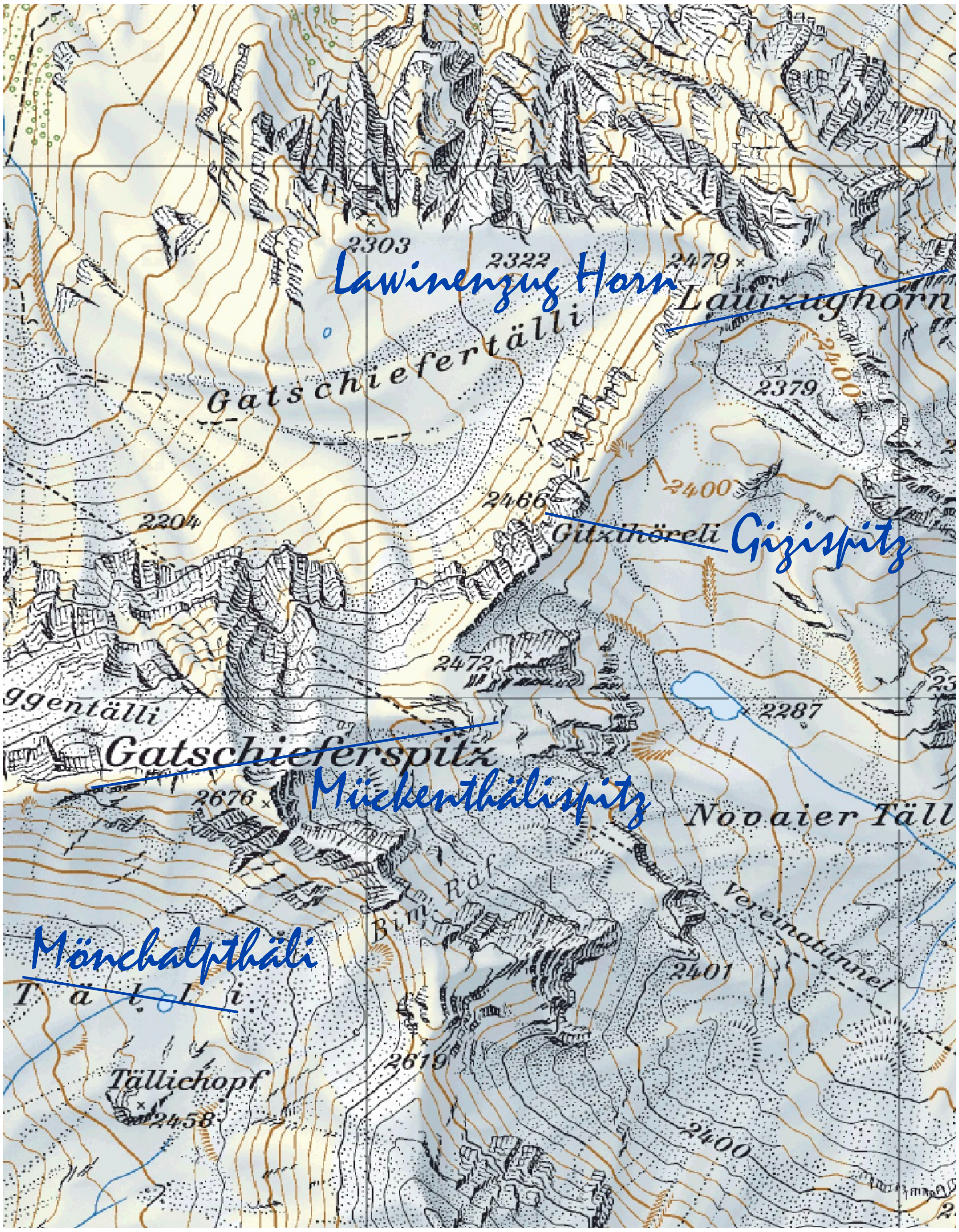
Identifying Unregistered Vernacular Toponyms and Alternative Spellings of Placenames in Switzerland

Named Entity Recognition in the Text+Berg Corpus



Master Thesis by Linda Ettlin

Image on Title Page
Source of underlying map: Swiss Federal Office of Topography

Universität Zürich^UZH

Department of Geography
Geographic Information Systems

Master Thesis GEO 511 Summer 2011

Identifying Unregistered Vernacular Toponyms and Alternative Spellings of Placenames in Switzerland

Named Entity Recognition in the Text+Berg Corpus

Linda Ettlin
Mittelweg 14
4142 Münchenstein
lettlin@geo.uzh.ch
05 823 703

Submission Date: August 31st 2011

Faculty Representative:	Prof. Dr. Robert Weibel
Adviser:	Dr. Ross Purves
Co-Adviser:	Prof. Dr. Martin Volk

Preface

This thesis may be signed with my name, but it would never have reached completion without the help and support of many others. Thank you to:

- Dr. Ross Purves, for his useful suggestions, continuous encouragement and most of all, his patience,
- Prof. Dr. Martin Volk, for his insights into a completely new area of research,
- Maya Bangerter, for her help with Perl and the gold standard,
- Mr. Robert Frey and the Rega employees, for taking an interest in my work and giving me a better understanding of how things work at the Rega control centre,
- Mr. Beat Dittli, for explaining about the research done by toponymists,
- Ms. Ramya Venkateswaran, for lending me her code,
- my co-students, for the excellent company,
- Oliver Burkhard and Kaspar Manz, for sharing countless meals of Bones and saving the day more than once,
- my other flatemates and friends, for the laughter and turning a blind eye to the dirty kitchen floor,
- Anna Belcher and Bethan Harries, for being there in the crucial moment,
- Lynne Theuns, for getting married at exactly the right time,
- Kaspar Meier, for the book and arranging the initial contact,
- Ricarda, for proofreading,
- my parents, for their continuous support, my GA and understanding that there just aren't enough hours to a day
- and finally to Theo, for luring me into the sun and reminding me of the important things in life.

Abstract

Toponyms, or as they are commonly known, placenames, are ever-present in day-to-day life. The names used by people to talk about places in everyday life are called vernacular toponyms and are typically not registered in administrative gazetteers. This can particularly cause problems for emergency services like the Swiss air-rescue organisation Rega who deal with toponyms in circumstances where successful communication about geographic space is often a matter of life or death. In order to improve this situation, more knowledge is needed on vernacular toponyms.

The aim of this thesis is to identify unregistered Swiss vernacular toponyms and alternatively spelled Swiss toponyms in the German articles of the Text+Berg corpus. The Text+Berg corpus is a collection of digitised and annotated yearbooks and magazines published by the Swiss Alpine Club. The toponyms are extracted using named entity recognition (NER) techniques. A lot of work in NER has been done for the English language, however, since the differences between the two languages are considerable, methods developed for English are not adaptable for German. New approaches are called for.

The rule-based NER system developed for this thesis project extracts normal nouns and named entities from the Text+Berg corpus. Three look-up lists are used to exclude normal German words, foreign toponyms and registered Swiss toponyms. In three data postprocessing steps the lists of extracted candidate toponyms are refined by taking into account alternative and antiquated spelling as well as hyphens and slashes in the tokens. In a final step, tokens and token sequences with characteristic toponym endings and beginnings respectively (e.g. *-horn* and *Piz*) are extracted from the lists of candidate placenames.

The results are analysed under various aspects, such as the number of hits a candidate toponym scores in an Internet search or in which year a candidate toponym was last mentioned in the corpus. Additionally, the results are checked for toponyms which are similar to registered Swiss toponyms (Levenshtein distance = 1). In general, the toponyms extracted from the Text+Berg corpus are not found very often on Swiss webpages, nor do they usually appear more than 20 times in the Text+Berg corpus. Also, about 80% of the toponyms were last mentioned in the corpus before the 1950s. Combined, these facts imply that the toponyms are of fine granularity (Zipf's law) and hence Rega could benefit from including fine granularity vernacular toponyms, such as those used by rock climbers and divers, in their gazetteer. Additionally, the search algorithm used by the Rega geographic information system should allow for slight spelling variations in toponyms because the various Swiss languages and dialects often complicate toponym spelling. This is highlighted by the fact that 13% of the extracted placenames were found to be within a Levenshtein distance of 1 from a toponym registered in SwissNames.

The results are evaluated by comparison to gold standard articles and the results of a NER approach designed to identify the names of mountains, glaciers and cabins in the Text+Berg corpus. The gold standard comparison shows that the thesis NER system preforms better at extracting unregistered or alternatively spelled toponyms than the NER approach for mountains, glaciers and cabins. Though both approaches achieved maximum precision values of 100%, the highest recall values were around 50% for the thesis NER system and 25% for the other NER approach. The evaluations also show, that using look-up lists has several drawbacks: Many unregistered Swiss toponyms are discarded because they match entries in the look-up lists (e.g. due to geo-non geo and geo-geo ambiguity). At the same time, false positives are included. Especially OCR mistakes pose a large problem in this respect. NER for compound toponyms is shown to be more difficult and less successful than for one-word toponyms. However, final precision values for both one-word and compound toponym NER are relatively good. 10% of the roughly 7'000 identified candidate toponyms are manually evaluated to estimate the NER system's overall precision, rendering 82% for one-word and 57% for compound toponyms. It is thus shown that NER for German can be accomplished using rules to extract unregistered and alternatively spelled toponyms and that satisfactory results can be achieved.

To build Swiss vernacular gazetteers, the work of Swiss toponymists should be used. The placenames used by certain interest groups, including rock climbers and divers, provide a further source of vernacular and particularly, fine granularity, toponyms.

It is hoped that the knowledge gained through this work may serve in the development of a state-of-the-art hybrid NER system for German. Future research should strive towards developing NER techniques for German which achieve good results comparable to those attained for English.

Zusammenfassung

Toponyme oder Ortsnamen, wie sie im Allgemeinen genannt werden, sind allgegenwärtig im täglichen Leben. Die von der Bevölkerung im Alltag verwendeten Ortsbezeichnungen werden umgangssprachliche Toponyme genannt und sind typischerweise nicht in administrativen Gazetteers enthalten. Insbesondere für Rettungsdienste wie die Schweizerische Rettungsflugwacht Rega kann dies Schwierigkeiten verursachen, da die möglichst genaue und rasche Identifikation eines Ortes entscheidend ist, um Leben zu retten. Mehr Wissen über umgangssprachliche Toponyme ist erforderlich, um diesem Problem entgegen zu wirken.

Das Ziel dieser Arbeit ist es, nicht registrierte umgangssprachliche schweizerische Toponyme und alternative Schreibweisen von schweizerischen Toponymen in deutschen Artikeln des Text+Berg Korpus zu identifizieren. Das Text+Berg Korpus ist eine Sammlung von digitalisierten und annotierten Jahrbüchern und Monatsschriften, welche vom Schweizer Alpen-Club herausgegeben wurden. Die Ortsnamen werden mit einem Named Entity Recognition (NER) Verfahren aus den Texten extrahiert. Im Bereich NER wurde bisher viel Forschung für die englische Sprache gemacht. Da sich die beiden Sprachen jedoch zu stark unterscheiden, lassen sich Verfahren, die für englische Texte entwickelt wurden, nicht einfach auf deutsche übertragen. Für die deutsche Sprache werden neue NER Verfahren benötigt.

Das regelbasierte NER System, welches im Rahmen dieser Arbeit entwickelt wurde, erkennt normale Nomen und Named Entities im Text+Berg Korpus. Es werden drei Listen verwendet, um normale deutsche Wörter, fremde Toponyme und registrierte Schweizer Toponyme zu eliminieren. Die Listen der extrahierten potentiellen Toponyme werden darauf in drei Schritten nachbereitet, indem sowohl alternative und veraltete Schreibweisen als auch Binde- und Schrägstriche berücksichtigt werden. In einem letzten Schritt werden dann mögliche Toponyme mit charakteristischen Endungen (z.B. *-horn*, für aus einem Wort bestehende Toponyme), bzw. Toponyme mit charakteristischen Anfängen (z.B. *Piz* für aus mehr als einem Wort bestehende Toponyme) in den NER-Ausgabelisten identifiziert.

Die Resultate werden auf verschiedene Merkmale hin untersucht, so z.B. wie viele Treffer ein potentielles Toponym bei einer Internetsuche erzielt, oder in welchem Jahr ein potentielles Toponym zum letzten Mal im Korpus erwähnt wurde. Ausserdem werden die Resultate nach Toponymen untersucht, die grosse Ähnlichkeit mit einem registrierten Schweizer Ortsnamen aufweisen (Levenshteindistanz = 1). Im Allgemeinen treten identifizierte Toponyme sowohl auf einer Schweizer Internetseite als auch im Korpus selten auf. Ausserdem werden ungefähr 80% der erkannten Toponyme vor 1950 zum letzten Mal im Korpus erwähnt. In dieser Kombination lassen die Fakten vermuten, dass die gefundenen Toponyme zu kleinen geographischen Einheiten gehören (Zipf's Gesetz) und somit könnte die Rega ihr Gazetteer sinnvoll aufwerten, indem die Namen von kleinen Orten eingefügt werden, welche z.B. von Kletterern und Tauchern benannt wurden. Zusätzlich sollte der Suchalgorithmus des von der Rega benutzten geographischen Informationssystem auch mit kleinen Änderungen in der Schreibweise eines Toponyms zurechtkommen, da in der

Schweiz durch die vielen Sprachen und Dialekte die Schreibweise von Toponymen oft unklar ist. Dies wird noch unterstrichen durch die 13% der identifizierten Ortsnamen, welche sich nur durch eine Levenshteindistanz von 1 von einem registrierten Schweizer Toponym unterscheiden.

Die Resultate werden evaluiert, indem sie mit einem Goldstandard und den Resultaten eines anderen NER Verfahrens verglichen werden. Letzteres wurde entwickelt, um die Namen von Bergen, Gletschern und Hütten im Text+Berg Korpus zu erkennen. Der Vergleich mit dem Goldstandard zeigt, dass das NER Verfahren dieser Arbeit bei der Erkennung von nicht registrierten oder alternativ geschriebenen Toponymen besser abschneidet als das Verfahren, welches Berg-, Gletscher- und Hüttennamen identifiziert. Obwohl beide Systeme einen maximalen Präzisions-Wert von 100% erzielen, liegen die höchsten Recall-Werte bei 50% für das Verfahren dieser Arbeit bzw. bei 25% für das andere NER Verfahren. Die Evaluation zeigt ebenfalls auf, dass die Verwendung von Listen im NER Verfahren einige Nachteile mit sich bringt: Viele nicht registrierte Schweizer Toponyme werden verworfen, weil sie einem Listeneintrag entsprechen (z.B. aufgrund von Geo-Geo- oder Geo-NonGeo-Ambiguität). Auf der andern Seite werden Wörter, die nicht Toponyme sind, fälschlicherweise in die Listen der NER Resultate aufgenommen. Vor allem OCR Fehler bereiten in dieser Hinsicht Probleme. NER für Toponyme, die aus mehreren Wörtern bestehen, erweist sich als schwieriger und weniger erfolgreich als für aus einem Wort bestehende Toponyme. Trotzdem sind die endgültigen Präzisions-Werte für Toponyme aus einem wie auch für Toponyme aus mehreren Wörtern relativ gut. 10% der rund 7'000 identifizierten potentiellen Toponyme werden manuell ausgewertet, um eine Schätzung des Präzisions-Wertes für das NER Verfahren zu erhalten. Dabei werden Präzisions-Werte von 82% für einfache Toponyme erzielt und 57% für aus mehreren Wörtern bestehende Ortsnamen. Somit wird gezeigt, dass regelbasierte NER Verfahren für die deutsche Sprache verwendet werden können, um nicht registrierte oder alternativ geschriebene Ortsnamen zu erkennen und dass es möglich ist, dabei befriedigende Resultate zu erzielen.

Um Gazetteers mit Schweizer umgangssprachlichen Ortsnamen zu erstellen, könnte sich die Zusammenarbeit mit Schweizer Ortsnamenforschern als hilfreich erweisen. Eine weitere mögliche Quelle umgangssprachlicher Toponyme bilden die Ortsnamen, welche von bestimmten Interessensgruppen, wie z.B. Kletterern und Tauchern, verwendet werden.

Das Wissen, welches aus dieser Arbeit hervorgeht, soll als Hilfe dienen, um ein modernes Hybridsystem für deutsches NER zu entwickeln. Zukünftige Forschung sollte sich weiterhin zum Ziel setzen, NER Verfahren für Deutsch zu entwickeln, die ähnlich gute Resultate liefern wie sie für Englisch erreicht werden.

Contents

List of Figures

List of Tables

Notation

Abbreviation	**Meaning**
CoNLL-2002	Conference of Computation Natural Language Learning in 2002
CoNLL-2003	Conference of Computation Natural Language Learning in 2003
DARPA	Defense Advanced Research Projects Agency
GIR	geographic information retrieval
GIS	geographic information system
HAREM	Evaluation Contest of Named Entity Recognition Systems for Portuguese
HMM	Hidden Markov Model
IE	information extraction
IR	information retrieval
IREX	Information Retrieval and Extraction Exercise
MUC	Message Understanding Conference
MUC-6	the sixth Message Understanding Conference
MUC-7	the seventh Message Understanding Conference
NE	named entity (abbreviation and part of speech tag)
NER	named entity recognition
NN	normal noun (part of speech tag)
OCR	optical character recognition
pos	part of speech
Rega	Swiss Air-Rescue (Schweizerische Rettungsflugwacht)
SAC	Swiss Alpine Club (Schweizer Alpen-Club)

1 Introduction

1.1 Context

Geographic placenames, so-called toponyms, are ubiquitous in everyday life. They are indispensable to communication in many situations, such as when giving directions or talking about the latest world news, for example. Toponyms are used in conversation, correspondence, reporting and documentation (Hill, 2006). Sometimes placenames are used in a formal way, for example when writing down an address. More often, however, toponyms are used informally and subconsciously, for instance when talking about a weekend trip or arranging a meeting place. The placenames used by people to communicate about space in everyday life are called vernacular toponyms.

Knowledge about toponyms is essential for rescue services. The site of an emergency can only be located if the placename given by the caller is known to the rescue service. Geographic information systems (GIS) are a useful aid in locating places and have come to play a central role for emergency services (Burenhult and Levinson, 2008). GIS use gazetteers to ground a toponym on a map. However, since only a fraction of vernacular toponyms are registered in administrative gazetteers, rescue services often have to deal with unregistered placenames, in which instance the GIS becomes useless (Davies et al., 2009). Hollenstein and Purves (2010) state the need for GIS which can also deal with vernacular geographic terms. This is underlined by Goodchild (2007) who points out that, in the case of an unknown or unclear emergency site, precious minutes are lost trying to determine the unambiguous location of the incident.

One example of an emergency service experiencing problems with unregistered vernacular toponyms is the Swiss air-rescue organisation *Rega* (acronym for “*re*scue” and “*g*arde *a*érienne”). Rega flies to the aid of people who are in situations beyond what conventional ambulances and rescue services can manage alone. Often this means that they provide assistance in remote and hard to access areas, such as mountainous regions.

When a call for help reaches the Rega Operations Centre at Zürich Airport, one of the first questions the rescue flight coordinator will try to clarify is *where* the emergency has taken place. As soon as the incident has been located a helicopter with a rescue team can be dispatched. Details, such as the patient's condition, background information concerning the situation and the exact coordinates of the person in distress, are then transmitted to the rescue team once they are en route to the site of the accident. This course of action ensures a rapid response, which is essential to the successful rescue of a person in danger or in need of medical assistance.

A critical link in this chain of events, which constitutes a rescue mission, is the location of the site of the emergency. There are many possible obstacles that can interfere with this initial step such as a bad telephone connection or a disoriented person making the emergency call. A significant problem is also posed by the gazetteer which the rescue flight coordinator uses to pinpoint locations. At Rega the most current version of a gazetteer called *SwissNames* is used. SwissNames is issued by the Federal Office of Topography swisstopo and has a good reputation concerning both quality and coverage. However, the smallest scale of the national maps made by swisstopo is 1:25'000. Since SwissNames contains only the toponyms which are used on these maps, this limit in resolution logically imposes a boundary for the granularity of the toponyms themselves: the smaller the scale of a map, the fewer toponyms can be displayed for a particular area if the map is to stay legible. This means that very local placenames, such as for example the name of a clearing in the forest or the shoulder of a mountain, are omitted (Piotrowski et al., 2010). Furthermore, despite the fact that Switzerland has four official languages, only few toponyms are included in more than one language on the national maps. Other names by which a geographic feature may be known are equally not included and are hence not listed in SwissNames. In other words, to find a place with the help of SwissNames, two conditions must be fulfilled: the place must be registered and the person conducting the search must know its *official* name. If, in any emergency situation, one of these conditions is not or cannot be satisfied, the rescue flight attendant must resort to other, time-consuming options to deduce the caller's location. Such delays can be dangerous - time is essential for all emergency services.

There are several reasons why toponyms exist which are not registered in administrative gazetteers. Some of these have already been mentioned in the preceding paragraphs. Subsequently, the four main sources of unregistered and alternatively spelled toponyms are summarised.

1.1.1 Vernacular Geography

Vernacular toponyms encompass the names of places which are not listed in an administrative gazetteer such as SwissNames. This could be, for instance, because the named feature is too small to be named in a map (e.g. *Grüenwändli* is the name of a limestone cliff close to St. Antönien in Eastern Switzerland which is used for rock climbing). Also, a place is often known by more than one name. Lay people know and use these alternative toponyms and are possibly not even aware of the official name as it is registered in SwissNames. Examples of such vernacular toponyms are *Alt-Münchenstein*, *Am Berg* and *Münchenstein Dorf* which are alternative names for the town of Münchenstein to distinguish it from the neighbouring town called Neumünchenstein.

The following two problems can also be regarded as subtypes of vernacular toponyms. They are mentioned separately here because they are particularly relevant to Switzerland with its multi-language culture and the toponym reforms which have changed the spellings and names of places several times over the past 60 years.

Language and Dialects

With four official languages, it is not unusual for a geographic feature in Switzerland to have as many names. Also, the large variety of dialects are cause for confusion, especially in the Swiss German speaking part of the country (e.g. *Burgdorf* is called *Burdlef* in the Bernese dialect and *Berthoud* in French). The discrepancy of these dialects to High German creates additional problems. For example, a location pronounced *Stöcklichrüz* in Swiss German could theoretically have the official name *Stöcklikreuz* or *Stöckchenkreuz*, which are High German variations. In this manner, different versions of just one toponym may exist in the various languages and dialects .

Old Versus New

As language and spelling evolve through time, so do the names of places (Hill, 2000). Only recently, the canton Thurgau decided to change the spelling of all names from their more or less High German version to the Swiss German pronunciations. In the course of this initiative, *Rotbühl* became *Roopel*, for example. Old versions of toponyms are not simply forgotten from one day to the next. Reforms and changes take time to establish themselves, thus creating a period where both new and old versions of placenames are part of the vernacular language.

1.1.2 Alternative Spellings

The way a certain toponym is spelled or punctuated can differ between documents (Hill, 2006). For example, *Wyssi Frau* could also be written as *Wissi Frau*. Rescue flight coordinators have to deal with the possibility of different spellings when entering a toponym in their GIS. Spelling variations are closely connected to the language and dialect problems as well as differences between old and new toponyms.

In this master thesis, I consider possibilities to improve Rega's toponym search and gazetteer by examining how lay people communicate about geographic places and by comparing the results to the swisstopo gazetteer SwissNames. To this end, I use named entity recognition (NER) techniques and string comparison programmes to extract and examine the toponyms that are mentioned in the yearbooks of the Swiss Alpine Club (SAC).

1.2 Motivation

Aside from the humanitarian aspect involved in doing research with the aim of aiding a rescue service, there is also significant scientific interest to motivate this thesis.

The SAC yearbooks are available as a digital text corpus[1]. Hence methods from the field of computational linguistics may be used for a systematic toponym search. This involves information extraction, which is closely related to information retrieval. Information retrieval concerns itself with finding material which holds useful information from a large collection of items. Especially because of the booming wealth of knowledge available on the Internet today, this field of study has great significance for the whole of the computerised world. On the website of the text engineering software GATE, which was developed at the

[1] A corpus is a collection of digitally available text, processed for a linguistic task. A word in a corpus is referred to as a token

University of Sheffield, this fact is expressed with the following words: “If information is power and riches, then it is not the amount that gives the values, but access at the right time and in the most suitable form” (GATE, sa).

Though much work has been done in the fields of both information extraction and information retrieval, most of the research has been done for the English language (Faruqui and Padó, 2010). Due to the capitalisation of all nouns and the complexity of the grammar, information extraction in German is more challenging and calls for a different approach (Volk and Clematide, 2001). As the SAC texts are written in German, this thesis poses an excellent oppurtunity to contribute to the development of new technologies for the recognition of toponyms in German.

1.3 Aims and Research Questions

The goal of this thesis is to extract unregistered vernacular and alternatively spelled Swiss toponyms from the Text+Berg corpus. The extracted toponyms are then analysed by examining the results under various aspects. This leads to the following three research questions:

Question 1: How can rule-based NER techniques be used to extract unregistered vernacular and alternatively spelled toponyms from a German corpus?

The fact that NER still uses rule-based approaches suggests that these methods are not without merit. However, in most cases lists of the names of persons, organisations and places are used to identify named entities. This will not be possible for unregistered or alternatively spelled placenames. In addition, little work has been done on NER for the German language. Since precisely the unregistered and alternatively spelled toponyms in a German corpus are of interest in this case, different rules must be implemented to extract these named entities.

Question 2: What are the characteristics of unregistered vernacular and alternatively spelled toponyms extracted from the corpus?

A characterisation of the results could present Rega with insight into how the problems caused by unregistered toponyms and alternative spellings can be tackled.

Question 3: What are the implications of the rules used to extract toponyms and the characteristics of these extracted placenames?

By analysing the rules used in the NER approach and combing this with the knowledge gained by characterising the NER results, the thesis NER system will be evaluated. The aim is to make suggestions for how the system could be adjusted to deliver results with certain desired characteristics.

1.4 Structure

Following this introduction, key terms such as information retrieval, information extraction and vernacular geography will be explained in more depth. An outline of the work that has been done in these fields of research is given in chapter 2, which is followed by a description of the data used (chapter 3). The methods applied to recognise toponyms in the SAC yearbooks are explained in chapter 4. The description of the results follows in chapter 5, after which both the results and the methodology are discussed and evaluated (chapter 6). Finally, conclusions are drawn in chapter 7.

2 Scientific Background

This chapter gives an overview of the areas of research which are relevant to the thesis. The significant terminology is explained and the important literature is briefly summarised. After an introduction to information retrieval (IR, section 2.1) and geographic information retrieval (GIR, section 2.2) the link is made to information extraction (IE, section 2.3) and named entity recognition (NER, section 2.4). Both GIR and NER are of direct relevance to the work done in this thesis. IR and IE are briefly introduced to give the reader a solid background and an overview of the importance and role of GIR and NER in the fields of information sciences and computational linguistics. A schematic overview of the relationships between these various fields of study is given in figure 2.1. IR holds the subfield GIR while geoparsing is a GIR subtask and corresponds to NER for toponyms in in the field of computational linguistics. NER is a subtask of IE.

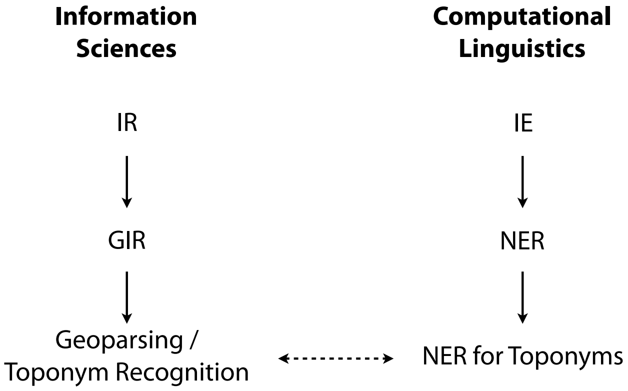


Figure 2.1: Schematic showing relationships between IR, GIR and geoparsing as well as, IE and NER.

The focus of the current chapter lies on the different approaches used in NER, which are central to this work. These are the rule-based approaches (subsection 2.4.3) and the machine learning based approaches (subsection 2.4.4). Techniques that combine these two methods are called hybrid approaches (subsection 2.4.5). Following the explanation of the various NER approaches, measures for the evaluation of NER are introduced (subsection 2.4.6). The NER section of the chapter closes with an outline of the work which has been done on NER for languages other than English (subsection 2.4.7).

Following this, details are given about the history, structure and purpose of gazetteers (section 2.5). Finally, the concept of vernacular geography is examined (section 2.6). This is set into perspective with the daily events at the Rega control centre (subsection 2.6.1), using information gained during a visit to the Rega base at Zurich airport and interviews with several rescue operators. The section closes with an insight into Swiss vernacular geography and Swiss toponyms (see subsection 2.6.2).

The research gaps and the relevance of the research questions are summarised at the end of this chapter in section 2.7.

2.1 Information Retrieval (IR)

As mentioned in the introduction, IR is an academic field of study, which is concerned with recovering documents containing certain desired information. Manning et al. (2008, p. 1) define IR more precisely as follows:

> "Information retrieval ... is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)."

