title is the name of the war in which the battle was fought.

[1] `https://www.library.ethz.ch/de/ms/DigiCenter`



Kappel ❙ 1531 ❙ Swiss Religious Wars

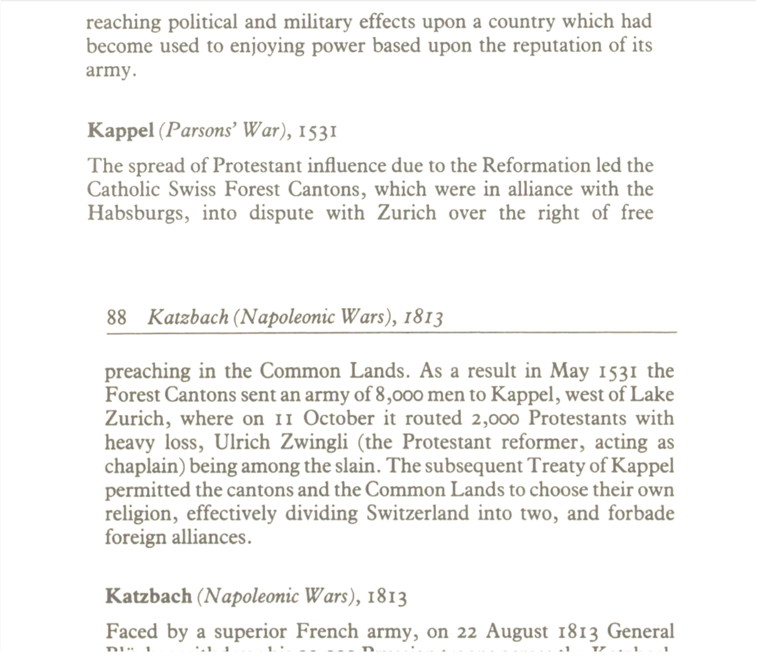
Amid open warfare between Catholics and Protestants in Switzerland, a large Catholic army marched on Zurich. Ten miles to the south at Kappel, a heavily outnumbered Protestant force was routed, the dead including the great Reformation leader Ulrich Zwingli. Following a further Protestant loss at **Zug** (24 October) Switzerland was permanently divided along religious lines (11 October 1531).

FIGURE 3.3: Example entry for the *Battle of Kappel* (Jacques 2007)

The title is followed by a descriptive text, which is often only a few sentences long. It describes how the battle took place in more detail, from where to where the armies moved and which commanders were important. In some cases, information is also given on the number of casualties and injured or prisoners. Who the battle won is also mentioned. Sometimes links are made to following battles or to battles that had already been fought. The battle names of those are written in bold. In the example, entry **Zug** emphasizes the before mentioned. Almost all entries in Jacques have the exact date of the battle in brackets at the end of the text. If a battle has lasted for a longer period, a time span is given. Otherwise, the date on which the battle took place is given. If this is not clear, sometimes only the month and the year are provided.
The dictionary entry by Sweetman on the *Battle of Kappel* is very similar to that of Jacques (see Figure 3.4). First, the name of the battle is given in the title. Then the name of the war is given in brackets, separated from the year by a comma. This information is very structured and can be extracted with appropriate methods. In the entries of Sweetman there is no structured date information. The details on when the battle took place, must therefore be taken from the continuous text.

3.2.2 Preprocessing

The goal of the books' preprocessing was to create a CSV file from the books that were initially in PDF format. Each row should represent one battle entry. The mentioned information which is available in a structured way (title of the battle, year, war name, date) should be extracted in a separate column. To achieve this, the text information first had to be read out of the PDF and then divided into the individual battle entries. This procedure of dividing texts into smaller pieces is known as *chunking*. Reading the text information from the PDF was possible with the module PDFminer by Shinyama (2019) from the Python programming language. This allowed extracting the text from the books in such a way that it could be further edited with this high-level programming language. For the whole project presented, Python was chosen as programming language. It is an open source software with a suitable environment for processing text data. As the processing of text data is widely used, the open source software community has been



reaching political and military effects upon a country which had become used to enjoying power based upon the reputation of its army.

Kappel *(Parsons' War)*, 1531

The spread of Protestant influence due to the Reformation led the Catholic Swiss Forest Cantons, which were in alliance with the Habsburgs, into dispute with Zurich over the right of free

88 *Katzbach (Napoleonic Wars), 1813*

preaching in the Common Lands. As a result in May 1531 the Forest Cantons sent an army of 8,000 men to Kappel, west of Lake Zurich, where on 11 October it routed 2,000 Protestants with heavy loss, Ulrich Zwingli (the Protestant reformer, acting as chaplain) being among the slain. The subsequent Treaty of Kappel permitted the cantons and the Common Lands to choose their own religion, effectively dividing Switzerland into two, and forbade foreign alliances.

Katzbach *(Napoleonic Wars)*, 1813

Faced by a superior French army, on 22 August 1813 General

FIGURE 3.4: Example entry for the *Battle of Kappel* (Sweetman 2004)

working on providing suitable tools for this purpose.

Even if it was possible to read the text with PDFminer, the structure had to be kept as best as possible. For Jacques's work, this meant that the text, which is in two columns, was read in accordingly. Initially, the system read beyond the line margin and sometimes mixed up the entries. Therefore, suitable settings had to be tried out. Another problem was that the headers and footers, which were also read in, had to be recognized and deleted. The *regex* module by Barnett (2020) proved to be helpful in this context and many other problems of the project. The module makes it possible to search for patterns in text. So in order to recognize the pattern of the header shown in Figure 3.4, for example, the regular expression shown in the code snippet in Figure 3.5 can be used.

```
headerPattern = re.findall(r’\d{2}\s\w*\s\(.*\),\s\d{4}’, text)
```

FIGURE 3.5: *regex* example to find header pattern from Figure 3.4

First, it recognizes a two-digit number, so the page number 88 is recognized. Then it identifies a space and any word following it. *Katzenbach* is thus recognized. Then a further space, as well as the bracket and its content is matched. The rest perceives the comma, the space and the year. Since there are of course also headers with a slightly different structure, which for example only contained a one-digit page number, several patterns had to be created. Due to the headers and footers often being followed by several line breaks, these are also used to indicate headers and footers to be deleted. The recognition of individual battle entries and their titles worked similarly. To recognize individual entries as the one for the *Battle of Kappel* (Figure 3.3) from Jacques, the regular expression shown in the code snippet in Figure 3.6 could be used.

Again, several different types of patterns had to be created to find all variations. For example, there are titles of battles that are composed of several words or dates at the end of the entries

```
battleEntry = re.findall(r'\n\w*\sy\s\d{4}\sy\s.*\(\d.*\d\)\.\n', text)
```

FIGURE 3.6: *regex* example to find battle entry pattern from Figure 3.3

that would not match the pattern of Figure 3.6. It proved to be very helpful that the vertical black line, which was used in the titles as a stylistic instrument, was read in as " y ". This made it easier to create suitable patterns that did not give other incorrect results from the text, which might accidentally have the same pattern.

Due to the use of regular expressions and the application of these patterns, it was possible to create two CSV files for the books, in which each row contained an extracted battle entry. In order to extract the structured information from the entries and store it in separate columns, various regular expression patterns were created. Before the CSV table entry looked like in Table 3.1 for the example of the *Battle of Kappel*, some steps had to be taken. For example, the date was converted to the format of dd.mm.yyyy.

battle_text	war_name	battle_title	battle_location_2nd	reference_modern_loc	battle_date_start	battle_date_end
Amid open warfare between Catholics and Protestants in Switzerland, a large Catholic army marched on Zurich. Ten miles to the south at Kappel, a heavily outnumbered Protestant force was routed, the dead including the great Reformation leader Ulrich Zwingli. Following a further Protestant loss at Zug (24 October) Switzerland was permanently divided along religious lines (11 October 1531).	*Swiss Religious War*	*Kappel*	*NA*	*NA*	*11.10.1531*	*11.10.1531*
Recovering after defeat at Zloczow and Soczawa, a reputed 200,000 Turks and Tatars under Ibrahim Shetan besieged John III Sobieski of Poland in his fortified camp at Zurawno (modern Zhuravno), on the Dniester east of Stryy. The Turks withdrew after costly losses, but they returned the following year to make a final attempt on the Ukraine at Chigirin (September–October 1676).	*Turkish Invasion of the Ukraine*	*Zurawno*	*Zurakow*	*Zhuravno*	*01.09.1676*	*01.10.1676*

TABLE 3.1: Example CSV row: preprocessed dictionary entry for the *Battle of Kappel* and the *Battle of Zurawno*. For the latter, it can be shown how several possible place names could be extracted (from the continuous text) or added (from the cross references) to a battle entry. Zuravno is the historical place name for the city of Zhuravno in today's Ukraine. Zurakow is another possible name given by Jaques.

Additionally hyphenated words like *"Refor- mation"* had to be merged to avoid distortions in later processing steps (due to lack of space, this is not shown in the given example). In addition, some letters which were incorrectly represented, whether due to an encoding error, the OCR process or errors in the process of extracting the text with PDFminer, were replaced. Examples are "*~o*" or "*’ s*" which should actually represent "*õ*" and "*’s*" respectively. It was also discovered that the OCR process resulted in the presence of several hyphens, all of which were made equal. As can be seen in the example of Sweetman, some battle entries were split in two by page breaks. Battle entries that were divided by line breaks also had to be merged accordingly, as this could negatively affect later steps.

As mentioned above (see Figure 3.2), Jaques has partly created cross references to other known variations of place names. The same applies to the use of historical place names. When they were used in continuous text, he put the nowadays common name of this place name in brackets after the historical names and called it *modern*. Both information could be extracted and assigned to the corresponding entries in such a way that, if available, there are several names for the location of a battle. This information was stored in the CSV tables under *battle_location_2nd* or *reference_modern_loc*. This data could only be taken from Jacques and not from Sweatman. Through pattern recognition, an attempt was also made to extract information on fatalities. However, since only very few entries contained such information, their extraction was not considered from early on in the project.

Chapter 4

Method

Now that the reader is familiar with the internal structure of the dictionaries and the first preprocessing steps have been introduced, the methodology of how the extracted battles are assigned to their coordinates will be presented. The method was chosen based on of the presented approaches, introduced in the second chapter. The method was selected to use as much contained spatial information as possible, which is mentioned in the battle entries, to disambiguate the battle place names.

4.1 General Overview

As noted above, several different candidate names can exist for a battle location, all of which can potentially contain ambiguities and therefore have multiple entries in gazetteers with different coordinates. Which place is meant by the authors? The aim of the method presented here is to determine this without any doubt through assigning coordinates.

While various implementations were examined, the core idea was the following: In the center stood a entity recognizer, which tried to recognize named entities. The main focus was on the recognition of all the geographical references mentioned, but also on the recognition of war actors. The former were used to disambiguate the place names from the title or the additional candidates (see Figure 5, battle_location_2nd and reference_modern_loc) using a map-based approach. For this purpose, the location names were compared with a gazetteer and the candidate with the shortest overall distance to the toponyms recognized by the entity recognizer was selected. Since the recognized toponyms may also had ambiguities, many candidates could appear. This method assumes that only places which are autocorrelated to the battle locations are mentioned in the battle entry text.

Again, the *Battle of Kappel* will serve as an example. The entity recognizer searches for all mentioned toponyms in the battle entry of Figure 3.3. If the system is able to identify all candidates correctly, these are the following: *Switzerland, Zurich, Kappel, Zug* and once more *Switzerland*. Duplicates have been removed from the list and place names that already appear in the title have been deleted. The list was then matched against a specially created gazetteer, which will be discussed later. As soon as a name was listed as a country in this gazetteer, it was treated as such and removed from the list, as it is not appropriate to measure distances from countries, as one does not know from which point to measure. In the example, *Zurich* and *Zug* remain as well as the place name *Kappel* from the title of the entry, which is to be disambiguated. A comparison of these three toponyms with the gazetteer resulted in a list of candidates, which can be viewed in Appendix A. The list contains 26 candidate locations for *Kappel*, eight for *Zurich* and five for *Zug*. The basic idea was now to calculate the distances of all combinations of each *Kappel* to one *Zug* and one *Zurich*. The *Kappel* which was part of the combination with the shortest overall distance was then chosen as the correct one. Figure 4.1 shows a map with all European ambiguous locations involved. The dots show those candidates who form the combination with the shortest overall distance, which is 21.73 km (Location Name, Lat/Long: *Kappel*, 47.22811/8.52727; *Zurich*, 47.36667/8.55; *Zug*, 47.17242/8.51745). All other candidates can therefore be rejected. Since there were many ambiguous place name candidates for some battles, and the calculation of all distances was very CPU intensive, candidates, which were certainly not the right ones, had to be sorted out first. How this was implemented will be explained once the reader has learned more about the Named Entity Recognizer and the gazetteer created specifically for this project.

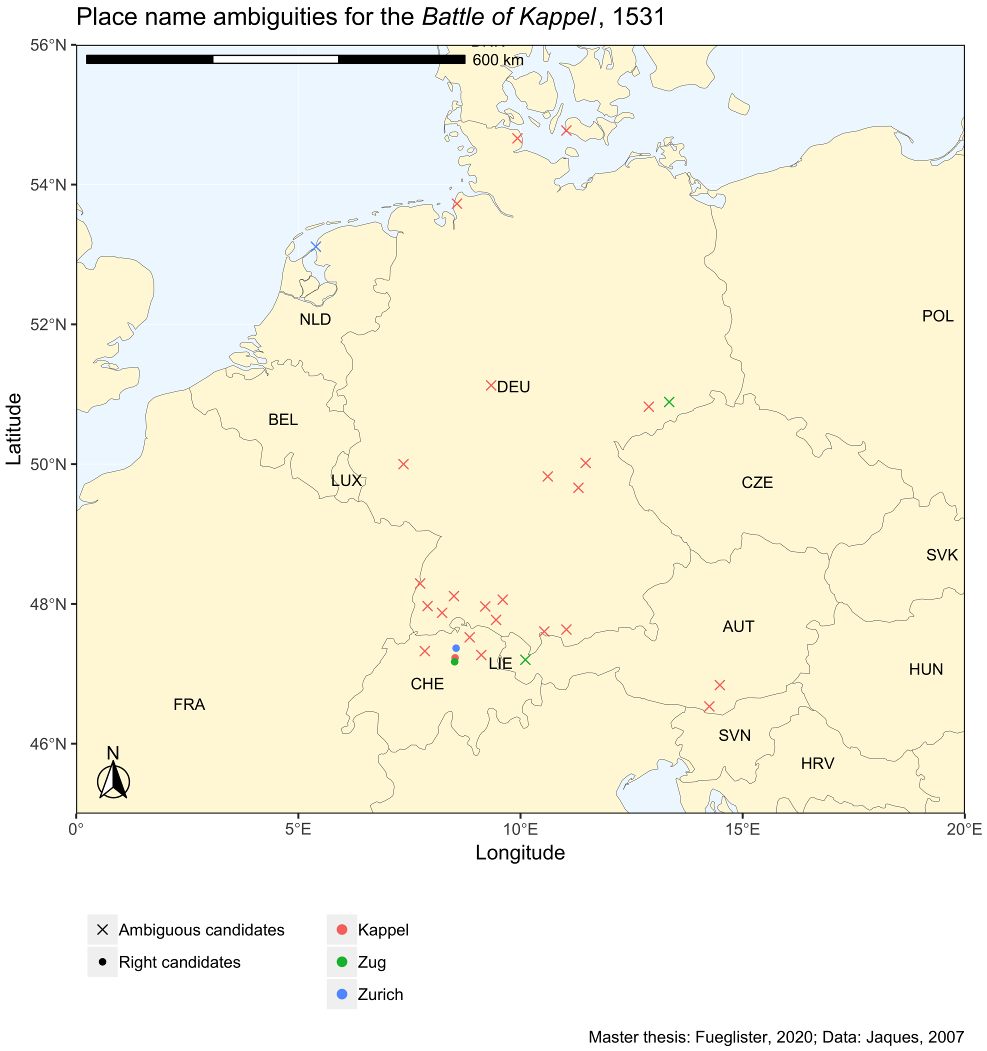


FIGURE 4.1: Map showing all European ambiguous locations for the entry of the *Battle of Kappel* from Jacques (2007)

4.2 Entity Recognizer

As mentioned above, an entity recognizer was at the heart of the system set up. The free, open-source library *spaCy* by Honnibal and Montani (2017) was used to implement such a recognizer. As discussed, it claims to deliver results in the range of other state of the art entity recognizers. The model needed to identify entities in text is based on a neural network architecture, which is an established approach in many NLP systems (Goldberg 2017). To get assigned labels like in Figure 2.2, a model is asked for a prediction. To be able to calculate with words, word embeddings, respectively their syntactic and semantic meaning, is determined from very large data sets. This approach was first implemented by Mikolov et al. (2013) and is widely used. *SpaCy* also calculates with these word vectors and provides several models for usage. The basis for these models was a very large text corpus created by Weischedel and Consortium (2013). This consists mainly of news texts. For this project the model *en_core_web_lg* was chosen as basis model. Since the processed data was not news texts and the type of language used may be somewhat different, this existing model was trained further to get better results. Applied on the battle dictionary from Jacques (2007), the performance of the *en_core_web_lg* model can be examined in Table 4.1.