With "material of an *unstructured* nature" Manning et al. refer to information which has not been brought into a format easily processed by computers. A classical example of structured material is a relational database, while unstructured information such as textual documents are less rigorously organised, with titles and paragraphs, for instance. Most texts have, of course, also the latent structure imposed by the grammar of the language in which the document is written (Manning et al., 2008). These subtle rules of language structure are used to process text in the field of information extraction (section 2.3).

IR emerged in the 1950s in response to the ever-growing amount of digitally accessible data (Singhal, 2001). During the first few decades, IR was only relevant to several specialised professions such as librarians and information experts. This changed with the arrival of the World Wide Web in the early 1990s. Enormous amounts of information became available to everyone with a computer and an Internet connection. This revolutionary development brought IR into the spotlight - efficient access to relevant information became a necessity for lay people as well (Baeza-Yates and Ribeiro-Neto, 1999).

Today, IR systems are as important as ever and an integral part of everyday life in human society, be it for searches on the web, on a private computer or within an enterprise, institution or domain. IR is used to launch a search via Apple's *Spotlight* or Window's *Instant Search*, for example. IR has also become valuable to people who deal with large amounts of information in their jobs, such as journalists, doctors and lawyers (Manning et al., 2008).

All the digital information which is available today, - be it on the Web, on a personal computer or in the files of an enterprise - would be worthless if it could not be accessed efficiently when the need arises. As any user of Internet search engines can confirm, IR generally works, but is far from perfect. Geography has become increasingly valuable to searches on the Web. To improve IR results, search queries are often refined by toponyms, for instance (Gan et al., 2008). If a parent is looking for good schools in a certain area, for example, it would be of little use to simply launch a Web search for *schools*. By adding the name of the locality, like *Basel*, to the query, the search results become much more pertinent. In such cases, placenames play an important role in IR. Generally, IR dealing with spatial information is referred to as *geographic* IR.

2.2 Geographic Information Retrieval (GIR)

As research on IR became a matter of public interest with the introduction of the World Wide Web, a new field of study began to form, which specialised IR for information with a spatial component. This relatively young area of research is called GIR. According to Kunz (2008), the first definition of GIR was the one given by Larson (1996, p. 83) in 1996:

> "Geographic Information Retrieval is concerned with providing access to georeferenced information sources. (...) It includes all of the areas that have traditionally formed the core of IR research with an emphasis, or addition, of spatially- and geographically- oriented indexing and retrieval."

Geographic information, that is information with a spatial reference, is omnipresent and one of the most significant types of information in this society (Lin and Ban, 2008). In fact, McDonald and Di (2003) estimated that as much as 80% of all information is geospatial in nature.

Purves and Jones (2006, p. 376) identify several challenges which GIR is still faced with today:

- "identification of geographic terms in documents and associating these terms with appropriate geographic locations;
- ways of indexing large collections efficiently for search on both thematic and geographic content;
- development of search engines and algorithms which can exploit such indexing systems;
- techniques to combine geographic and thematic relevance in appropriate ways;
- methods to allow users to formulate queries to such search systems; and
- design of interfaces and visualisations which allow users to effectively explore and assess returned document sets."

The first point mentioned in this list refers to what is known as *geotagging* (Lieberman et al., 2010), which is composed of two important GIR tasks often referred to as *geoparsing* and *geocoding* (Zubizarreta et al., 2009).

Geoparsing, also known as toponym recognition (Lieberman et al., 2010), is the process of recognising placenames in a text (Zubizarreta et al., 2009). One difficulty geoparsing contends with is geo-non geo ambiguity. This term refers to normal words which are used as toponyms. Examples of such geo-non geo ambiguous Swiss toponyms are *Mönch* (monk), *Jungfrau* (virgin) and *Berg* (mountain). In computational linguistics, geoparsing is called NER (Leidner, 2007), although NER is not only concerned with the identification of locations, but also with finding references to organisations and persons in a document. Geoparsing, or NER for locations, is the focus of this thesis. Since the techniques used to approach this GIR task originate in the field of NER (Brunner, 2008), the topic will be approached from the NER point-of-view later on in this chapter (see section 2.4).

Geocoding can be referred to as toponym resolution (Buscaldi and Rosso, 2008). This GIR subtask deals with assigning the correct set of geographic coordinates to the toponyms recognised during geoparsing (Zubizarreta et al., 2009). The main challenge in geocoding is dealing with ambiguous toponyms, that is placenames which can refer to various locations. This problem is called geo-geo ambiguity. There are, for instance, six different locations in Switzerland called *Wangen*, located in the cantons of Schwyz, Glarus and Zurich, not including the two other towns called *Wangen* colloquially but with the official names *Wangen an der Aare* (canton of Bern) and *Wangen bei Olten* (canton of Solothurn). Since geocoding requires the information supplied by a gazetteer and this thesis particularly looks at toponyms, which are *not* known to a gazetteer, geocoding will not be treated in further detail.

2.3 Information Extraction (IE)

Natural language, as opposed to programming languages, can be ambiguous and often the same concept could be communicated in several ways, using different words. This poses a large problem to systems which rely on understanding and interpreting natural language text, such as IR systems, for example (Vallez and Pedraza-Jimenez, 2007). Both Brants and Google Inc. (2004) and Bear et al. (1998) acknowledge that there is potential for improving IR with the help of IE, a field of research belonging to computational linguistics.

The essence of IE was captured in the words of the great Sherlock Homes: "It is of the highest importance in the art of detection to be able to recognise, out of a number of facts, which are incidental and which vital" (Doyle, 1894, p. 407). IE can be described as the automatic, digital version of detective work.

Sarawagi (2008, p. 263) defines IE as

> "... the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources."

IE is not to be confused with IR. While IR identifies documents that hold the desired information in response to a query, IE picks out salient bits of information from documents: "*Information Retrieval* gets sets of relevant documents - you analyse the *documents*. *Information Extraction* gets facts out of the documents - you analyse the *facts*" (GATE, sa).

As with IR, the arrival of the Internet promoted research on IE at a hitherto unknown scale. Many and various different applications require IE to retrieve structured information from unstructured documents. Examples of such applications include automatic news tracking, personal information management and various web oriented applications (Sarawagi, 2008).

The field of study became an area of greater interest in the late 1980s. In 1987, the US Defense Advanced Research Projects Agency (DARPA) supported the first of seven Message Understanding Conferences (MUCs) (Leidner, 2007). The original aim of these conferences was to encourage research on automatic text analysis, the special focus being on military communication. The MUCs have been said to play a central role in the field of IE by defining the research programme and advancing its progress significantly (Grishman and Sundheim, 1996).

These conferences are also referred to as contests, since the participants were expected to develop systems, which were then evaluated and compared to each other (Leidner, 2007). The systems were designed to solve a predefined IE task. Examples of such IE tasks are *template relationships* and *scenario template extraction* (Borthwick, 1999). A system preforming the template relationships task should be able to detect the relationship between *Calvin* and *Hobbes* from the phrase *Calvin's best friend Hobbes*. For the scenario template task, the programme is expected to answer text comprehension questions (given an appropriate text), such as:

> Where did Prince Charles go skiing?
>
> Who owns Apple Inc.?

Another of these IE tasks was the *named entity (NE) task*, which was first assigned for the sixth MUC (MUC-6) in 1995 (Stevenson and Gaizauskas, 2000). The NE task is also known as NER (see section 2.4) (Rössler, 2007).

2.4 Named Entity Recognition (NER)

2.4.1 What Is a NE?

The MUC-6 coined the term NE and it has since remained an expression singular to automatic language processing (Rössler, 2007). For the NE task, the MUC-6 (MUC-6 Appendix, 1995) defined NEs as

> "... proper names, acronyms, and perhaps miscellaneous other unique identifiers, which are categorized ... as follows:

> ORGANIZATION: named corporate, governmental, or other organizational entity [e.g. *Novartis*, *Lindt & Sprüngli*, *UBS*, *Ascom AG*, *Bank Coop*, *Stiftung für junge Auslandschweizer*, *Betty Bossi*]
>
> PERSON: named person or family [e.g. *Ursula Andress*, *Professor Euler*, *Fr. Spyri*, *Globi*, *Winkelried*, *von Haller*]
>
> LOCATION: name of politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.) [e.g. *Piz Ault*, *Seealpsee*, *Berner Oberland*, *Aletschgebiet*, *Rhein*, *Confederatio Helvetica*]"

Although this classification gives a good idea of what a NE might be, it is far from a precise definition. In fact, an unambiguous, universally applicable description of NEs is still lacking (Rössler, 2007). Borthwick (1999) and Rössler (2007) have attempted to circumscribe NEs with *proper names* (or *Eigennamen* in German) but this, too, captures only part of what NEs are. Mikheev et al. (1999) and Rössler (2007) settle on a more open definition and suggest that what a NE is depends on the context it is used in. Most of the literature, however, seems to adhere to the original definition stated in the MUC-6 (Nadeau and Sekine, 2007). Likewise, in this thesis the term NE will be used to refer to names of organisations, persons and locations.

2.4.2 What Is NER?

The NE task of the MUC-6 went a step beyond the MUC-6 definition of NEs as names of organisations, persons and locations. This IE subtask was described as the finding and classifying of all expressions in a document which could be assigned to one of the following seven categories: organisation (e.g. Rega), person (e.g. Florence Nightingale), location (e.g. Piz Bernina), time (e.g. eight o'clock), date (e.g. 1^{st} October), monetary amounts (e.g. € 5.20) and percent expressions (e.g. 99%) (MUC-6 Appendix, 1995).

The NE task can, ultimately, be described as *recognising NEs, temporal expressions and expressions of quantities* in a text. Hence the NE task is also called NE *recognition* (NER). In the words of Kozareva (2006, p. 15),

> "NER consists in detecting the most silent and informative elements in a

> text such as names of people, company names, location, monetary currencies, dates."

Since the NE term is not clearly defined, there logically is also a lack of consensus regarding the concept of NER. Some, like Kozareva, adhere to the original NE task as it was laid out by the MUC-6. Borthwick (1999, p. 1) adds to this the category "... none-of-the-above ..." while Leidner (2007, p. 57) includes a NE which he refers to as "... other proper names ..." in order to include for example names of ships (e.g. Titanic) and pets (e.g. Garfield). Others, like Chieu and Ng (2003), reduce NER to four categories: person, organisation, location and miscellaneous. Rössler (2007), along with the majority of the literature, considers only the MUC-6 entities - that is organisations, persons and locations - for NER. In accordance with the previously stated definition of NEs (see section 2.4.1), this is also the understanding that will be adopted for this thesis.

NER is a necessary building block for many IE tasks, such as the mentioned template relationships and scenario template tasks: before answering questions concerning the relationship between NEs or the content of a document, it is necessary to recognise the NEs themselves. Being a prerequisite for IE tasks, NER is consequently also of significance for effective IR (Mikheev et al., 1999). Sekine and Isahara (1999, p. 1) call NER "one of the basic techniques in IR and IE".

There are various methods used to preform the task of NER. Two main approaches can be distinguished: the *rule-based* (section 2.4.3) and the *machine learning based approach* (section 2.4.4) (Rössler, 2007), (Chieu and Ng, 2003). The so-called *hybrid approach* (section 2.4.5), which combines both rules and machine learning, is sometimes listed as a third category (Liu et al., 2011).

2.4.3 Rule-based Approaches

Here, it will be distinguish between the list-based approach and other rule-based approaches. The list-based approach is treated separately, since it differs substantially from the other rule-based approaches and plays an important role in many hybrid approaches.

List-based Approach

The list-based method is probably the simplest way of accomplishing NER. It works by comparing the words in a text to the entries in precompiled lists of NEs such as first names, surnames, companies or locations. A list of locations is usually called a *gazetteer*. In GIR and NER concerned specifically with the recognition of toponyms, the list-based approach is referred to as the *gazetteer lookup approach* (Brunner, 2008). It is the oldest method to detect geographic names in text (Jones et al., 2001). Although such list- or gazetteer-based methods are straightforward in theory, they are, in fact, rather problematic in the application. Firstly, the compilation of the lists themselves presents a problem. Which names should be included? How many names should be included? Should all variations of a name (e.g. Monte Rosa, Monte-Rosa, Monterosa) be incorporated? It is clear that compiling an extensive list is a time-consuming task.

A further problem is the actual detection of the NEs in the text. Small variations such as the change due to grammatical cases (e.g. *Pischahorns* as the genitive form of *Pischahorn*) may be enough to cause a NE to be overlooked and any NEs not recorded in the lists will not be recognised at all. McDonald (1996) argues, that using only lists for NER is not sufficient, since one can never presume to include the names of *all* organisations, persons and locations. Furthermore, virtually any normal word can be used as part of a name, as is demonstrated by the following example: *"Her name was equally preposterous. April Wednesday, she called herself ..."* (MacLean, 1976, p.68). This means, that either *all* words are included in such lists, resulting in many false recognitions, or it must be accepted that many NEs will not be detected. Names of firms that look like person names are also problematic and cannot be resolved with the help of lists alone. Such an example is *Betty Bossi*, the name of a Swiss cook book publisher.

A purely list-based approach is naïve and has many serious drawbacks, as was illustrated previously. However, lists such as gazetteers are still useful tools when applied in combination with other approaches, and are frequently used for NER (Lin and Ban, 2008).

Other Rule-based Approaches

The rule-based approach is also referred to as the *handcrafted* (Borthwick, 1999) or *knowledge engineering approach* (Chieu and Ng, 2003). Each of these terms hints at a characteristic of the NER method discussed here. Systems using this approach use the regularities in a language to formulate certain *rules* for the detection of NEs. These rules have to be assembled and composed *by hand* (that is to say that this must be done by a person as opposed to a machine) and the author of these rules must possess *knowledge of the patterns and the grammar of a language.*

Even though the term NE was not established until 1995, the *Proper Name Facility* proposed by McDonald (1996), which is part of a system presented at the Workshop on Acquisition of Lexical Knowledge from Text in 1993, has been called *the* prototype of rule-based NER from which all other rule-based approaches have been derived (Rössler, 2007). In his approach, McDonald (1996) used internal and external evidence to complement the basic list-based method, which, at that time, was the hitherto applied approach (Rössler, 2007). McDonald (1996) introduced the terms *internal evidence* and *external evidence* to distinguish between the knowledge to be gained from the structure *within* a proper name and that which can be surmised from the *surrounding* text in which the name appears. For example, in *Ben & Jerry's Homemade Holdings, Inc.*, the abbreviation *Inc.* is internal evidence, indicating that *Ben & Jerry's Homemade Holdings* is a company. What, however, if *Ben & Jerry's* appears without this additional information? The company name could just as well be referring to two people called Ben Franklin and Jerry Goldsmith, for example. In such cases, the external evidence may be of some help. Consider for example the following sentence:

> *Ben & Jerry's is the best ice cream company in the world.*

Here, the context *ice cream company* makes it evident that *Ben & Jerry's* is the name of a firm.

Internal evidence is used to recognise temporal and quantity expressions such as monetary currencies and dates. Palmer and Day (1997) solved these two MUC-6 NER subtasks with very satisfactory results for the six languages Chinese, English, Spanish, Portuguese, French and Japanese by using less than 5 rules for detecting expressions of quantities and less than 30 rules to find temporal expressions.

Using internal and external evidence to detect names of organisations and persons is more complex since they don't always appear with indicators like Inc., & Co., Ltd., Mr., Dr. or Miss, and the internal structure of a person name (first name followed by a surname) is not always conclusive. To detect locations is arguably even more challenging, however, since there are many possible patterns (e.g. for mountains alone there are numerous indicators: *Mount* St. Helens, Albula *Range*, Rocky *Mountains*, Allalin *Group*, *Ben* Nevis, Gotthard *Massive*, Snowmass *Mountain*, Broad *Peak* etc.) and often placenames don't even have a specific internal structure (e.g. Oslo, Nile, Eiger).

The Proper Name Facility used by McDonald (1996) executes NER in three steps. First, the sequence of words, which make up the NE, is delimited, i.e. the beginning and the end of the NE are found. This is done by using the basic rule that a NE consists of a continuous sequence of capitalised words. A punctuation mark or a non-capitalised word end the NE. This rule covers most cases and the few exceptions can be treated by including several additional rules. Next, the NE is allotted to one of the NE categories by considering first the internal evidence and then, if the information is not conclusive, also the external evidence. In the third and final step, the recognised NE is recorded to be used also as external evidence for the detection of further references in the text. All of the rule-based approaches that were presented at the MUC-6 and the following, seventh MUC (MUC-7), also feature these three stages of NER as proposed by McDonald (1996) (Rössler, 2007).

In general, the rule-based approaches can be used to construct strong NER systems, but these methods call for a lot of skill and resources (Borthwick, 1999). The disadvantages of rule-based methods are summarised by Borthwick (1999) in 4 points:

- Ruled-based approaches are expensive, since a qualified computational linguist is required to construct the patterns.

- They are not universally applicable, but must be modified to each new type of text.

- They are not language independent - the rules and lists have to be rewritten for every language.

- The quality of the NER system depends heavily on the proficiency of the computational linguist building the system and the amount of time spent on its development.

For these reasons, Borthwick (1999) sees greater promise in the machine learning-based approaches.

2.4.4 Machine Learning-based Approaches

The idea behind machine learning-based approaches is to let the computer do the job for less money and in less time than the computational linguist, who would be needed to build a handcrafted NER system (Borthwick, 1999). A prerequisite of machine learning methods is generally a large set of training data. This is a document on which NER has been carried out manually and is assumed to be correctly annotated, hence also called a gold standard or the ground truth (Overell and Rüger, 2006). From this training text the system can "learn" how to annotate a text using statistical analysis (Brunner, 2008). How the learned information is then used to annotate a text depends on the specific approach.

Like the writing of a rule-based NER system, the manual annotation of a large set of training data requires time and competence in linguistics. However, Borthwick (1999) suggests that the production of a human-annotated text of about 100'000 words costs only between one and three person-days and is a task that can be done by undergraduates, whereas it took one person-month to build the weakest rule-based NER system of the MUC-7. Hence it can be said that machine learning-based methods are relatively quick to develop and in addition they are also more robust (Rössler, 2007).

There exists a multitude of machine learning-based approaches. In this thesis, the two most fundamental methods will be described to illustrate the concept of machine learning: the maximum entropy approach and the hidden markov model (Brunner, 2008).

Maximum Entropy

Entropy is a measure for disorder or, in the case of information theory, for the uncertainty of a piece of information (Shannon, 1984). The higher the entropy, the less sure the information is. The basic idea of maximum entropy modeling is to find a model which correctly represents all that is known and makes no assumptions about the unknown, i.e. models a uniform distribution for all the scenarios on which no facts are available. This is also the reason the approach is called *maximum* entropy modeling: if all possible scenarios given the current state of knowledge are equally likely, then they are also all equally unlikely, i.e. there is an overall *maximum* uncertainty. In other words, the more uniform a distribution, the higher the entropy (Berger, 1996).

Applied to NER, the maximum entropy approach works with the intelligence acquired from the training data as well as certain given constraints to assign probabilities to the possible outputs of the model. In other words, the *history* and a set of *features* are used to predict the *future* of a model (Borthwick, 1999). Mathematically, this is expressed with the conditional probability $p(f|h)$: given the history h of a word (i.e. the knowledge gained about the word from the training data), what is the probability of the future f, i.e. the probability that the word belongs to a certain NE class?

To calculate $p(f|h)$, it is necessary to use the information supplied by the features $g_i(h, f)$ connected to h and f. These are each associated with a parameter α_i. Finally, "... the conditional probability of the future given the history is the product of the weightings for all features which are active on the $\langle h, f\rangle$ pair, normalized over the products for all the futures" (Borthwick, 1999, p. 19):

$$p(f|h) = \frac{\prod_i \alpha_i^{g_i(h,f)}}{\sum_f \prod_i \alpha_i^{g_i(h,f)}}$$

Examples of NER systems that use the maximum entropy approach are described by Borthwick (1999), Chieu and Ng (2003) and Rössler (2007).

Hidden Markov Model

A hidden Markov model (HMM) possesses a finite number n of possible states. The model starts at a randomly chosen state and emits a symbol before moving on to a next state (which could also be the same as the previous state). The model continues in this fashion, moving from state to state and emitting a symbol in every state. At any time in this sequence of states, the transition to the next state is given by a probability distribution that depends solely on the current state and ignores how this present state was reached - such a process is referred to as a Markov process. Equally, which symbol is emitted by the model at any given time depends entirely on the current state of the model. Since a user of the model sees only the symbols produced by the model and not the states, the sequence of states is *hidden* (Carstensen et al., 2010).

In NER, the observed symbols are the words and the hidden states are the NE classes (Rössler, 2007). In other words, a HMM emitted the given succession of words in a text according to a sequence of NE classes, which is not visible to the reader. The idea of approaches using HMMs is to find the sequence of states that is most likely to have produced this text. Mathematically this is expressed using the joint probability $p(w, s)$: what is the probability that at a given time the model is in state s and emits the word w?

$p(w, s)$ at a given time t is calculated as the product of products for every time step i between 1 and t of the probability that the current state is s_i given the previous state s_{i-1} and the probability that the word w_i is emitted given the current state s_i:

$$p(w, s) = \prod_{i=1}^{t} p(s_i|s_{i-1})p(w_i|s_i)$$

With the system *Indentifinder*^{TM} developed by the firm BBN for the MUC-7, it was demonstrated for the first time that NER approaches using HMMs have the potential to compete with the traditional rule-based approaches (Bikel et al., 1997; Miller et al., 1998; Bikel et al., 1999).

2.4.5 Hybrid Approaches

NER approaches which use both manually assembled rules *and* machine learning methods are called hybrid approaches. As there are many machine learning approaches to chose from and multiple ways to incorporate rules, the amount of hybrid contributions to NER is immense.

Mikheev et al. (1999) propose a NER system that may be called a classical example of a hybrid approach. The system unites rule-based grammars with a maximum entropy model in five steps. In the first step, so-called *sure-fire rules* are applied. These are grammar rules which consider both the internal and external evidence and ensure that a word or sequence of words is only tagged with a NE category if the context is conclusive. Next, partial matches of the detected NEs from step one are looked for in the rest of the text and marked as possible NEs (e.g. if *Ben & Jerry's Homemade Holdings, Inc.* was tagged in the first step, then the partial match *Ben & Jerry's* will be marked as a possible company in the second step). The partial matches are then considered by a maximum entropy model. In a third step, the system reapplies the grammar rules, using the information gathered in the previous stages. The penultimate step is similar to the second, looking for partial matches and using a maximum entropy model to make probabilistic assignments. Finally, the words in the title of the text (Mikheev et al. (1999) applied NER to new wires) are classified using the knowledge gained in annotating the text (Mikheev et al., 1999). This approach showed that small or no lists are needed to achieve good NER results. Rössler (2004) believes, however, that this is not true for languages, such as German, where capitalisation is not conclusive evidence of a NE.

Further examples of hybrid NER apporaches are given by Dutta et al. (2005), Piskorski et al. (1999), Srihari et al. (2000), Stevenson and Gaizauskas (2000) and Volk and Clematide (2001).

2.4.6 Evaluation of NER

The quality of NER is measured and described using the benchmarks which are also applied in IR: *precision* and *recall*. These standards were introduced by Kent et al. (1955) to compare ranking algorithms. Precision (P) and recall (R) are defined as follows (Rössler, 2007):

$$P = \frac{\text{number of correctly tagged NEs}}{\text{number of tagged NEs}}$$

$$R = \frac{\text{number of correctly tagged NEs}}{\text{total number of NEs in gold standard}}$$

In other words, precision is a measure of how many positively identified NEs are in fact falsely tagged (hence also the term *false positives* (Brunner, 2008)) and recall shows the portion of words correctly marked as NEs in relation to the total number of NEs in the text.

Both these indicators are necessary to measure and compare the qualities of NER systems. It is, for example, simple to create a system with high precision by applying very strict rules so that few NEs are tagged but are all correctly tagged. This would, however, result in a low recall value.

In this thesis, NER is implemented only for toponyms, hence precision and recall are here defined as follows:

$$P = \frac{\text{number of correctly tagged toponyms}}{\text{number of tagged toponyms}}$$

$$R = \frac{\text{number of correctly tagged toponyms}}{\text{total number of toponyms in gold standard}}$$

While precision can always be calculated directly from one's results, recall can only be defined where an attempt is made to retrieve *all* toponyms and a gold standard exists.

Sometimes NER is also evaluated with the *F-measure*. This is calculated from the precision (P) and recall (R) values as follows (Rössler, 2007):

$$\textit{F-measure} = \frac{2 \cdot P \cdot R}{P + R}$$

2.4.7 NER for German and Languages other than English

Like for many other natural language processing tasks, in the field of NER most systems have been developed for the English language (Faruqui and Padó, 2010). The fundamental difference between English and German is that while in English, generally only proper names are capitalised, in German *all* nouns start with a capital letter. This means that the number of potential NEs is much larger (Rössler, 2004). Also, German is morphologically complex (Faruqui and Padó, 2010; Rössler, 2004). This makes NER in German a complicated task (Volk and Clematide, 2001). The outputs of the two maximum entropy systems constructed by Chieu and Ng (2003) for use on both English and German texts demonstrate this difficulty: precision and recall were lower for the German texts in both cases, despite certain adjustments to the systems (see table 2.1).

	Model 1	Model 2
English		
Precision	86.83%	88.12%
Recall	86.84%	88.51%
German		
Precision	77.05%	76.83%
Recall	51.73%	57.34%

Table 2.1: Precision and recall values of two maximum entropy models for English and German (Chieu and Ng, 2003).

Aside from the impediments presented by the language itself, NER research for German is additionally complicated by the scarcity of German training corpora (Faruqui and Padó, 2010). This makes development of state-of-the-art machine learning techniques difficult.

Volk and Clematide (2001) used precompiled lists as well as lists generated by the NER programme itself, which learns company names and last names from the corpus. Piskorski et al. (1999) use machine learning-based approaches (weighted finite-state automata and generic dynamic tries) as do Didakowski et al. (2007), who combine weighted transducers with a German semantic noun classification. Rössler (2007) and Faruqui and Padó (2010) both work with semantic generalisation, a machine learning-based approach that exploits the similarity between words such as for example *Deutschland*, *Ostdeutschland* and *Westdeutschland*.

Research has also been done on language-independent NER systems. In 2002 and 2003, the Conference of Computation Natural Language Learning (CoNLL-2002, CoNLL-2003) gave the participants the task of developing systems which could do NER without prior knowledge of the language. In 2002 the NE data consisted of Spanish and Dutch files while in 2003 the data was in English and German. For each language training data was supplied with which the learning methods could be trained, a development file was included to allow for adjustment of the system parameters and a file of test data was provided on which the developed and fine-tuned NER system could be tested. The CoNLL-2003 NE data also included a large file of unannotated data for each language. In 2002 twelve systems participated while sixteen took part in the CoNLL-2003 shared task. All systems used machine learning-based approaches to accomplish NER. The best results for both conferences are listed in table 2.2 (the F-measure was added to allow for comparison with NER results for Japanese, which are mentioned later on in the text). For German the results of the system with the highest precision is listed. Several systems achieved a higher recall value, the highest being 66.21%. The corresponding precision value is 69.37%. For the other languages, the same system attained highest precision and highest recall (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003).

	Precision	Recall	F-measure
Spanish	81.38%	81.40%	81.39
Dutch	77.83%	76.29%	77.05
English	88.99%	88.54%	88.76
German	83.87%	63.71%	72.41

Table 2.2: The best preforming systems for the CoNLL-2002 shared task (Spanish and Dutch) (Tjong Kim Sang, 2002) and the CoNLL-2003 shared task (English and German) (Tjong Kim Sang and De Meulder, 2003).

Like the results published by Chieu and Ng (2003) (see table 2.1), the precision and recall values achieved for German during the CoNLL-2003 are lower than those attained for English. In fact, the recall value for NER on German texts is the lowest of all four languages.

Some work has also been done exclusively for other languages. Examples include the Information Retrieval and Extraction Exercise (IREX) project for Japanese and the Evaluation Contest of Named Entity Recognition Systems for Portuguese, abbreviated HAREM.

IREX was held as a competition for IR and IE for the Japanese language. The IREX project consisted of two tasks: IR and NER. Fifteen systems participated in the IREX exercise. The best achieved F-measure value was 83.86 for the unrestricted domain formal run (Sekine and Isahara, 1999), which is comparable to results achieved for the CoNLL-2002 and CoNLL-2003 (see table 2.2).

HAREM took place in 2005 and 2008 (Linguateca, 2006, 2010). The HAREM held in 2005 created the first state-of-the-art NER systems for Portuguese (Santos et al., 2006). Ten participants took part in the competition. No results could be found for the individual systems. Santos et al. (2006) list precision and recall values according to text genres and

NE categories. The highest values for precision and recall of approximately 90% were achieved for fictional texts. Equally no results could be found for the HAREM of 2008 since most documents are in Portuguese.

2.5 Gazetteers

Hill (2000, p. 1) calls gazetteers "...geospatial dictionaries of geographic names...". Instead of just being a list of names, a gazetteer entry gives additional information about each named location. The most basic gazetteer links placenames to geographic locations (Axelrod, 2003), also called footprints (e.g. longitude and latitude coordinates) (Hill, 2006). The Alexandria Digital Library Project additionally includes the type or category of the geographic feature (e.g. river, forest, city, house, continent, etc.) in their definition of a gazetteer entry (Hill, 2000). In figure 2.2 an example of a gazetteer entry with these three core elements is shown. Additional information about the geographic feature, such as for example its extent or altitude, can also be part of a gazetteer entry (Hill, 2006).

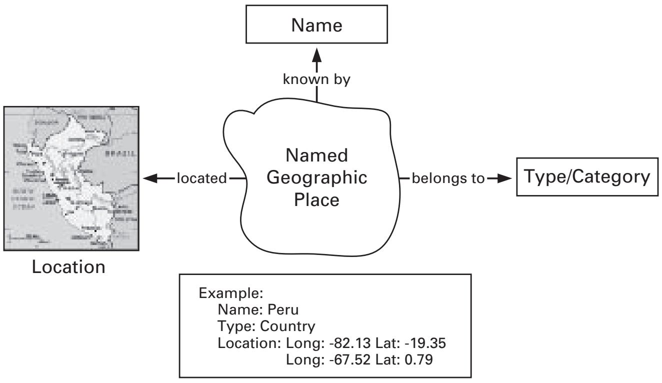


Figure 2.2: The core elements of a gazetteer entry as defined by the Alexandria Digital Library Project (Hill, 2006).

Lists of placenames with associated information have been made for centuries: fragments of the gazetteer of Stephen of Byzantium exist, which date back to the 6^{th} century. In the 19^{th} century geographic knowledge had grown considerably, as had the need to easily access it. The availability and demand for geographic information led to the production of many gaztteers during this period. Examples include Blackie's (Scotland, 1850), Ritter's (Germany, 1874) and Lippincott's (United States, 1865) (Columbia Electronic Encyclopedia, 2003). These early gazetteers were printed as alphabetical or hierarchical lists of toponyms enriched with additional information and often did not include geospatial coordinates (Hill, 2006).

Modern digital gazetteers play a central role in searches with a geographic compontent (Davies et al., 2009) and find various applications in information management (Hill, 2006). In the context of a GIS, Hill (2006, p. 97) describes gazetteers as translators "...between formal and informal means of georeferencing...", meaning they supply an unambiguous geographic location (coordinates as an example of a formal means of georefencing) for every placename (informal means of georeferencing). This quality is what makes gazetteers so important to rescue services such as Rega.

In the past several years, a need for more advanced gazetteers has surfaced (Vestavik, 2004). Gazetteers should, for example, be capable of dealing with vernacular geographic terms (Hollenstein and Purves, 2010), give information about spatial and semantic relationships between places (Vestavik, 2004) and be customisable and support time (Goodchild, 1999). Especially the attempts made to integrate vernacular geography into a gazetteer are of interest to this work and are discussed in the following section.

2.6 Vernacular Geography

Vernacular geography is the way people communicate about places in everyday life. Wilke (2009, p. 1) gives the following definition:

> "Vernacular geography is concerned with place names and their relations as they are used in people's everyday ... language."

Hence, vernacular geography can be seen as part of what Egenhofer and Mark (1995) refer to as *common-sense* or *naïve geography*. This "... is the body of knowledge that people have about the surrounding geographic world" (Egenhofer and Mark, 1995, p. 4). While naïve geography is a broad field, encompassing how people think and communicate about both the spatial and temporal component of geography, vernacular geography is more specifically concerned with what names people use to refer to geographic entities and which boundaries they set these places. Such vernacular toponyms are central to this thesis. Twaroch et al. (2009) also speak of *commonplace* names and define vernacular toponyms as

> "... everyday placenames, which may or may not correspond to administrative gazetteers."

The description of vernacular placenames given by Twaroch et al. (2009) as well as others (Hollenstein (2008), Wilke (2009)) does not clarify whether the term *vernacular placename* applies only to speech or also to the written language. For example, the official name of the Swiss valley *Val Maighels* can be found on the Internet as *Val Maigels* (in fact a Google query for *Val Maigels* receives almost 1'500 results more than the official name) - should *Val Maigels* be regarded as an separate vernacular toponym? Neither is it clear if dialect names, such as *Eisidle* for *Einsiedeln* count as an independent vernacular name. Since Davies et al. (2009, p. 177) mention not only multiple names for a geographic location but also "... variations on one name ..." and no definition of vernacular toponyms was found excluding dialect names, the term *vernacular placename* is here expanded to include dialect versions of an official toponym. Alternative spellings of toponyms, however will be treated separately and are not counted as vernacular placenames.

In their definition, Twaroch et al. (2009) mention an important fact: geographic names used in an everyday context are not always the ones that appear in the official gazetteers. Moreover, Jones and Purves (2008) suggest that this is generally the case - vernacular toponyms typically do not correspond to the entries in administrative gazetteers. Consequently, official gazetteers are "incomplete". A place may have multiple names, for example, which are not all registered in the administrative gazetteer, or a geographic feature that is nameless in official documents has a name in the local populace. Additional complications arise due to geo-geo ambiguity when a name is unofficially used for one place, while it is, in fact, the name of another place in the administrative records. Davies et al. (2009) point out that in times when maps and gazetteers were bound to paper, the extent to which alternative vernacular toponyms could be included for one entry was limited. Digital data has introduced new possibilities in this field, allowing many different names and a larger amount of additional information to be associated with one named geographic feature.

However, the way people communicate about space is inherently imprecise. References like *near Davos* or *Albulagebiet* contain no information about an exact distance or a clear boundary. The problems invoked by vernacular and naïve geography for geographic information sciences have been a topic in scientific literature for more than half a century

(Egenhofer and Mark, 1995) and various attempts have been made to improve administrative gazetteers and render vernacular geography less vague. These include Jones and Purves (2008) and Twaroch et al. (2009), who gather vernacular names and helpful information from the Internet. Hollenstein (2008) more particularly focuses on Flickr. Gan et al. (2008) study geographic Internet search queries, investigate their properties and propose a new taxonomy for search queries with a geographic component. Montello et al. (2003) look at various methods to deal with vague spatial queries while Davies et al. (2009) present a study of user needs and suggest solutions for the modelling of vague placenames. Evans and Waters (2007) use a web based mapping system which is intended to function as a link between professional and popular geographic conceptions. Galton and Hood (2005) focus on the modeling of vague vernacular geographic entities and introduce a method called *Anchoring* which preserves the vagueness of a feature by only modeling what is known about it. As far as is known, there have been no attempts to use a text corpus such as Text+Berg for the exploration of vernacular placenames.

2.6.1 Rega and Vernacular Geography

Like other emergency response services, Rega uses a geographic information system (GIS) to coordinate their efforts. This GIS is comprised of the national maps issued by swisstopo of the scales 1:500'00, 1:100'000, 1:50'000 and 1: 25'000 along with several layers of information containing data on the SAC mountain cabins, hospitals and ski regions as well as the flight paths of the Rega helicopters. An important features of the GIS is the gazetteer, which enables the rescue flight coordinator to locate a toponym on the map. Rega depends heavily on their gazetteer to locate emergency sites quickly and reliably and has added individual toponyms themselves such as well-frequented starting and landing sites for paragliders. These toponyms are generally added when they have been used in the context of some rescue mission.

Unregistered vernacular toponyms pose a problem to real-time emergency response services (Davies et al., 2009). A sad example of how misunderstandings and confusion concerning a placename can cost a rescue team precious minutes is described in BBC (2007): a ten-year old boy was drowning in a lake. The police and an ambulance were called but help arrived too late. The news report (BBC, 2007) explains: "The initial call ... gave the wrong location. This was no-one's fault, as the lake is known by several different names locally and there are other similar lakes nearby." This is just one of many similar examples (Davies et al., 2009).

The role GIS play in emergency services is a current topic in research (Pettersson et al., 2004; Davies et al., 2010). Only recently, work has begun on dealing with the problems posed by placenames that arise for rescue services (Lang, 2010). Lang (2010) differentiates between five main types of problems, which may occur in relation with the use of toponyms in real-time emergency response: data-related, topological, semantic, orthographic and human-related problems. Data-related problems arise with data that is out-of-date or has not yet been included in the system. Topological difficulties include problems caused by ambiguous placenames. Semantic problems encompass the confusion caused by vernacular toponyms while orthographic problems are due to misunderstandings during an emergency call and the various possibilities of spelling a name. Lastly, human-related issues may arise during call-reception and communication within the rescue service.

Like other emergency services, Rega faces such challenges in dealing with toponyms:

Data-related problems

The placename search algorithm of Rega's GIS is simple and does not tolerate much discrepancy. For example, the place *Hooraa* would not be found, if it were spelled *Hora*. This strict filter algorithm is necessary, however, to prevent the rescue flight coordinator from being overwhelmed by a long list of possible placenames. Since time is precious and the coordinator must divide his/her attention among several tasks at once, a short, manageable toponym list is more serviceable than a long, exhaustive list. Only recently (at the time of writing about 18 months ago) a full text search was integrated into the system, allowing the rescue flight coordinator to search for any toponyms containing the entered sequence of letters, which must be at least three letters in length.

A big problem, which contains both data-related, semantic and orthographic aspects, is rooted in a federal decree proclaimed in 1938. In the course of the mental national defense movement during World War II, it was decided by the Swiss Federal Council that all regional localities should henceforth be referred to with their vernacular Swiss German name. This move away from the hitherto official German names was one of many attempts to dissociate Switzerland from Nazi-Germany (Widmer, 2009). The changing of placenames can be observed when comparing old maps with new ones. For example, a mountain in the Valais that goes by *Tschajetuhorn* nowadays, is labeled with *Zayettazhorn* on the Siegfried maps, which were produced in the late 19th and early 20th century. Other examples are *Öugschtchummuhorn* (Augstkummenhorn), *Wyssi Flue* (Weissenfluh) and *Chrinnulücku* (Krinnenlücke) (Swiss Federal Office of Topography, sa).

Despite its more than 70-year old history, it is only in recent years that this change of toponyms has been cause for debate in the population and has created problems for the rescue services. This is probably due to the fact that it wasn't until 1998 that the first series of maps with the Swiss German toponyms were issued by swisstopo (Widmer, 2009). The new nomenclature of the canton of Thurgau was especially prominent in the media during the past several years, because Swiss German names and words were used even more consistently than was originally intended by the Confederacy (Schoch, 2010). This led to names such as *Hooraa* (Hohrain), *Holpmishus* (Holzmannshaus) and *Roopel* (Rotbühl) (Schoch, 2009; Widmer, 2009).

Such radical name changes cause problems for rescue services for several reasons. For one, this means that the GIS data, which is only updated every so often by the systems provider, is quickly obsolete (Knöpfel, 2009). Another reason is that it adds to the confusion, which is already inherent to dealing with placenames and emergencies (Knöpfel, 2009; Schoch, 2009). For instance, while the rescue service may already be in possession of the new names, an accident victim calling for help is likely to still be carrying an old map or tour book and hence can only use the old names to describe his/her location. Likewise, it takes time to adapt all the hiking trail signalisations.

In addition, there remains the question of spelling, which is much less evident for Swiss German words and names since no official orthography exists. Even for locals, names such as *Holpmishus* are complicated to read, write and pronounce, let alone the difficulty they present to people who do not speak Swiss German. Problems can arise on both ends of a telephone call to a rescue service: a person with little or no knowledge of Swiss German will struggle to communicate where he or she is and equally, rescue coordinators who come from another language background or do not have sufficient local knowledge will have trouble understanding and correctly interpreting the Swiss German toponym. Since many rescue services depend on foreign employees and employees who are not form the local area, this is a serious setback (Knöpfel, 2009).

Rega also grapples with such problems as becomes evident in a statement given by Robert Frey, head of the Rega control centre for helicopters, concerning the change of geographic names in Switzerland (Frey, 2007) (own translation):

> "It is of the utmost importance to the Rega operations centre that ... [geographic] names stay as stable as possible. The geographic names must be easily legible, otherwise a rapid transmission by telephone is not ensured in the case of an emergency. (...) The geographic names should only be changed in exceptions, on no account, however, due to new rules of spelling."

More recently, the canton of Thurgau has decided to change many of the vernacular terms back to their old German names. This back and forth has complicated matters even more: the next release of the national map series by swisstopo will not take place before 2016 and until then maps, books and path signalisations will continue to use the Swiss German names, which, despite being new, have already become antiquated (Christo, 2011).

Topological problems

Rega also needs to deal with the ambiguity of placenames. For example, there exist two locations called *Vest* and one named *Plän Vest*, all in the same valley. The aspect of ambiguous toponyms in the context of emergency response is treated in depth by Lang (2010) and is hence only briefly mentioned in this thesis.

Semantic problems

Switzerland is rich in languages and dialects. There are 4 official languages - German, French, Italian and Romansh - while the spoken language in the northern part is in reality Swiss German as opposed to standard German. This variety of languages leads to different names for places: the mountain town *Sankt Moritz* in Eastern Switzerland is called *Saint-Moritz* in French and *San Murezzan* in Romansh while Switzerland's largest city is known by the names *Zürich* (German), *Zurigo* (Italian) and *Turitg* (Romansh). The famous *Matterhorn* (German) is called *Mont Cervin* in French and *Monte Cervino* in Italian.

In addition to the languages themselves, there is such a diversity of Swiss German dialects that often Swiss from different cantons have difficulty understanding each other. This is not only the case for words but also names: *Mechiuche* is Bernese for *Meikirch* and locals may refer to the afore mentioned *Plän Vest* with variations such a *Plan Vest*, *Pluin Vest* and *Plain Vest*.

Finally, there are also those vernacular toponyms, which are not linguistic alternatives. Examples include *Münchenstein Dorf*, *Altmünchenstein* or simply *Am Berg*, which are vernacular names for the town known as *Münchenstein*.

Orthograhic problems

As mentioned previously in the context of Swiss German toponyms (section 2.6.1), the fact that there aren't any official orthographic rules in Swiss German makes it challenging to spell Swiss German toponyms. For example, on hearing *Heeje Hubel*, it is hard to discern the spelling because *Heeje* is a word from a local dialect.

Furthermore, in a Swiss German conversation it is unclear if the term used is the colloquial or the official name. So a rescue flight coordinator may for example incorrectly assume that *Stöcklichrüz* is the Swiss German variation of a place called *Stöcklikreuz*.

In the case of orthographic problems, even asking the caller for the spelling may not be helpful, since he/she may be using an old map or give a vernacular, unofficial spelling.

Human-related problems

Since Rega is often called by people stranded in the mountains or otherwise out-of-doors, factors such as wind and a bad telephone connection add to problems of dialect and unknown placenames by complicating effective communication. Rega has tried to counter this problem with various technological solutions, for example by developing an iPhone App which sends the user's coordinates directly to the Rega operations centre.

2.6.2 Swiss Vernacular Geography

Beat Dittli is a toponymist, that is a scientist who researches toponyms. On behalf of the canton of Zug he compiled the book *Zuger Ortsnamen* (Placenames of Zug), which comprises more than 12'000 placenames in 5 volumes (Dittli, 2007). This is not the only work of its kind. Many other cantons have commissioned a similar inventory and the results are available to the public in libraries as well as online (Bickel et al., 2011). The researchers use old maps, texts and other written references as well as knowledge gain in fieldwork and interviews with locals to discover vernacular toponyms. In their research toponymists also include toponyms which are no longer in use but mark them as out-of-date.

In an interview with the author of this thesis, Dittli explained that toponyms are created where they are needed. This means that certain groups of people develop their own toponyms to refer to places which are important to them. Climbers and divers are such groups, giving cliffs and segments of a lake shore a name by which to identify them. These group-specific toponyms may superimpose or simply fine-tune official placenames. An example of this phenomenon is the name *Eldorado*, which is famous in rock climber circles and refers to a south-facing cliff face of the Brünberg near the Grimselpass.

Swiss Toponyms

An insight into how certain geographic features in the Swiss Alps came by their names is given by Werlen (2008) and Schorta (1999).

Werlen (2008) gives an overview of the primary words which appear in the names of peaks in the upper part of the Valais. In general, the names of Swiss mountain peaks are relatively new compared to other toponyms: peaks were not systematically labeled until around the early 19^{th} century when alpinism became popular in Switzerland and the first national maps were produced. The origin of several names in the region of the upper Valais have been traced to early mountaineers and prominent figures such as Ludwig von Welden (1780-1853), Louis Agassiz (1807-1873) and the canon Joseph-Arnold Berchtold (1780-1859) (Werlen, 2008).

Due to the diverse culture groups who appear in the settlement history of the upper part of the Valais, placenames may have a Celtic, Gallo-Roman, Franco-Provençal, German or Italian linguistic background. It is sometimes difficult to trace this linguistic origin of a toponym because up to the 19^{th} century official documents were written in Latin and as a consequence many names were translated into this language, thereby obscuring the original placename (Werlen, 2008).

In his work, Werlen (2008) shows that many primary words which appear in Valaisan mountain names originate in everyday life: parts of the human body (e.g. *Kopf* (head)), animal parts (e.g. *Horn* (horn)), common objects (e.g. *Sattel* (saddle)) and parts of a church (e.g. *Chrits* (cross)) are examples of such primary words. Werlen (2008) closes his work by pointing out a Swiss idiosyncrasy - it was uncommon to name mountains after people. Though some exceptions exist (e.g. *Dufourspitze*, *Agassizhorn*), this particularity has made for a diverse and colourful collection of toponyms (Werlen, 2008).

Schorta (1999) presents a collection of about 2'500 toponyms from the Grisons with explanations regarding their origins and meanings. His book starts with an introduction to how today's landscape of names was formed. The toponyms are divided into two main classes: toponyms originating in features of nature and toponyms which come from a cultural background. These categories are divided into various subclasses, examples of which are listed in table 2.3 (Schorta, 1999).

Nature names	**Culture names**
Characteristic vegetation	Viniculture
Forests	The Romans
Rivers	The Church
Waterfalls	Castles
Lakes	Climbing
Hunting grounds	The mining industry

Table 2.3: Examples of the subcategories of the two main classes of toponyms, according to Schorta (1999)

Schorta (1999) states that every field name, every name of a body of water and every name of a part of the terrain originates in a common word or common compound word with a meaning. He names examples such as *Valbella* (pretty valley), *Surselva* (above the forest) and *Rhein* (river). In the Grisons, toponyms originate from various cultural groups such as the Romans, the Walser and Italian speaking peoples (Schorta, 1999).

2.7 Research Gaps and Research Questions

IR is an essential tool in today's knowledge-rich and knowledge-dependent world - information which can't be accessed when the need arises is worthless. Geography has become increasingly important in the improvement of IR on the Web (Gan et al., 2008). Therefore, research on GIR is of relevance to improve IR as a whole. Toponym recognition is a GIR subtask which still poses a challenge in the field (Purves and Jones, 2006). In IE, toponym recognition corresponds to NER, which is of basic significance to both IR and IE (Sekine and Isahara, 1999).

Most of the research on NER has been done for the English language (Faruqui and Padó, 2010). NER systems for English are not easily adjusted for languages which are substantially different from English, such as German with its capitalised nouns and more complex grammar (Volk and Clematide, 2001). NER systems which participated in the MUC-6 and MUC-7 achieved precision and recall values of over 90% for English within a restricted domain (Rössler, 2004) while results for German are still significantly lower (best precision and recall results are around 80 and 60% respectively). Research on NER for the German language is needed to bring German NER systems up to standard.

Digital gazetteers are employed in various capacities of information management (Hill, 2006) and are relied on for searches with a geographic component (Davies et al., 2009), such as for the work of emergency services. A need to construct gazetteers which are capable of more advanced services (Vestavik, 2004), like dealing with vernacular geography, has surfaced. This is also relevant for rescue services, who deal with vernacular toponyms in almost every emergency situation and struggle with the data-related, topological, semantic, orthographic and human-related problems caused thereby.

The portrayed research gaps give rise to the following research questions which set the focus of this thesis:

Question 1: How can rule-based NER techniques be used to extract unregistered vernacular and alternatively spelled toponyms from a German corpus?

Rule-based methods set the starting grounds for NER and are still employed today in hybrid systems. Hence manually composed rules are a valuable part of state-of-the-art NER. This fact as well as the lack of a gold standard for the Text+Berg corpus, which could be used as training data, are the main reasons why rule-based techniques were chosen over machine learning-based methods for this thesis.

The lack of work on German NER as well as NER accomplished without the use of lists and gazetteers are the prominent research gaps which are addressed here.

Question 2: What are the characteristics of unregistered vernacular and alternatively spelled toponyms extracted from the corpus?

The aim is to characterise the extracted unregistered toponyms. The knowledge thus gained may help towards the vernacular expansion of gazetteers such as SwissNames. This would be of value to Rega and other rescue services.

Question 3: What are the implications of the rules used to extract toponyms and the characteristics of these extracted placenames?

By linking the rules used for the NER process to the results, a basis is created for the construction of a NER system which is capable of extracting toponyms with certain desired characteristics. This knowledge will be particularly valuable to research conducted on the Text+Berg corpus.

3 Data

For this work of research, NER was performed on a digital collection of texts in order to study the use of vernacular placenames. Several other data sets, such as lists of German words, registered Swiss toponyms and foreign toponyms, were necessary to accomplish the NER task. In this chapter all the data relevant to the thesis are described.

3.1 Text+Berg Corpus

A corpus is "... a collection of spoken or written statements which are digitally available, i.e. they are stored on computers and are machine-readable, and have been processed for a linguistic or computational linguistic task" (Carstensen et al., 2010, p. 482) (own translation). A corpus differs from a text archive, which is also a collection of digitally available texts, in that a corpus holds additional information relevant to linguistic analysis. Such information may, for example, include the identification of the part of speech and normalised form of each word in a corpus. The words in a corpus are also called *tokens*, where this term can be expanded to include any symbol or string forming a linguistic unit in a text (Oxford University Press, sa).

The Text+Berg corpus is a result of *Text+Berg digital*, a joint project of the Institute of Computational Linguistics and the German Seminar of the University of Zurich. *Text+Berg digital* was launched by Dr. Noah Bubenhofer (German linguist) and Prof. Dr. Martin Volk (computational linguistics) in 2008. This project consists in the digitalisation and linguistic processing of all the articles which have been published by the Swiss Alpine Club (SAC) in its yearbooks and members' journals (Bubenhofer et al., sa).

The SAC was founded in 1863 and released a yearbook (*Jahrbuch des Schweizer Alpenclub*) nearly every year thereafter until 1923. Starting in 1925 the annual yearbook was then replaced by the monthly members' journal called *Die Alpen/Les Alpes/Le Alpi/Las Alps* (The Alps), which is still being printed today. From 1957 onwards, the SAC published both a French and a German issue of the members' journal. The articles and other contributions

Schweizer Alpenclub 1866
JAHRBUCH DES SCHWEIZER ALPENCLUB
JAHRBUCH des S. A. C.
XXXII 1896
Jahrbuch des Schweizer Alpenclub
1903-04
IXL
JAHRBUCH DES SAC
LII 1917
DIE ALPEN LES ALPES LE ALPI
XV 1939
1982
LVIII
Quartalshefte 1990 Jahrgang 66

Figure 3.1: Yearbooks and members' journals of the SAC (Volk, 2009)

were mainly written by club members of the SAC as well as editorial staff. The corpus consists of several volumes where each volume corresponds either to one yearbook or the collection of journal issues published in one calendar year in one language (French or German). At the time of writing, the corpus has been expanded to include all such SAC publications from 1864 up to the year 2009. Since no yearbooks were released in the years 1870, 1915 and 1924, the corpus counts 196 volumes (90 volumes between 1864 and 1956 and 53 volumes between 1957 and 2009 in both German and French), containing 35'750'466 words in total or about 5,6 million different words on approximately 87'000 pages. The volumes vary in length between roughly 150 and 760 pages and also contain some articles in Italian, Romansh, English and even Swiss German, though the vast majority of texts is in German or French (Bubenhofer et al., 2011).

These books and journals comprise various types of text such as expedition reports, articles on flora, fauna, folklore and geology and even poems. Nevertheless, all texts have a common topic: alpinism. Thanks to its continuity and focus, this vast collection of nearly 150 years of homage to mountaineering presents a valuable store of texts. By processing the material and making it digitally available, the Text+Berg project has created a source of information which is of interest to many branches of research such as linguistics and cultural science, history, literature and philosophy. The inherent nature of the texts also makes the Text+Berg corpus attractive for geographic sciences: the articles abound with references to places, descriptions of landscapes and accounts of glaciers, mountains and other geographic elements (Bubenhofer et al., 2011, sa).

However, it is the frequent mentions of toponyms which makes the corpus uniquely interesting to this thesis. Volk et al. (2009) state that the richness of these texts in terms of geographic references takes toponym recognition to a new level, since most of the previous work in NER was done using newspaper articles where toponyms are much less frequently used. In their description of Text+Berg, Piotrowski et al. (2010, p. 15) address a further aspect of the SAC publications, which makes them valuable to this work: "... [the mountaineering accounts] reflect the reality of the time and its contemporary perception." The articles of the digitised SAC yearbooks and journals were generally written by alpinism enthusiasts for alpinism enthusiasts. It may therefore be assumed that these authors used not only contemporary but also vernacular toponyms in their descriptions. Piotrowski et al. (2010, p. 15-16) confirm this: "... most of the accounts ... mention many "small" features (e.g., mountain huts, foot passes, ridges) not commonly contained in gazetteers." It is the aim of this thesis to find precisely such vernacular, unregistered toponyms. Since this study is particularly concerned with Swiss toponyms in the German language, the French volumes and articles are not included in the analysis. This leaves 143 volumes and roughly 9'700 articles which are of interest.

The following steps illustrate how the analog material of the SAC yearbooks and members' journals was processed and transformed to the digital, annotated form (Bubenhofer et al., sa) (own translation):

1. The texts are scanned using a document scanner with a paper feeder.

2. The scanned pages are processed with an OCR[1] programme.

3. The amount of OCR mistakes is reduced using automatic methods, for example by merging the outputs of two separate OCR programmes.

4. The structure of the documents is annotated with XML tags (according to the recommendations made by the Text Encoding Initiative). Certain meta data is also included.

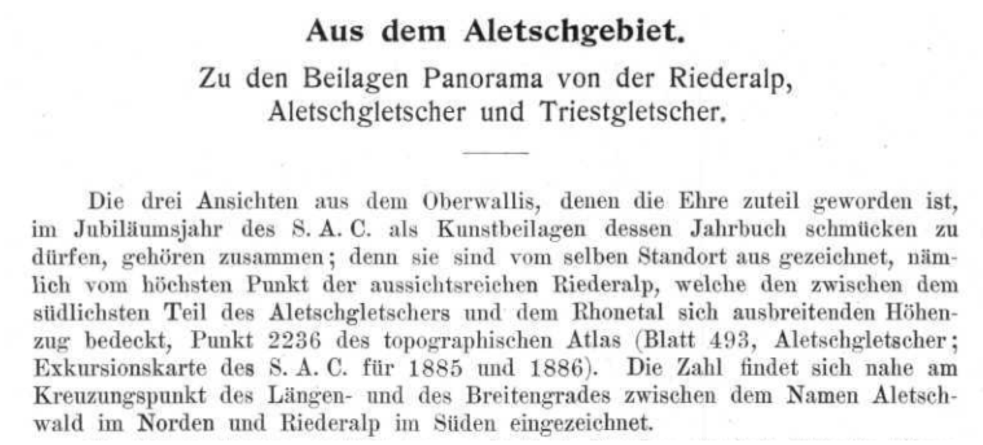
5. The language of each sentence is identified (German, French, Italian, English, Romansh, Swiss German).

6. The text is segmented into tokens and sentences.

[1] OCR is the abbreviation for optical character recognition

7. Each token is classified according to its part of speech (part of speech tagging).
8. German and French articles which are translations of one another are aligned.
9. NER is preformed for the names of mountains, glaciers and cabins (see section 3.1.2). The recognised names are saved in a separate document.
10. The annotated corpus is saved in various formats: TEI-XML, CWB Open Corpus Workbench, PDF.

In figures 3.2 and 3.3, an excerpt from the SAC yearbook of 1912 and a digitised, annotated part of the same text are shown. In this second excerpt, which is from an XML file of the Text+Berg corpus, the typical structure and meta data of the processed text can be observed. The extent of each article (*article*), paragraph (*div*), sentence (*s*) and word (*w*) is defined by the corresponding delimiters where "<*delimiter...*" marks the beginning and "<\delimiter>" the end of that unit. Also, every article, sentence and word is numbered (*n=...*) so as to be uniquely identifiable in a specific file. Each file corresponds either to one yearbook or the collection of members' journal issues published in one calendar year. At the beginning of an article the entry in the table of contents (*tocEntry*) for that article is given, typically stating the title, author, main article language (*lang=...*) and if applicable also a text category (e.g. club news, SAC chronicles, essays, etc.). Aside from the identifying number each sentence is also marked with the language in which the sentence is written. The smallest elements of the text, the words, are equipped with the most data: id number, part of speech tag (*pos=...*), lemma (normalised version of the word) and the word as it appears in the text. Where the word is not recognised by the programme, the lemma is marked as unknown (*unk*). For this thesis, the pos-tags *NN* for *normal noun* and *NE* for *named entity* are of importance.