NER statistics

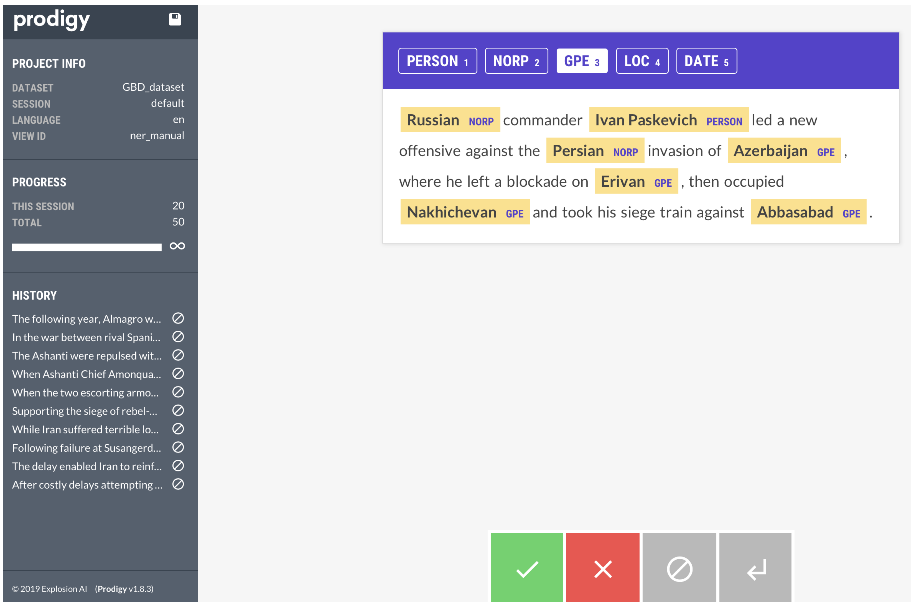
	GPE	*NORP*	*LOC*	*PERSON*	*DATE*
Precision	24.47	56.49	35.48	42.05	15.61
Recall	51.00	75.94	41.25	76.72	21.32
F-score	33.07	64.78	38.15	54.32	18.03

TABLE 4.1: NER statistics (in %) for the *en_core_web_lg* model of *spaCy*

As spaCy distinguishes between GPE (countries, cities, states) and LOC (non-GPE locations, mountain ranges, bodies of water), these two types of entities were selected to extract toponyms from the battle texts. The overall performance of the model applied to the battle data was rather low, suggesting there was room for improvement. Especially with regard to these two labels the training of a new model based on the existing one should lead to better results.

4.2.1 Training New Model with *prodigy*

The annotation software prodigy was used to improve the model and to create my own model tailored to the battle data. Prodigy is an annotation tool from Explosion AI, a company that has also been significantly involved in the development of *spaCy* and is therefore perfectly compatible. Prodigy was probably the missing piece in the coding pipeline of many NLP applications. Although the software is not free of charge, it closes the gap to generate efficiently training data for machine learning applications. To get training data, one has to label named entities in sentences with the corresponding labels. This can be very time consuming. But the environment provided by prodigy reduces this significantly. Figure 4.2 shows the interface where entities could be highlighted and therefore be stored as training data. Training data was collected based on the existing entity classes of the base model. However, I restricted myself to those entities I was interested in. These were the following: GPE (countries, cities, states) and LOC (non-GPE locations, mountain ranges, bodies of water), from which I hoped to get the necessary geographical information for the geocoding step. PERSON (Person names) have been considered as it could have been possible to get some exciting information about the battle actors. NORP (Nationalities or religious or political groups) was coded to get the involved actors from the battle entries. The entity class DATE (Absolute or relative dates) was also considered because not all date information could be extracted from the structured data extraction. See here for all entity classes of *spaCy*.



PROJECT INFO

DATASET GBD_dataset
SESSION default
LANGUAGE en
VIEW ID ner_manual

PROGRESS

THIS SESSION 20
TOTAL 50

HISTORY

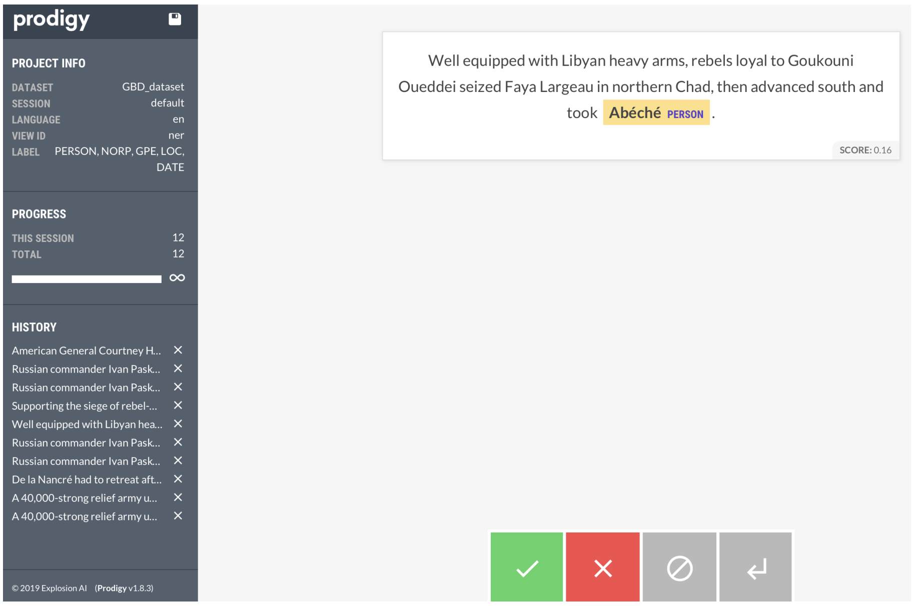
The following year, Almagro w... ⊘
In the war between rival Spani... ⊘
The Ashanti were repulsed wit... ⊘
When Ashanti Chief Amonqua... ⊘
When the two escorting armo... ⊘
Supporting the siege of rebel-... ⊘
While Iran suffered terrible lo... ⊘
Following failure at Susangerd... ⊘
The delay enabled Iran to reinf... ⊘
After costly delays attempting ... ⊘

© 2019 Explosion AI (Prodigy v1.8.3)

Russian NORP commander Ivan Paskevich PERSON led a new offensive against the Persian NORP invasion of Azerbaijan GPE , where he left a blockade on Erivan GPE , then occupied Nakhichevan GPE and took his siege train against Abbasabad GPE .

FIGURE 4.2: Prodigy interface, where the user manually highlights named entities and thereby generates training data

This was the first step, which can be called manual step. After a first basic set of training data was collected, my first own model could be trained. Afterwards, an additional way of annotating the texts was applied. The prodigy interface for this step can be inspected in Figure 4.3. The software now used the trained model to make suggestions for annotated labels to the user. The user could then decide if the labels given by the model were correct or not. This allowed a very efficient generation of even more training data. 512 sentences were annotated in total. The training data was then split in an evaluation (20%) dataset and a training dataset (80%). The extent to which the existing model has been improved is shown in chapter 5 (Results).



prodigy

PROJECT INFO

DATASET	GBD_dataset
SESSION	default
LANGUAGE	en
VIEW ID	ner
LABEL	PERSON, NORP, GPE, LOC, DATE

PROGRESS

THIS SESSION	12
TOTAL	12

∞

HISTORY

- American General Courtney H... ✕
- Russian commander Ivan Pask... ✕
- Russian commander Ivan Pask... ✕
- Supporting the siege of rebel-... ✕
- Well equipped with Libyan hea... ✕
- Russian commander Ivan Pask... ✕
- Russian commander Ivan Pask... ✕
- De la Nancré had to retreat aft... ✕
- A 40,000-strong relief army u... ✕
- A 40,000-strong relief army u... ✕

© 2019 Explosion AI (Prodigy v1.8.3)

Well equipped with Libyan heavy arms, rebels loyal to Goukouni Oueddei seized Faya Largeau in northern Chad, then advanced south and took Abéché PERSON .

SCORE: 0.16

FIGURE 4.3: Prodigy interface, which gives the user suggestions for possible annotations.

4.3 Gazetteer

A gazetteer had to be created so that distances could be calculated, as in the introductory example of the *Battle of Kappel*. Since, as mentioned above, there is no worldwide historical gazetteer yet, a separate gazetteer had to be created. As GeoNames contains probably the most place names, it was taken as basis. In order to cover a wider range of historical place names, Getty Thesaurus of Geographic Names (TGN) was also used and merged with GeoNames to form an even larger gazetteer. Since GeoNames provides the contained entries with different feature codes, it had to be decided which of them to consider. (To view the whole list of feature codes of GeoNames, see www.geonames.org/codes). It was taken into account that no area-like features were included. For those features it is unclear from which point to measure a distance. Administrative units were therefore not considered in the same way as forests, deserts or lakes. A similar procedure was chosen for the integration of the TGN, from which all inhabited places were taken. With this half a million additional locations a gazetteer could be created, which contained more than 12 million entries worldwide. GeoNames also contains some entries with the feature code *BTL battlefield*. However, these are limited in number and have been omitted to avoid being influenced by another kind of georeferencing of battles. To avoid duplicates and to make sure that a physical location only appeared once in my gazetteer, every time I added a new place name, it was compared as if such an entry already existed. If a place with the same spelling was already present, it was checked if these two places were more than 15 km apart. If this was the case, they were evaluated as geo/geo ambiguities and both were listed under the same place name. If the distance was smaller, the entry was not considered and it was assumed that it was the same location. Since GeoNames partly lists several alternative names and also lists one in *ascii* normalized format, a gazetteer was created, which stores separate entries for each of these variations in the spelling of the place name of its alternative name. Therefore, if a search was made for the hometown of the university of Zurich, The Swiss city of *Zurich*, the entries listed in Table 4.2 were returned (the second entry shows the correct one). But if one searched for the common Swiss spelling with an umlaut, only one *Zürich* was returned, which is according to the used gazetteers the only one written with an umlaut - the Swiss *Zürich* (Table 4.3).

	Place name	Country	Alternative names	Latitude	Longitude
1	Zurich	CA	None	43.421630	-81.627730
2	Zurich	CH	Cirihe,Cirikh,Ciurichas,Cjurikh,Cjurikh khot,C...	47.366670	8.550000
3	Zurich	DZ	Sidi Amar,Zurich	36.542120	2.305400
4	Zurich	NL	Surig	53.111340	5.394370
5	Zurich	US	None	39.234450	-99.438160
6	Zurich	US	None	43.150620	-77.043300
7	Zurich	US	Alvord,Zurich	37.182710	-118.260100
8	Zurich	US	None	48.584440	-109.030440

TABLE 4.2: Search result for the place name *Zurich*

	Place name	Country	Alternative names	Latitude	Longitude
1	Zürich	CH	Cirihe,Cirikh,Ciurichas,Cjurikh,Cjurikh khot,C...	47.366670	8.550000

TABLE 4.3: Search result for the place name *Zürich*

4.4 Precision Codes

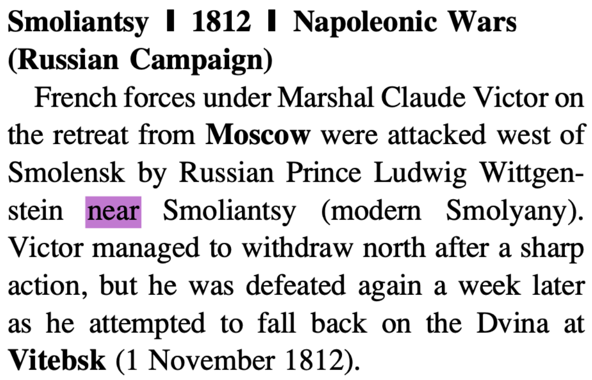
Before the battles could be geocoded in space using additional context (see the next section), two different precision codes were added to the dataset.

4.4.1 "*battle_date_precision*"

Since not all dates, which were generated by the structured data extraction, were exact to the day, a measure was introduced to communicate this imprecision to the user. If the exact date was known (e.g. 11.10.1531), the value 1 was assigned under *battle_date_precision*. If only for example "September-October 1676" could be extracted, the first day of the month (01.09.1676 and 01.10.1676) was saved and the precision code was set to 2. If only the year was known, then only this was considered and a 3 was given under *battle_date_precision*, for the least precise dates.

4.4.2 "*battle_loc_vagueness*"

In order to assign exact coordinates to the battles at all, one had to assume that the battles were actually fought in the locations listed in the titles. But this was certainly not always the case. Sometimes Jacques (2007) even stated that the battle took place only "near" or "southwest of" the assumed battle location. An example is the entry about the *Battle of Smoliantsy* where Napoleonic troops returned from Russia.



Smoliantsy I 1812 I Napoleonic Wars (Russian Campaign)

French forces under Marshal Claude Victor on the retreat from **Moscow** were attacked west of Smolensk by Russian Prince Ludwig Wittgenstein near Smoliantsy (modern Smolyany). Victor managed to withdraw north after a sharp action, but he was defeated again a week later as he attempted to fall back on the Dvina at **Vitebsk** (1 November 1812).

FIGURE 4.4: Example of a battle described by spatial vagueness.

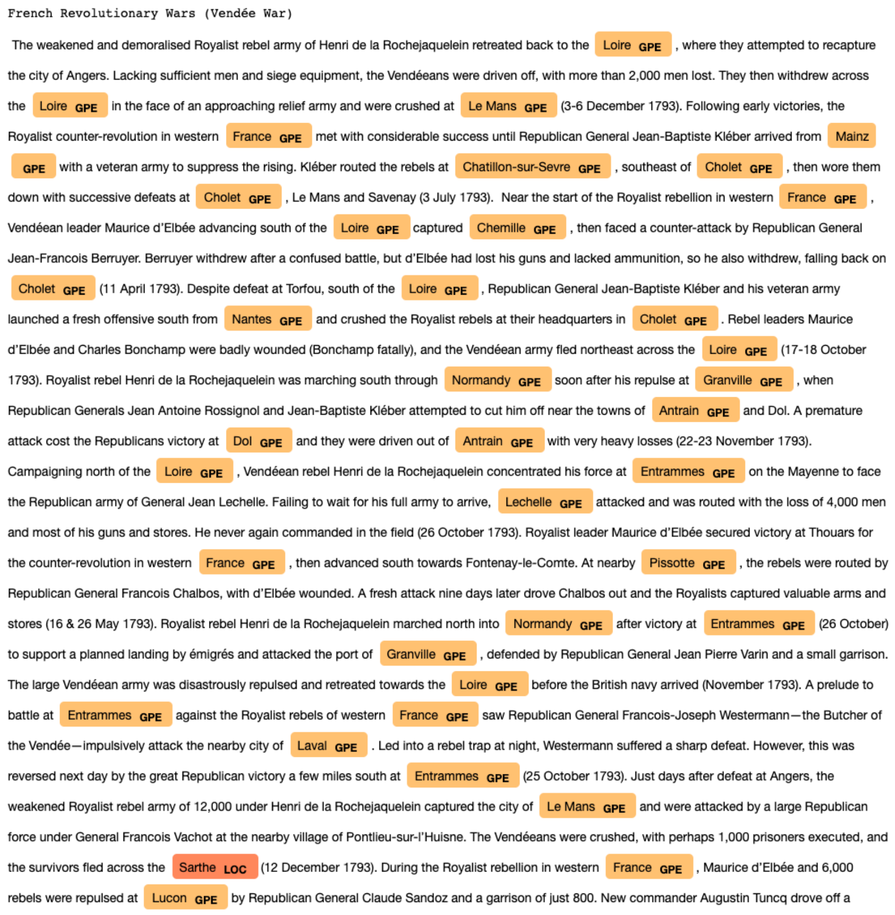
This spatial vagueness was searched by using regex. In Figure 4.4 one gets to know for example, that the battle did not actually took place in the named village of *Smoliantsy* itself. To find similar cases, where spatial vagueness was involved and therefore influenced the geocoding accuracy, each battle description was checked if the battle location or its additional candidates were mentioned together with "near" (see regex in Figure 4.5) or one of the cardinal directions "north" or "northeast" and so forth. If this was the case, a integer value was assigned (1: "near", 2: cardinal direction e.g. "north of").

```
re.search(r" near " + battle_location, battle_description,re.DOTALL)
```

FIGURE 4.5: *regex* example to find involved spatial vagueness

4.5 Use of War Context

In the first version of my map-based approach I created all combinations of the different candidate pots. For the *Battle of Kappel* example and the gazetteer results (see table in appendix A) this means that *1. Kappel* was calculated together with *1. Zurich* and *1. Zug*. Then *1. Kappel* with *1. Zurich* and *2. Zug* respectively *1. Kappel* with *1. Zurich* and *3. Zug* and so on. With a small number of possible location candidates involved, this was no problem and gave some good results. However, if several toponyms were recognized in the text and each of them had several geo/geo ambiguous candidates, there were simply too many combinations to calculate them efficiently. Inspired by Buscaldi and Rosso (2008a), more context was included and unlikely candidates were sorted out early on. To achieve this, all battles fought in the same war were grouped together and the text entries were merged into one large war text. This would imply that the *Battle of Kappel* was joined together with the *Battle of Zug*, which was also fought in the *Swiss Religious War*. Since this example does not contain too much ambiguity, I demonstrate the approach applied to the battles which together formed the *Vandée War* fought in de *French Revolutioanry Wars* from 1793 until 1795. The war is described in Jacques (2007) by the aggregate of all the individual battles. An excerpt of the text, which consists of all the text descriptions of the individual battles, can be read in Figure 4.6 (see Appendix A for the full text).