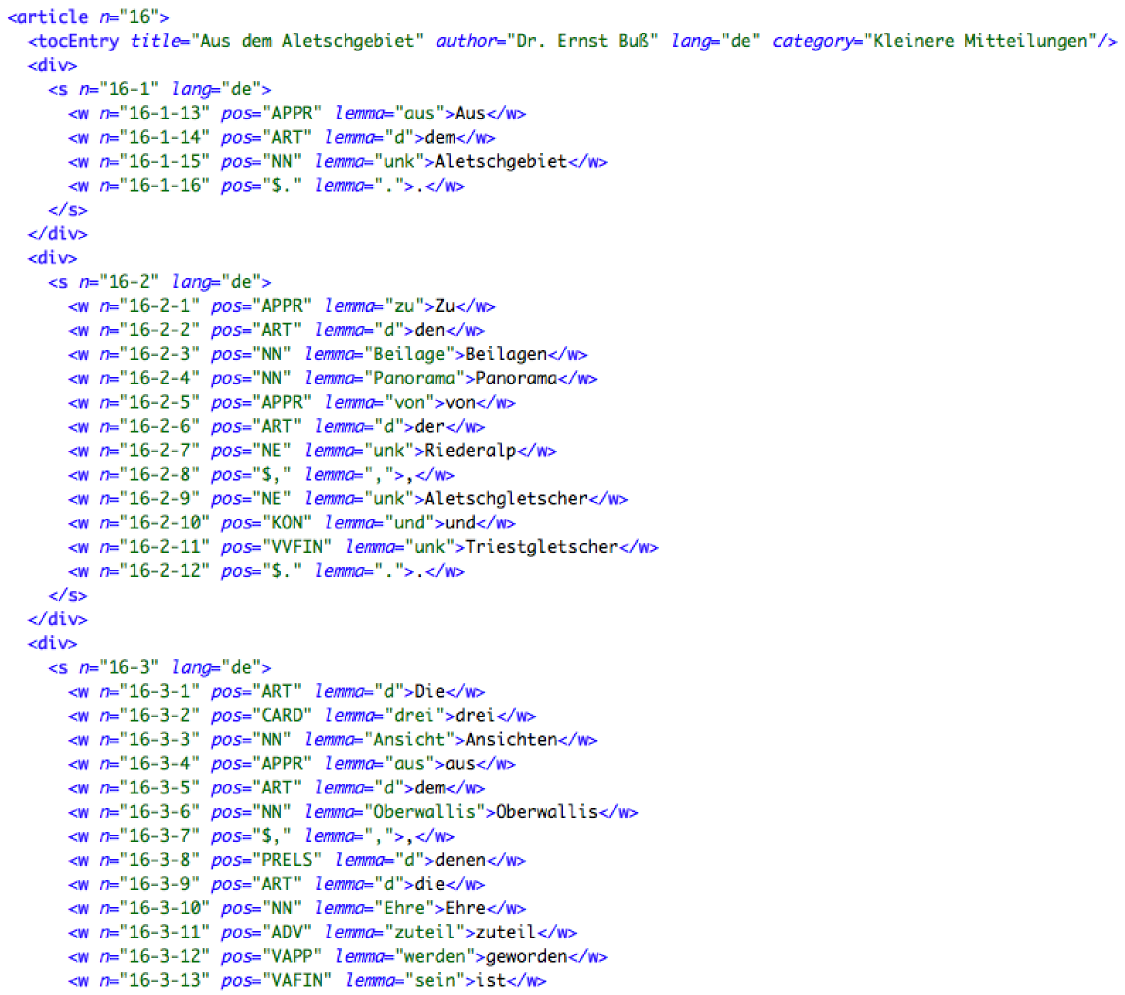


Aus dem Aletschgebiet.

Zu den Beilagen Panorama von der Riederalp,
Aletschgletscher und Triestgletscher.

Die drei Ansichten aus dem Oberwallis, denen die Ehre zuteil geworden ist, im Jubiläumsjahr des S. A. C. als Kunstbeilagen dessen Jahrbuch schmücken zu dürfen, gehören zusammen; denn sie sind vom selben Standort aus gezeichnet, nämlich vom höchsten Punkt der aussichtsreichen Riederalp, welche den zwischen dem südlichsten Teil des Aletschgletschers und dem Rhonetal sich ausbreitenden Höhenzug bedeckt, Punkt 2236 des topographischen Atlas (Blatt 493, Aletschgletscher; Exkursionskarte des S. A. C. für 1885 und 1886). Die Zahl findet sich nahe am Kreuzungspunkt des Längen- und des Breitengrades zwischen dem Namen Aletschwald im Norden und Riederalp im Süden eingezeichnet.

Figure 3.2: Beginning of the article *Aus dem Aletschgebiet* from the SAC yearbook of 1912 (Schweizer Alpen-Club SAC, sa)



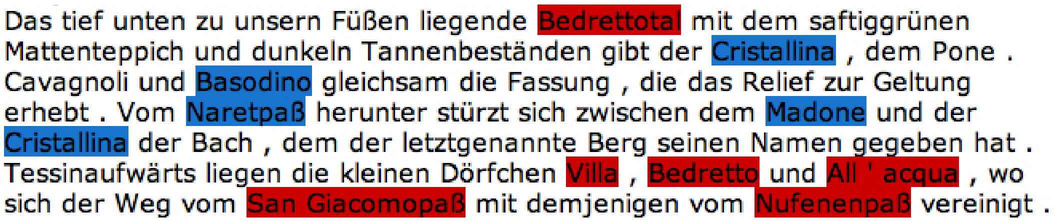
```
<article n="16">
  <tocEntry title="Aus dem Aletschgebiet" author="Dr. Ernst Buß" lang="de" category="Kleinere Mitteilungen"/>
  <div>
    <s n="16-1" lang="de">
      <w n="16-1-13" pos="APPR" lemma="aus">Aus</w>
      <w n="16-1-14" pos="ART" lemma="d">dem</w>
      <w n="16-1-15" pos="NN" lemma="unk">Aletschgebiet</w>
      <w n="16-1-16" pos="$." lemma=".">.</w>
    </s>
  </div>
  <div>
    <s n="16-2" lang="de">
      <w n="16-2-1" pos="APPR" lemma="zu">Zu</w>
      <w n="16-2-2" pos="ART" lemma="d">den</w>
      <w n="16-2-3" pos="NN" lemma="Beilage">Beilagen</w>
      <w n="16-2-4" pos="NN" lemma="Panorama">Panorama</w>
      <w n="16-2-5" pos="APPR" lemma="von">von</w>
      <w n="16-2-6" pos="ART" lemma="d">der</w>
      <w n="16-2-7" pos="NE" lemma="unk">Riederalp</w>
      <w n="16-2-8" pos="$," lemma=",">,</w>
      <w n="16-2-9" pos="NE" lemma="unk">Aletschgletscher</w>
      <w n="16-2-10" pos="KON" lemma="und">und</w>
      <w n="16-2-11" pos="VVFIN" lemma="unk">Triestgletscher</w>
      <w n="16-2-12" pos="$." lemma=".">.</w>
    </s>
  </div>
  <div>
    <s n="16-3" lang="de">
      <w n="16-3-1" pos="ART" lemma="d">Die</w>
      <w n="16-3-2" pos="CARD" lemma="drei">drei</w>
      <w n="16-3-3" pos="NN" lemma="Ansicht">Ansichten</w>
      <w n="16-3-4" pos="APPR" lemma="aus">aus</w>
      <w n="16-3-5" pos="ART" lemma="d">dem</w>
      <w n="16-3-6" pos="NN" lemma="Oberwallis">Oberwallis</w>
      <w n="16-3-7" pos="$," lemma=",">,</w>
      <w n="16-3-8" pos="PRELS" lemma="d">denen</w>
      <w n="16-3-9" pos="ART" lemma="d">die</w>
      <w n="16-3-10" pos="NN" lemma="Ehre">Ehre</w>
      <w n="16-3-11" pos="ADV" lemma="zuteil">zuteil</w>
      <w n="16-3-12" pos="VAPP" lemma="werden">geworden</w>
      <w n="16-3-13" pos="VAFIN" lemma="sein">ist</w>
```

Figure 3.3: Beginning of the article *Aus dem Aletschgebiet* from the Text+Berg corpus XML file of the SAC yearbook of 1912 (Bubenhofer et al., 2011)

Various research and analyses have been done on the Text+Berg corpus. Examples include several gold standard articles, on which toponym recognition was preformed manually (subsection 3.4) as well as the NER performed for the whole corpus, where only the names of mountains, glaciers and cabins were extracted (subsection 3.1.2). Both these sets of data will be used in the evaluation of the thesis' NER process. Also useful for this work is the analysis of the Text+Berg corpus with the system GERTWOL, which recognises German word forms (subsection 3.1.3). These materials are described in more detail in the following subsections.

3.1.1 Gold Standard

The Text+Berg corpus comes with a collection of 39 articles from the years 1900, 1903, 1911, 1939, 1968 and 1986 which have been manually processed for the mention of toponyms by Maya Bangerter, a member of the *Text+Berg digital* project team. Five of these articles are used to evaluate the NER process developed in this thesis. An excerpt of one of the five articles used for the system evaluation is included in figure 3.4. The toponyms marked in blue refer to mountains, all other toponyms are highlighted in red.



Das tief unten zu unsern Füßen liegende Bedrettotal mit dem saftiggrünen Mattenteppich und dunkeln Tannenbeständen gibt der Cristallina , dem Pone . Cavagnoli und Basodino gleichsam die Fassung , die das Relief zur Geltung erhebt . Vom Naretpaß herunter stürzt sich zwischen dem Madone und der Cristallina der Bach , dem der letztgenannte Berg seinen Namen gegeben hat . Tessinaufwärts liegen die kleinen Dörfchen Villa , Bedretto und All ' acqua , wo sich der Weg vom San Giacomopaß mit demjenigen vom Nufenenpaß vereinigt .

Figure 3.4: Excerpt from the gold standard article *Bergfahrten in der Zentralschweiz*, SAC yearbook 1903 (Bangerter, 2010)

The gold standard articles were processed with the annotation tool *MMAX2* (Müller and Strube, 2006), using the work done on the markup language *SpatialML* (Doran, 2009; Mani et al., 2008) as a guideline. The annotated articles are available online (Bangerter, 2010).

3.1.2 NER for Mountains, Glaciers and Cabins

Basic NER has already been performed on the Text+Berg corpus. References to Swiss mountains, glaciers and cabins where recognised in the text with the help of the gazetteer *SwissNames 25* (Bubenhofer et al., 2011).

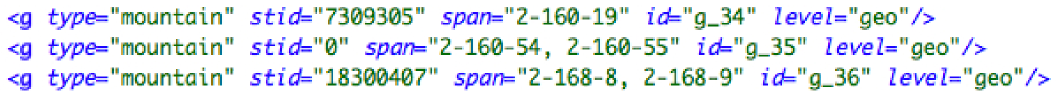
To identify the names of mountains in the text the toponyms of the categories *Massiv* (important massive), *HGipfel* (main alpine peak), *GGipfel* (lesser alpine peak), *KGipfel* (small peak) and *Grat* (ridge) were extracted from SwissNames 25. This new list was revised to exclude placenames which are also common German nouns such as *Burg* (castle), *Esel* (donkey), *Hengst* (stallion) and *Turm* (tower). Furthermore, the list was expanded to include 4'239 additional toponyms which fall into one of the categories mentioned subsequently (Bubenhofer et al., 2011):

- The genitive form of the toponyms extracted from SwissNames 25 (e.g. *Aletschhorns* is the genitive of *Aletschhorn*).
- Toponyms which appear in the Text+Berg corpus and consist of two words where the first part of the name is *Aiguille*, *Aiguilles*, *Cime*, *Cima*, *Dent*, *Dents*, *Mont*, *Monte*, *Piz*, *Pizzo* or *Vanil*.
- Toponyms which appear in the Text+Berg corpus and end in one of the following three suffixes, which are typical for the names of mountains: *-horn*, *-stock*, *-grat*. Also, toponyms ending in the plural forms *-hörner* and *hörnern*. (To increase precision, these potential toponyms were compared to a list of common German nouns and manually checked before adding them to the toponyms extracted from SwissNames 25.)
- Variations of important toponyms in other languages (e.g. *Cervin*, which is French for *Matterhorn* and *Pilate*, which is French for *Pilatus*).

This resulted in a list of mountain names with more than 10'000 entries, scoring 158'025 hits in the corpus (Bubenhofer et al., 2011).

For the detection of the names of glaciers and cabins the toponyms of the categories *Gletscher* (glacier) and *Huette* (cabin) where extracted from SwissNames 25. The so gathered 367 names of glaciers and 519 entries for cabins resulted in 6649 and 7856 hits respectively (Bubenhofer et al., 2011).

The recognised toponyms are stored in separate files of the corpus - one for every volume. Figure 3.5 shows an excerpt of such a NER file. The tag *type* shows whether the toponym is the name of a mountain, glacier or cabin. In this example, all three entries are mountians. The number following *stid* is the swisstopo ID, that is the number associated with this toponym in SwissNames 25. This makes it possible to access all the information in the gazetteer concerning the geographic feature in question. If this ID is zero, as is the case in the second entry, this means that either the toponym is not in the gazetteer or that this toponym is listed more than once in SwissNames 25. Without a disambiguation module it is not possible to decide which is the correct reference. *span* gives the ID of the toponym in the Text+Berg corpus and shows how many tokens the toponym consists of. For example, in both the second and third entry, two corpus IDs are listed signifying that the toponym consists of two parts. In the example (figure 3.5), the entries refer to the toponyms *Surettahorn*, *Piz Ferrera* (which is not listed in SwissNames 25) and *Punta Nera* respectively.



```
<g type="mountain" stid="7309305" span="2-160-19" id="g_34" level="geo"/>
<g type="mountain" stid="0" span="2-160-54, 2-160-55" id="g_35" level="geo"/>
<g type="mountain" stid="18300407" span="2-168-8, 2-168-9" id="g_36" level="geo"/>
```

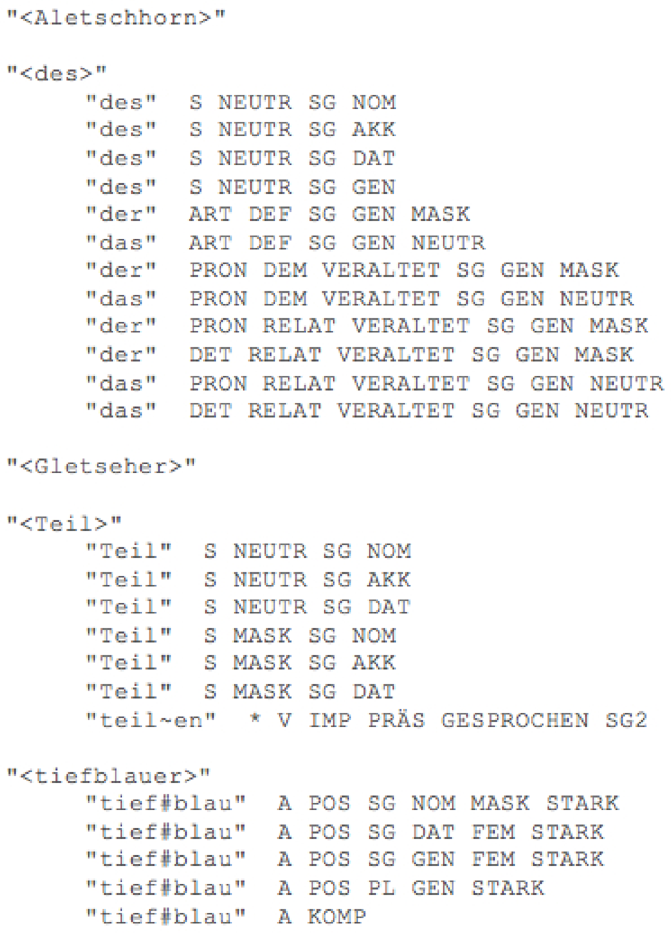
Figure 3.5: Excerpt from the Text+Berg corpus NER file of the SAC yearbook of 1912 (Bubenhofer et al., 2011)

The contents of these NER files will be compared to the results of the thesis' NER process in chapter 6.

3.1.3 GERTWOL

"GERTWOL is a system for automatic recognition of German word forms" (Lingsoft Language Solutions, sa). It is based on a model for morphological analysis called the two-level model, which was developed by Prof. Kimmo Koskenniemi in 1983. This method for the analysis of the internal structure of words is language independent and has been implemented to describe roughly 30 languages. The Collins German Dictionary (Second Edition, Copyright HarperCollins Publishers) serves as the basic lexicon for GERTWOL. To this lexicon material another 6'300 common nouns and 11'000 proper nouns were added which had not been recognised by GERTWOL during initial test runs. With its 85'000 words and its tools to derive the morphology of words and compound words, GERTWOL is reported to accurately recognise 99% of German words in a correctly spelled text. For unrestricted texts the coverage is at about 98% (Lingsoft Language Solutions, sa).

The Text+Berg corpus was analysed using GERTWOL. In the results of this analysis, each distinct word which appears in the corpus files is listed, followed by the possible word forms. An excerpt from this list is shown in figure 3.6.



```
"<Aletschhorn>"

"<des>"
     "des"  S NEUTR SG NOM
     "des"  S NEUTR SG AKK
     "des"  S NEUTR SG DAT
     "des"  S NEUTR SG GEN
     "der"  ART DEF SG GEN MASK
     "das"  ART DEF SG GEN NEUTR
     "der"  PRON DEM VERALTET SG GEN MASK
     "das"  PRON DEM VERALTET SG GEN NEUTR
     "der"  PRON RELAT VERALTET SG GEN MASK
     "der"  DET RELAT VERALTET SG GEN MASK
     "das"  PRON RELAT VERALTET SG GEN NEUTR
     "das"  DET RELAT VERALTET SG GEN NEUTR

"<Gletseher>"

"<Teil>"
     "Teil"  S NEUTR SG NOM
     "Teil"  S NEUTR SG AKK
     "Teil"  S NEUTR SG DAT
     "Teil"  S MASK SG NOM
     "Teil"  S MASK SG AKK
     "Teil"  S MASK SG DAT
     "teil~en"  * V IMP PRÄS GESPROCHEN SG2

"<tiefblauer>"
     "tief#blau"  A POS SG NOM MASK STARK
     "tief#blau"  A POS SG DAT FEM STARK
     "tief#blau"  A POS SG GEN FEM STARK
     "tief#blau"  A POS PL GEN STARK
     "tief#blau"  A KOMP
```

Figure 3.6: Excerpts form the morphological analysis of the Text+Berg corpus using GERTWOL (Lingsoft Language Solutions, sa)

The three words *Teil* (part), *tiefblauer* (deep blue) and *des* (the) were recognised by GERTWOL which is why they are followed by the possible word forms. For example, *Teil* could be either a noun (`S`) or a verb (`V`) which is capitalised because it is at the beginning of a sentence. `NEUTR` (neuter) and `MASK` (masculine) give the gender of the noun, `SG` shows that the noun is in its singular form and `NOM` (nominative), `AKK` (accusative) and `DAT` (dative) show the possible cases. Similarly the abbreviations for the verb show that the word could be the imperative form (`IMP`) of *teilen* in the present tense (`PRÄS`) for the second person singular (`SG2`) in direct speech (`GESPROCHEN`).

Both *Aletschhorn* and *Gletseher* are not followed by abbreviations. This means that neither was recognised by GERTWOL. In the first case this is because *Aletschhorn* is an unknown toponym while *Gletseher* is an OCR mistake and hence not identified as the noun *Gletscher* (glacier).

This clear structural distinction between recognised words, which are followed by further lines of information, and unrecognised words, which stand alone, is used to filter out known German words in the NER method developed for this thesis (see section 4.1.1).

3.2 SwissNames

In the following section, the federal gazetteer of Swiss toponyms is introduced. To give the reader a better understanding of the nature of this gazetteer's entries, a brief insight into Swiss toponym legislation is provided first.

3.2.1 Swiss Legislation on Geographic Names

The spelling of placenames in Switzerland can be described as an unresolved problem with an eventful past. Until 1948, topographical surveyors were the ones who decided how to spell the toponyms they recorded. They were even free to chose between the dialect or High German form of a placename. In 1948, the directive *Weisung für die Erhebung und Schreibweise der Lokalnamen bei Grundbuchvermessungen in der deutschsprachigen Schweiz*, which had its origins in the Swiss mental defense movement of World War II, became effective. This directive assigned jurisdiction concerning the spelling of toponyms to the cantons, who were expected to appoint nomenclature commissions. Since the directive only stated guidelines, which were free to interpretation, large differences still prevailed between the spellings implemented by the various cantonal nomenclature commissions. This situation has caused Swiss toponyms to be a mix of both High German and different dialect forms (Werlen, 2008).

Today, the Swiss Federal Office of Topography swisstopo gives recommendations to the cantonal nomenclature committees and takes on a coordinating role (Swiss Federal Office of Topography, 2009) in an attempt to standardise the naming of geographic features in Switzerland. Details on the current state of legislation regarding Swiss toponyms is given in the subsequent text.

Article 7 of the federal law on geoinformation states the responsibilities of the Swiss Federal

Council concerning geographic names (Federal Authorities of the Swiss Confederation, 2009) (own translation):

1. "The Federal Council makes regulations concerning the coordination of the names of municipalities, townships and roads. The Federal Council regulates the other geographic names, the competences, the process and the meeting of the costs.
2. The Federal Council acts as the last level of jurisdiction in the case of disputes caused by the first clause."

The Swiss Federal Council complied with these responsibilities by composing the ordinance governing geographic names.

Ordinance Governing Geographic Names (GeoNV)

On July 1, 2008 the GeoNV came into operation. The new ordinance states that toponyms must be easy to understand, write and copy, not only for the locals but for visitors as well. For this reason the authorities move away from dialect names and reinforce High German toponyms. The ordinance also declares that the change of existing names will only be authorised if it is in the interest of the general public. The responsibilities involved in the governance of Swiss toponyms are defined and alloted:

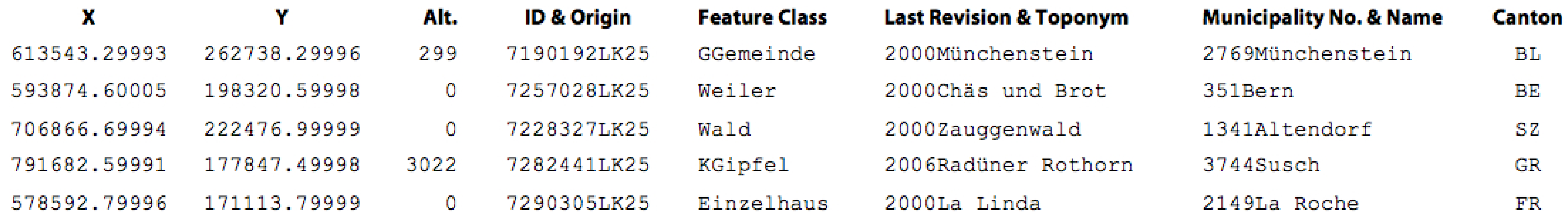
1. The Swiss Federal Office of Topography swisstopo is put in charge of making rules for the geographic names of the national and cadastral survey.
2. The Federal Council delegates the task of making guidelines concerning building addresses and the spelling of the names of municipalities, townships and roads to swisstopo.
3. In the case of guidelines for the names of stations, the jurisdiction lies with the Federal Office of Transport.

Recommendations for building addresses and the spelling of the names of municipalities, townships and roads have been devised. In the case of the building addresses and the spelling of the road names, the recommendations are only available for the German speaking part of Switzerland. Guidelines for the spelling of the names of stations have also been published (Cadastral Surveying in Switzerland, 2010b). The rules for the names used by the national and cadastral survey are still pending. At the time of writing the directive *Weisung für die Erhebung und Schreibweise der Lokalnamen bei Grundbuchvermessungen in der deutschsprachigen Schweiz* dating back to 1948 is the latest document of this kind. This directive was only composed for the German speaking part of Switzerland and is no longer valid. A task group is currently working on a revised version (Cadastral Surveying in Switzerland, 2010a).

3.2.2 What Is SwissNames?

SwissNames is the name of the official Swiss national gazetteer, which is based on the content of the national maps. It was compiled and is maintained by the Federal Office of Topography swisstopo. The gazetteer contains approximately 193'000 georeferenced entries and comprises six subsets corresponding to the sets of toponyms used for the national maps in different scales. *SwissNames 25*, for example, is the set of all toponyms which appear on the national maps scaled 1:25'000. Similarly, there are sets for the other maps issues by swisstopo, which have the scales 1:50'000, 1:100'000, 1:200'000 and 1:500'000. Finally, there is also *SwissNames Ortschaften* (SwissNames settlements), which is the set of names of all municipalities, cities, towns, villages and hamlets that appear on any one of the national maps (all scales) (Swiss Federal Office of Topography, 2002).

Each SwissNames entry consists of a toponym, the coordinates of the object it is referring to and several other attributes such as an ID number and a feature class. There are 76 feature classes which fine-tune the rougher categories *settlements*, *valleys*, *areas*, *bodies of water and lakes*, *mountains*, *passes*, *streets and facilities* and *single objects* (Swiss Federal Office of Topography, 2002). An excerpt from SwissNames 25 is shown in figure 3.7. Explanations are provided in table 3.1.



X	Y	Alt.	ID & Origin	Feature Class	Last Revision & Toponym	Municipality No. & Name	Canton
613543.29993	262738.29996	299	7190192LK25	GGemeinde	2000Münchenstein	2769Münchenstein	BL
593874.60005	198320.59998	0	7257028LK25	Weiler	2000Chäs und Brot	351Bern	BE
706866.69994	222476.99999	0	7228327LK25	Wald	2000Zauggenwald	1341Altendorf	SZ
791682.59991	177847.49998	3022	7282441LK25	KGipfel	2006Radüner Rothorn	3744Susch	GR
578592.79996	171113.79999	0	7290305LK25	Einzelhaus	2000La Linda	2149La Roche	FR

Figure 3.7: Excerpt from the SwissNames file with explanatory column titles Swiss Federal Office of Topography (2008)

Column Title	Detailed Explanation
X	x-coordinate of toponym (Swiss National coordinate system CH1903)
Y	y-coordinate of toponym (Swiss National coordinate system CH1903)
Alt.	altitude in meters above sea level (optional attribute, hence here *0* does not necessarily mean altitude = 0 m.a.b.s.l.)
ID & Origin	toponym ID and origin of the data, *LK25* being short for *Landeskarte* (national map) *1:25'000*
Feature Class	toponym feature class (e.g. *Weiler* = hamlet, *Wald* = forest)
Last Revision & Toponym	year of last data revision followed directly by the toponym
Municipality No. & Name	unique municipality number and name of the municipality
Canton	abbreviation of canton name

Table 3.1: Explanations to the structure of the SwissNames gazetteer

3.2.3 Quality of SwissNames

SwissNames is one of the most comprehensive and detailed sets of Swiss toponyms. The gazetteer covers all of Switzerland homogeneously and is updated yearly. Updates are made according to information received from federal and cantonal authorities and relying on the knowledge gathered by topographers in the field for the overall update of the national maps, which is done every six years. The non-ambiguity and stability of the data make such updates possible. Furthermore, the simple structure of the gazetteer allows for an implementation of SwissNames on various systems (Swiss Federal Office of Topography, 2005). The excellent quality of the data makes SwissNames a very helpful tool for rescue services.

Despite this carefully composed and kept gazetteer, there are several drawbacks to be noted.

- SwissNames is not flawless. There are several erroneous entries in the release (SwissNames 25, dated from 2008) which was used for this thesis. For example *Hofer H!tta* should read *Hofer Hütta* and *LÉpine* should be *L'Epine*.

- Only the *abbreviations* of the cantons are included. Hence a canton whose name doesn't otherwise appear in SwissNames as the name of a town (e.g. *Bern*, *Glarus*, *Schwyz*) or some other geographic feature (e.g. *Uri Rotstock*, *Port-Valais*, *Cabane du Jura*) are not included in SwissNames. *Basel-Landschaft*, *Graubünden* and *Obwalden* are not registered in SwissNames 25, for example.

- SwissNames and the national maps don't always match. For instance, *Rotbüel* is the toponym registered in SwissNames (both the version from 2008 and the set used at Swiss Federal Office of Topography (2010)) while on the national maps available online (Swiss Federal Office of Topography, 2010) this name has been replaced with *Roopel*. Similarly, there are differences between scales. On the national maps scaled 1:25'000 and 1:50'000 the landmark mountain of Wildhaus is called *Wildhuser Schofberg* whereas on the map scaled 1:100'000 it is the *Wildhuser Schafberg* (Swiss Federal Office of Topography, 2010). Such disaccord between the maps and the gazetteer as well as between the maps of different scales may only be slight but can nonetheless cause confusion.

- Since the largest scale of the national maps is 1:25'000, the amount and granularity of toponyms is limited: although SwissNames is extensive, it is by no means exhaustive.

3.3 Foreign Toponyms

The articles in the Text+Berg corpus do not always treat Swiss alpinism. In several cases the texts are about foreign places. In order to better judge the geographic focus of an article (i.e. whether it contains a lot of foreign toponyms, which would reduce NER precision values since only Swiss toponyms are of interest), a list of foreign toponyms was assembled by hand using Internet sources. This list contains all foreign countries and their capitals (Welt-Blick.de, sa) as well as foreign eminent mountains and mountain ranges (Welt-Blick.de, sa; Winkler, sa; Kilcher and Soutar, sa; Wikipedia, 2009, 2011b,a,c). The countries and capitals are listed both in English and German whereas for the rest of the toponyms only the German versions were included. Where no German version was found, the native form was added to the list (e.g. *Rocky Mountains*, *Nanga Parbat*). The further processing of this raw list of foreign toponyms is described in chapter 4.

4 Methodology

In this chapter, the NER system developed in the course of the thesis is described to illustrate how lists of unregistered Swiss toponyms were extracted from the Text+Berg corpus. The methods used to analyse and evaluate the results generated by this NER process are also explained in detail. The entire methodology is structured in four parts. The second step constitutes the actual extraction of NEs (NER). Together with the data postprocessing, the NER step is the central part of the NER system where the candidate toponyms are extracted and the selection is refined.

1. Data preprocessing (section 4.1)
2. NER (section 4.2)
3. Data postprocessing (section 4.3)
4. Result analysis (section 4.4)

Of the various NER approaches described in chapter 2, mainly rule-based methods (both list-based and other) are relevant to this work. Since the data used was prepared with machine learning-based approaches, however, the process as a whole, from digital but unannotated text to NER results, can be classified as a hybrid approach. The rule-based system developed in the course of this thesis consists of several programmes, splitting NER into various steps. A schematic overview of these NER steps and the methods used for result analysis is included for orientation (see figure 4.1).

Subsequently, each step of the NER system and the corresponding programmes are explained in more detail. All programmes were written in the programming language *Perl* with the exception of the code used for the Internet search, which was written in Java.

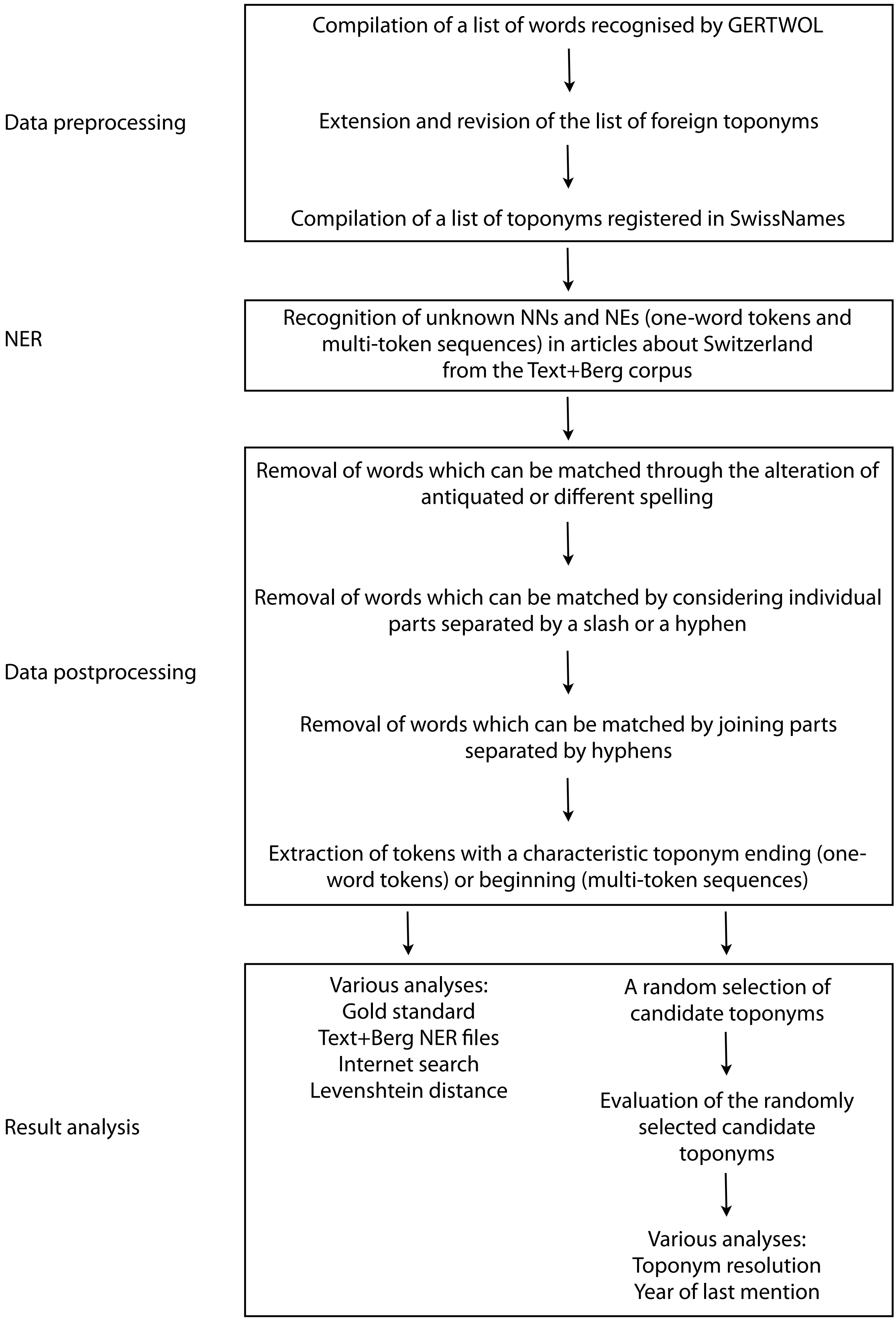

Data preprocessing
Compilation of a list of words recognised by GERTWOL
Extension and revision of the list of foreign toponyms
Compilation of a list of toponyms registered in SwissNames
NER
Recognition of unknown NNs and NEs (one-word tokens and multi-token sequences) in articles about Switzerland from the Text+Berg corpus
Data postprocessing
Removal of words which can be matched through the alteration of antiquated or different spelling
Removal of words which can be matched by considering individual parts separated by a slash or a hyphen
Removal of words which can be matched by joining parts separated by hyphens
Extraction of tokens with a characteristic toponym ending (one-word tokens) or beginning (multi-token sequences)
Result analysis
Various analyses:
Gold standard
Text+Berg NER files
Internet search
Levenshtein distance
A random selection of candidate toponyms
Evaluation of the randomly selected candidate toponyms
Various analyses:
Toponym resolution
Year of last mention

Figure 4.1: Schematic overview of the methodology

4.1 Data Preprocessing

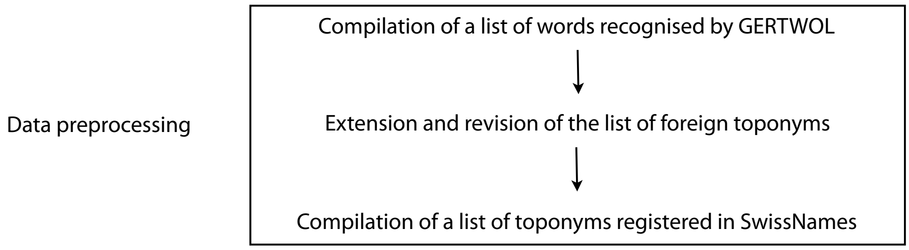

Data preprocessing
Compilation of a list of words recognised by GERTWOL
Extension and revision of the list of foreign toponyms
Compilation of a list of toponyms registered in SwissNames

Figure 4.2: Schematic overview of the data preprocessing

In the following section it is explained how the three lists which are predominant in the main NER steps are assembled. These are the lists containing recognised German words, foreign toponyms and registered Swiss toponyms. These three lists will henceforth be referred to as the look-up lists. More information on the exact content of the look-up lists can be found in appendix A

4.1.1 Words Recognised by GERTWOL

The results of the GERTWOL analysis on the tokens in the Text+Berg corpus have a distinct structure: all words which were recognised by GERTWOL are followed by lines of grammatical information, while words which were not recognised stand alone (see section 3.1.3). Using this structure makes it possible to easily identify all known German words. Recognised German words beginning with an upper case letter are stored in a list. Only capital words are considered since the aim of this NER system is to detect toponyms, which inherently begin with an upper case letter (exceptions are OCR mistakes which are ignored). Some toponyms consist of multiple words of which one or more may begin with a lower case word, as in *Piz dal Teo* or *Alpe dell'Efra*. However, in this work only compound toponyms of the latter type, where the lower case word is joined to a capital word by an apostrophe (*dell'Efra*), or compound toponyms where all components begin with a capital letter are considered (e.g. *Vadret Calderas*). GERTWOL did not recognise lower case words of the type *dell'Efra* and hence these cases are neglected during the compilation of a list of words recognised by GERTWOL.

4.1.2 Foreign Toponyms

The list of foreign toponyms which was manually assembled requires the addition and removal of certain entries before being employed in the main part of the NER system. One such adjustment is the addition of the individual parts of toponyms consisting of more than one word (e.g. *Rocky Mountains*). This is necessary because the structure of the Text+Berg corpus requires the NER system to consider one-word tokens as well as sequences of tokens. The thus extended list of foreign toponyms is then manually checked to eliminate common nouns and names such as *Ben* (e.g. from the toponym *Ben Nevis*), *Halbinsel* (peninsula) and *Wald* (forest) as well as parts of toponyms which are not conclusively foreign such as *Alpen* (alps), *Grosser* (large) or *San* (saint).

Entries in the list which also appear in SwissNames are removed to avoid a false classification due to widespread terms like *Aiguille*, *Col* and *Corno.* Several of these entries which are listed in both the SwissNames gazetteer and the list of foreign toponyms are not deleted, however, since placenames such as *Canada*, *Rome* and *Paris* are much more likely to appear in a foreign context.

Finally, for foreign toponyms which are written with the German letter *ß*, called *scharfes s* or *Eszett*, variations are added where the *ß* is substituted with *ss*. This is done because nowadays *ß* is generally not used in Switzerland. Examples of such foreign toponyms include *Großglockner*, *Großbritannien* (Great Britain) and *Weißrussland* (Belarus).

4.1.3 Toponyms Registered in SwissNames

The registered Swiss toponyms are extracted from the last three columns of the gazetteer SwissNames. As with the foreign toponyms, the Swiss toponyms are not simply extracted one-to-one as they appear in the gazetteer (often as multi-word toponyms). In addition to the original SwissNames entries several variations of the toponyms are added to allow for one-to-one word comparison during NER:

1. If the actual toponym was followed by a clarifying term in brackets (like the abbreviation of a canton, for example), the toponym and the term in brackets are also included separately in the extended list of SwissNames. E.g. the original entries in the SwissNames gazetteer *Beinwil (Freiamt)* and

Schmitten (GR) are included as well as the variations *Beinwil*, *Freiamt*, *Schmitten* and *GR*.

2. If the toponym contains a slash ("/"), the toponym parts on either side of the slash are also included separately in the list. E.g. the original entry *Breil/Brigels* is included as well as the variations *Breil* and *Brigels*.

3. If the toponym contains a hyphen ("-"), that part of the toponym is also added separately to the list. E.g. the original entry *Val Bos-chetta* is included as well as the variation *Bos-chetta*. This is done because in the next step the hyphens are used to split a toponym into parts and so only segments like *Bos* and *chetta* are included while *Bos-chetta* would be lost.

4. Finally, the toponym is split into parts along whitespaces, hyphens and slashes and each part is added separately to the list. E.g. the original entry *Kom. Reckingen-Gluringen/Grafschaft* is included as well as the variations *Kom.*, *Reckingen*, *Gluringen* and *Grafschaft*.

These four steps are applied for both the names of all the geographic features registered in SwissNames as well as the municipalities they belong to. As a result of the splitting of the original entries, toponym parts are generated such as *della*, *"Im* and *L'A*. Such characters and words are not distinctly Swiss and for this reason all parts of toponyms consisting only of lower case letters, beginning with a non-word character (such as a hyphen, bracket, etc.) or containing less than two letters directly following each other are excluded from the compiled list of Swiss toponyms. Finally, the abbreviations for the cantons are also added to the list (e.g. *GR* for *Graubünden*).

Geo-Non Geo Ambiguity in SwissNames

To prevent false identification of tokens in the Text+Berg corpus as Swiss toponyms or parts of Swiss toponyms, all geo-non geo ambiguous toponyms and parts of toponyms are removed from the list of extracted SwissNames. To do this, the extended list of registered Swiss toponyms is compared to the list of words recognised by GERTWOL. Any words which appear in both files are saved. The resultant list is manually scrutinised for terms which, though recognised by GERTWOL, are most likely to appear as Swiss toponyms or as part of such.

To clarify future terminology: unless mentioned otherwise, *list of Swiss toponyms* will refer to the *ambiguous* list containing all SwissNames toponyms and parts thereof. This is also the list of Swiss toponyms which is one of the three look-up lists.

For facilitated word comparison later on, the list compiled from SwissNames is sorted into 27 different files according to their first letter (the 26 letters of the alphabet and one file for all words beginning with an Umlaut). Also, a further list of registered Swiss toponyms and toponym parts is generated from which all identified geo-non geo ambiguous words have been removed.

4.2 NER in the Text+Berg Corpus

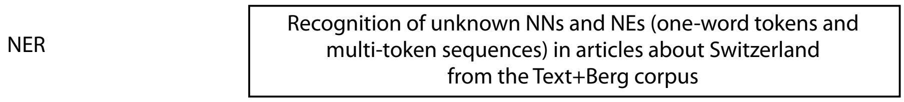

NER
Recognition of unknown NNs and NEs (one-word tokens and multi-token sequences) in articles about Switzerland from the Text+Berg corpus

Figure 4.3: Schematic overview of NER in the Text+Berg corpus

Many toponyms, such as *Basel* for instance, consist of just one word. As has been mentioned, toponyms can also be comprised of multiple words (e.g. *Radüner Rothorn*). The aim of this NER system is to extract unregistered toponyms of both types. To simplify the task and keep programmes short and clear, two separate series of programmes were written for NER as well as data postprocessing - one to look for toponyms composed of a single word and a second version to find compound toponyms. In this and the next section (4.3), methods will first be explained for the case of one-word toponyms, followed by the adjustments made to account for compound toponyms. In each case the programmes are similar and only slight alterations were necessary. These programmes could also easily be adapted to extract and filter for other types of tokens.

To illustrate the NER and data postprocessing steps of the system, the implemented rules will be explained at the end of each subsection (step) using a short text written by the author for this purpose. The text is entirely fictional and despite the use of several existing toponyms, makes no pretence at being geographically accurate. To simplify matters, only the rules for the one-word tokens will be discussed in this manner.

NER for One-Word Tokens

This step is the heart of the NER system: it is where the actual recognition of NEs in the Text+Berg corpus takes place.

Since the focus of this thesis lies on the *German* versions of Swiss toponyms, it is important to restrict NER to articles and sentences in German. For this reason, each yearbook file[1] in the Text+Berg corpus is divided into articles, using the delimiting XML tags. Articles which are tagged to indicate that they are not written in German are skipped, the other articles are split into lines and processed further.

A prerequisite to the actual NER process is distinguishing between articles written about places in Switzerland and articles which treat foreign regions, since for this thesis the interest lies in *Swiss* toponyms. The articles are classified into three categories: *Swiss*, *foreign* and *unknown*. The category *unknown* is for articles where both Swiss and foreign topoynms are lacking and no decision can be made. For the classification of the articles the titles are searched for foreign toponyms. If a token in the title matches a foreign toponym the article is immediately classified as *foreign* and skipped. This rule is not implemented for Swiss toponyms since Swiss placenames often appear in titles without giving an indication of the the actual topic of the article. For example, in the article *Akademischer Alpenclub Bern: IX.-XI. Jahresbericht* (in the SAC yearbook of 1916) the Swiss toponym *Bern* is part of a club name while the subsequent text talks of the club's international expeditions, among other topics.

If no foreign toponyms are detected in the title, the sentences of the articles are scrutinised next. Only sentences written in German are considered. (N.B. Even if an article is marked as German, it may contain short passages (e.g. quotes) in another language.) Within a German sentence each token labeled as a NN or NE is compared to the lists of unambiguous Swiss and foreign toponyms. If a count of ten[2] foreign toponyms is reached, the article is classified as *foreign* and skipped. If, however, the article is not skipped and ten or more unambiguous Swiss toponyms have been recognised by the end, it is classified as *Swiss.* If neither of these cases occurr, the article is classified as *unknown* and not processed further.

[1] Between 1957 and 2009 only the German editions are considered.

[2] Limit set by the author. Of 100 evaluated titles, 85 were classified correctly (i.e. an article about Switzerland was not classified as *foreign* or *unknown* and an article with a foreign or non-geographic content was not classified as *Swiss*), resulting in an accuracy of 85%.

Swiss articles are evaluated again line by line from the beginning. All NNs and NEs which fulfill certain criteria are stored to the output file. In each case the corresponding lemma, the volume of the yearbook and the identification number of the token are also included. The mentioned criteria for extracting NNs and NEs are implemented to reduce the number of OCR mistakes and exclude any registered toponyms. The rules for the extraction of NNs and NEs are listed subsequently:

- The token begins with a capital letter.
- The token contains only letters, hyphens, apostrophes and slashes. The punctuation marks are allowed because they often appear in toponyms (e.g. *S-chanf*, *Grava d'Laisch* and *Waltensburg/Vuorz*). Periods, which also appear as part of toponyms (e.g. *Sta. Maria*) are not included in the pattern search since such tokens are generally either OCR mistakes, parts of names and titles or an abbreviation belonging to a compound toponym, which is not of interest yet.
- The token ends with a letter.
- The token does not contain a lower case letter followed by an upper case letter.
- The token does not match any entry in the look-up lists.
- If the token ends in *-es*, *-n* or *-s* (but not in *-ss*), the last or last two letters are removed, depending on the ending (removal of one letter for forms like *-stöcklis* (→ *-stöckli*) and *-hörnern* (→ *-hörner*), removal of two letters for endings like *-joches* (→ *-joch*)). The thus normalised version of the token does not match any entry in the look-up lists. This is done to eliminate case forms of known words or toponyms.

Examples of OCR mistakes which are avoided thanks to these rules are, for instance, *Hinter-Hühnerstoek.129* and *.,Graubünden* and *Hnrter-Trifth8rner*Gletscher.*

NER for Sequences of Multiple Tokens

To create a list of compound candidate toponyms, the same procedure is executed as for the single-word candidate toponyms. Instead of simply looking for one-word tokens, however, series of two or more NNs and NEs directly following each other in a sentence are targeted. (This simple rule is similar to how McDonald (1996) extracted NEs by looking for a continuous sequence of capitalised words.) Like the single NNs and NEs, each part of the sequence has to fulfill the specified criteria to reduce OCR mistakes. Additionally, these criteria are extended by four new rules:

- Tokens may begin with a sequence of lower case letters followed by an apostrophe followed by a capital letter. This condition is to include parts of toponyms such as *l'Obergabelhorn* and *dell'Isra*.
- A period may be part of a token. Abbreviations such as *St.* for *Sankt* are often part of compound toponyms.
- Tokens may not start with an upper case letter followed by a period since this is generally the abbreviation of a first name. The only exception is the letter *S* since *S.* is used as an abbreviation for *San*, as in *San Bartolomeo*.
- Tokens may not contain a lower case letter proceeded by a period since this is most likely an OCR mistake (e.g. *Pz.B.adile*).

Example

The NER step will be illustrated using the following text (for simplicity's sake the text is not displayed in the XML format of the corpus). The tokens annotated as NEs or NNs are printed in bold.

"Wir erreichten endlich den ***Stockhornpaß. Alphonse*** *bewunderte das* ***Thal*** *des* ***Findelbachs****, welches wir nun vor uns sehen konnten und in seiner* ***Begeisterung*** *war sein französischer* ***Accent*** *markanter denn je. Wir setzten unsere* ***Stockhorn-Besteigung*** *fort. Das* ***Marschiren*** *war anstrengend, doch wir hielten öfters an, um vom* ***Thernnometer*** *ablesen zu können, wie kalt es war. Das* ***Haupthinder-niss*** *im Aufstieg bildete die steile* ***Breitwang****. Auf dem Gipfel angekommen, genossen wir die schöne Aussicht und die* ***Sti/le****. Lange blieben wir jedoch nicht, denn die Vorstellung einer warmen* ***Mahlzeit*** *in der* ***Wirthschaft*** *war sehr verlockend und wenige* ***Stunden*** *später erreichten wir auch die* ***Hirtensteinalp****."*

The tokens annotated as NEs or NNs are listed subsequently, divided into the tokens which were matched to an entry in one of the look-up lists and those which remain as candidate unregistered toponyms.

Matched

Findelbachs, Begeisterung, Aufstieg, Breitwang, Gipfel, Aussicht, Vorstellung, Mahlzeit, Stunden

Unmatched

Stockhornpaß, Alphonse, Thal, Accent, Stockhorn-Besteigung, Marschiren, Thernnometer, Haupthinder-niss, Sti/le, Wirthschaft, Hirtensteinalp

Findelbachs was matched with the SwissNames entry *Findelbach* because the genetive *-s* was removed for comparison. The toponym *Breitwang* can also be found in SwissNames. The rest of the matched tokens are recognised German words.

The list of unmatched tokens is worked on in the following data postprocessing steps.

4.3 Data Postprocessing

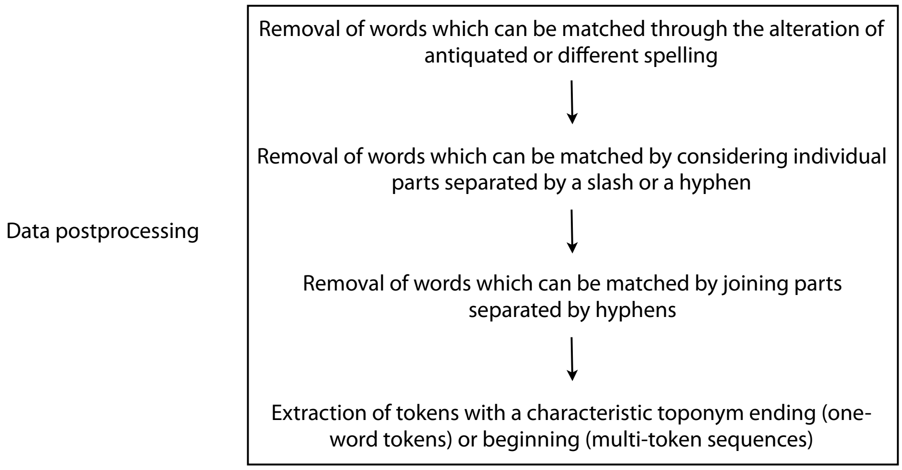

Data postprocessing
Removal of words which can be matched through the alteration of antiquated or different spelling
Removal of words which can be matched by considering individual parts separated by a slash or a hyphen
Removal of words which can be matched by joining parts separated by hyphens
Extraction of tokens with a characteristic toponym ending (one-word tokens) or beginning (multi-token sequences)

Figure 4.4: Schematic overview of Data postprocessing

Despite an initial filtering step during the main NER step which removed NNs and NEs identified as known Swiss or foreign toponyms or normal German words, there are still several variations of the extracted tokens which should be tried for a match. This is done in the following three data postprocessing steps. In the last step, candidate toponyms with characteristic beginnings or endings are extracted.

4.3.1 Antiquated and Different Orthography

The history of German orthography did not really begin until the late 19th century when Konrad Duden published his works on German spelling (Martin, 2011). Changes to the hitherto orthography were made in the early 20th century. These changes included replacing *th* with a *t* and *cc* with *kz* (Bundesverfassungsgericht (BVerfG), 2011). The German speaking part of Switzerland followed the example of their German speaking neighbours and still adheres to approximately the same orthographic rules that are used in Germany and Austria. One exception is the Eszett, the use of which has not been taught in schools since the 1930s (Siebenhaar and Wyler, 1997).

Such antiquated and different orthography appears especially in the older SAC publications. In this postprocessing step several such cases are considered.

Orthography in One-Word Tokens

In this first step of postprocessing, NNs and NEs which have been extracted from the Text+Berg corpus are checked for classical patterns of antiquated orthography. These old patterns are replaced with the corresponding modern pattern. Also, the difference between Swiss and German orthography is taken into account by replacing the Eszett (*ß*) with a double s (*ss*) and vice versa. As already mentioned, the Eszett is used in Germany but generally not in Switzerland. The letter patterns which are looked for and replaced are listed along with the corresponding changes in the following list:

- The sequence of letters *th*, is replaced with *t*. E.g. *Diemtigthal* is changed to *Diemtigtal*.
- The sequence of letters *cc*, is replaced with *kz*. E.g. *Accent* is changed to *Akzent*.
- The sequence of letters *ire*, is replaced with *iere*. E.g. *Botanisiren* is changed to *Botanisieren*.
- The sequence of letters *irt*, is replaced with *iert*. E.g. *Delegirte* is changed to *Delegierte*.
- The letter *ß*, is replaced with *ss*. E.g. *Allalinpaß* is changed to *Allalinpass*.

- The sequence of letters *ss*, is replaced with *ß*. E.g. *Edelweissblüten* is changed to *Edelweißblüten*.

These new versions of the tokens are then compared to the entries in the look-up lists. Normalised versions of tokens ending in *-es*, *-n* or *-s* are also included in the match search. If no matches are found, the tokens are kept in the NER output list in their original form.

Orthography in Sequences of Multiple Tokens

Each token in each sequence is checked for alternate spellings, following the same rules as stated for the one-word tokens. The sequence is added to the output file only if for each of its tokens one version is matched by an entry in the look-up lists.

Example

Here it will be illustrated how antiquated and alternative orthography are treated using the example text. The hitherto unmatched NEs or NNs are printed in bold.

"Wir erreichten endlich den ***Stockhornpaß****.* ***Alphonse*** *bewunderte das* ***Thal*** *des Findelbachs, welches wir nun vor uns sehen konnten und in seiner Begeisterung war sein französischer* ***Accent*** *markanter denn je. Wir setzten unsere* ***Stockhorn-Besteigung*** *fort. Das* ***Marschiren*** *war anstrengend, doch wir hielten öfters an, um vom* ***Thernnometer*** *ablesen zu können, wie kalt es war. Das* ***Haupthinder-niss*** *im Aufstieg bildete die steile Breitwang. Auf dem Gipfel angekommen, genossen wir die schöne Aussicht und die* ***Sti/le****. Lange blieben wir jedoch nicht, denn die Vorstellung einer warmen Mahlzeit in der* ***Wirthschaft*** *war sehr verlockend und wenige Stunden später erreichten wir auch die* ***Hirtensteinalp****."*

The unmatched NEs and NNs are listed subsequently, divided into the tokens which could be matched to an entry in one of the look-up lists by this data postprocessing step and those which remain candidate unregistered toponyms.

Matched

Stockhornpaß, Thal, Accent, Marschiren, Wirthschaft

Unmatched

Alphonse, Stockhorn-Besteigung, Thernnometer, Haupthinder-niss, Sti/le, Hirtensteinalp

The Eszett in *Stockhornpaß* was replaced with a double s. Since *Stockhornpass* is listed in SwissNames, this toponym was matched. The remaining matched tokens were recognised in the GERTWOL list after the antiquated spelling patterns where replaced, resulting in the German words *Tal*, *Akzent*, *Marschieren* and *Wirtschaft*.

Among the unmatched tokens both the variations *Ternnometer* and *Hiertensteinalp* were checked for matches without success and hence not excluded from the list of candidate unregistered toponyms. Since none of the other unmatched tokens contain an antiquated or alternative spelling pattern, they were ignored during this data postprocessing step.

4.3.2 Slashes and Hyphens as Separators in Tokens

Separators in One-Word Tokens

A next step in postprocessing is to consider the parts of tokens separated by a slash ("/") or hyphen ("-"). The tokens with slashes are split at the slashes and then each part is compared to the look-up lists. If such a part contains one of the previously mentioned alternative spelling patterns, the sequence of letters in question is replaced accordingly (see 4.3.1). This altered version of the token part is also compared to the look-up lists, as are normalised forms of tokens and token parts ending in *-es*, *-n* or *-s*. The token is only removed from the list of NER results if at least one version of every part of the token is recognised.

The same procedure is followed for tokens with hyphens, except in this case an additional version of each token part is included in the comparison to the words from the GERTWOL list: a hyphen is added to the end of the token part (e.g. *Engelberger-*).

Separators in Sequences of Multiple Tokens

As with the single NNs and NEs, the tokens in each sequence are checked for matches when split at slashes and hyphens.

Example

By help of the example text it will be illustrated how the rules for tokens containing slashes and hyphens (seen as separating elements) were implemented. The unmatched NEs or NNs are printed in bold.

"Wir erreichten endlich den Stockhornpaß. ***Alphonse*** *bewunderte das Thal des Findelbachs, welches wir nun vor uns sehen konnten und in seiner Begeisterung war sein französischer Accent markanter denn je. Wir setzten unsere* ***Stockhorn-Besteigung*** *fort. Das Marschiren war anstrengend, doch wir hielten öfters an, um vom* ***Thernnometer*** *ablesen zu können, wie kalt es war. Das* ***Haupthinder-niss*** *im Aufstieg bildete die steile Breitwang. Auf dem Gipfel angekommen, genossen wir die schöne Aussicht und die* ***Sti/le****. Lange blieben wir jedoch nicht, denn die Vorstellung einer warmen Mahlzeit in der Wirthschaft war sehr verlockend und wenige Stunden später erreichten wir auch die* ***Hirtensteinalp****."*

The hitherto unmatched NEs and NNs are listed subsequently. The parts of one token could be matched to entries in the look-up lists by this data postprocessing step. The rest remain candidate unregistered toponyms.