French Revolutionary Wars (Vendée War)

The weakened and demoralised Royalist rebel army of Henri de la Rochejaquelein retreated back to the Loire GPE , where they attempted to recapture the city of Angers. Lacking sufficient men and siege equipment, the Vendéeans were driven off, with more than 2,000 men lost. They then withdrew across the Loire GPE in the face of an approaching relief army and were crushed at Le Mans GPE (3-6 December 1793). Following early victories, the Royalist counter-revolution in western France GPE met with considerable success until Republican General Jean-Baptiste Kléber arrived from Mainz GPE with a veteran army to suppress the rising. Kléber routed the rebels at Chatillon-sur-Sevre GPE , southeast of Cholet GPE , then wore them down with successive defeats at Cholet GPE , Le Mans and Savenay (3 July 1793). Near the start of the Royalist rebellion in western France GPE , Vendéean leader Maurice d'Elbée advancing south of the Loire GPE captured Chemille GPE , then faced a counter-attack by Republican General Jean-Francois Berruyer. Berruyer withdrew after a confused battle, but d'Elbée had lost his guns and lacked ammunition, so he also withdrew, falling back on Cholet GPE (11 April 1793). Despite defeat at Torfou, south of the Loire GPE , Republican General Jean-Baptiste Kléber and his veteran army launched a fresh offensive south from Nantes GPE and crushed the Royalist rebels at their headquarters in Cholet GPE . Rebel leaders Maurice d'Elbée and Charles Bonchamp were badly wounded (Bonchamp fatally), and the Vendéean army fled northeast across the Loire GPE (17-18 October 1793). Royalist rebel Henri de la Rochejaquelein was marching south through Normandy GPE soon after his repulse at Granville GPE , when Republican Generals Jean Antoine Rossignol and Jean-Baptiste Kléber attempted to cut him off near the towns of Antrain GPE and Dol. A premature attack cost the Republicans victory at Dol GPE and they were driven out of Antrain GPE with very heavy losses (22-23 November 1793). Campaigning north of the Loire GPE , Vendéean rebel Henri de la Rochejaquelein concentrated his force at Entrammes GPE on the Mayenne to face the Republican army of General Jean Lechelle. Failing to wait for his full army to arrive, Lechelle GPE attacked and was routed with the loss of 4,000 men and most of his guns and stores. He never again commanded in the field (26 October 1793). Royalist leader Maurice d'Elbée secured victory at Thouars for the counter-revolution in western France GPE , then advanced south towards Fontenay-le-Comte. At nearby Pissotte GPE , the rebels were routed by Republican General Francois Chalbos, with d'Elbée wounded. A fresh attack nine days later drove Chalbos out and the Royalists captured valuable arms and stores (16 & 26 May 1793). Royalist rebel Henri de la Rochejaquelein marched north into Normandy GPE after victory at Entrammes GPE (26 October) to support a planned landing by émigrés and attacked the port of Granville GPE , defended by Republican General Jean Pierre Varin and a small garrison. The large Vendéean army was disastrously repulsed and retreated towards the Loire GPE before the British navy arrived (November 1793). A prelude to battle at Entrammes GPE against the Royalist rebels of western France GPE saw Republican General Francois-Joseph Westermann—the Butcher of the Vendée—impulsively attack the nearby city of Laval GPE . Led into a rebel trap at night, Westermann suffered a sharp defeat. However, this was reversed next day by the great Republican victory a few miles south at Entrammes GPE (25 October 1793). Just days after defeat at Angers, the weakened Royalist rebel army of 12,000 under Henri de la Rochejaquelein captured the city of Le Mans GPE and were attacked by a large Republican force under General Francois Vachot at the nearby village of Pontlieu-sur-l'Huisne. The Vendéeans were crushed, with perhaps 1,000 prisoners executed, and the survivors fled across the Sarthe LOC (12 December 1793). During the Royalist rebellion in western France GPE , Maurice d'Elbée and 6,000 rebels were repulsed at Lucon GPE by Republican General Claude Sandoz and a garrison of just 800. New commander Augustin Tuncq drove off a

FIGURE 4.6: The first half of the text describing the battles of the war in the French *Vandée*. For the full text, check the appendix A.

To geocode the individual battles, the following steps were taken:

1. Comparison of all the recognized toponyms from the war text as well as the battle place names from the titles (also *battle_location_2nd* and *reference_modern_loc*) with my gazetteer and extraction of all coordinates.

2. Calculation of the spatial median (the map in Figure 4.7 shows all ambiguous candidates involved worldwide for the *Vandée War* and their spatial median).

3. Removal of all locations further than 2000 km away from the spatial median.

4. Calculation of all combinations within the individual battles (similarly as already explained for the example of the *Battle of Kappel*).

5. Selection of the location that belongs to the combination with the shortest overall distance.

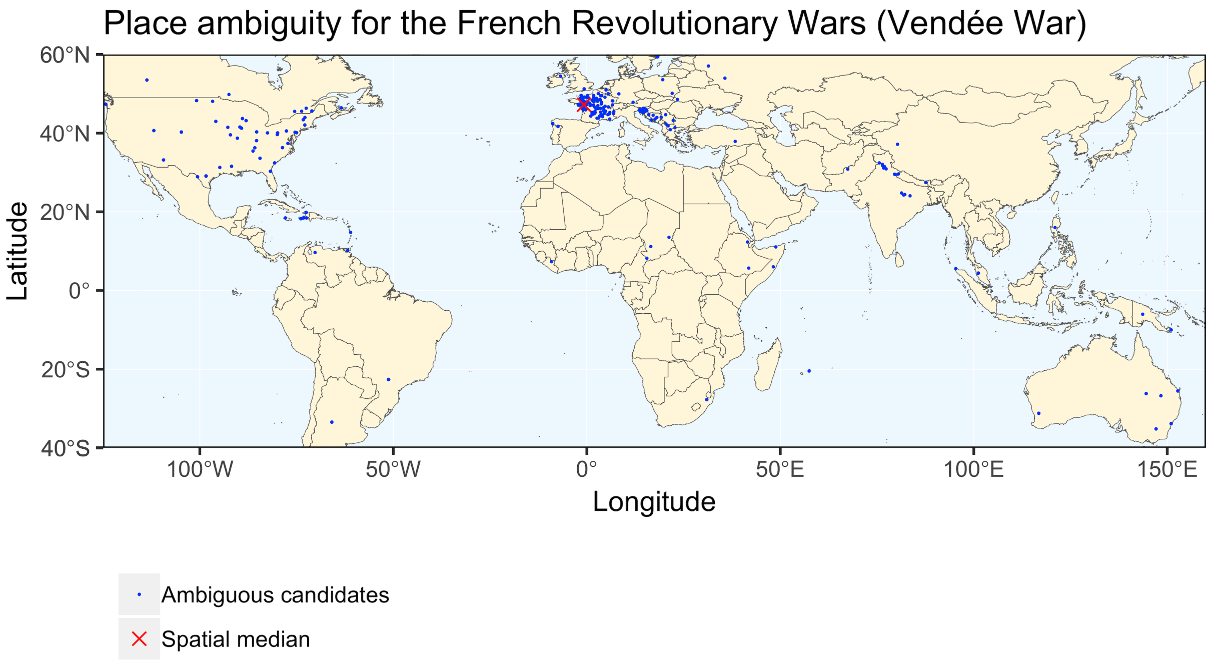
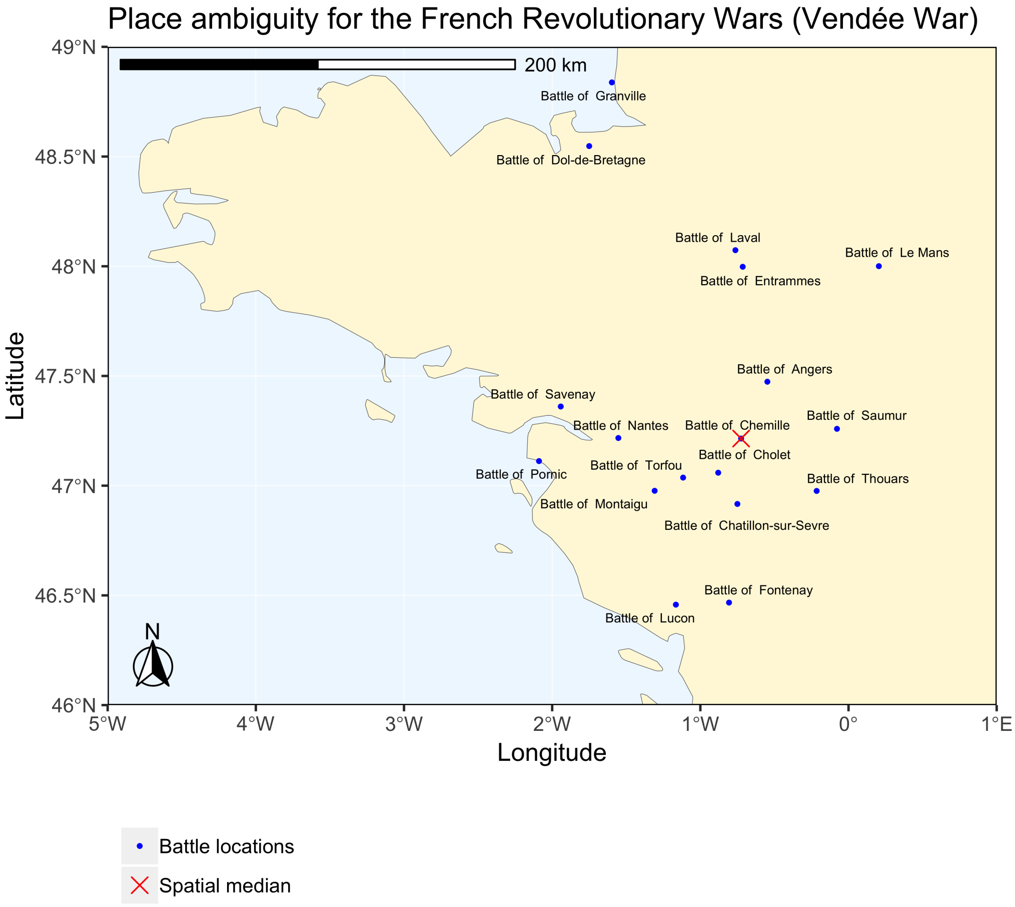


FIGURE 4.7: The map shows all ambiguous location candidates that appear in the text entries from battles of the *Vandée War* (see Appendix A).

By putting the individual battles together and thereby looking first at the larger war context, the spatial median can be used to determine the approximate region of battle events. By reducing the possible candidate locations so that each was within 2000 km of the spatial median (step 3.), it became possible to calculate the remaining combinations per battle and thus find the correct battle locations. The distance of 2000 kilometres in which the locations were still considered was chosen somewhat randomly, although different distances were tried out. However, the distance had to be as large so that the entire war was covered and no locations were excluded in advance. For battles that had a place name in the title that was free of ambiguity, meaning only one entry with that name in the gazetteer, those coordinates were assigned to that battle.

The map in Figure 4.8 shows the disambiguated battle locations for the *Vandée War*. A map with all the ambiguous locations in France, together with the disambiguated locations can be found in the Appendix A. In the course of this geocoding process, the participating war actors were also extracted by the NER recognizer (NORP) and saved to the battle entries.



Master thesis: Fueglister, 2020

FIGURE 4.8: Map showing the resulting locations for the battles from the *Vandée War*.

4.6 Transferability of the Method

The presented method was applied to all battles of Jacques. It also became quickly clear that the method could also be applied to Sweetman's work. The transferability of the method to another dictionary was therefore given. However, this opened the difficulty of combining the resulting two sets of battles. When looking through the dictionaries manually, there were many entries that existed twice, because the two dictionaries had listed the same battles. By searching for a method to sort out these candidates, however, no method was found yet. The difficulty is that there is very little information about the battles that could be used to match the same battles. For example, if one wanted to match the battles using the war names and the date, it became obvious that different names were used. When using the date, the problem was that the necessary precision could not be extracted from the entries of Sweetman, neither by structured extraction, nor by NER. Also, the place names were too rarely written exactly the same, so this would have helped to match the battles. In the sign of this thesis it was therefore decided to concentrate on the work of Jacques and to invest more time in the evaluation of the dataset obtained. The achievement of a comprehensive battle dataset was not endangered by this, since Jacques's dictionary already

appeared to be very extensive. How many battles could be coded and how well the geocoding system could locate the battles in space is discussed in the next chapter.
The described difficulties arose also when trying to compare the resulting dataset with the existing one from Dincecco and Onorato (2016). This dataset could therefore not be used for validation. Therefore, a separate validation dataset was created from hand, which served as a basis for evaluating the spatial error of my dataset. How this was done and how large the deviation is, will follow in the next chapter.

Chapter 5

Results

The entire resulting dataset, which was geocoded by machine using the described method, contains 4575 battles worldwide. Taking the same study area as Iyigun, Nunn, and Qian (2017), and the same time period (1401-1900), my dataset counts 2589 battles, which is 112 battles more than the one form Iyigun, Nunn, and Qian (2017). Within the study area selected by me for the entire period under consideration, 2217 battles were georeferenced. Compared to Dincecco and Onorato (2016) from 1100 on (not 800) until 1799, my datasets contains an additional 1192 battles.

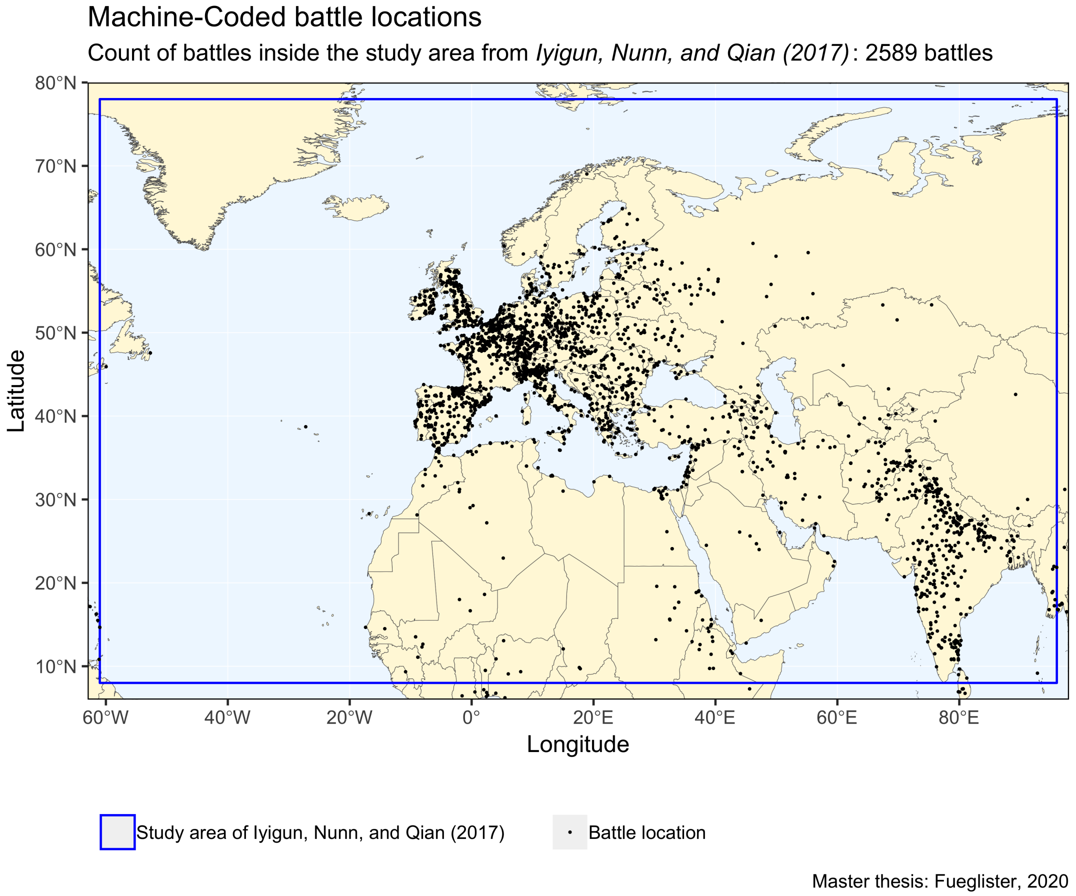

Machine-Coded battle locations
Count of battles inside the study area from *Iyigun, Nunn, and Qian (2017)*: 2589 battles
80°N
70°N
60°N
50°N
40°N
30°N
20°N
10°N
Latitude
60°W
40°W
20°W
0°
20°E
40°E
60°E
80°E
Longitude
Study area of Iyigun, Nunn, and Qian (2017)
Battle location
Master thesis: Fueglister, 2020

FIGURE 5.1: Spatial distribution of all machine coded European battles.

The map in Figure 5.1 shows all European battles, which could be coded with the presented method (a map with all battles worldwide can be found in the A). The purpose of this chapter is to present the results of the different validation steps that were undertaken to determine whether the presented method was suitable and therefore an acceptable answer to the research question. First, the model of the NER recognizer was validated. This allows statements to be made at the level of toponym recognition, respectively how good geoparsing worked. Furthermore, the coordinates for a random selection of battles were set by hand, which made it possible to determine the spatial error of the created system. In addition, a selection of non-georeferenced battles was examined to find out what the causes were that the system could not locate them in space. The main reason for this was that there were no corresponding entries in the gazetteer. However, the rationale behind this was manifold and has been analyzed. The interpretation of the results follows in chapter 6 (Discussion).

5.0.1 Validation on the Level of Toponym Recognition

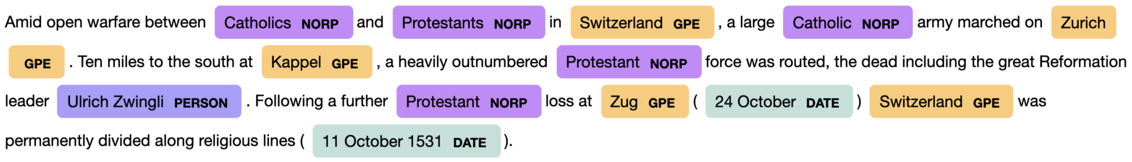
To find out how well my NER model could recognize toponyms and other entities, 20% of the annotated training data was used for validation. The results for Precision, Recall and F-score are listed in Table 5.1. The values can be compared with the initial results of the base model from Figure 4.1.

NER statistics

	GPE	*NORP*	*LOC*	*PERSON*	*DATE*
Precision	37.95	57.66	72.92	51.35	18.99
Recall	90.17	88.46	85.37	90.15	23.62
F-score	53.41	69.82	78.65	65.43	21.05

TABLE 5.1: NER statistics for the extended model to extract data from battle dictionaries.

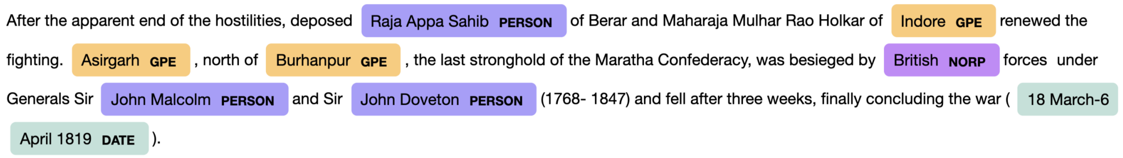
One can observe that the whole model has been improved for all trained entities. Because my NER model's precision lags a little behind the recall, it recognizes many entities as the right ones, but also classifies many entities that are not actually the correct ones. However, in terms of toponyms, this is better than if the system were to classify less. This is because with a higher recall there are more potential toponym candidates, which are checked again by the gazetteer, if those are potential geographic references. Thus, the system works with more information, but also runs less risk of omitting toponyms, which are important for geocoding.
As can be seen in Figure 5.2, the model succeeded without error in predicting the entities of the text about the *Battle of Kappel*.



Amid open warfare between Catholics NORP and Protestants NORP in Switzerland GPE , a large Catholic NORP army marched on Zurich GPE . Ten miles to the south at Kappel GPE , a heavily outnumbered Protestant NORP force was routed, the dead including the great Reformation leader Ulrich Zwingli PERSON . Following a further Protestant NORP loss at Zug GPE (24 October DATE) Switzerland GPE was permanently divided along religious lines (11 October 1531 DATE).

FIGURE 5.2: NER visualization of the *Battle of Kappel*

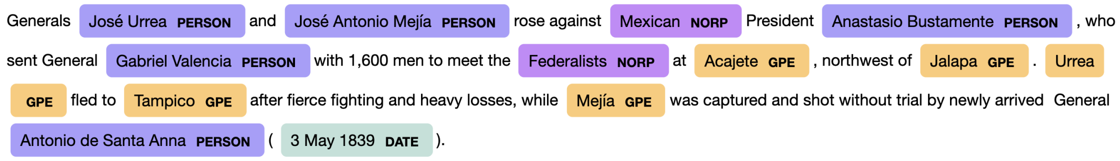
However, if one looks at the model's prediction for the *Battle of Asirgarh* (Figure 5.3) in the *3rd British-Maratha War*, which was fought in today's India, one can find a weakness of the model. While Asirgarh and Burhanpur were correctly recognized as place names (GPE), Indore was also recognized as such. This is not wrong, since it is also an Indian city. However, the place name is equally part of the name of Maharaja Mulhar Rao Holkar of Indore. Since these place names are not necessarily related to the description of the location and therefore not necessarily autocorrelating, they may distort the geocoding. This kind of name, including a place name, appears relatively often in the book of Jacques. Although I annotated these entities as *PERSON*, the base model could not be adapted accordingly.



After the apparent end of the hostilities, deposed Raja Appa Sahib PERSON of Berar and Maharaja Mulhar Rao Holkar of Indore GPE renewed the fighting. Asirgarh GPE , north of Burhanpur GPE , the last stronghold of the Maratha Confederacy, was besieged by British NORP forces under Generals Sir John Malcolm PERSON and Sir John Doveton PERSON (1768- 1847) and fell after three weeks, finally concluding the war (18 March-6 April 1819 DATE).

FIGURE 5.3: NER visualization of the *Battle of Asirgarh*

Another classification which classified an actual PERSON entity as GPE and thus could falsify the map-based geocoding is the example of the *Battle of Acajete* (see Figure 5.4). *Urrea* and *Mejia* were categorizes as GPE although it should be labeled as person names (PERSON). This is relevant because the wrongly classified entities have geo/non-geo ambiguity. For these entities there are locations named accordingly. Examples are: *Urrea, Mexico* (lat/long: 26.82101, -109.61888) or three different *Mejias* (*Mejia, Cuba* (lat/long: 20.51667, -75.73333); *Mejia, Boliivia* (lat/long: -17.05, -61.71667); *Mejia, Colombia* (lat/long: 1.12763, -77.39254)). How good the geocoding is, despite some possible wrong classifications, is covered in the next subchapter.

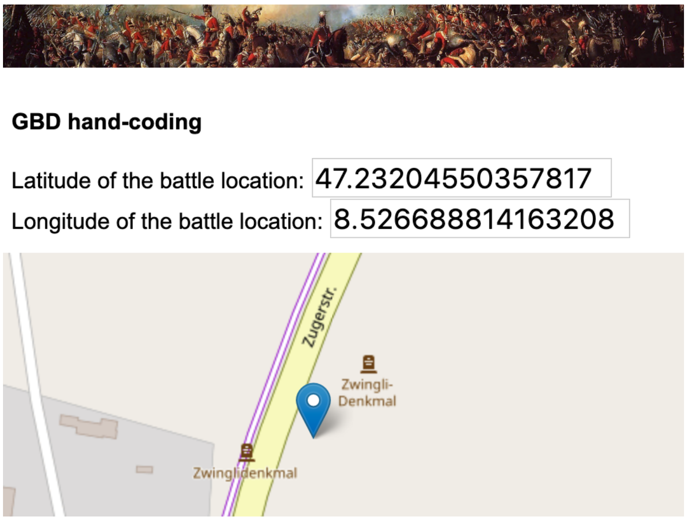


Generals José Urrea PERSON and José Antonio Mejía PERSON rose against Mexican NORP President Anastasio Bustamente PERSON , who sent General Gabriel Valencia PERSON with 1,600 men to meet the Federalists NORP at Acajete GPE , northwest of Jalapa GPE . Urrea GPE fled to Tampico GPE after fierce fighting and heavy losses, while Mejía GPE was captured and shot without trial by newly arrived General Antonio de Santa Anna PERSON (3 May 1839 DATE).

FIGURE 5.4: NER visualization of the *Battle of Acajete*

5.0.2 Validation on the Level of Toponym Resolution

In order to find out how well the system and the chosen geocoding method located the battles in space, a comparison was made with hand coded battles. While the hand coded battles were considered as ground truth (the location where the battle took place), the spatial deviation to the machine coded coordinates could be calculated. This deviation is called spatial error in the following. A similar approach was taken by Weidmann (2015) while calculating the spatial error of the UCDP GED dataset. To represent the totality of the geocoded battles, a representative validation sample of 360 battles was randomly selected. Battle after battle, these were looked through manually and the location where the battle took place was researched. Mainly online research and other battle dictionaries were used to find the right place without any ambiguity. In order not to be dependent on a gazetteer again like the set-up system, a simple interface was created for the manual assignment of the coordinates (see Figure 5.5).



GBD hand-coding

Latitude of the battle location: 47.23204550357817

Longitude of the battle location: 8.526688814163208

FIGURE 5.5: Interface of the hand coding tool

The marker contained in it could be placed on the map in the way it was considered appropriate based on the conducted research. If no more than the place name of the location, where the battle was fought, could be found, the marker was placed in the center of that location. For the researched battles, which were located in my study area, the map in Figure 5.6 shows both the machine-coded and the hand-coded coordinate. If they are far apart, a line connects them. One can see that the system has coded many battles to the same location, as was also discovered during later investigation.

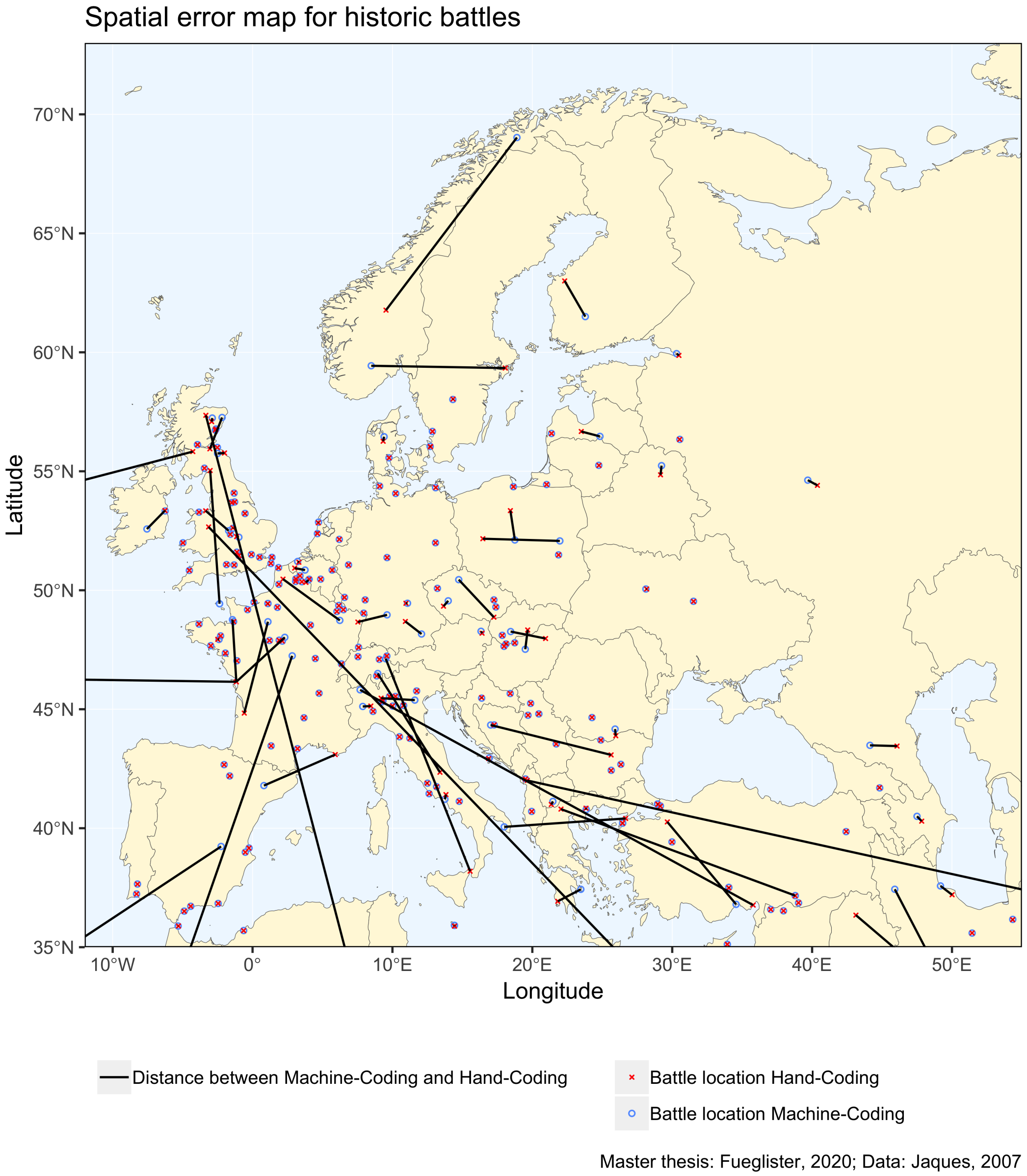

Spatial error map for historic battles
70°N
65°N
60°N
55°N
50°N
45°N
40°N
35°N
Latitude
10°W
0°
10°E
20°E
30°E
40°E
50°E
Longitude
Distance between Machine-Coding and Hand-Coding
Battle location Hand-Coding
Battle location Machine-Coding
Master thesis: Fueglister, 2020; Data: Jaques, 2007

FIGURE 5.6: Spatial error map

Nevertheless, one can also see that some battles were coded to a completely wrong location. That this happened could have several reasons. If the place name in the title had geo/geo ambiguity, but the correct location was not stored in the gazetteer and only the ambiguous ones were available, the system assigned the one which was part of the combination with the shortest overall distance. This happened, for example, for the *Battle of Kringen*. The battle took place at a place called *Kringen*, three kilometres from the Norwegian town of *Otta* (lat/long: 61.77214, 9.53956). In my gazetteer there is only one entry for a hill called *Kringen* (lat/long: 69.02625, 18.90064), but it is located about nine hundred kilometres further north. The battle was therefore assigned with these wrong coordinates. It also happened that there was a wrong assignment of coordinates, because an even smaller overall distance was found for wrong ambiguous candidates. This was the case, for example, for the *Battle of Bordeaux*, 1453. In this battle entry, *Calais* in northern France was recognized as a GPE entity and therefore used for the map-based approach to find the smallest overall distance. For the battle location *Bordeaux*, from the title, several entries were found in the gazetteer. Also, for the place name *Calais*, several entries were found in the gazetteer. The combination with the shortest overall distance was now formed by *Calais* (lat/long: 50.95194, 1.85635) and *Bordeaux* (TGN ID765988, lat/long: 49.08826, 0.324). While the first one was the correct one meant by the author, the used *Bordeaux* was not the one where the battle took place. The battle namely took place in Bordeaux (lat/long: 44.83785,-0.59188) on the Atlantic coast. It was also shown that a battle, which was the only one fought in a war, was rather wrongly georeferenced, because lesser toponyms were available to calculate the spatial median to make a pre-selection. This was the case, for example, for the *Battle of Abukir*, 1801, where an error occurred when chunking this battle while extracting the war name. Because of this error the text could not be merged with the other battles of the French Revolutionary Wars (Middle East). This resulted in a much smaller text, which led to much less recognized toponyms. For this reason, the spatial median was not calculated according to Egypt as it was for the other battles of this war. This induced the selection of an Aboukir (lat/long: 18.2509, -77.34327) in Jamaica, which is only ten kilometres from an Alexandria (lat/long: 18.30411, -77.35311), which is also in Jamaica. The observation that lesser context is more likely to lead to wrong coordinates being assigned to a battle, was made several times. For wars with many battles and also correspondingly long texts and therefore a lot of context, the opposite was true. Thus, it barely came to wrong assignments.

The question is, how large the spatial error in the whole hand-coded validation sample is, in order to draw conclusions about the complete battle dataset. The distribution of this spatial error for Europe can be seen in Figure 5.7. 80 % of the machine-coded battles have a spatial error of less than 22.49 kilometres. If one looks at the 75th percentile, it is 6.66 kilometres. Much larger spatial errors occur after the 85th percentile, where deviations of about 88 kilometres and more were measured. If one compares these values with the measurements for the spatial error of those battles outside of Europe (see Figure A.3 in the Appendix), one finds that already from the 75th percentile on, the spatial error is larger as 412 kilometres. This measurement therefore suggested that the dataset should be limited to Europe due to the greater percentage of battles with a large spatial error. The decision will be discussed in the following chapter 6 "Discussion".