Matched
Stockhorn-Besteigung

Unmatched
Alphonse, Thernnometer, Haupthinder-niss, Sti/le, Hirtensteinalp

The token *Stockhorn-Besteigung* was split at the hyphen into its two parts. Since *Stockhorn* is a toponym registered in SwissNames and *Besteigung* is a normal German noun recognised by GERTWOL, this token can be excluded from the list of candidate unregistered toponyms.

Both *Haupthinder-niss* and *Sti/le* were also split at the hyphen and slash respectively. Since at least one of the two resulting parts per word (*Haupthinder*, *niss* and *Sti*, *le*) did not match any entry in the look-up lists, these tokens are retained as candidate unregistered toponyms. As none of the other unmatched tokens contain a hyphen or slash, they were ignored during this data postprocessing step.

4.3.3 Hyphens as Joining Elements in Tokens

Joining Elements in One-Word Tokens

Certain tokens with hyphens are subjected to a further round of alteration and filtration: all tokens containing a hyphen followed by a lower case letter are split along the hyphens. Starting from the end of the token, each part is inspected. If the part begins with an upper case letter, it is stored separately. If, however, the part begins with a lower case letter, it is joined to the part proceeding it in the token. By joining the two parts, a new token part is created. This replaces the proceeding part to which the lower case string was joined. An example is included to illustrate this procedure:

Step 1 Split the token *Lawinen-Ver-schütteten-Suchgerätes* along the hyphens into the parts *Lawinen*, *Ver*, *schütteten* and *Suchgerätes*.

Step 2 Start at the end of the token: The fourth and last part, *Suchgerätes*, begins with an upper case letter and is hence stored separately.

Step 3 The third part, *schütteten*, starts with a lower case letter and is hence joined to the second part of the token, *Ver*. *Verschütteten* is saved as the new second part.

Step 4 The second part of the token is now *Verschütteten* and starts with an upper case letter. Hence it is stored separately.

Step 5 The first part of the token, *Lawinen* also begins with an upper case letter and is saved separately.

The resultant new tokens are then subjected to the same procedure as in the last post-processing step: where applicable, alternatively spelled variations are added as well as normalised versions and versions ending with a hyphen. Then all possible forms of the token parts are compared to the entries in the look-up lists. As before, the token is not removed from the list of NER results unless at least one version of every part of the token is recognised. In the afore mentioned example, the candidate toponym *Lawinen-Ver-schütteten-Suchgerätes* can be correctly identified as a non-toponym and is excluded from the list of NER results because the parts *Lawinen*, *Verschütteten* and *Suchgerätes* match entries in the look-up lists.

Joining Elements in Sequences of Multiple Tokens

The sequence parts containing slashes followed by a lower case letter are treated separately, following the method described previously. The sequence is not removed from the list of NER outputs unless a match is found for every token and every token part.

Example

By help of the example text it will be illustrated how the rules for tokens containing hyphens (viewed as joining elements) were implemented. The hitherto unmatched NEs or NNs are printed in bold.

"Wir erreichten endlich den Stockhornpaß. ***Alphonse*** *bewunderte das Thal des Findelbachs, welches wir nun vor uns sehen konnten und in seiner Begeisterung war sein französischer Accent markanter denn je. Wir setzten unsere Stockhorn-Besteigung fort. Das Marschiren war anstrengend, doch wir hielten öfters an, um vom* ***Thernnometer*** *ablesen zu können, wie kalt es war. Das* ***Haupthinder-niss*** *im Aufstieg bildete die steile Breitwang. Auf dem Gipfel angekommen, genossen wir die schöne Aussicht und die* ***Sti/le****. Lange blieben wir jedoch nicht, denn die Vorstellung einer warmen Mahlzeit in der Wirthschaft war sehr verlockend und wenige Stunden später erreichten wir auch die* ***Hirtensteinalp****."*

The unmatched NEs and NNs are listed subsequently. One token could be matched to an entry in the look-up lists by this data postprocessing step. The rest remain candidate unregistered toponyms.

Matched

Haupthinder-niss

Unmatched

Alphonse, Thernnometer, Sti/le, Hirtensteinalp

The token *Haupthinder-niss* was split at the hyphen into its two parts. Because it begins with a lower case letter, the latter part *niss* was joined to *Haupthinder*, resulting in the new word *Haupthinderniss*. This form of the original token is a German noun and was matched to an entry in the GERTWOL list.

Since none of the other tokens contain hyphens, they are not considered in this step.

4.3.4 Characteristic Toponym Beginnings and Endings

Despite the various steps of postprocessing there are still many OCR mistakes and false positives such as person names in the final lists of NER results. In this case McDonald's (1996) internal evidence can be used to get a more condensed and relevant list. Many toponyms have a characteristic ending or beginning, often also signaling to a reader what type of geographic feature the placename is referring to. Endings such as *-gletscher* (glacier), *-tobel* (ravine) and *-grätli* (little ridge) are examples of common toponym endings, while many compound Alpine placenames begin with words like *Piz* (peak), *Alp* and *Fuorcla* (pass). Such characteristic parts of toponyms are gathered from three sources:

- the documentation of the Text+Berg NER files, where several toponym beginnings and endings were determined for the recognition of mountain, glacier and cabin names,
- the most frequent occurrences in SwissNames and
- the most frequent occurrences in the postprocessed list of NER outputs.

This resulted in a list of 94 typical toponym endings and 329 typical toponym beginnings (see appendix A).

Typical Endings of One-Word Tokens

The list containing the one-word tokens is filtered for words with a characteristic ending. Also included are genitive and dative forms like *-stöcklis*, *-joches* and *-hörnern*. These tokens and their normalised forms have already been compared to the entries in the look-up lists in a previous step and since the the tokens are not altered in any way, there is no need to check for matches again.

Typical Beginnings of Sequences of Multiple Tokens

Similarly, the list containing the token sequences is filtered for entries containing a characteristic toponym beginning. This characteristic word must either be the first token of the sequence or, if it is one of the following tokens, it must be preceded by a space or a non-word character, such as a hyphen for example. In each case the characteristic beginning must be followed by a space, since it should be a seperate token. These rules make it possible to also recognise partial sequences such as *Piz Forbisch* from *Kleinen Piz Forbisch* or *Alp Albeina* from *Saaser-Alp Albeina.* The compound candidate toponyms are also checked for grammatical case forms and are normalised, since in several cases only part of the sequence is considered and hence may be matched in the look-up lists.

Example

By help of the example text it will be illustrated how this last data postprocessing step was implemented. The unmatched NEs or NNs are printed in bold.

"Wir erreichten endlich den Stockhornpaß. ***Alphonse*** *bewunderte das Thal des Findelbachs, welches wir nun vor uns sehen konnten und in seiner Begeisterung war sein französischer Accent markanter denn je. Wir setzten unsere Stockhorn-Besteigung fort. Das Marschiren war anstrengend, doch wir hielten öfters an, um vom* ***Thernnometer*** *ablesen zu können, wie kalt es war. Das Haupthinder-niss im Aufstieg bildete die steile Breitwang. Auf dem Gipfel angekommen, genossen wir die schöne Aussicht und die* ***Sti/le****. Lange blieben wir jedoch nicht, denn die Vorstellung einer warmen Mahlzeit in der Wirthschaft war sehr verlockend und wenige Stunden später erreichten wir auch die* ***Hirtensteinalp****."*

The hitherto unmatched NEs and NNs are listed subsequently. One token was extracted due to its ending, which is suggestive of a toponym.

Extracted
Hirtensteinalp

Discarded
Alphonse, Thernnometer, Sti/le

The token *Hirtensteinalp* was extracted due to the characteristic ending *-alp.* This is likely an unregistered Swiss vernacular toponym or an alternative spelling of a placename registered in Switzerland.

The remaining three tokens are not processed further. Two (*Thernnometer*, *Sti/le*) are OCR mistakes, the correct words being *Thermometer* and *Stille.* *Alphonse* is a French name which GERTWOL does not recognise.

4.4 Result Analysis

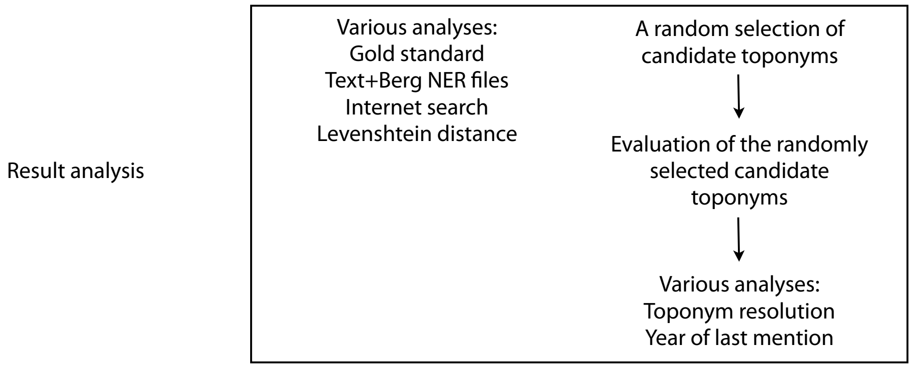


Figure 4.5: Schematic overview of result analysis

The results are analysed under various aspects. Comparison to several gold standard articles should give an insight into how well the NER system performs at each step and where the problems lie. Comparison to the Text+Berg NER files are intended to highlight the differences between the thesis NER system and the approach used to extract the names of mountains, glaciers and cabins. The search for the candidate toponyms on Swiss webpages may give an idea of which candidate toponyms are Swiss and which aren't. Filtering for candidate toponyms which are similar to an entry in SwissNames is done using the Levenshtein distance. To learn something about the overall performance of the NER system a random selection of candidate toponyms is manually evaluated. Toponym resolution is carried out for 100 of the randomly selected tokens which are classified as Swiss. Finally, all randomly selected tokens which are identified as Swiss are assessed for their last mention in the SAC yearbooks. This is done to get an estimate of each toponym's age.

4.4.1 Gold Standard

To get an idea of the performance of the here developed NER system, the results from NER and data postprocessing are compared to five Text+Berg gold standard articles. The articles were chosen so as to have articles of different ages and written on different topics:

- *Bergfahrten in der Zentralschweiz* (Ascents in Central Switzerland), 1903
- *Bianco-Bernina im Winter* (Bianco-Bernina in Winter), 1939
- *Wohin die Neugier führen kann* (Where Curiousity Can Lead), 1968
- *Blockgletscher im Weissmies und Aletsch und ihre photogrammetrische Kartierung* (Rock Glaciers in the Weissmies and Aletsch Areas and Their Photogrammetric Field Mapping), 1968
- *Ausgewählte Klettertouren in den Waadtländer Alpen* (Selected Climbing Tours in the Alps of Vaud), 1986

In table 4.1 the lengths (number of sentences and number of tokens) of these five gold standard articles are listed. The number of toponyms is also given for each article.

Article	**Sentences**	**Tokens**	**Toponyms**
Bergfahrten	116	2’415	146
Bianco-Bernina	101	1’863	45
Neugier	146	2’875	76
Blockgletscher	331	5’571	104
Klettertouren	149	2’745	61

Table 4.1: Characterisation of the chosen gold standard articles

The main NER processes are repeated specifically for each of the five articles. The results are then manually compared to the gold standard article and evaluated for each step of NER and data postprocessing. However, since all registered toponyms are ignored by this NER system, the gold standards have to be adjusted accordingly. To get an accurate picture, the toponyms marked in each gold standard article are checked for matches in SwissNames. All direct matches as well as matches to an entry in the SwissNames list after modifications due to different spelling, hyphens, slashes and normalisation are discarded. Also removed from the list of gold standard toponyms are very general placenames since they are not interesting to the purpose of this thesis and would be recognised by GERTWOL (e.g. *Zentralschweiz*, *Westalpen*, *Voralpen*, etc.).

4.4.2 NER for Mountains, Glaciers and Cabins

The Text+Berg NER files hold information on the names of mountains, cabins and glaciers which were recognised in the Text+Berg corpus. To see how many of these toponyms are detected by the thesis' NER system, the entries of the Text+Berg NER files are compared to the NER output. However, since the toponyms themselves are not included in the NER files of the Text+Berg corpus, it is first necessary to obtain the corresponding tokens from the yearbook and member journal files.

Token Extraction

The Text+Berg NER files contain the corpus IDs of the recognised tokens as well as the toponyms' swisstopo ID (*stid*). As only unregistered toponyms are of interest, specifically the entries with a swisstopo ID set to zero are considered[3]. The whole corpus is scanned for matches to the corpus IDs of the toponyms referenced in the Text+Berg NER files. To exclude any registered toponyms, the extracted tokens are compared to the entries in the SwissNames gazetteer before being saved to a list of Text+Berg NER toponyms.

Token Comparison

The so compiled list should now contain only unregistered names of mountains, cabins and glaciers. All matching entries in the data postprocessing output files of one-word tokens and token sequences are identified. To get an idea of how many of the toponyms which were not matched by entries in the data postprocessing output files are Swiss, a random selection (using a programme similar to the one described in subsection 4.4.5) of 10% of these toponyms is assessed manually.

4.4.3 Internet Search

Using the Internet for toponym disambiguation or the comparison of toponyms is not a new concept (Philip Smart, 2009; Naaman et al., 2006; Pasley et al., 2008). The idea is to see how many hits a query (in this case a candidate Swiss toponym) receives. The

[3] Pro memoria: *stid=0* signifies that either the toponym is not registered in SwissNames or the toponym is ambiguous and therefore no swisstopo ID can be assigned

search is restricted to Swiss pages since the focus lies on Swiss toponyms. The more hits a candidate toponym achieves, the higher the probability that it is, in fact, a Swiss and not a foreign toponym. The Internet search is done using a Java code[4] which reads the entries in a list and records how many hits each entry scores for a search with the *yahoo* search engine. The hits on the internet are set into perspective with the amount of times the toponym is mentioned in the corpus.

4.4.4 Levenshtein Distance

The Levenshtein distance, often also called the edit distance, is used as a measure for the similarity of two strings and is defined as the minimal number of transformations which are needed to turn one of the strings into the other. One *transformation* is for instance the deletion or insertion of a letter or the replacement of a letter with another one (Bernstein et al., 2005). The concept is illustrated by means of an example, showing how *Bernau* is turned into *Bärnalp* by the following transformations:

1. Replacement of *e* with *ä* (→ *Bärnau*).
2. Replacement of *u* with *l* (→ *Bärnal*).
3. Insertion of *p* at the end of the string (→ *Bärnalp*).

Hence the Levenshtein distance between the two strings is three.

In this thesis, the Levenshtein distance is used to detect words and sequences of words among the NER results which are similar to entries in the SwissNames gazetteer, thus suggesting a possible identification. Since a Levenshtein distance of two can already match a pair of completely different strings (e.g. *Aletschalp* and *Aletschwald*), the Levenshtein distance is set to 1 for the result analysis. To increase efficiency, only words beginning with the same letter and of approximately the same length (maximum one letter difference) are compared. Two exceptions are made:

1. Words beginning with an Umlaut are first compared to toponyms beginning with an Umlaut. In a second step, the Umlaut is replaced with the corresponding regular letter and an *e* (*Ä* → *Ae*, *Ö* → *Oe*, *Ü* → *Ue*). This

[4] This Java programme was written and kindly loaned for use in this thesis by Ramya Venkateswaran, a PhD student at the University of Zurich's Departement of Geography.

is an alternative way of writing an Umlaut, used especially in the case of a capital Umlaut. The altered words are then compared to toponyms beginning with the regular letter (A, O or U) and of lengths one letter longer than those considered before (due to the added e).

2. If a word contains an Eszett, it is replaced with a double s since toponyms in SwissNames are not written with an Eszett and $\textit{ß} \rightarrow ss$ already implies a Levenshtein distance of two. Here, the length of the word with *ss* is taken as a reference.

For efficient access to toponyms beginning with a certain letter, the files of alphabetically sorted SwissNames toponyms and partial toponyms are used (see 4.1.3). All toponyms which are sufficiently similar to a candidate toponym are saved to an output file. The same steps are followed for both the list with one-word tokens and the list containing multi-token sequences. Sequences were more generally checked for a capital Umlaut in any token to account for all parts of the sequence which may begin with an Umlaut. Depending on how many capital Umlaut are replaced, the length of the considered toponyms is adjusted accordingly.

The algorithm used to calculate the Levenshtein distance was written by a software engineer and is available online (Bendersky, 2003).

4.4.5 A Random Selection of Candidate Toponyms

To get an idea of the quality of the lists obtained by the end of the data postprocessing, 10% of the output lists are evaluated. Using a *Perl* function called *shuffle* these candidate toponyms are randomly selected from the NER output files. For each randomly selected candidate toponym, one reference is also selected randomly. Using this reference, the title of the article and the sentence in which the candidate toponym appears in the Text+Berg corpus are extracted. This additional information is included to facilitate the manual evaluation of the randomly selected candidate toponyms.

Evaluation of the Randomly Selected Candidate Toponyms

The randomly selected candidate toponyms are manually assessed and classified into one of the three categories *Swiss* (for Swiss toponyms), *foreign* (for foreign toponyms) and *non* (for words which are not toponyms, e.g. *Auberge* (French for *hostel*) or *Sildwestwand* (OCR mistake, should be *Südwestwand*)). In each case the decision is made using the information provided by the article title and the sentence context as well as the Internet. In difficult cases the Text+Berg file is consulted to get a better understanding of the context the candidate toponym appears in.

Toponym Resolution

For 100 toponyms (50 toponyms consisting of one word, 50 compound toponyms) which have been identified as Swiss, the official corresponding SwissNames entry was determined by hand. The municipality to which the toponym's geographic feature belongs is also saved. Using the official toponym and the municipality for disambiguation if necessary, the coordinates of the toponym are mined from the SwissNames gazetteer with a *Perl* programme. The coordinates are then imported into ArcGIS to create a map of the located toponyms.

Year of Last Mention of a Toponym

The Text+Berg corpus holds texts which are more than one century old. Consequently, it is to be expected that toponyms are mentioned which may no longer be in use today. Also, rules of orthography have changed, creating several spelling variations of the same toponym. To get an idea of the impact which this large span of years has on the results, the randomly extracted Swiss toponyms are evaluated with respect to their "age": in each case the year of the most recent mention in the Text+Berg corpus is determined.

5 Results and Interpretations

The results generated by the NER system explained in the previous part of the thesis are described and interpreted in the following chapter. First, the outcome of the data preprocessing steps is briefly presented, after which the NER and data postprocessing results are summarised. The main section of this chapter is devoted to the output of the result analysis programmes.

5.1 Results from Data Preprocessing

The size of the lists produced at each step of the data preprocessing are shown in table 5.1.

Data preprocessing step	**Number of entries**
Words recognised by GERTWOL	254’346
Foreign toponyms	1’796
Toponyms registered in SwissNames	122’233
Overlap of the GERTWOL list and SwissNames	5’605
... of which are geo-non geo ambiguous	2’875

Table 5.1: An overview of the lists generated during data preprocessing

The number of entries in the list containing the registered Swiss toponyms and the individual parts of these toponyms is smaller than that of the SwissNames gazetteer (ca. 193’000 entries (Swiss Federal Office of Topography, 2005)) because each toponym or toponym part was only included once, whereas there are many ambiguous toponyms in SwissNames.

The words which are both found in the SwissNames look-up list and recognised by GERTWOL were analysed manually for geo-non geo ambiguity. Examples of toponyms which were classified as geo-non geo ambiguous are *Zug*, *Dürrenmatt* and *Seebad* (see appendix A for a larger excerpt of the list of geo-non geo ambiguous toponyms). *Jungfrau*, *Schiahorn* and *Bernina*, on the other hand, are examples of toponyms which appear both in the

GERTWOL list and SwissNames but are not included in the list of geo-non geo ambiguous words, since they are likely to be used in a Swiss context in the SAC volumes.

5.2 Results from NER and Data Postprocessing

During NER, 11'560 articles were analysed, 9'689 of which were identified as written in German. Of these 4'680 were classified as *Swiss*, 1'635 as *foreign* and 3'472 as *unknown.* The number of one-word tokens and sequences of tokens recognised during NER and in each data postprocessing step are listed in tables 5.2 and 5.3. The numbers refer to distinct entries - if a token or sequence of tokens was recognised multiple times, the hits were merged to one entry in such a way as to preserve the metadata (corpus IDs, lemmata, etc.) of each find. The percentages included in the last two columns are to give a qualitative idea of the impact of each step. The column $\%_{NER}$ shows the impact in terms of the original list of NER results, while the the column *Reduction* lists the percentage by which the proceeding number of entries was reduced.

5.2.1 One-Word Tokens

Main NER step	One-word tokens	$\%_{NER}$	Reduction
NER	111'678	100%	-
Orthography check	109'529	98.1% (of 111'678)	1.9% (of 111'678)
Separating elements	95'599	85.6% (of 111'678)	12.7% (of 109'529)
Joining elements	91'246	81.7% (of 111'678)	4.6% (of 95'599)
Characteristic endings	4'273	3.8% (of 111'678)	95.3% (of 91'246)

Table 5.2: An overview of the output for NER and each data postprocessing step (one-word tokens)

As can be gathered from table 5.2, each step reduced the number of one-word tokens by a relatively small, but distinct amount. The exception is of course the last step, which was very selective by only recognising toponyms with a distinct ending. Other than this outlier, the largest impact was made by splitting the tokens at slashes and hyphens and considering the individual parts. This suggests that a large amount of the data is prone to OCR mistakes. The least effect was achieved by the replacement of alternative spelling patterns. Though relatively insignificant, this reduction is nonetheless surprising, considering the simplicity of the algorithm and the small number of patterns searched for.

5.2.2 Sequences of Multiple Tokens

Main NER step	Token sequences	$\%_{NER}$	Reduction
NER	36'884	100%	-
Orthography check	35'139	95.3% (of 36'884)	4.7% (of 36'884)
Separating elements	35'068	95.1% (of 36'884)	0.2% (of 35'139)
Joining elements	34'445	93.4% (of 36'884)	1.8% (of 35'068)
Characteristic beginnings	2'740	7.4% (of 36'884)	92.0% (of 34'445)

Table 5.3: An overview of the output for NER and each data postprocessing step (token sequences)

About 70% less token sequences than one-word tokens were recognised. Also, the post-processing steps did not have an equally strong impact on the number of entries in the list of NER results. Interestingly enough, the replacement of alternative spelling patterns reduced the list most of all in the case of token sequences (not counting, of course, the restrictive search for toponyms with characteristic beginnings). The reason for this may be the simplicity of the code. Treating all parts of a token sequence containing slashes and hyphens and checking all possible versions produced a convoluted programme. Additionally, *all* parts of *all* tokens of a sequence had to be matched with either a registered toponym or a German word before the sequence was deleted from the list of NER results. While the first three data postprocessing steps have limited success, almost double the relative amount of token sequences with a characteristic beginning than tokens with a special ending are recognised. This is likely due to the fact, that many more characteristic beginnings than endings were included in the search.

5.3 Result Analysis

5.3.1 Gold Standard

The results of the gold standard analysis are summarised in the five tables 5.4, 5.5, 5.6, 5.7 and 5.8. Each table shows the precision and recall values achieved during NER and the ensuing postprocessing steps for one article, as well as the number of unregistered or alternatively spelled toponyms in the gold standard (designated by the letter N, recall values were evaluated using this number). The values vary from article to article. In every case the precision lies at 100% for the step where tokens with characteristic toponym endings or beginnings were identified. It will be seen, however, that as a consequence recall values often decrease.

Also included in each table are the precision and recall values achieved by the NER approach used to detect the names of mountains, glaciers and cabins (referred to as the corpus NER approach subsequently). Here, too, only the unregistered and alternatively spelled toponyms are considered, so the recall values are calculated using the same N as for the thesis NER system results.

Bergfahrten in der Zentralschweiz

The precision and recall values achieved for this text are the highest of all five articles. Considering the age of the article (over 100 years old) it is interesting to note the improvement in precision by the orthography check. Although the recall value drops slightly, it is clear to see that a fair number of tokens could be resolved as non-toponyms by altering the spelling to a more modern version. The following two postprocessing steps neither decreased nor increased precision and recall, while more than 50% of gold standard toponyms are detected with a precision of 100% during the last step. The relatively high precision and recall values in the previous steps are limited by several OCR mistakes and toponyms recognised by GERTWOL respectively.

The corpus NER approach also achieves a precision of 100% but the recall value is only about half the recall value achieved by the thesis NER system in its last data postprocessing step.

Main NER step	**Precision**	**Recall** (N=67)
NER	68.8%	71.6%
Orthography check	75.4%	68.7%
Separating elements	75.4%	68.7%
Joining elements	75.4%	68.7%
Characteristic beginnings/endings	100%	50.7%
Corpus NER approach	100%	25.4%

Table 5.4: Gold standard comparison for the article *Bergfahrten in der Zentralschweiz*, 1903

Bianco-Bernina im Winter

Both precision and recall values are lower for this article than for the previous one but they are still relatively high and can be increased to over 50 and 60% respectively by the first three postprocessing steps. Precision is limited by OCR mistakes and unrecognised German words, while recall is not higher because several toponyms were recognised by GERTWOL, e.g. *Berninascharte* and *Berninagruppe*. Again, considering orthography, but this time also altering tokens containing slashes and hyphens, have a good effect on precision. Recall is not decreased during these processes. The last postprocessing step achieves a rather low recall value of 30%.

The corpus NER approach also achieves a precision of 100% but, again, the recall value is much lower than the recall value achieved by the thesis NER system in its last data postprocessing step.

Main NER step	**Precision**	**Recall** (N=20)
NER	44.8%	65.0%
Orthography check	48.1%	65.0%
Separating elements	52.0%	65.0%
Joining elements	52.0%	65.0%
Characteristic beginnings/endings	100%	30.0%
Corpus NER approach	100%	5.0%

Table 5.5: Gold standard comparison for the article *Bianco-Bernina im Winter*, 1939

Wohin die Neugier führen kann

Precision and recall values start poorly for this article. Low precision is caused mainly by an unrecognised family name which is mentioned several times in different combinations, while low recall is due to the fact, that many words are recognised by GERTWOL, e.g. *Jungfraubahn* and *Rottal-Route*. Precision is only slightly increased by the first three steps of data postprocessing, while recall is not affected. The last postprocessing step achieves a low recall value at around 30%, comparable to the values of the second article.

Since no unregistered or alternatively spelled toponyms are extracted by the corpus NER approach for this article, both the precision and recall value are at 0%.

Main NER step	**Precision**	**Recall** (N=7)
NER	15.4%	28.6%
Orthography check	15.4%	28.6%
Separating elements	16.7%	28.6%
Joining elements	16.7%	28.6%
Characteristic beginnings/endings	100%	28.6%
Corpus NER approach	0%	0%

Table 5.6: Gold standard comparison for the article *Wohin die Neugier führen kann*, 1968

Blockgletscher im Weissmies und Aletsch und ihre photogrammetrische Kartierung

These are the lowest precision values of all five articles, because the article contained an unusually large amount of OCR mistakes and technical terms, which GERTWOL did not recognise. There is only a slight increase in precision during the first three data postprocessing steps while recall values decrease noticeably. Again, these low values are mainly due to GERTWOL - too many toponyms and toponym parts are recognised. As in the previous article, recall for the last step is at around 30%.

No unregistered or alternatively spelled toponyms are extracted by the corpus NER approach for this article. Hence, both the precision and recall value are at 0%, like in the preceding articles.

Main NER step	**Precision**	**Recall** (N=12)
NER	9.9%	66.7%
Orthography check	9.9%	66.7%
Separating elements	9.0%	50.0%
Joining elements	10.0%	50.0%
Characteristic beginnings/endings	100%	33.3%
Corpus NER approach	0%	0%

Table 5.7: Gold standard comparison for the article *Blockgletscher im Weissmies und Aletsch und ihre photogrammetrische Kartierung*, 1968

Ausgewählte Klettertouren in den Waadtländer Alpen

This article sports the lowest recall values and also dropping (!) precision values for the first three steps of data postprocessing. The recall value achieved by looking for tokens with characteristic toponym beginnings or endings is also poor at below 5%. Low precision values are caused mainly by names of persons and climbing terminology which GERTWOL did not recognise. Low recall is largely due to wrong pos-tags for parts of sequences. For example, *Paroi* in *Paroi du Diamant* was wrongly tagged as an *attributive adjective* (probably because this is a French word and hence is unknown to GERTWOL). Since this NER system only considers NNs and NEs, the first two words in the toponym (*Paroi*, *du*) could not be detected. In any case recognition of such compound toponyms is problematic since they contain a part in lower case letters, as in this example, *du*.

Main NER step	**Precision**	**Recall** (N=21)
NER	20.0%	28.6%
Orthography check	20.0%	28.6%
Separating elements	17.2%	23.8%
Joining elements	14.8%	19.0%
Characteristic beginnings/endings	100%	4.8%
Corpus NER approach	0%	0%

Table 5.8: Gold standard comparison for the article *Ausgewählte Klettertouren in den Waadtländer Alpen*, 1986

Since no unregistered or alternatively spelled toponym is extracted by the corpus NER approach for this article, both the precision and recall value are at 0%.