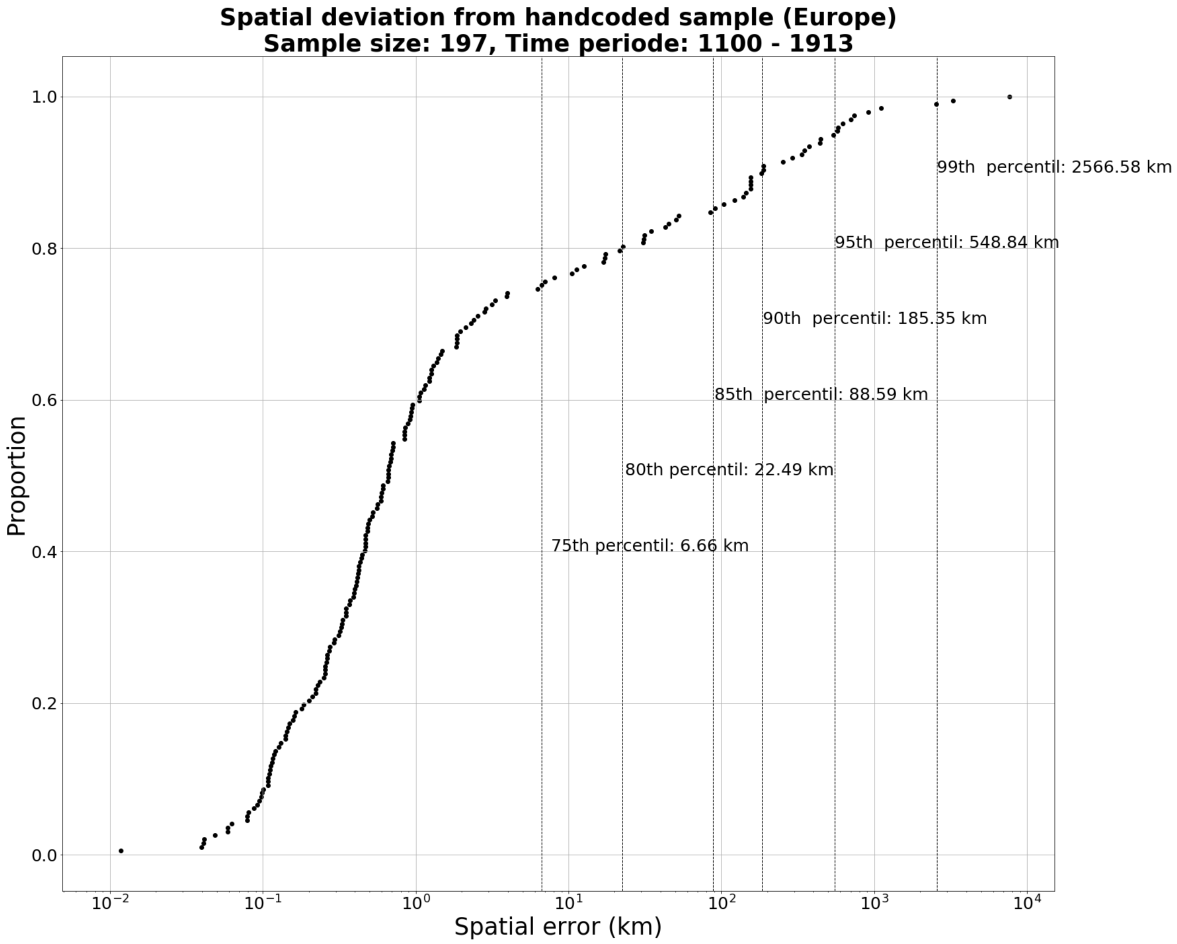


FIGURE 5.7: Spatial error map for the battles inside Europa

5.0.3 Investigation of the Reasons for Non-Geocoded Battles

The 2589 geocoded European battles already represent a greater number of battles than Dincecco and Onorato (2016), nevertheless, the system was unable to geocode some battles. For Europe, this means that 462 battles did not receive any coordinates. That these battles have been fought in Europe is known because of the spatial median of the wars. There was no continent that had significantly more non-geocoded battles (see Figure 5.8). In this subsection, the reasons why some battles could not be geocoded are presented.

To find out the different reasons, a sample of 250 battles was selected and looked through manually. I searched for reasons why no coordinates could be assigned to the battle titles respectively the places mentioned in them. A review of the sample entailed that the reasons could be divided roughly into five categories. The pie chart in Figure 5.9 shows the proportions of the different categories I created. Each category will be discussed below.

Different Spelling

Most of the battles did not receive coordinates because the battle location searched for in the gazetteer was written differently. Therefore, the comparison of the supposed toponym with

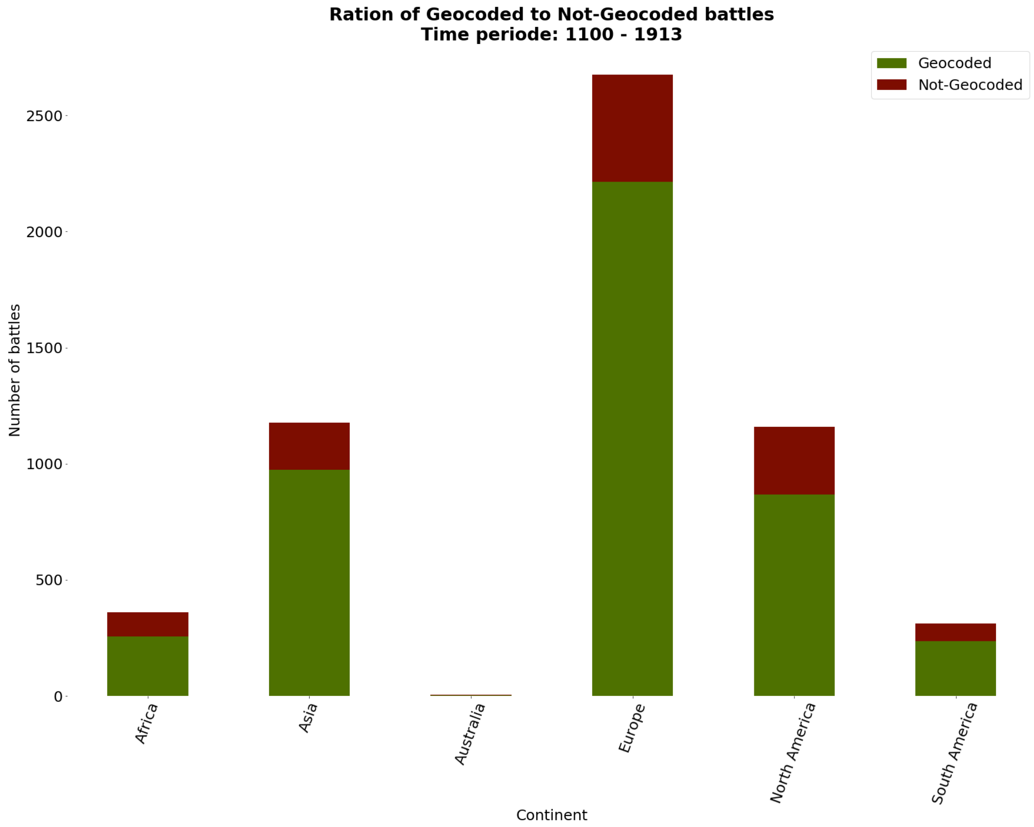


FIGURE 5.8: Barplot showing the number of geocoded battles per continent. Non-geocoded battles have not received coordinates.

the gazetteer did not yield any result. For example, Jacques sometimes used slightly different spellings than in the entries taken from GeoNames and TGN. An example is the name of a village which Jacques calls *Bourgtherolde* and is supposed to be in the French Normandy. Searching GeoNames for this entry, no results are found. With further research, however, one finds out that it must be the village *Bourgtheroulde-Infreville* (lat/long: 49.3, 0.88333). The same applies to the Austrian village *Frastanz* (lat/long: 47.21735, 9.62995), for which Jacques uses the place name *Frastenz*, and no entry exists in my gazetteer. While one gets many search results for *Parwan*, one does not find any for *Pirvan*. If one continues to search for a place in the Netherlands which is called *Gertruydenberg* by Jaques, one will only find an entry if one searches with a slightly adapted spelling (*Geertruidenberg*, lat/long: 51.70167, 4.85694). The fact that this problem could occur when comparing the names of places with the gazetteer was recognized early on. To solve this problem different kinds of approximate string matching also called fuzzy matching methods were tried out to get not only the result for the exact search query. This would also have yielded results with slightly different spellings. However, this could not be implemented successfully, as in some cases a change to a single letter could already lead to a completely different search result.

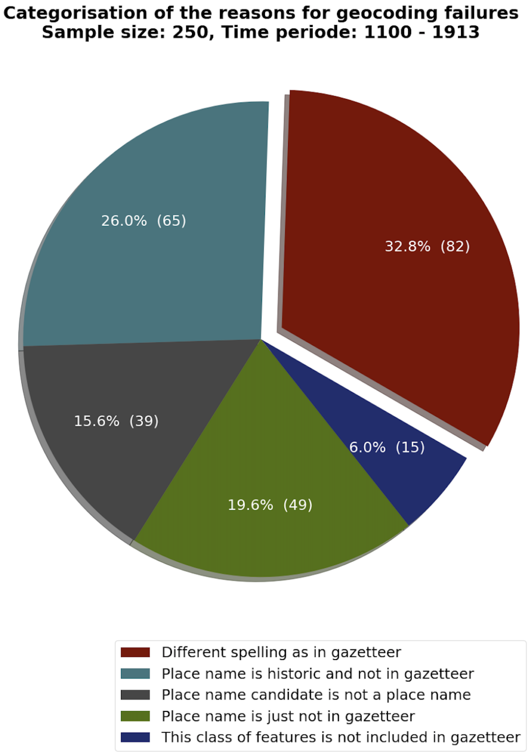


FIGURE 5.9: Results of the inspection of the reasons, why a battle was not geocoded.

Historic Place Name

For 65 of the non-geocoded battles investigated, the reason why they did not receive any coordinates was that it was a historical place name that is not used in today's gazetteers. An example of such a place is the locality, where a well known tram junction with the name *Stauffacher* in *Zürich, Switzerland* is located today. The locality was formerly known as *St Jakob on the Sihl*. This is due to the chapel of St. Jakob and the leprosarium which were located there. Since this place no longer has this name today, it has no entry in a non-historical gazetteer. The fact that the place is not in TGN, which claims to contain historical places, is simply because it is not extensive enough. The same applies, for example, to a village in today's Mali with the name *Anfao*. The village, which was located near the present-day *Gao* (lat/long: 16.27167, -0.04472), was the location of a battle that took place during the *Wars of the Songhai Empire*. There is no entry for this historical village in my gazetteer and therefore the battle received no coordinates. That the percentage of non-geocoded battles was not higher, due to the use of historical place names, was probably because Jaques translated many of them into today's usage and created cross-references.

Not a Toponym

As Jacques has already been quoted, the majority of all battles are named after locations. This makes my georeferencing method possible. But Jacques also lists battles that do not have a toponym in their title. This was the case for almost 16 percent of sampled battles. The *Battle* *of the Shirts* is one that was not named by its location, but rather after the fact, that the fighting soldiers fought not in their armour but in their shirts due to the heavy heat. Other examples of battles that are not named after locations are naval battles. These have names like: *Spanish Armada* or *Virginius Incident*, where Virginius was the name of an American ship, on which Spain executed 53 American and British passengers in 1873. Even if in the descriptions of the battles, which were not named after locations, place names were mentioned, it was not attended to assign the battle to these locations. This was because it was not certain whether the mentioned place name was the one where the battle took place.

Missing in Gazetteer

For some battles that were named after their location, there was simply no Gazetteer entry available, although the location still had the same name used by Jacques on common online maps or in other sources. A battle fought in *Taraori* (lat/long: 29.79791, 76.93858) serves as an example. This city is located in the Indian state of *Haryana* and has no entry with the feature code of "populated place" in TGN or GeoNames. Also other common names for this location are searched in vain. If one uses the online search mask of GeoNames, one may not get any results, but one will sometimes be referred to a corresponding Wikipedia articles. However, this information is not available when downloading GeoNames.

Another Feature Class

The smallest amount of non-geocoded battles did not get any coordinates because there was no entry in my Gazetteer although there was information about it in GeoNames or TGN. This is because I have decided not to include some toponym feature classes in my gazetteer. As an example serves a battle which was fought in the War of the American Revolution. This took place at Stono Ferry. If one searches for this toponym in GeoNames, one will get an entry with the name Stono Ferry Golf Course (ID: 7265679) with the feature class: "golf course". The same applies to the *Battle of Hopton Heath*, for which only one entry with the feature class "railroad station" can be found, which was not included like "bay" or "bus station". That these features were not included can certainly be discussed. But the main reason for this is that a certain amount of control should be maintained.

Chapter 6

Discussion

That the chosen method to georeference battles in space gave promising results becomes apparent when comparing the resulted geocoded battle locations from the *Vandée War* with a simple overview map showing the course of the war. (see Figure 6.1). All 18 battles which were geocoded by my system can also be found on this overview map and were assigned to the same locations.

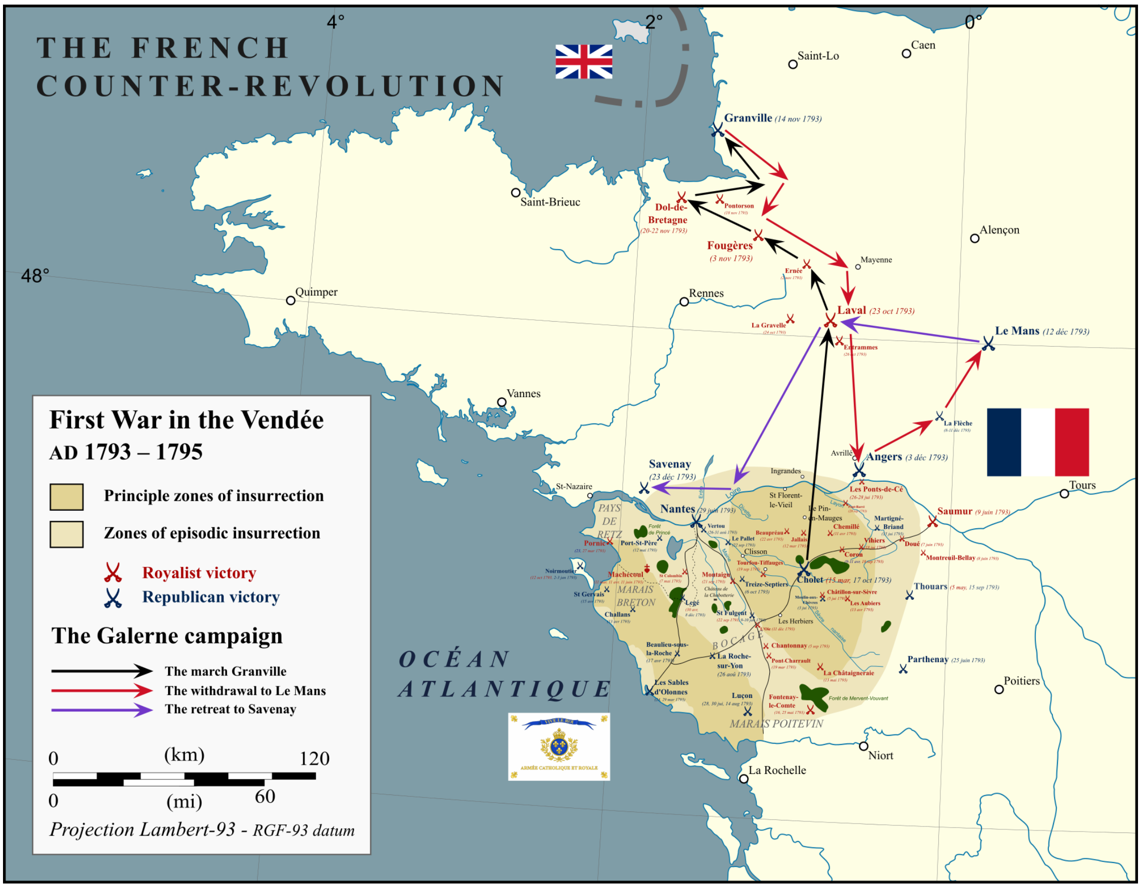


FIGURE 6.1: Overview map of the war in the French *Vandée* by Gouachevalier (2015)

The overview map also suggests that many, probably smaller battles were not listed in Jacques

(2007), and therefore are not in my dataset. One can find 30 battles more on the overview map. Although many battles were georeferenced very successfully, the lack of these battles raises the question whether a suitable data source was used. These and other questions will be discussed below. An answer to my research questions will also be given. A very interesting observation, shall already be mentioned at this point. If one looks at the spatial median, which was calculated for all toponyms involved and their ambiguous location candidates, it is noteworthy that this median is located directly in the zone called "principle zones of insurrection" on the overview map. This examination, that the spatial median was located in the presumed centre of war, was discovered for many other wars too. In order to locate wars in space in future research projects this could be of interest.

6.1 Used Data

Initially, the statement made by Iyigun, Nunn, and Qian (2017) was recorded, stating that Clodfelter (2017) would have 115 battles more than Jacques (2007) in his battle dictionary. Although the dictionary of Jacques was geocoded in this work, the resulting dataset contains 112 battles more for the same spatial as well as temporal coverage. One could assume that this difference is due to the fact that many battles were wrongly coded in this study area. But if one considers the achieved rate of spatial error, this is unlikely. In addition, there are 462 battles that would have been coded to Europe but the system was unable to assign coordinates and are therefore not in the dataset. If one could also code those battles, the discrepancy to Iyigun, Nunn, and Qian (2017) would increase even further. The fact that such a high number of battles could be coded certainly, legitimized the choice of the corresponding dictionary of Jacques and makes it a work not to be ignored when creating a historic battle dataset.

However, there is also reason to expect that the use of a single dictionary or the limitation to a single type of data source can lead to biased end results. If one looks, for example, at Southern Africa on the map with the worldwide distribution of battles (see A.4), it is striking that there must have been a great number of battles in South Africa, whereas not a single battle can be found in Namibia or Botswana. This is simply because Jacques does not contain battles from these countries. In Namibia, for example, the *Battle of Otjimbinge*[1] in 1863, where Herero fought aginst the Nama, is missing. Just like this battle, those that were part of the *Vandée War* and likely many more were missing. The same selection bias as Weidmann (2015) sees in the contemporary datasets therefore also results when compiling a dataset for historical battles. The fact that Iyigun, Nunn, and Qian (2017), as well as Dincecco and Onorato (2016), and the work in this project, obtained very different numbers of battles, shows, as assumed in the introduction, that the creation of a dataset on historical battles can not yet be declared as complete.

Looking back, it would certainly have been profitable if a system had been created from the very beginning which could function even more independently of the respective input data. Because the place where the battle took place is also mentioned in the descriptive text, it would have made more sense if a system could read the place of interest (battle location) out of a continuous text and then disambiguate it. This way, a greater abundance of different text sources could serve as

[1] For more information on this battle, see Tonchi, Lindeke, and Grotpeter (2012)

a data basis, which would counteract a possible bias. The experience gained in this project could certainly be used to try to build such a system, which is even more complex as the built one.

6.2 NER Performance

Although the existing NER model could achieve a significantly better performance by training and adapting to the processed text data, the question arises whether the choice of the base model was appropriate. As mentioned, this had been trained based on news articles. In retrospect, this choice is not considered a weakness of the system, as the characteristics are similar. In addition, a performance was achieved that comes close to state of the art values mentioned by Gritta et al. (2018). When using a base model as a starting point, two additional points are particularly striking. On the one hand, an externally determined annotation scheme must be retained when using a base model as a starting point. Respectively the existing is very difficult to "overtrain". On the other hand, it was not possible to pay special attention to the distinction of metonomy. As Gritta et al. (2018) mentions, this is generally to be seen as a major limitation when using NER for the geoparsing step. Nevertheless, what seems to be important is the fact that it was possible to extract a lot of geographical references which could be used for the map-based method to geocode the battles.

6.3 Gazetteer Quality

For the study area in which I was ultimately interested in, the used gazetteer appears to have been of reasonable quality, as many battles could be correctly geocoded. However, it turned out that for a quarter of the non-geocoded battles, the reason was a non-existent gazetteer entry, due to the fact that the place was a historical place. So, the lack of a worldwide historical gazetteer certainly has an impact. The fact that it is not larger is because the author has translated the historical places into their present names.

In this work it could be shown that the assignment of coordinates within Europe is associated with fewer errors. This correlates on the one hand with the findings of Smith and Crane (2001), who showed that there are less ambiguities for European place names. On the other hand, it probably also reflects the quality of the gazetteer used, for which Acheson, De Sabbata, and Purves (2017) found out that the quality for Europe is better than for other world regions.

For those other parts of the world, it will be interesting to see with what precision one could geocode historic events if a worldwide historical gazetteer would be available. If one had only dealt with Europe from the beginning and wanted to achieve an even higher coverage of place names, the creation of a gazetteer specifically for Europe could have been discussed. Possibly, with the help of many small national gazetteers, an even higher coverage would have been achieved, but this would also have implied a tremendous amount of work. To make better use of the already existing gazetteer entries, the tried and tested fuzzy matching would certainly be conceivable, but requires meticulous control, because as mentioned, it can very quickly lead to a match that differs only in one letter but means a completely different location. If this is used together with a map-based geocoding approach, this is not ideal, as additional false candidates may quickly have an influence.

6.4 Importance of the Use of Context

The experiences made during the georeferencing of the battles, some of which could be brought closer to the reader with the help of examples, show impressively that the presence of extensive context led to a more robust spatial median. Through this, a corresponding pre-selection could be made. If only very few toponyms were available, the map-based approach was less successful in assigning correct coordinates. The importance of existing context, which is considered as eminent by Gritta et al. (2018), as well as by Leetaru (2012), is therefore shared in any case. In relation to one of my sub-questions, one can answer that the more context there is, the easier it is to disambiguate autocorrelated places. This statement is supported by every correctly geocoded battle in my dataset, which contains a high degree of geographical ambiguity. This is because the origin for the correct assignment is in this case always the spatial median, which is more robust the more autocorrelated toponyms close to the "true" location candidate could be extracted.

The answer to my first sub-question of my research question has already been presented in the results and will be discussed here. For 75% of the European battles it can be assumed that the spatial error is below 6.66 km. The 80th percentile is 22.49 km. This results in a smaller spatial error as measured by Weidmann (2015) for the UCDP GED dataset (80th percentil, 50km). 80 percent of the data are therefore of the same quality as hand-coded contemporary conflict event data. If one investigates the quality of the data, it is important to reflect on the usage of the data. As mentioned above, in conflict research, raster cells are used to calculate the influence of conflict events together with other data. The size of these grid cells, which stands for the scale of analysis, therefore also determines the tolerance to the spatial error of the georeferenced battles. If a 100×100 kilometre grid is chosen, larger deviations play a smaller role. But even in this case, incorrect georeferencing can lead to a battle being aggregated to a neighbouring cell. If, on the other hand, a finer grid is chosen (e.g. 1×1 km), large spatial errors will have a corresponding influence on the result of the method. At this point, it should be mentioned that the influence of spatial precision and the size of the spatial error and its effect on the result of the chosen method is given little to no consideration in the literature so far.

The obtained results show that I was able to create a spatial dataset containing the most important information from the battle dictionaries. A large number of battles could be provided with very precise coordinates. In addition, I was able to extract exact time information, as well as the desired information about which war a battle belongs to, and the actors involved. The usefulness of the actor data still needs to be investigated. Nevertheless, my leading research question can be answered to the effect that such a task can be tackled with a map-based approach and a named entity recognizer as its core, as well as a gazetteer as comprehensive as possible.

However, the dataset created in this work has a weakness: there are battles which, apart from the respectable values for the spatial error, are simply wrongly georeferenced and can therefore be several hundred kilometres away. Certainly, the worst encoded events in other conflict datasets also show very high spatial errors, but I still consider entries that have a spatial error of several hundred kilometres as intolerable for all research projects. Methods to filter completely wrong coded battles out or to locate them correctly has not yet succeeded. Before the dataset can be used in research, this problem must therefore be solved.

One must also ask oneself whether it made sense for the existing task to leave the coding to a machine. Certainly, the work carried out so far has shown that existing datasets are still very incomplete. If I were to be asked again how I would tackle a similar task, I would probably try to implement a semi-machine-based procedure. The system would suggest the best possible candidate to the user, and would also locate it graphically on a map with all the available related information. Based on this, the user could take over the decision and adjust the coordinates if necessary. A completely wrong assignment of coordinates on the other side of the globe would be much less likely, because the user could also include other context information in the decision, which were not contained in this project, apart form the geographical one. Knowing that Turkish troops fought in the *Battle of Aboukir*, which was wrongly located in Jamaica, a user would not be likely to locate in the Caribbean Sea if contextual information was visually highlighted and geographic information was mapped accordingly. But this example shows impressively how important context information is. In this work the spatial context could be used profitably very often. However, the inclusion of additional (non-spatial) context would have been desirable at times too. This kind of semi-machine-based system could also be used again to control the created dataset and the assigned coordinates. Completely wrongly located battles could be identified and corrected. Such a system, like the one presented here, would probably not take as much time to correct or completely recode the battles as Iyigun, Nunn, and Qian (2017) indicated for the creation of their dataset. A conclusion that can be drawn after this project is therefore that, with the help of machines, it is possible to create a spatial dataset in significantly less time.

A further limitation, which has been identified and which has been dealt with in a transparent manner, is the fact that, as with other known event datasets (see Eck (2012)), the existing spatial vagueness is not sufficiently reproduced in the assigned coordinates. However, this weakness also exists in hand-coded datasets, even though significantly more human working time is spent on coding. Nevertheless, the introduction of the geoprecision code for this dataset can make users aware of this inaccuracy.

Chapter 7

Conclusion and Further Work

This chapter concludes this master thesis. At first, a short summary about the achieved work will be given. After that, the most important results of the thesis will be summarized before an outlook for upcoming tasks will be given.

7.1 Summary

In order to compile a machine-based spatial dataset containing historical battles, suitable data sources were searched first. Using Jacques's work as the main source of data, I found a machine-readable book that listed the individual historical battles in independent text entries. Much work had to be invested in a pre-processing step to chunk the battles into machine-readable pieces. Thereby, some structured data, such as the dates of the battles, could be extracted. Then a system was set up which contained a named entity parser as its core. A specifically trained model made sure that enough toponyms could be extracted from the texts. These were then compared with a custom-built gazetteer. Ambiguous candidates for a location where a battle took place were then disambiguated using a map-based approach. It was assumed that the locations mentioned in the description of a battle text are autocorrelated. This resulted in a dataset which contains more battles than the existing ones of Dincecco and Onorato (2016) and Iyigun, Nunn, and Qian (2017). The dataset was then extensively evaluated. On the one hand, the reasons why no coordinates could be assigned to a battle were investigated and on the other hand, the spatial error of the battles which received coordinates was calculated using a self-produced hand-coded battle dataset.

7.2 What Has Been Achieved?

It has been shown that a machine-based approach for creating such a dataset can significantly reduce the high human workload of hand coding. For 80 percent of the coded battles, a similarly high accuracy as widely used contemporary hand-coded datasets was achieved (see Weidmann (2015)). For the majority of the battles the involved geographical ambiguity could therefore be successfully disambiguated. It could be shown that context information is very important for successful disambiguation. Even when working with very short texts, which Gritta et al. (2018)

consider challenging, a way could be found to use additional context to pre-select possible toponyms.

It had to be accepted that a worldwide historical gazetteer is still missing. The fact that this lack of a historical gazetteer can hinder research projects such as this one, but also other applications, was shown by the fact that a not disregardable part of the battles could not be successfully geocoded due to the historical name of the location.

Apart from the very accurate geocoded battles, the created dataset also has two limitations. First, like all known conflict event datasets, the geographical vagueness involved in the written description of a conflict event is not taken into account when assigning coordinates. But what seems to be more obstructive for the decision to use the created dataset for the NASTAC project, is the fact that the created dataset also contains battles that were assigned with completely wrong coordinates due to unresolved ambiguities. This is often because too little context information could be used.

What are, nevertheless, two more important achievements of this thesis are on the one hand that it could be shown that the existing datasets of Dincecco and Onorato (2016) and Iyigun, Nunn, and Qian (2017) are still far from complete, since they do not contain many battles. Therefore, if one wants to test the initially mentioned analogy of Tilly (1985) quatitatively with spatial conflict data, even more effort is required to obtain a more complete dataset. Furthermore, it could be shown that the used battle dictionaries are biased, which can have a consequences on the quality of the created datasets.

7.3 Outlook and Future Work

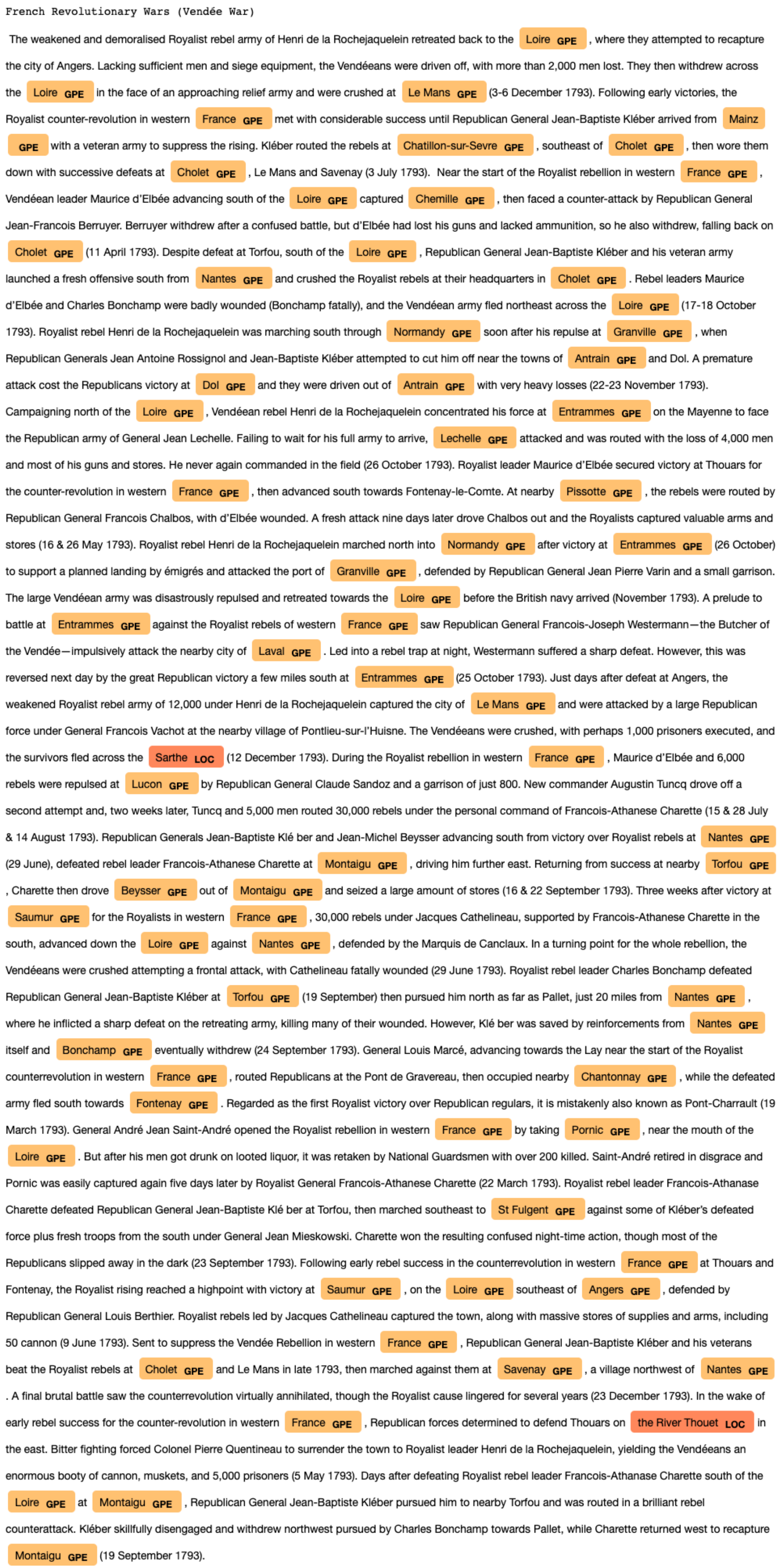
In connection with the NASTAC project, it must first be decided how to proceed with the generated data. On the one hand, a way should be found to correct the completely wrong geocoded battles. Furthermore, the battles which have not received coordinates should be processed accordingly. If the dataset of Iyigun, Nunn, and Qian (2017) should be published after all, it should be compared with the one created here.

The observation that the calculated spatial median often represents a possible centre of war very accurately can be used to create, like Hallberg (2012) did for more recent conflict, a dataset containing the spatial extent for historical conflicts. In a broader context, I see a need to explore how geographic vagueness can be included in event datasets created by machine. In addition, and because gazetteers are so central to the automatic processing of geographic text data, work on them should be continued and intensified. For the processing of historical data it is eminent that a comprehensive historical gazetteer will soon be available to ensure that even forgotten place names are remembered.