In all five cases, precision values were limited mainly by OCR mistakes and unknown words such as names of persons, specific terminology or words in other languages. Recall, on the other hand, was adversely affected by comparison to the words recognised by GERTWOL as well as wrong pos-tagging in the corpus and the difficulty in detecting compound toponyms. Many unregistered Swiss toponyms were excluded because of this. The implications of this will be discussed in chapter 6.

The corpus NER approach extracted unregistered or alternatively spelled toponyms only in two cases (see tables 5.4 and 5.5). For both these articles, the precision value was 100% but the recall value was significantly lower than that achieved by the thesis NER system. In the remaining three cases no corresponding toponyms were extracted by the corpus NER, resulting in precision and recall values of 0%. This strongly suggests, that the thesis NER approach performs better at detecting unregistered or alternatively spelled toponyms than the corpus NER system. This is most likely due to the restrictive search for just the names of mountains, glaciers and cabins.

5.3.2 NER for Mountains, Glaciers and Cabins

Token Extraction

A list of 3'820 distinct one-word and compound toponyms resulted from the corpus extraction of tokens identified in the Text+Berg NER files. If the extraction is limited to German sentences in German articles about Switzerland (same selection as the thesis NER system), this is reduced to 2'497 distinct toponyms. This value of nearly 2'500 toponyms is the number against which the results of the thesis NER system are compared in the next step of result analysis, token comparison.

Token Comparison

By looking for matches in the output files of the penultimate data postprocessing step (before selective filtering for tokens with a particular ending or beginning), 1'428 (57% of 2'497) toponyms are detected while 1'069 (43% of 2'497) remain unmatched. The unmatched toponyms can be explained by the restrictive rules which were stipulated in the thesis NER process: 1'026 of these missed toponyms match entries in the lists of foreign toponyms or words recognised by GERTWOL and were hence excluded during the thesis NER process. The remaining 43 are explained by genitive and dative case forms of registered toponyms or toponyms containing an old spelling pattern, an Eszett or a hyphen which could also be matched to a Swiss or foreign toponym or a GERTWOL word after the steps of data postprocessing.

The analysis of 10% of the 1'069 undetected toponyms revealed that about 60% are unregistered or alternatively spelled Swiss toponyms, while the remaining 40% are foreign toponyms. This means that more than half of the names of mountains, glaciers and cabins which were excluded by the thesis NER system were wrongly discarded.

5.3.3 Internet Search

The results of the Internet search are summarised in graphs 5.1 (one-word toponyms) and 5.2 (compound toponyms). The x-axes display the number of times a token appears in the Text+Berg corpus while the y-axes show the logarithm of the number of Internet hits plus one (this increment of the count by one is necessary since several candidate toponyms did not return any Internet search results). Some of the outlying points are labeled to give an impression of the kind of candidate toponyms which are involved.

One-word Candidate Toponyms

Twelve tokens are mentioned more than 100 times in the Text+Berg corpus and were excluded from the graph to reduce the span. These tokens are *Blümlisalp*, *Fellenberg*, *Rhonetal*, *Göscheneralp*, *Linththal*, *Rhonethal*, *Verstanklahorn*, *Konkordiahütte*, *Weissfluhjoch*, *Wyttenbach*, *Bovalhütte* and *Märjelensee*. Both *Fellenberg* and *Wyttenbach* are surnames. The remaining 10 are variations of well-known Swiss toponyms.

The remaining tokens, mentioned less than 100 times in the Text+Berg corpus and dsiplayed in the graph, show no clear trend. Most tokens are mentioned less than 20 times in the corpus. An intriguing fact is that the tokens which achieved the highest number of Internet hits do not appear often in the corpus. This can be explained by the fact that while the tokens in question, like *Metal*, *Métal* and *Auberge*, are fairly common words (not

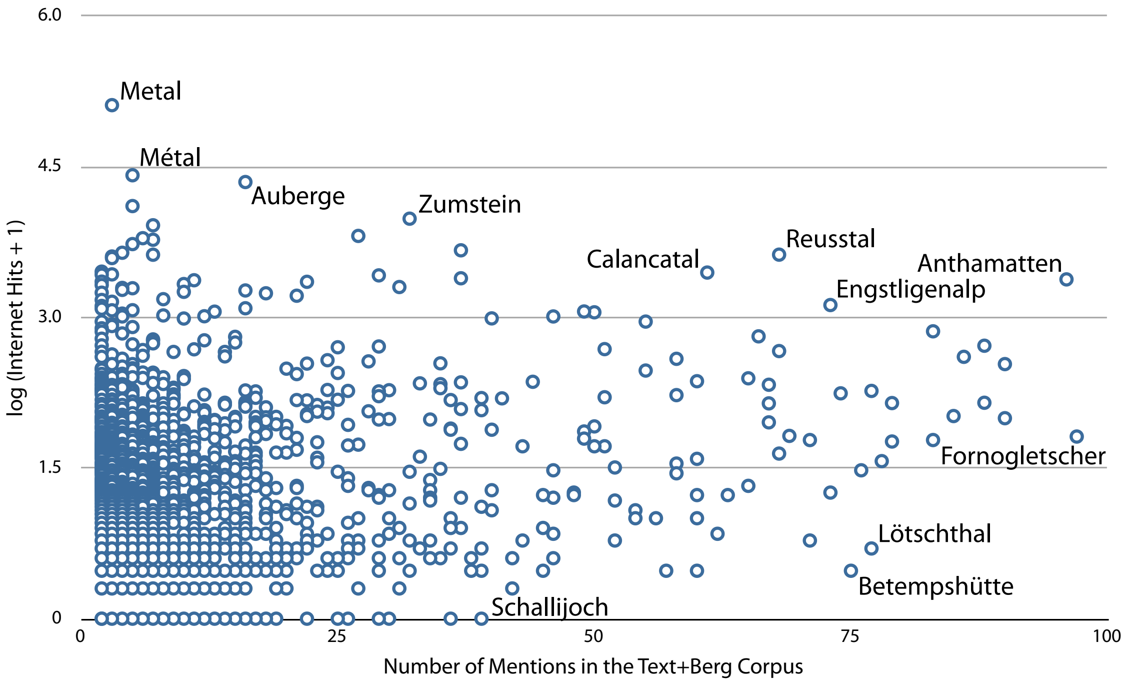


Figure 5.1: One-word candidate toponyms: Internet hits vs. mentions in the corpus

toponyms!) in everyday life (and are consequently found in abundance on the Internet), they are not central to mountaineering. In addition these three words are from foreign languages (English, French) and for this reason are less likely to appear many times in German articles.

Zumstein and *Anthamtten* are surnames and hence appear relatively often in both the corpus and the Internet[1]. The rest of the labels are unregistered Swiss vernacular toponyms (e.g. *Fornogletscher* referring to *Vadrec del Forno*) or spelling variations of registered toponyms (e.g. *Schallijoch* from *Schalijoch*).

Compound Candidate Toponyms

Four token sequences are mentioned more than 50 times in the Text+Berg corpus and were removed from the list to reduce the span of the graph. These sequences are *Piz Rusein*, *Monte Viso*, *Piz Tumbif* and *Piz Uertsch*. *Monte Viso* is an Italian mountain while the remaining three are spelling variations of well-known Swiss mountains (*Piz Russein*, *Piz Tumpiv* and *Piz Üertsch*).

[1] Additionally, *Zumstein* is also part of a toponym: the *Zumsteinspitze* is situated in the well-known Monte Rosa area.

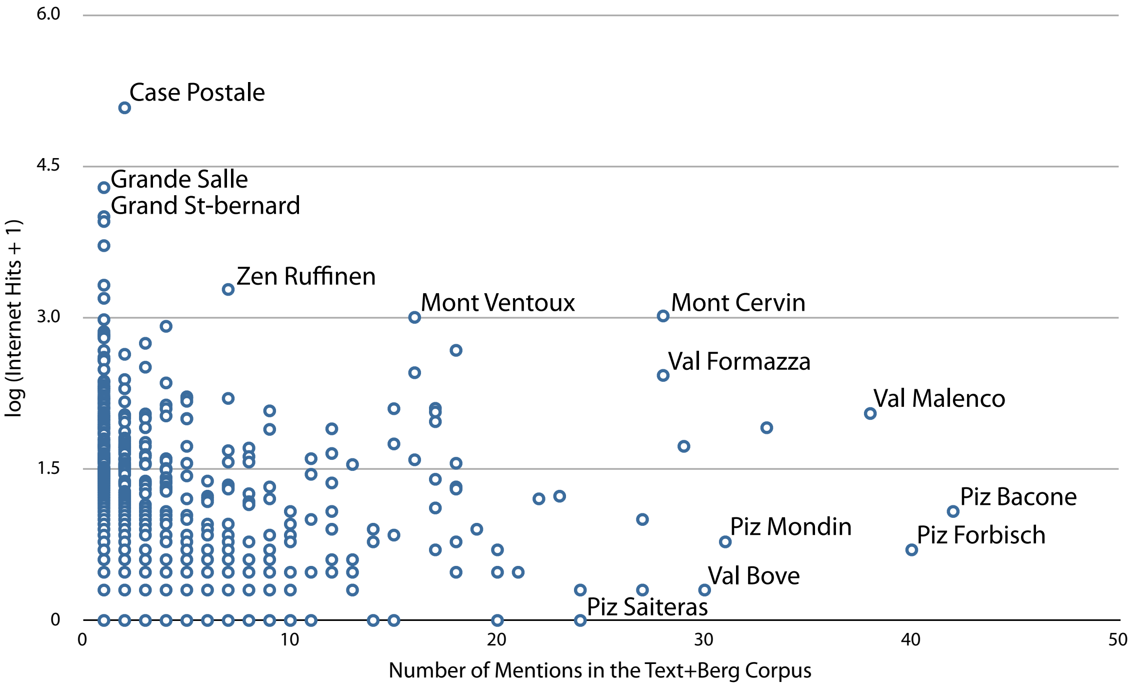


Figure 5.2: Compound candidate toponyms: Internet hits vs. mentions in the corpus

There are several similarities between these results for compound candidate toponyms and those received for the one-word candidate toponyms (see figure 5.1). Again, no overall trend is discernible. Most sequences appear less than 10 times in the corpus. As with the one-word candidate toponyms, the highest number of Internet hits are coupled with few mentions in the corpus. These token sequences are common words in foreign languages (*Case Postale* and *Grande Salle*), special spelling variations of well-known toponyms (*Grand St-bernanrd*) and surnames (*Zen Ruffinen*).

Among the remaining labels there are several foreign toponyms (*Mont Ventoux*, *Val Formazza*, *Val Malenco* and *Val Bove*). *Piz Forbisch* and *Piz Mondin* are spelling variations of the Swiss mountain names *Piz Forbesch* and *Piz Mundin*. *Piz Bacone* and *Mont Cervin* are unregistered Swiss vernacular toponyms, being the Italian name for *Piz Bacun* and the French name for the famous *Matterhorn*. *Piz Saiteras* appears to be an unregistered vernacular toponym referring to the Swiss mountain *Piz Salteras* but could also be an OCR mistakes in some cases.

In summary it can be said that the Internet search did not yield conclusive findings. What has worked for more well-known toponyms fails in this case because most of the candidate toponyms found in the thesis NER system are too specific and are rarely mentioned even in the Text+Berg corpus. The fact that most candidate toponyms returned few Internet hits is in itself useful information, however. It suggests that only few people know these toponyms, that they are not used anymore or that the granularity of these toponyms is fine. This latter explanation is in accordance with Zipf's law which states that "the frequency of a word [(here, more specifically, a toponym)] is inversely proportional to the rank" (Tague-Sutcliffe, 1992, p. 2). In this sense, rank is interpreted as the granularity of a given toponym.

5.3.4 Levenshtein Distance

Of the 4'273 identified one-word candidate toponyms 544 were found to be within a Levenshtein distance of one to a SwissNames entry. Of the 2'740 compound candidate toponyms 344 fulfilled this condition. This is equivalent to about 13% in both cases, corresponding to more than every eighth candidate toponyms found by the thesis NER system. The similarity between candidate toponyms and SwissNames entries could be a manifestation of the confusion concerning the correct spelling of toponyms in Switzerland. Considering the age of some of the analysed articles and the few Internet hits scored by most of the candidate toponyms, another plausible explanation is that these are old spelling variations of placenames still used today.

5.3.5 A Random Selection of Candidate Toponyms

The 10% of candidate toponyms which were randomly selected for manual evaluation correspond to 428 one-word tokens and 274 token sequences. The results of the evaluation of these roughly 700 candidate toponyms is shown in table 5.9.

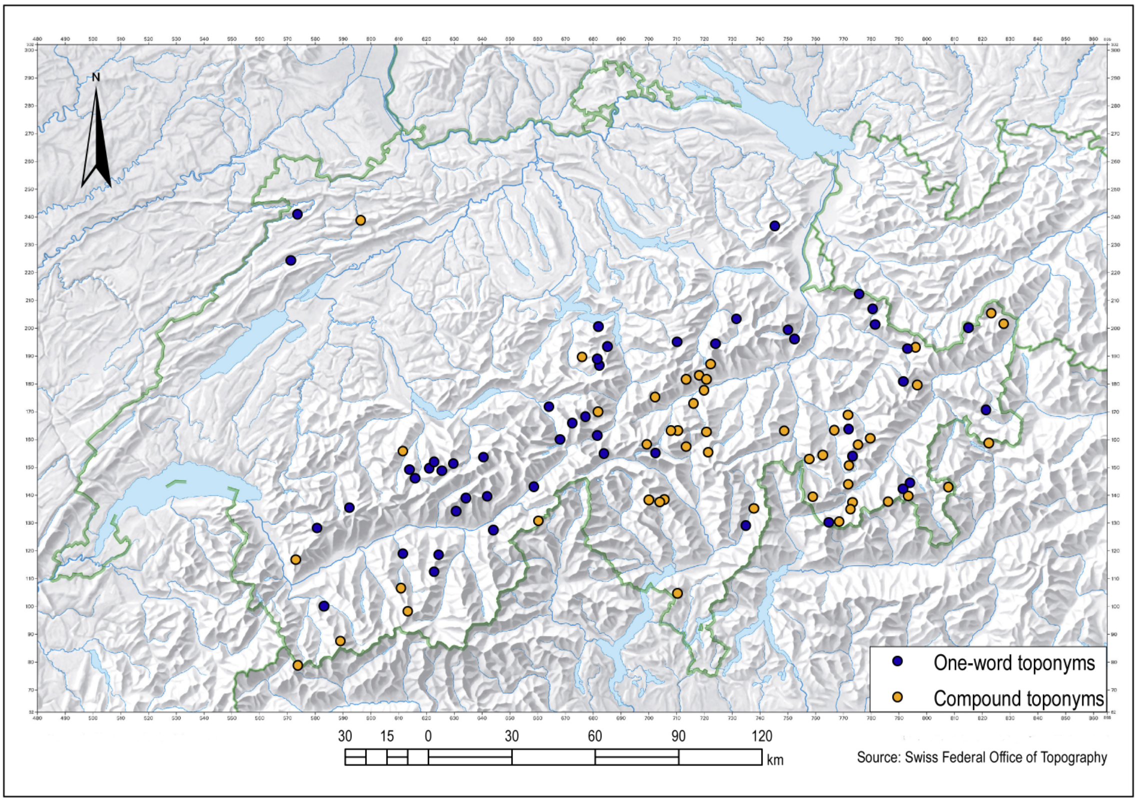
	Swiss toponyms	**Foreign toponyms**	**Non**
One-word tokens	82%	13%	5%
Token sequences	57%	36%	7%

Table 5.9: Evaluation of randomly selected candidate toponyms

There is a striking difference between the performance for one-word and compound candidate toponyms. While 82% of the one-word candidate toponyms are, in fact, unregistered

Swiss toponyms, only a little over half of the token sequences achieve this classification. This is counter-balanced by a greater number of foreign toponyms and a marginally larger amount of sequences classified as *non* (i.e. not toponyms). This can be explained by the beginnings and endings which are considered. Many of the endings, like *-hörnli*, *-furka* and *-flue*, are almost exclusive to Switzerland. The beginnings, on the other hand, include words which are found more in other countries. Examples of such toponym beginnings are the Italian *Cima* and *Rifugio* and the French *Mont*. Since both these languages are spoken in Switzerland, such toponym beginnings can also be found in SwissNames. The slightly larger percentage of sequences classified as *non* in comparison to the one-word tokens is once more due to the coarse rule which recognises all sequences of NNs and NEs containing at least one unrecognised word. This means that sequences like *Mont Blanc Pocket Book* are identified as candidate compound toponyms, which, although they may contain a toponym, can not be considered as a placename on the whole.

Toponym Resolution



Source: Swiss Federal Office of Topography

Figure 5.3: Toponym resolution for 100 Swiss toponyms not registered in SwissNames

Toponym resolution was done for 50 one-word and 50 compound toponyms of the candidate toponyms identified as Swiss. The resulting map is shown in figure 5.3.

There is an evident concentration of the toponyms to the mountainous regions of Switzerland, namely the Alps, along with three outliers in the Jura. Also discernible is a better coverage of the Grisons and the German speaking part of Switzerland, leaving few resolved toponyms in Western Switzerland and the Ticino.

This distribution was to be expected, considering the nature of articles published by an Alpine club. The concentration of the toponyms in the German speaking part of Switzerland and the Grisons could be explained by the fact that there are more German-speaking Swiss and they are more likely to frequent places closer to home. This is mere speculation, however. The map pertinently shows that unregistered vernacular toponyms exist for the Alpine and Jurassic region.

Last Mention of a Toponym

Graphs 5.4 and 5.5 show the last time one-word and compound toponyms were mentioned in the Text+Berg corpus. The x-axes display the years from 1864 to 2009 while the number of toponyms which were last mentioned in a certain year can be determined from the y-axes.

An accumulation of last mentions of one-word toponyms is discernible between the mid-1880's and late-1940's. There is a distinctive gap between the early-1970's and the beginning of the 21st century. In the last years included in the evaluation, more unregistered toponyms are mentioned again.

For compound toponyms the balance is even more clearly shifted to the first half of the SAC's history. Peaks occur at the beginning and around the turn of the 19th century while only few unregistered toponyms appear after the mid-1940's.

This concentration of unregistered toponyms in the more distant past can also be observed when considering the numbers in table 5.10. The 146 years from the beginning of SAC yearbooks to 2009 are spilt into a first group of 29 and 4 following groups of 30 years. For each group the number of toponyms last mentioned in the corresponding time span was calculated.

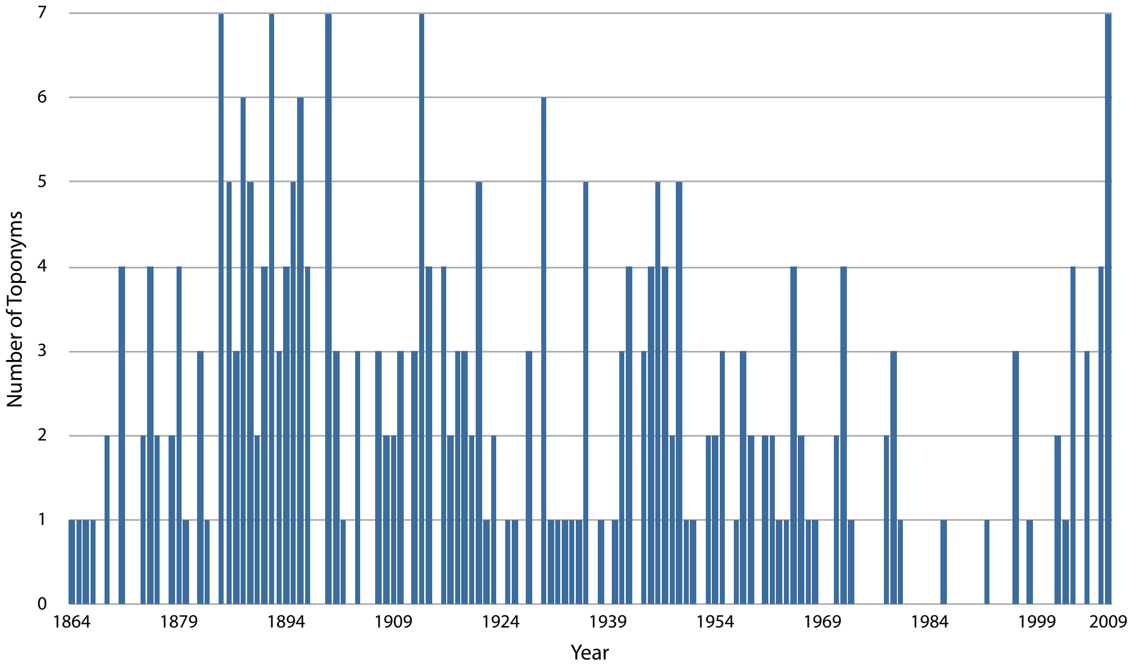


Figure 5.4: Years of last mention of randomly chosen one-word candidate toponyms classified as Swiss

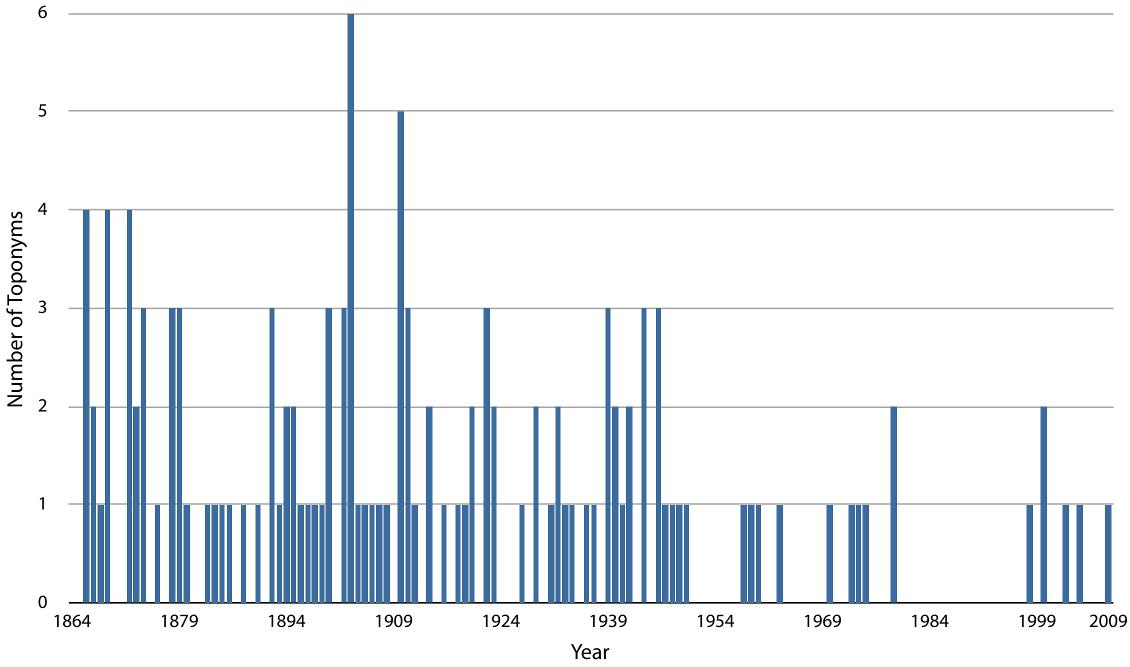


Figure 5.5: Years of last mention of randomly chosen compound candidate toponyms classified as Swiss

Span of years	One-word toponyms	Compound toponyms
1864-1892	68	37
1893-1921	79	42
1922-1950	57	33
1951-1979	40	10
1980-2009	28	6

Table 5.10: Number of toponyms last mentioned in each span of years

For both one-word and compound toponyms, there is a peak at the turn of the 19^{th} century and decreasing values for the following time spans.

Toponyms last mentioned many years ago are not necessarily out-of-date, however. As was seen in the results of the Internet search, most candidate toponyms are mentioned only a few times in the Text+Berg corpus. This could also suggest that they are fine granularity toponyms and hence not used very often.

6 Discussion

This chapter discusses the results described in the previous part of the thesis. First, the thesis NER system is evaluated and suggestions are made for its improvement. The system is also compared to the NER approach used on the Text+Berg corpus to recognise names of mountains, glaciers and cabins. In the second section, the research questions are answered and the possible implications of the results for Rega are discussed.

6.1 Used Approach

The gold standard comparison has shown the main problems encountered by this approach. The system performed best for articles like the first two, which were straightforward mountaineering accounts with few OCR mistakes, unrecognised names of persons or specific terminology. On the other hand, the articles with the worst precision and recall had very specific topics such as climbing tours and rock glaciers and were rich with scientific terms and names of persons who had established climbing routes.

The NER system's main problems can be categorised into data related and system related and are summed up by the following points:

Data related problems

- Despite the good quality of the Text+Berg corpus there are still OCR mistakes to contend with. These are a large source of interference when looking for unregistered toponyms.
- Wrong pos-tags in the corpus also adversely affect the results. Wrong tagging can be caused by words from languages other than German.
- GERTWOL does not recognise all German words, hence the list of known German words which is used as a look-up list is incomplete.

System related problems

- The system naïvely considers all NNs and NEs which are not in the look-up lists as candidate toponyms. The problem of thereby recognising many false positives which are names of persons or words in other languages, for example, is not treated.
- The system ignores any tokens recognised by GERTWOL. This eliminates many unregistered Swiss toponyms from the NER output lists, as was seen not only in the gold standard comparison but also when comparing the results to the Text+Berg output files. For compound candidate toponyms the individual parts are compared to all look-up lists. This means that one part of the compound may be recognised by SwissNames, another by GERTWOL and a third may be part of the list of foreign toponyms. As a whole, however, the compound could well be an unregistered Swiss toponym.
- By only considering NNs and NEs as potential parts of toponyms, the system cannot recognise compound toponyms which contain words of a different pos.
- The system considers any series of capitalised NNs and NEs as a candidate toponym. As a result, this crude method creates many false positives.

Little can be done about data related problems except keeping them in mind when constructing or improving an NER system and trying to account for them as best as possible. This was attempted with some success for the OCR mistakes. A lot of OCR mistakes could be avoided by rejecting any NNs and NEs which did not fit specified criteria:

One-word tokens

- The token begins with a capital letter.
- The token contains only letters, hyphens, apostrophes and slashes.
- The token ends with a letter.
- The token does not contain a lower case letter followed by an upper case letter.

Token sequences

- Tokens either begin with a capital letter or with a sequence of lower case letters followed by an apostrophe followed by a capital letter.

- Tokens contain only letters, hyphens, apostrophes, slashes and periods.
- Tokens end with a letter.
- Tokens do not contain a lower case letter proceeded by a period or followed by an upper case letter.

There are, on the other hand, many possibilities to address the system related problems and thereby improve this rule-based NER system without having to revert to a complex state-of-the-art machine learning-based or hybrid approach. Most of the problems encountered are caused by compound toponyms. The search for compound toponyms is certainly a more challenging task than NER for one-word toponyms and constitutes a weakness of this NER system. The search could be refined in a number of ways. By analysing known toponyms, common lower case words which appear in compound toponyms could be determined, such as *dal*, *la* and *an*. These words are often coupled to one of the characteristic toponym beginnings, such as in *Piz dal Teo* and *Chateau de la Rosière*, or come after a known toponym, as in *Wangen an der Aare*. These regularities can be exploited to compose new rules for the recognition of these types of compound toponyms.

The crude rule which lets the NER system consider almost any sequence of NNs or NEs as a compound toponym could also be made more selective by working with characteristic beginnings, registered toponyms and known German words. For example, only sequences should be considered, which are composed solely of unknown tokens, characteristic beginnings and/or registered Swiss toponyms.

The comparison to the GERTWOL list is necessary in this NER system since a large proportion of the NNs and NEs which are extracted are actually normal German words. However, it may be argued, that if GERTWOL recognises a one-word toponym, which is not registered in SwissNames, then this toponym cannot be very unknown. *Berninagruppe* is such an example. Although not registered, this toponym would not pose any great problem to a rescue flight coordinator at Rega. The matter is slightly different for compound toponyms. A better result may be achieved by stipulating that *all* parts of a compound toponym be contained in the same look-up list for it to be excluded from the list. This way, undue exclusion of unregistered Swiss toponyms could be avoided.

Despite these many shortcomings, the system also performs well in some aspects. The first three data postprocessing steps increase precision in four out of five gold standard articles, for example, and at the same time they don't or only marginally decrease recall in three out of these five cases. Also, looking for tokens and sequences with characteristic toponym endings and beginnings respectively, yielded 100% precision for all five articles. This result is a lucky coincidence, of course, but the evaluation of the randomly chosen one-word candidate toponyms also shows that an encouraging 82% are unregistered Swiss toponyms. Token sequences are, again, a different matter, where only 57% of the randomly chosen candidate toponyms could be classified as Swiss.