Appendix A

Appendix

	Place name	Country	Alternative names	Latitude	Longitude
0	Kappel	AT	Kappel,Kappel an der Drau	46.533330	14.250000
1	Kappel	AT	Kappel,Kappel am Krappfeld	46.838610	14.486390
2	Kappel	CH	None	47.268910	9.116220
3	Kappel	CH	Kappel,Kappel am Albis	47.228110	8.527270
4	Kappel	CH	None	47.521670	8.855390
5	Kappel	CH	Kappel,Kappel SO,Kappel i Sveits,Kappel’,ka pe...	47.324750	7.846470
6	Kappel	DE	Kapellen,Kapeln,Kappalen,Kappel,Kappel’n,ka pe...	54.661220	9.931300
7	Kappel	DE	None	50.821330	12.887910
8	Kappel	DE	Kappel	50.000000	7.366670
9	Kappel	DE	None	49.661580	11.303700
10	Kappel	DE	None	48.292220	7.740980
11	Kappel	DE	None	48.112070	8.501960
12	Kappel	DE	None	48.060470	9.599910
13	Kappel	DE	Kappel	47.963430	9.206290
14	Kappel	DE	Kappel	47.969140	7.908460
15	Kappel	DE	None	47.872550	8.235200
16	Kappel	DE	None	47.771250	9.449500
17	Kappel	DE	None	47.605120	10.535290
18	Kappel	DE	Cappel,Kappel	53.726480	8.570300
19	Kappel	DE	Cappel,Kappel	51.127030	9.339370
20	Kappel	DE	None	49.823910	10.613750
21	Kappel	DK	None	54.773610	11.032700
22	Kappel	EE	Kabala,Kappel	59.356670	26.655830
23	Kappel	RU	Bukhta Nore-Kapel’lakht,Bukhta Nore-Kapel’lakh...	60.016670	27.866670
24	Kappel	None	None	50.016667	11.466667
25	Kappel	None	None	47.633333	11.033333
26	Zurich	CA	None	43.421630	-81.627730
27	Zurich	CH	Cirihe,Cirikh,Ciurichas,Cjurikh,Cjurikh khot,C...	47.366670	8.550000
28	Zurich	DZ	Sidi Amar,Zurich	36.542120	2.305400
29	Zurich	NL	Surig	53.111340	5.394370
30	Zurich	US	None	39.234450	-99.438160
31	Zurich	US	None	43.150620	-77.043300
32	Zurich	US	Alvord,Zurich	37.182710	-118.260100
33	Zurich	US	None	48.584440	-109.030440
34	Zug	AT	None	47.200560	10.109400
35	Zug	CH	Cug,Tugium,ZLM,Zoug,Zug,Zugo	47.172420	8.517450
36	Zug	DE	Zug	50.889200	13.346040
37	Zug	EH	Sug,Zoug,Zoûg,Zug	21.565230	-14.103880
38	Zug	IR	Zaug,Zowk,Zug,Zuk,Zūg,Zūk,zwg,zwk	33.695070	58.999200



French Revolutionary Wars (Vendée War)

The weakened and demoralised Royalist rebel army of Henri de la Rochejaquelein retreated back to the Loire GPE , where they attempted to recapture the city of Angers. Lacking sufficient men and siege equipment, the Vendéeans were driven off, with more than 2,000 men lost. They then withdrew across the Loire GPE in the face of an approaching relief army and were crushed at Le Mans GPE (3-6 December 1793). Following early victories, the Royalist counter-revolution in western France GPE met with considerable success until Republican General Jean-Baptiste Kléber arrived from Mainz GPE with a veteran army to suppress the rising. Kléber routed the rebels at Chatillon-sur-Sevre GPE , southeast of Cholet GPE , then wore them down with successive defeats at Cholet GPE , Le Mans and Savenay (3 July 1793). Near the start of the Royalist rebellion in western France GPE , Vendéean leader Maurice d'Elbée advancing south of the Loire GPE captured Chemille GPE , then faced a counter-attack by Republican General Jean-Francois Berruyer. Berruyer withdrew after a confused battle, but d'Elbée had lost his guns and lacked ammunition, so he also withdrew, falling back on Cholet GPE (11 April 1793). Despite defeat at Torfou, south of the Loire GPE , Republican General Jean-Baptiste Kléber and his veteran army launched a fresh offensive south from Nantes GPE and crushed the Royalist rebels at their headquarters in Cholet GPE . Rebel leaders Maurice d'Elbée and Charles Bonchamp were badly wounded (Bonchamp fatally), and the Vendéean army fled northeast across the Loire GPE (17-18 October 1793). Royalist rebel Henri de la Rochejaquelein was marching south through Normandy GPE soon after his repulse at Granville GPE , when Republican Generals Jean Antoine Rossignol and Jean-Baptiste Kléber attempted to cut him off near the towns of Antrain GPE and Dol. A premature attack cost the Republicans victory at Dol GPE and they were driven out of Antrain GPE with very heavy losses (22-23 November 1793). Campaigning north of the Loire GPE , Vendéean rebel Henri de la Rochejaquelein concentrated his force at Entrammes GPE on the Mayenne to face the Republican army of General Jean Lechelle. Failing to wait for his full army to arrive, Lechelle GPE attacked and was routed with the loss of 4,000 men and most of his guns and stores. He never again commanded in the field (26 October 1793). Royalist leader Maurice d'Elbée secured victory at Thouars for the counter-revolution in western France GPE , then advanced south towards Fontenay-le-Comte. At nearby Pissotte GPE , the rebels were routed by Republican General Francois Chalbos, with d'Elbée wounded. A fresh attack nine days later drove Chalbos out and the Royalists captured valuable arms and stores (16 & 26 May 1793). Royalist rebel Henri de la Rochejaquelein marched north into Normandy GPE after victory at Entrammes GPE (26 October) to support a planned landing by émigrés and attacked the port of Granville GPE , defended by Republican General Jean Pierre Varin and a small garrison. The large Vendéean army was disastrously repulsed and retreated towards the Loire GPE before the British navy arrived (November 1793). A prelude to battle at Entrammes GPE against the Royalist rebels of western France GPE saw Republican General Francois-Joseph Westermann—the Butcher of the Vendée—impulsively attack the nearby city of Laval GPE . Led into a rebel trap at night, Westermann suffered a sharp defeat. However, this was reversed next day by the great Republican victory a few miles south at Entrammes GPE (25 October 1793). Just days after defeat at Angers, the weakened Royalist rebel army of 12,000 under Henri de la Rochejaquelein captured the city of Le Mans GPE and were attacked by a large Republican force under General Francois Vachot at the nearby village of Pontlieu-sur-l'Huisne. The Vendéeans were crushed, with perhaps 1,000 prisoners executed, and the survivors fled across the Sarthe LOC (12 December 1793). During the Royalist rebellion in western France GPE , Maurice d'Elbée and 6,000 rebels were repulsed at Lucon GPE by Republican General Claude Sandoz and a garrison of just 800. New commander Augustin Tuncq drove off a second attempt and, two weeks later, Tuncq and 5,000 men routed 30,000 rebels under the personal command of Francois-Athanese Charette (15 & 28 July & 14 August 1793). Republican Generals Jean-Baptiste Klé ber and Jean-Michel Beysser advancing south from victory over Royalist rebels at Nantes GPE (29 June), defeated rebel leader Francois-Athanese Charette at Montaigu GPE , driving him further east. Returning from success at nearby Torfou GPE , Charette then drove Beysser GPE out of Montaigu GPE and seized a large amount of stores (16 & 22 September 1793). Three weeks after victory at Saumur GPE for the Royalists in western France GPE , 30,000 rebels under Jacques Cathelineau, supported by Francois-Athanese Charette in the south, advanced down the Loire GPE against Nantes GPE , defended by the Marquis de Canclaux. In a turning point for the whole rebellion, the Vendéeans were crushed attempting a frontal attack, with Cathelineau fatally wounded (29 June 1793). Royalist rebel leader Charles Bonchamp defeated Republican General Jean-Baptiste Kléber at Torfou GPE (19 September) then pursued him north as far as Pallet, just 20 miles from Nantes GPE , where he inflicted a sharp defeat on the retreating army, killing many of their wounded. However, Klé ber was saved by reinforcements from Nantes GPE itself and Bonchamp GPE eventually withdrew (24 September 1793). General Louis Marcé, advancing towards the Lay near the start of the Royalist counterrevolution in western France GPE , routed Republicans at the Pont de Gravereau, then occupied nearby Chantonnay GPE , while the defeated army fled south towards Fontenay GPE . Regarded as the first Royalist victory over Republican regulars, it is mistakenly also known as Pont-Charrault (19 March 1793). General André Jean Saint-André opened the Royalist rebellion in western France GPE by taking Pornic GPE , near the mouth of the Loire GPE . But after his men got drunk on looted liquor, it was retaken by National Guardsmen with over 200 killed. Saint-André retired in disgrace and Pornic was easily captured again five days later by Royalist General Francois-Athanese Charette (22 March 1793). Royalist rebel leader Francois-Athanase Charette defeated Republican General Jean-Baptiste Klé ber at Torfou, then marched southeast to St Fulgent GPE against some of Kléber's defeated force plus fresh troops from the south under General Jean Mieskowski. Charette won the resulting confused night-time action, though most of the Republicans slipped away in the dark (23 September 1793). Following early rebel success in the counterrevolution in western France GPE at Thouars and Fontenay, the Royalist rising reached a highpoint with victory at Saumur GPE , on the Loire GPE southeast of Angers GPE , defended by Republican General Louis Berthier. Royalist rebels led by Jacques Cathelineau captured the town, along with massive stores of supplies and arms, including 50 cannon (9 June 1793). Sent to suppress the Vendée Rebellion in western France GPE , Republican General Jean-Baptiste Kléber and his veterans beat the Royalist rebels at Cholet GPE and Le Mans in late 1793, then marched against them at Savenay GPE , a village northwest of Nantes GPE . A final brutal battle saw the counterrevolution virtually annihilated, though the Royalist cause lingered for several years (23 December 1793). In the wake of early rebel success for the counter-revolution in western France GPE , Republican forces determined to defend Thouars on the River Thouet LOC in the east. Bitter fighting forced Colonel Pierre Quentineau to surrender the town to Royalist leader Henri de la Rochejaquelein, yielding the Vendéeans an enormous booty of cannon, muskets, and 5,000 prisoners (5 May 1793). Days after defeating Royalist rebel leader Francois-Athanase Charette south of the Loire GPE at Montaigu GPE , Republican General Jean-Baptiste Kléber pursued him to nearby Torfou and was routed in a brilliant rebel counterattack. Kléber skillfully disengaged and withdrew northwest pursued by Charles Bonchamp towards Pallet, while Charette returned west to recapture Montaigu GPE (19 September 1793).

FIGURE A.1: The entire text describing the battles of the war in the French *Vendée*.

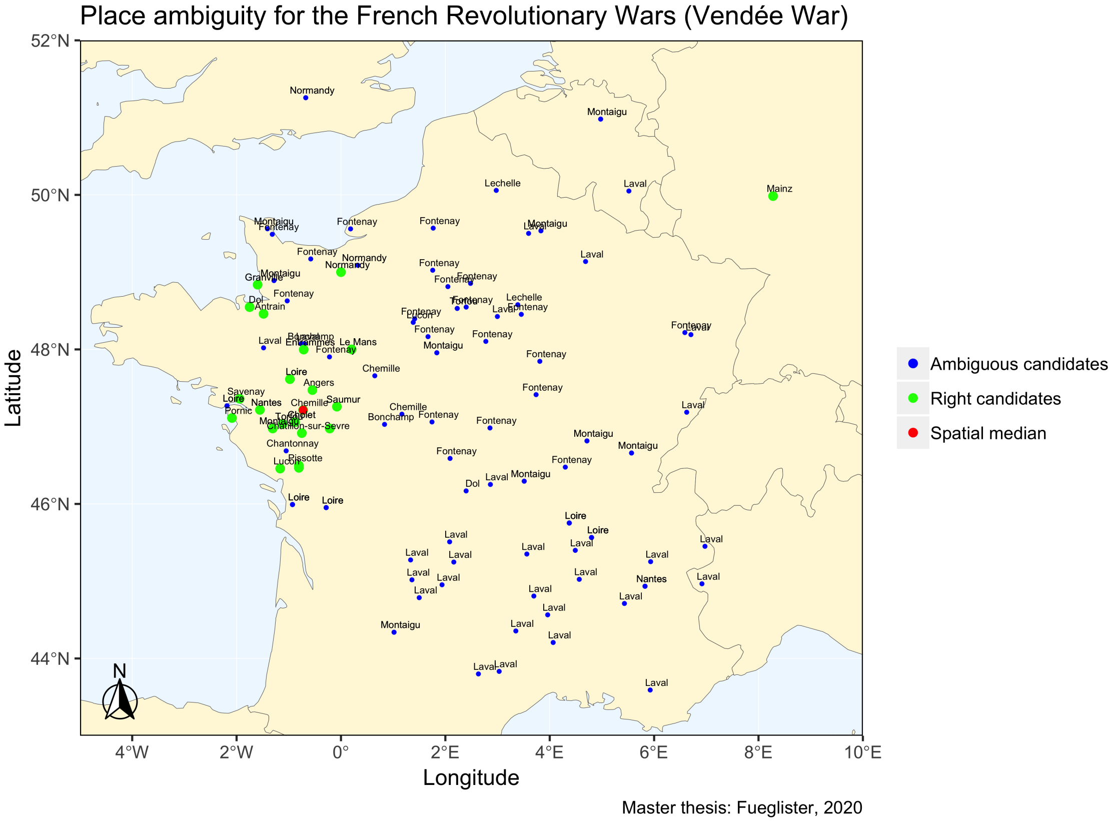


FIGURE A.2: Map showing the resulting locations together with the ambiguous locations within France, as well as the spatial median for the *Vandée War*.

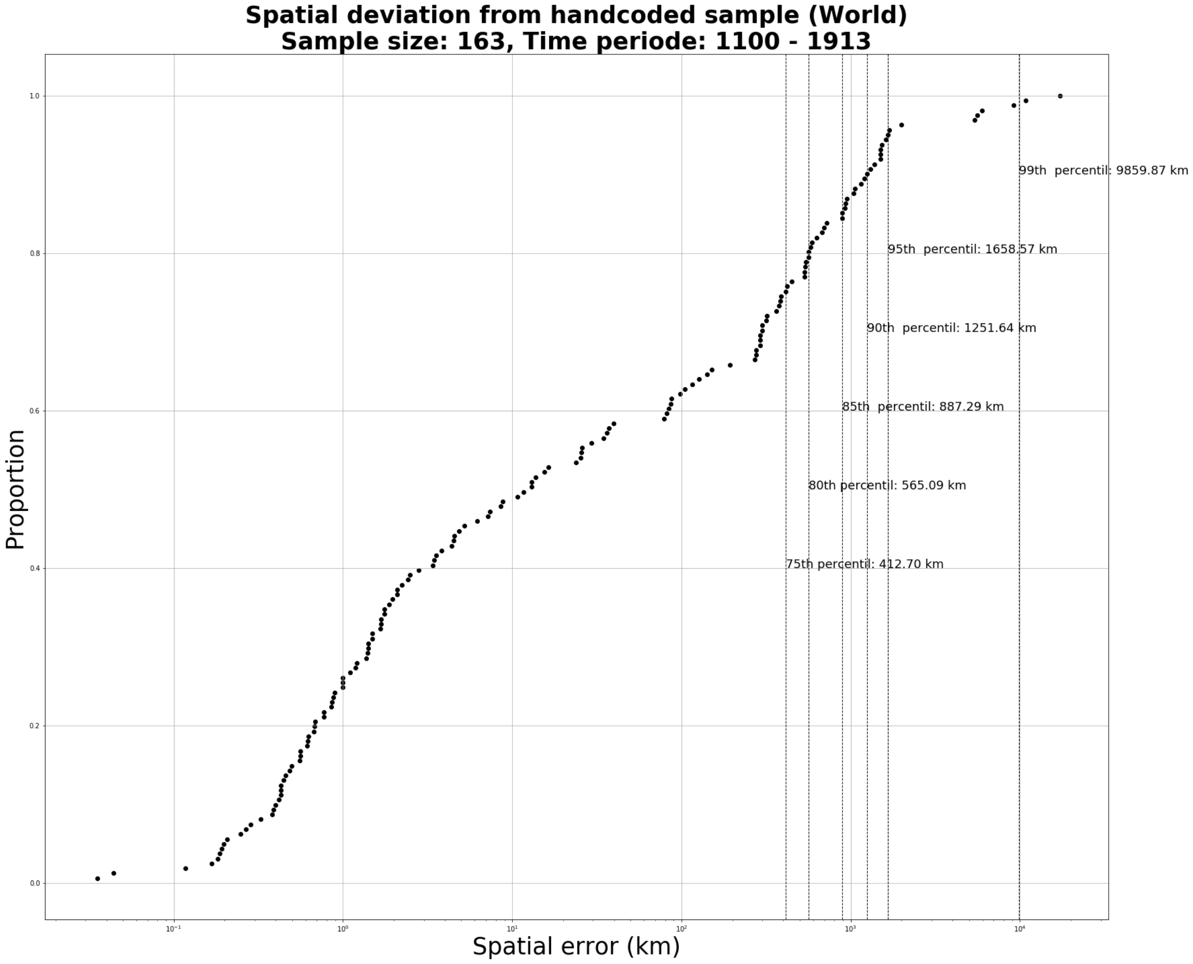


FIGURE A.3: Spatial error of hand-coded battles in Europe with battles outside Europe.

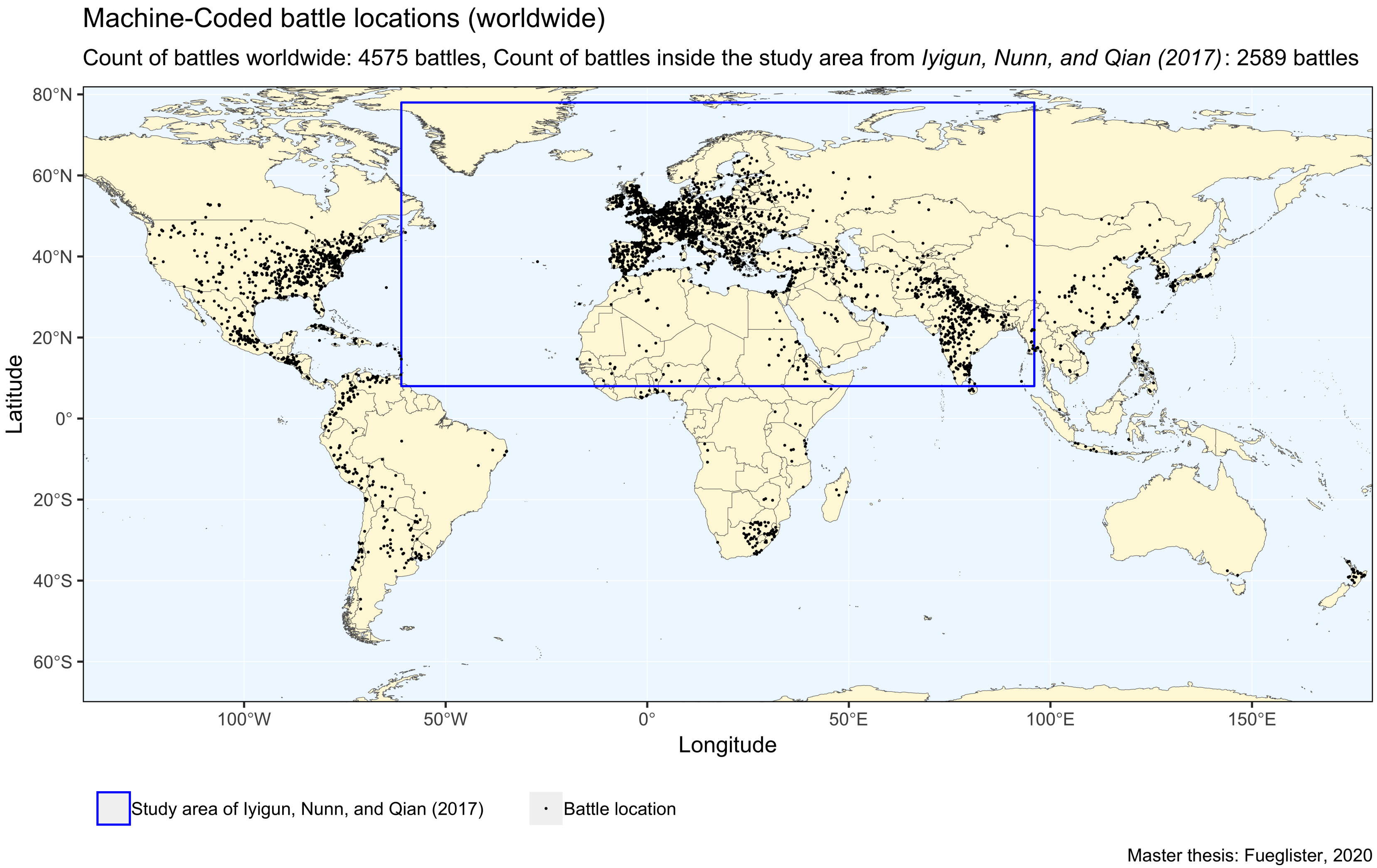

Machine-Coded battle locations (worldwide)
Count of battles worldwide: 4575 battles, Count of battles inside the study area from *Iyigun, Nunn, and Qian (2017)*: 2589 battles
80°N
60°N
40°N
20°N
0°
20°S
40°S
60°S
Latitude
100°W
50°W
0°
50°E
100°E
150°E
Longitude
Study area of Iyigun, Nunn, and Qian (2017)
Battle location
Master thesis: Fueglister, 2020

FIGURE A.4: All geocoded battles worldwide

Bibliography

Acheson, E., De Sabbata, S., and Purves, R. S. (2017). “A quantitative analysis of global gazetteers: Patterns of coverage for common feature types”. *Computers, Environment and Urban Systems* 64, pp. 309–320.

Ahlers, D. (2013). “Assessment of the accuracy of GeoNames gazetteer data”. *Proceedings of the 7th workshop on geographic information retrieval*. ACM, pp. 74–81.

Amitay, E., Har’El, N., Sivan, R., and Soffer, A. (2004). “Web-a-where: geotagging web content”. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 273–280.

Azar, E. E. (1980). “The Conflict and Peace Data Bank (COPDAB) Project”. *The Journal of Conflict Resolution* 24.1, pp. 143–152.

Backer, D., Bhavnani, R., and Huth, P. (2016). *Peace and Conflict 2016*. Peace and Conflict. Taylor & Francis.

Barnett, M. (2020). *regex*. URL: `https://pypi.org/project/regex/` (visited on 01/30/2020).

Bennett, B. (2010). “Spatial Vagueness”. *Methods for Handling Imperfect Spatial Information*. Ed. by R. Jeansoulin, O. Papini, H. Prade, and S. Schockaert. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 15–47.

Bonacker, T. and Imbusch, P. (1999). “Begriffe der Friedens-und Konfliktforschung: Konflikt, Gewalt, Krieg, Frieden”. *Friedens-und Konfliktforschung*. Springer, pp. 73–116.

Braithwaite, A. (2010). “MIDLOC: Introducing the Militarized Interstate Dispute Location dataset”. *Journal of Peace Research* 47.1, pp. 91–98.

Branch, J. (2016). “Geographic information systems (GIS) in international relations”. *International Organization* 70.4, pp. 845–869.

Brecke, P. (1999). “Violent conflicts 1400 AD to the present in different regions of the world”. *1999 Meeting of the Peace Science Society, unpublished manuscript*.

Brunner, T. J. and Purves, R. S. (2008). “Spatial Autocorrelation and Toponym Ambiguity”. *Proceedings of the 5th Workshop on Geographic Information Retrieval*. GIR ’08. Napa Valley, California, USA: Association for Computing Machinery, 25–26.

Buscaldi, D. (2011). “Approaches to disambiguating toponyms”. *Sigspatial Special* 3.2, pp. 16–19.

Buscaldi, D. and Magnini, B. (2010). “Grounding toponyms in an Italian local news corpus”. *Proceedings of the 6th workshop on geographic information retrieval*, pp. 1–5.

Buscaldi, D. and Rosso, P. (2008a). “Map-based vs. knowledge-based toponym disambiguation”. *Proceedings of the 5th Workshop on Geographic Information Retrieval*, pp. 19–22.

Buscaldi, D. and Rosso, P. (2008b). “A conceptual density-based approach for the disambiguation of toponyms”. *International Journal of Geographical Information Science* 22.3, pp. 301–313.

Clausewitz, C. (1883). *Vom kriege: Hinterlassenes werk des generals Carl von Clausewitz*. Vol. 1. R. Wilhelmi.

Clodfelter, M. (2017). *Warfare and Armed Conflict: A Statistical Reference to Casualty and Other Figures, 1500-20*. 4th ed. Jefferson, NC: McFarland.

Coventry, K. R. and Garrod, S. C. (2004). *Saying, seeing and acting: The psychological semantics of spatial prepositions*. Psychology Press.

Creasy, E. (1851). *The Fifteen Decisive Battles of The World from Marathon to Waterloo*. Burt, New York. ISBN: 9783734030154.

Derungs, C. and Purves, R. S. (2016). “Mining nearness relations from an n-grams Web corpus in geographical space”. *Spatial Cognition & Computation* 16.4, pp. 301–322.

Dincecco, M. and Onorato, M. G. (2016). “Military conflict and the rise of urban Europe”. *Journal of Economic Growth* 21.3, pp. 259–282.

Earl, J., Martin, A., McCarthy, J. D., and Soule, S. A. (2004). “The use of newspaper data in the study of collective action”. *Annu. Rev. Sociol.* 30, pp. 65–80.

Eck, K. (2012). “In data we trust? A comparison of UCDP GED and ACLED conflict events datasets”. *Cooperation and Conflict* 47.1, pp. 124–141.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). “One sense per discourse”. *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, pp. 233–237.

Gleditsch, K. S., Metternich, N. W., and Ruggeri, A. (2014). “Data and progress in peace and conflict research”. *Journal of Peace Research* 51.2, pp. 301–314.

Goldberg, Y. (2017). “Neural network methods for natural language processing”. *Synthesis Lectures on Human Language Technologies* 10.1, pp. 1–309.

Gouachevalier, M. (2015). *The french counter revolution*.

Gritta, M., Pilehvar, M. T., Limsopatham, N., and Collier, N. (2018). “What’s missing in geographical parsing?” *Language Resources and Evaluation* 52.2, pp. 603–623.

Grossner, K., Janowicz, K., and Kessler, C. (2016). “Place, period, and setting for linked data gazetteers”. *Placing names: Enriching and integrating gazetteers*, pp. 80–96.

Hallberg, J. D. (2012). “PRIO Conflict Site 1989-2008: A Geo-Referenced Dataset on Armed Conflict”. *Conflict Management and Peace Science* 29.2, pp. 219–232.

Hammond, J. and Weidmann, N. B. (2014). “Using machine-coded event data for the micro-level study of political violence”. *Research & Politics* 1.2.

Harbottle, T. B. (1904). *Dictionary of Battles from the Earliest Date to the Present Time*. Vol. 10. S. Sonnenschein & Company, Limited.

Hill, L. L. (2000). “Core elements of digital gazetteers: placenames, categories, and footprints”. *International Conference on Theory and Practice of Digital Libraries*. Springer, pp. 280–290.

Hirschberg, J. and Manning, C. D. (2015). “Advances in natural language processing”. *Science* 349.6245, pp. 261–266.

Honnibal, M. and Montani, I. (2017). “spaCy: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing”. *To appear*. URL: `https://spacy.io`.

Hulaman, L., Kathman, J., and Shannon, M. (2014). “Beyond Keeping Peace: United Nations Effectiveness in the Midst of Fighting”. *American Political Science Review* 108.4, 737–753.

Iyigun, M., Nunn, N., and Qian, N. (2017). “The Long-run Effects of Agricultural Productivity on Conflict, 1400-1900”. Working Paper Series No. 24066.

Jacques, T. (2007). *Dictionary of Battles and Sieges: A Guide to 8,500 Battles from Antiquity Through the Twenty-first Century*. Greenwood Press Westport.

Lacina, B. and Gleditsch, N. P. (2005). “Monitoring trends in global combat: A new dataset of battle deaths”. *European Journal of Population/Revue Européenne de Démographie* 21.2-3, pp. 145–166.

Laffin, J. (1986). *Brassey’s battles: 3,500 years of conflict, campaigns, and wars from AZ*. Potomac Books Inc.

Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. University of Chicago press.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). “Neural architectures for named entity recognition”. *arXiv preprint arXiv:1603.01360*.

Langacker, R. W. (1986). *Foundations of cognitive grammar: Theoretical prerequisites*. Vol. 1. Stanford university press.

Larson, R. R. (1996). “Geographic information retrieval and spatial browsing”. *Geographic information systems and libraries: patrons, maps, and spatial information [papers presented at the 1995 Clinic on Library Applications of Data Processing, April 10-12, 1995]*.

Leetaru, K. (2011). “Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space”. *First Monday* 16.9.

– (2012). “Fulltext Geocoding Versus Spatial Metadata for Large Text Archives: Towards a Geographically Enriched Wikipedia”. *D-Lib Mag.* 18.9/10.

Leetaru, K. and Schrodt, P. A. (2013). “GDELT: Global data on events, language, and tone, 1979-2012”. *International Studies Association annual conference, San Francisco, CA*. Vol. 2. 4, pp. 1–49.

Leidner, J. L. and Lieberman, M. D. (2011). “Detecting Geographical References in the Form of Place Names and Associated Spatial Natural Language”. *SIGSPATIAL Special* 3.2, 5–11.

Leveling, J. and Hartrumpf, S. (2008). “On metonymy recognition for geographic information retrieval”. *International Journal of Geographical Information Science* 22.3, pp. 289–299.

Levinson, S. C. (2003). “Frames of reference”. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Language Culture and Cognition. Cambridge University Press, 24–61.

Manguinhas, H., Martins, B., and Borbinha, J. (2008). “A geo-temporal web gazetteer integrating data from multiple sources”. *2008 Third international conference on digital information management*. IEEE, pp. 146–153.

McCurley, K. S. et al. (2001). “Geospatial mapping and navigation of the web.” *WWW* 1, pp. 221–229.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). “Efficient estimation of word representations in vector space”. *arXiv preprint arXiv:1301.3781*.

Münkler, H. (2002). *Über den Krieg: Stationen der Kriegsgeschichte im Spiegel ihrer theoretischen Reflexion*. Velbrück Wiss.

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., Murdock, V., et al. (2018). “Geographic information retrieval: Progress and challenges in spatial search of text”. *Foundations and Trends in Information Retrieval* 12.2-3, pp. 164–318.

Raleigh, C., Linke, A., Hegre, H., and Karlsen, J. (2010). “Introducing ACLED: an armed conflict location and event dataset: special data feature”. *Journal of peace research* 47.5, pp. 651–660.

Richardson, L. F. (1960). *Statistics of deadly quarrels*. Vol. 1960. Boxwood Pr.

Roberts, M. (1995). “The military revolution, 1560-1660”. *The military revolution debate : readings on the military transformation of early modern Europe*. Ed. by C. J. Rogers. Westview Press, pp. 13–36.

Roller, S., Speriosu, M., Rallapalli, S., Wing, B., and Baldridge, J. (2012). “Supervised text-based geolocation using language models on an adaptive grid”. *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics, pp. 1500–1510.

Shinyama, Y. (2019). *PDFMiner*. URL: `https://pypi.org/project/pdfminer/` (visited on 01/30/2020).

Singer, J. D. and Small, M. (1982). *Resort to Arms: International and Civil War,1816–1980*. Beverly Hills, CA: Sage.

Singer, P. (2001). “Corporate Warriors: The Rise of the Privatized Military Industry and Its Ramifications for International Security”. *International Security* 26.3, pp. 186–220.

Smith, D. A. and Crane, G. (2001). “Disambiguating geographic names in a historical digital library”. *International Conference on Theory and Practice of Digital Libraries*. Springer, pp. 127–136.

Smith, D. A. and Mann, G. (2003). “Bootstrapping toponym classifiers”. *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, pp. 45–49.

Southall, H., Mostern, R., and Berman, M. L. (2011). “On historical gazetteers”. *International Journal of Humanities and Arts Computing* 5.2, pp. 127–145.

Stina, H. (2019). “UCDP GED Codebook version 19.1”. *Department of Peace and Conflict Research*.

Sundberg, R. and Melander, E. (2013). “Introducing the UCDP georeferenced event dataset”. *Journal of Peace Research* 50.4, pp. 523–532.

Sweetman, J. (2004). *A dictionary of European land battles: from the earliest times to 1945*. Spellmount, Limited Publishers.

Talmy, L. (1983). “How language structures space”. *Spatial orientation*. Springer, pp. 225–282.

Teitler, B., Lieberman, M., Panozzo, D., Sankaranarayanan, J., Samet, H., and Sperling, J. (2008). “NewsStand: A new view on news”. *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. ACM, p. 18.

The Getty Research Institute (2017). “Getty thesaurus of geographic names”. URL: `https://www.getty.edu/research/tools/vocabularies/tgn/`.

Themnér, L. and Wallensteen, P. (2014). “Armed conflicts, 1946-2013”. *Journal of Peace Research* 51.4, pp. 541–554.

Tilly, C. (1985). “War making and state making as organized crime”, pp. 121–139.

Tonchi, V., Lindeke, W., and Grotpeter, J. (2012). *Historical Dictionary of Namibia*. Historical Dictionaries of Africa. Scarecrow Press.

United Nations. Dept. of Economic and Social Affairs (1974). *United Nations Conference on the Standardization of Geographical Names*. Tech. rep.

Urdal, H. and Hoelscher, K. (2012). “Explaining urban social disorder and violence: An empirical study of event data from Asian and sub-Saharan African cities”. *International Interactions* 38.4, pp. 512–528.

Valentino, B., Huth, P., and Balch-Lindsay, D. (2004). ““Draining the sea”: mass killing and guerrilla warfare”. *International Organization* 58.2, pp. 375–407.

Weber, M. (1921). *Wirtschaft und gesellschaft: Grundriss der verstehenden Soziologie*. Mohr Siebeck.

Weidmann, N. B. (2015). “On the accuracy of media-based conflict event data”. *Journal of Conflict Resolution* 59.6, pp. 1129–1149.

Weischedel, R. M. and Consortium, L. D. (2013). *OntoNotes release 5.0*.

Wick, M. (2012). “GeoNames”. URL: `http://www.geonames.org/`.

Wright, Q. (1942). *A study of war*. University of Chicago Press.

Yang, J. and Zhang, Y. (2018). “NCRF++: An Open-source Neural Sequence Labeling Toolkit”. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Declaration of Authorship

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.

Signed: B. Füglister

Date: 25.04.2020