The thesis NER system has potential to be expanded beyond toponym extraction to include toponym resolution. In this work, toponym resolution was done manually, but it could be taken a step further and automised using the Text+Berg corpus. Automatic toponym resolution for toponyms mentioned in a text is not a new idea: summaries and comparisons of different approaches are presented by Leidner (2007) and Patullo (2008). Often, toponym resolution is done using extensive gazetteers but, as was the case with NER, such methods are of little use when dealing with toponyms which are not registered in a gazetteer. For the Text+Berg corpus, evidence supplied in the text itself could be used to ground such unregistered toponyms. By considering the vicinity of an unregistered or alternatively spelled toponym in an article, the four registered toponyms which appear closest to the unregistered placename in the text can be used to create a polygon, the hypothesis being that the unregistered toponym is contained within this polygon. Andogah et al. (2008, p. 2) state that "[p]laces of the same type ... or near/adjacent to each other are more likely to be mentioned in a given discourse". For example, *Piz Bernina* is likely to be mentioned in context with other mountains (e.g. *Aletschhorn*) and/or with other geographic features which are close by in the physical world (e.g. *Vadret da Morteratsch* - the Morteratsch Glacier). Since the articles in the SAC yearbooks are often mountaineering accounts which can be assumed to have a local scope, using textual context to ground toponyms seems reasonable. Also, spatial information has been used by different toponym resolution approaches to *disambiguate* toponyms (Smith and Crane, 2001; Rauch et al., 2003; Pouliquen et al., 2004; Leidner et al., 2003; Overell and Rüger, 2006) - in a similar manner, spatial information could be used to roughly ground an unknown topoynm. Once such polygons have been created, it would be informative to then compare the polygons constructed for the toponyms which were manually resolved in this thesis with their actual location. In the same way, the polygons determined for toponyms which are similar to a

SwissNames entry (Levenshtein distance of one) could be compared to the location of the corresponding SwissNames entry.

During the manual grounding of 100 toponyms, generally only one reference of each placename was analysed for clues about its location. However, the different appearances in the Text+Berg corpus may, in fact, be referring to different geographic features with the same name. In fact, this is often the case for toponyms of fine granularity (Derungs et al., 2011) and as the analysis of the NER results suggests, unregistered toponyms used in the Text+Berg corpus are frequently of fine granularity. This problem of geo-geo ambiguity must be addressed in the case of automatic toponym resolution.

6.1.1 Comparison to NER for Names of Mountains, Glaciers and Cabins

In the gold standard comparison, the thesis NER system was shown to preform better at extracting unregistered or alternatively spelled toponyms than the NER approach used to detect mentions of mountains, glaciers and cabins in the Text+Berg corpus (referred to as the corpus NER approach subsequently). There are three main differences between the thesis NER system and the corpus NER approach. These are summarised in table 6.1.

Thesis NER	**Corpus NER**
Toponyms should be Swiss	Toponyms can be foreign
Toponyms should not be registered in SwissNames	Toponyms may be registered in SwissNames
System is not tailored to look for toponyms referring to any particular kind of geographic feature	System is tailored to look for toponyms referring to mountains, glaciers or cabins

Table 6.1: Comparison of the thesis NER system to the corpus NER approach

By looking for candidate toponyms with characteristic toponym beginnings and endings in the last step of data postprocessing, the thesis NER system becomes more similar to the corpus NER approach by looking for particular geographic features such as lakes (*Lai*, *-see*) and glaciers (*Vadret*, *-gletscher*). This last step yielded relatively good precision values so it would be worth considering an adaption of the thesis NER approach which would lead more in this direction of looking for selective geographic features. This could easily be done and would, in fact, simplify the system. By investing more work into the distinction between foreign and Swiss articles and toponyms, precision could be augmented as well. Recall would be limited however, since there are many toponyms with neither typical beginnings or endings.

6.2 Research Questions

After presentation and evaluation of the results, the research questions posed at the beginning of this work can now be answered:

Question 1: How can rule-based NER techniques be used to extract unregistered vernacular and alternatively spelled toponyms from a German corpus?

The rule-based NER system developed for this thesis project extracted NNs and NEs from annotated German texts. Three look-up lists were used to exclude normal German words, foreign toponyms and registered Swiss toponyms. In three data postprocessing steps the lists of extracted candidate toponyms were refined by taking into account alternative and antiquated spelling as well as hyphens and slashes in the tokens. In a final step, tokens and token sequences with characteristic toponym endings and beginnings respectively were extracted from the lists of candidate placenames. To identify unregistered or alternatively spelled toponyms, gazetteers were of no use since the interest of this thesis lies in precisely those placenames which are not registered. NER was done without the direct use of lists and gazetteers to identify unregistered or alternatively spelled toponyms. Instead, the look-up lists were used to exclude known entities. This proved to be problematic in that several toponyms which are not registered in SwissNames but were recognised by GERTWOL were excluded, thus reducing recall. At the same time, by only excluding sequences in which *all* parts were registered in a look-up list, many false positives were extracted, thus reducing precision. Nonetheless, this approach would not have worked without the use of these lists and good results were achieved for texts with few unknown German words (100% precision and over 50% recall).

By gold standard comparison, the thesis NER system was shown to preform better at extracting unregistered or alternatively spelled toponyms than the NER approach used on the same corpus which aimed at identifying all the names of mountains, glaciers and cabins using a gazetteer and certain characteristic toponym endings and beginnings. The evaluation of the thesis NER system by comparison to five gold standard articles and this other NER approach has shown both its strengths and weaknesses, summarised in table 6.2.

Strengths	**Weaknesses**
Data postprocessing generally increases precision	Only NNs and NEs are considered
Data postprocessing generally does not decrease recall	Tokens recognised by GERTWOL are excluded, including many geo-non geo ambiguous toponyms.
Extraction of candidate toponyms with characteristic beginnings and endings yields high precision values	All NNs and NEs which are not in the look-up lists are considered candidate toponyms

Table 6.2: Overview of the main strengths and weaknesses of the thesis NER system

The best precision and recall values achieved by comparison to a gold standard article were 100% and just over 50% respectively. The precision value is perfect, while the recall value is comparable to the lower of the two values cited by Chieu and Ng (2003) (51.73%). Both the better recall given by Chieu and Ng (2003) (57.34%) and the best recall value achieved during the CoNLL-2003 shared task (63.71%) are not completely removed form this thesis NER recall value, either. A manual evaluation of 10% of the candidate toponyms revealed precision values of 82% for one-word candidate toponyms and 57% for compound candidate toponyms. This precision value for one-word toponyms is better than the highest one achieved by Chieu and Ng (2003) (77.05%) and comparable to the best precision value for German reached in the course of the CoNLL-2003 shared task (83.7%, Tjong Kim Sang and De Meulder (2003)).

This work has shown that rule-based NER for German texts, used to extract unregistered and alternatively spelled toponyms, can be accomplished with relatively good results. NER in German requires different approaches than those used for English NER (Volk and Clematide, 2001). This success of using rules for German texts could motivate increased use of rules in state-of-the-art hybrid techniques.

Question 2: What are the characteristics of unregistered vernacular and alternatively spelled toponyms extracted from the corpus?

The knowledge gained about the extracted unregistered and alternatively spelled toponyms is summarised in the following four points:

- The majority of unregistered and alternatively spelled toponyms which were extracted by the thesis NER system scored few hits during an Internet search, suggesting that the toponyms are either not known to many people, out-of-date or of fine granularity.
- Manual toponym resolution of 100 toponyms revealed a concentration in the Alpine and Jurassic regions of Switzerland, as is to be expected for a collection of articles such as the Text+Berg corpus.
- A majority of toponyms were found to be last mentioned before the 1950s. This does not necessarily mean they are out-of-date, however. If the toponyms are of fine granularity, they would not be used very often.
- The evaluation of the Levenshtein distance showed that, in about every 8^{th} case, the extracted toponyms are similar to a corresponding registered toponym (Levenshtein distance = 1).

Implications for Rescue Services like Rega

Toponyms of fine granularity which are not registered in SwissNames, such as rock climbing areas or specific hiking trails, are likely to be of use to Rega. The more precisely a location is known, the easier it is to find a person in need. The lists of one-word and compound candidate toponyms which were extracted by the thesis NER system provide such placenames. However, before these lists can be used to aid Rega in localising emergencies, they still require processing. False positives as well as Swiss toponyms which are out-of-date should be eliminated and more importantly for Rega, toponym resolution is necessary. The extensive work of Swiss toponym researchers may be of use here. Though the entire collection of toponyms is too large and too fine-grained to be of any use to Rega as a whole, it could be used as a reference to process the NER results for old toponyms and also to link the unregistered placenames to coordinates.

Implications can also be drawn from the results of the Levenshtein evaluation. The amount of toponyms found which are similar to registered toponyms encourages the implementation of a more lenient search algorithm in Rega's GIS. Such an algorithm could alleviate cases where the rescue flight coordinator cannot guess the correct spelling of a toponym since it is not always evident from the way a placename sounds or is pronounced. To preserve an initial short list of possible matches, this lenient algorithm should, however, be added merely as an additional option, not as the standard. Similar to the full-text search option which was added only recently, this could be activated when the standard search delivers no results. To improve usability and relevance, all the toponyms which were proposed in the first list should be excluded from this second list. Other rules could be added to reduce the number of hits delivered by such a second, more lenient search so the rescue flight coordinator is not overwhelmed by a long list of possible toponyms. One such rule could be that only toponyms beginning with the same letter are included. Allowances could be made for the discrepancies between Swiss German and High German by including simple translations such as *-horn* → *-hore*.

Question 3: What are the implications of the rules used to extract toponyms and the characteristics of these extracted placenames?

By combining the knowledge gained from the evaluation of the NER system and the characterisation of the results, conclusions can be drawn and suggestions are made for future NER systems developed for the extraction of unregistered and alternatively spelled toponyms. The implications are especially relevant to NER systems constructed for the Text+Berg corpus.

- OCR mistakes must be taken into consideration when developing an NER system for the Text+Berg corpus (in it's present state). An analysis of the OCR mistakes which appear in the lists of NER results may provide a better understanding of how to avoid them. Also, further antiquated spelling patterns could be identified this way and could then be included in data postprocessing.
- For the extraction of exclusively Swiss toponyms a better differentiation is needed between Swiss and non-Swiss articles and placenames. Also, if specifically Swiss toponyms are of interest, toponym beginnings and endings which are not character-

istically Swiss are problematic. In general, however, using characteristic toponym endings and beginnings to extract placenames provides results with good precision.

- NER must be improved for compound candidate toponyms. One such improvement would be to expanded the NER process to allow for tokens with other pos-tags, as toponyms are not only composed of NEs and NNs.
- Registered toponyms and identified unregistered or alternatively spelled toponyms could be used to find further placenames, for example by considering unknown tokens close to a recognised toponym in the text. This could increase recall.
- By adding a lists of names to the look-up lists, precision values could be improved. Also, the Levenshtein distance may prove useful when comparing candidate toponyms to the look-up lists. Tokens which are very similar to words recognised by GERTWOL could be OCR mistakes for instance. It must be noted, however, that by using look-up lists to exclude known entities, most geo-non geo ambiguous toponyms are also discarded.

In summary, it can be said that the thesis NER approach in its current state can be used to find especially Swiss unregistered and alternatively spelled toponyms consisting of one word which do not match an entry in a look-up list. Only compound toponyms which are composed of upper case words are extracted. Also, unregistered geo-non geo and geo-geo ambiguous toponyms are excluded from the results by comparison of the tokens to the entries in look-up lists. However, the structure of the NER system can be used as a starting ground to identify such ignored entities and other types of tokens in the Text+Berg corpus.

7 Conclusion

Unregistered vernacular toponyms and alternative spellings pose a large problem to rescue services such as Rega. The existence of unregistered toponyms can be explained by many factors: As official toponyms change over time, old versions of placenames may still be in use. The opposite can also occur, that the offical placenames are almost forgotten and a new, perhaps dialect version, of the toponym is used more often instead. Fine granularity and group-specific toponyms like the names of cliffs used for rock climbing, names of diving sites or names of starting and landing areas for paragliders also make up a part of the unregistered Swiss vernacular toponyms. Additionally, alternative spellings of toponyms are common in Switzerland due to its many languages and dialects. The aim of this thesis was to extract Swiss toponyms which are not registered in the official gazetteer SwissNames from the Text+Berg corpus. This aim was achieved with relatively good results by using a rules-based NER approach.

7.1 Achievements

The following points sum up what has been achieved with this work:

- An adaptable rule-based NER system has been constructed which:
    - classifies articles into different categories according to the tokens (toponyms in this particular case) present in the text,
    - recognises specifically pos-labeled tokens and sequences of such tokens (in this case NNs and NEs) without using look-up lists for identification (look-up lists were used to exclude known entities),
    - can exclude certain types of OCR mistakes (e.g. NNs and NEs which do not end in a letter are rejected) during identification of the tokens,
    - detects certain patterns in tokens and sequences (such as alternative spelling and case-specific endings) and replaces or deletes them,

  - detects certain characters in tokens and sequences (slashes and hyphens) and considers the token when split along these characters,

  - detects certain characters in tokens and sequences (in this case hyphens), removes these characters and patches the token parts together where appropriate and

  - extracts tokens and sequences with particular endings and beginnings respectively.

- About 7'000 different unregistered toponyms were extracted from the Text+Berg corpus with high (82% for one-word toponyms) and medium (57% for compound toponyms) precision.

- NER results were characterised by analysis of aspects (age, Internet hits, Levenshtein distance to SwissNames entries, etc.). The findings suggest that either the toponyms are of fine granularity (few Internet hits and few mentions in the corpus) or they are out-of-date (about 80% are last mentioned before the 1950s). 13% of the extracted toponyms appear to be spelling variations of registered Swiss toponyms.

- The results were evaluated by comparison to a gold standard and another NER system, which was designed to identify the names of all mountains, glaciers and cabins in the Text+Berg corpus. The gold standard comparison showed best precision and recall values of 100 and just over 50% respectively. Also, it was shown that the thesis NER system performs better at the identification of unregistered or alternatively spelled toponyms than the NER system which extracts the names of mountains, glaciers and cabins (the highest precision and recall values achieved during gold standard comparison were 100 and 25% respectively).

- Toponym resolution was done for 100 of the detected unregistered Swiss vernacular toponyms, revealing a concentration of the identified placenames in the Alpine and Jurassic regions of Switzerland, as was to be expected considering the nature of the analysed texts.

7.2 Insights

This NER method relied on look-up lists to exclude known entities, the idea being that unknown words are candidate unregistered or alternatively spelled toponyms. This approach has several drawbacks: Geo-non geo ambiguous toponyms are not identified due to the comparison with the list of German words recognised by GERTWOL. Similarly, unregistered geo-geo ambiguous toponyms which were matched to entries in SwissNames or the list of foreign toponyms are not included in the NER results. In addition, many toponyms which are not registered in SwissNames were unnecessarily removed from the list of results because parts matched entries in the other look-up lists. Working with look-up lists in this manner not only caused many unregistered and alternatively spelled toponyms to be excluded from the output, but also *included* a large amount of false positives. Especially OCR mistakes posed a large problem in this respect. During the developing phase, it was seen that to construct a sound NER system, it is important that the rules are adapted quite specifically to the type of data on which the NER system is used. For example, it is helpful to be aware of the type of OCR mistakes present in the data. Similarly, since over 100-year old articles were worked with, allowances needed to be made for antiquated orthography.

NER for compound toponyms was shown to be more difficult and less successful than for one-word toponyms. The possibility of false positives was increased because the rules used for the extraction of one-word toponyms were only slightly altered. Additionally, only compound toponyms composed of NNs and NEs were considered, excluding all compound toponyms containing tokens with other pos-tags and thus reducing recall. However, precision values for both one-word and compound toponym NER improved significantly using characteristic endings and beginnings to extract placenames: 82% was achieved for one-word toponyms and 57% for compound toponyms. In total, about 7'000 candidate unregistered or alternatively spelled toponyms with characteristic endings and beginnings were identified. It has thus been shown that NER for German can be accomplished using rules to extract unregistered and alternatively spelled toponyms and that satisfactory results can be achieved.

In general, the toponyms extracted from the Text+Berg corpus are not found very often on the Internet, nor do they usually appear more than 20 times in the Text+Berg corpus.

Additionally, most of the toponyms (about 80%) were last mentioned in the corpus before the 1950s. Combined, these facts imply that the toponyms are out-of-date or of fine granularity (Zipf's law). This latter possibility suggests that rescue services like Rega could benefit from expanding their gazetteers by including fine granularity vernacular toponyms such as are used by certain interest groups like rock climbers and divers. Additionally, the search algorithm used by the emergency service GIS should allow for slight spelling variations in toponyms. This suggestion is made because of the various Swiss languages and dialects which complicate official toponym spelling. This is highlighted by the fact that 13% of the extracted placenames were found to be within a Levenshtein distance of one from a toponym registered in SwissNames.

7.3 Future Work

Before the results of this thesis can be implemented into a rescue service's GIS, the data needs to be filtered for false positives and each toponym must be geo-referenced. This could be done with the help of the extensive research provided by Swiss toponymists. Additionally, the placenames used by interest groups who frequent the outdoors are a source of vernacular toponyms, which should be incorporated into a vernacular gazetteer. To prevent a flood of unregistered toponyms at Rega, certain high-risk groups should be identified which are involved in accidents more often than others. By including unregistered toponyms such as the names of starting and landing sites for paragliders, Rega has already started in this direction. Also, a rescue service's GIS search algorithm should allow for slight spelling variations in placenames since the various Swiss languages and dialects often complicate toponym spelling.

For the particular task of extracting toponyms which are not registered in a gazetteer, the rule-based NER system developed in this thesis can be used as starting basis for more enhanced rule-based techniques. However, the future of NER does not lie in strict rule-based approaches: machine-learning and hybrid methods are called for. It is hoped that the knowledge gained through this work may serve in the development of a state-of-the-art hybrid NER system for German. Future research should strive towards developing NER techniques for German which achieve good results comparable to those attained for English.

Bibliography

Andogah, G., G. Bouma, J. Nerbonne, and E. Koster (2008). Placename ambiguity resolution. In *LREC 2008 workshop on Methodologies and Resources for Processing Spatial Language*.

Axelrod, A. E. (2003). On building a high performance gazetteer database. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, HLT-NAACL-GEOREF '03, Stroudsburg, PA, USA, pp. 63–68. Association for Computational Linguistics.

Baeza-Yates, R. A. and B. Ribeiro-Neto (1999). *Modern Information Retrieval*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Bangerter, M. (2010). Toponym analysis. Goldstandard. http://www.cl.uzh.ch/people/alumni/maya/ner.html. (Date accessed: 14.07.2011).

BBC (2007). PCSOs 'did not watch boy drown'. BBC News website. http://news.bbc.co.uk/2/hi/7007081.stm. (Date accessed: 17.06.2011).

Bear, J., D. Israel, J. Petit, and D. Martin (1998). Using information extraction to improve document retrieval. In *Proceedings of the Sixth Text Retrieval Conference (TREC-6*, pp. 367–377.

Bendersky, E. (2003). Code: levenshtein.pl. http://eli.thegreenplace.net/programs-and-code/. (Date accessed: 03.05.2011).

Berger, A. (1996). A brief maxent tutorial. http://www.cs.cmu.edu/afs/cs/user/aberger/www/html/tutorial/tutorial.html. Carnegie Mellon University, School of Computer Science. (Date accessed: 03.06.2011).

Bernstein, A., E. Kaufmann, C. Bürki, and M. Klein (2005). How similar is it? towards personalized similarity measures in ontologies. In *7. Internationale Tagung Wirtschaftsinformatik*, pp. 1347–1366.

Bickel, H., M. H. Graf, and E. Nyffenegger (heads of project) (2011). ortsnamen.ch. http://www.ortsnamen.ch/. (Date accessed: 18.07.2011).

Bikel, D., S. Miller, R. Schwartz, and R. Weischedel (1997). Nymble: a high-performance learning name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP)*, Washington, D.C., pp. 194–201.

Bikel, D., R. Schwartz, and R. Weischedel (1999). An algorithm that learns what's in a name. *Machine Learning 34*(1–3), 211–231.

Borthwick, A. E. (1999). *A maximum entropy approach to named entity recognition*. Ph. D. thesis, New York University, New York, NY, USA. AAI9945252.

Brants, T. and Google Inc. (2004). Natural language processing in information retrieval. In *Proceedings of the 14th Meeting of Computational Linguistics in the Netherlands*, pp. 1–13.

Brunner, T. (2008). Geographic Information Retrieval: Identifikation der geographischen Lage von Zeitungsartikeln. Master's thesis, Universität Zürich.

Bubenhofer, N., M. Volk, A. Althaus, M. Jitca, M. Bangerter, and R. Sennrich (Eds.) (2011). *Text+Berg-Korpus (Release 145)*. Institut für Computerlinguistik, Universität Zürich. Digitale Edition des Jahrbuch des SAC 1864-1923 und Die Alpen 1925-2009.

Bubenhofer, N., M. Volk (heads of project), and Text+Berg Team (s.a.). Text+Berg digital. Projekt zur korpuslinguistischen Erschliessung alpinistischer Literatur. http://www.textberg.ch. (Date accessed: 22.06.2011).

Bundesverfassungsgericht (BVerfG) (2011). Entscheidungen. http://www.bverfg.de/entscheidungen/rs19980714_1bvr164097.html. (Date accessed: 03.08.2011).

Burenhult, N. and S. C. Levinson (2008). Language and landscape: a cross-linguistic perspective. *Language Sciences 30*(2/3), 135–150.

Buscaldi, D. and P. Rosso (2008). Map-based vs. knowledge-based toponym disambiguation. In *Proceeding of the 2nd international workshop on Geographic information retrieval*, GIR '08, New York, NY, USA, pp. 19–22. ACM.

Cadastral Surveying in Switzerland (2010a). Amtliche Vermessung. http://www.

cadastre.ch/internet/cadastre/de/home/topics/geonames/av.html. (Date accessed: 03.08.2011).

Cadastral Surveying in Switzerland (2010b). Dokumente zum Thema. http://www.cadastre.ch/internet/cadastre/de/home/topics/geonames/doku.html. (Date accessed: 03.08.2011).

Carstensen, K.-U., C. Ebert, C. Ebert, S. Jekat, R. Klabunde, and H. Langer (Eds.) (2010). *Computerlinguistik und Sprachtechnologie: Eine Einführung* (3 ed.). Heidelberg: Spektrum Akademischer Verlag.

Chieu, H. L. and H. T. Ng (2003). Named entity recognition with a maximum entropy approach. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, Stroudsburg, PA, USA, pp. 160–163. Association for Computational Linguistics.

Christo (2011). Wende im Namenstreit akzeptiert. Thurgauer Zeitung. http://www.thurgauerzeitung.ch/ostschweiz/thurgau/kantonthurgau/tz-tg/art123841,1689273. (Date accessed: 15.06.2011).

Columbia Electronic Encyclopedia (2003). 6th ed. Boston: Houghton Mifflin. Accessed through http://www.answers.com (Date accessed: 09.08.2011).

Davies, C., I. Holt, J. Green, and L. Diamond (2009). User needs and implications for modelling vague named places. *Spatial Cognition & Computation: An Interdisciplinary Journal 9*(3), 174–194.

Davies, C., C. Li, and J. Albrecht (2010). *Human understanding of space*, Chapter 2, pp. 19–36. Wiley Online Library.

Derungs, C., R. Purves, and B. Waldvogel (2011). Disambiguation of fine granularity toponyms using morphometric information. http://meridian.aag.org/callforpapers/program/AbstractDetail.cfm?AbstractID=37592. (Date accessed: 10.08.2011).

Didakowski, J., A. Geyken, and T. Hanneforth (2007). Eigennamenerkennung zwischen morphologischer Analyse und Part-of-Speech Tagging: ein automatentheoriebasierter Ansatz. *Zeitschrift für Sprachwissenschaft 2*, 157–186.

Dittli, B. (2007). *Zuger Ortsnamen: Lexikon der Siedlungs-, Flur- und Gewässernamen im Kanton Zug. Lokalisierung, Deutung, Geschichte. 5 Bde. und Kartenset.* Balmer Verlag.

Doran, C. (2009). SpatialML: Annotation scheme for marking spatial expressions in natural language. Technical Report, The MITRE Corporation.

Doyle, A. C. (1894). *The Memoirs of Sherlock Holmes*, Chapter The Adventure of the Reigate Squire, pp. 1–297. George Newnes. http://etc.usf.edu/lit2go/contents/1100/1168/1168.pdf (Date accessed: 03.06.2011).

Dutta, K., N. Prakash, and S. Kaushik (2005). Hybrid framework for information extraction for geographical terms in Hindi language texts. In *Proceedings of NLP-KE'05*, pp. 577–581.

Egenhofer, M. J. and D. M. Mark (1995). Naive geography. In A. U. Frank and W. Kuhn (Eds.), *Spatial Information Theory: A Theoretical Basis for GIS, International Conference COSIT*, Volume 988 of *Lecture Notes in Computer Science*, Semmering, Austria, pp. 1–15. Springer.

Evans, A. and T. Waters (2007). Mapping vernacular geography: Web-based gis tools for capturing "fuzzy" or "vague" entities. *International Journal of Technology, Policy and Management* *7*(2), 134–150.

Faruqui, M. and S. Padó (2010). Training and evaluating a German named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.

Federal Authorities of the Swiss Confederation (2009). Art. 7 Geografische Namen. http://www.admin.ch/ch/d/sr/510_62/a7.html. (Date accessed: 17.06.2011).

Frey, R. (2007). Stellungnahme zur Verordnung über geographische Namen (GeoNV). http://www.lokalnamen.ch/bilder/20070223_rega.pdf. (Date accessed: 17.06.2011).

Galton, A. and J. Hood (2005). Anchoring: A new approach to handling indeterminate location in gis. In A. Cohn and D. Mark (Eds.), *Spatial Information Theory*, Volume 3693 of *Lecture Notes in Computer Science*, pp. 1–13. Springer Berlin / Heidelberg.

Gan, Q., J. Attenberg, A. Markowetz, and T. Suel (2008). Analysis of geographic queries in a search engine log. In *Proceedings of the first international workshop on Location and the web*, LOCWEB '08, New York, NY, USA, pp. 49–56. ACM.

GATE (s.a.). GATE information extraction. http://gate.ac.uk/ie/. (Date accessed: 15.03.2011).

Goodchild, M. F. (1999). The future of the gazetteer. Presented at the digital gazetteer information exchange workshop, Oct 13-14. http://www.alexandria.ucsb.edu/~lhill/dgie/DGIE_website/Perspectives/Goodchild.htm. Transcribed and edited from audiotape. (Date accessed: 08.08.2011).

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal 69*(4), 211–221.

Grishman, R. and B. Sundheim (1996). Message understanding conference-6: a brief history. In *Proceedings of the 16th conference on Computational linguistics - Volume 1*, COLING '96, Stroudsburg, PA, USA, pp. 466–471. Association for Computational Linguistics.

Hill, L. L. (2000). Core elements of digital gazetteers: Placenames, categories, and footprints. In *J. Borbinha & T. Baker (Eds.), Research and Advanced Technology for Digital Libraries : Proceedings of the 4th European Conference, ECDL 2000*, pp. 280–290. Springer.

Hill, L. L. (2006). *Georeferencing : the geographic associations of information.* Digital libraries and electronic publishing. Cambridge, MA: MIT Press.

Hollenstein, L. (2008). Capturing vernacular geography from georeferenced tags. Master's thesis, University of Zürich, Institute of Geography.

Hollenstein, L. and R. S. Purves (2010). Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Inforation Sciene 1*, 21–48.

Jones, C. B., H. Alani, and D. Tudhope (2001). Geographical information retrieval with ontologies of place. In *Spatial Information Theory, LNCS 2205*, pp. 322–335. Springer.

Jones, C. B. and R. S. Purves (2008). Geographical information retrieval. *International Journal of Geographical Information Science 22*, 219–228.

Kent, A., M. M. Berry, F. U. Luehrs, and J. W. Perry (1955). Machine literature searching viii. operational criteria for designing information retrieval systems. *American Documentation 6*(2), 93–101.

Kilcher, S. and N. Soutar (s.a.). Faszination Gebirge. Gebirge der Welt. http://www.gebirge.mykilcher.ch/welt/welt.html. (Date accessed: 04.07.2011).

Knöpfel, M. (2009). Wenn Ortsnamen Verwirrung stiften. Thurgauer Zeitung. http://www.thurgauerzeitung.ch/thurgau-alt/ostschweiz/thurgau/kantonthurgau/tz-tg/art131331,1706795. (Date accessed: 15.06.2011).

Kozareva, Z. (2006). Bootstrapping named entity recognition with automatically generated gazetteer lists. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, EACL '06, Stroudsburg, PA, USA, pp. 15–21. Association for Computational Linguistics.

Kunz, R. (2008). Evaluation of spatial relevance in geographic information retrieval. Master's thesis, University of Zurich.

Lang, N. C. (2010). Toponyme in der Einsatzdisposition. Master's thesis, University of Zürich.

Larson, R. R. (1996). Geographic information retrieval and spatial browsing. In L. Smith and M. Gluck (Eds.), *GIS and Libraries: Patrons, Maps and Spatial Information*, pp. 81–124. University of Illinois.

Leidner, J. L. (2007). *Toponym Resolution in Text.* Ph. D. thesis, University of Edinburgh.

Leidner, J. L., G. Sinclair, and B. Webber (2003). Grounding spatial named entities for information extraction and question answering. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, HLT-NAACL-GEOREF '03, Stroudsburg, PA, USA, pp. 31–38. Association for Computational Linguistics.

Lieberman, M., H. Samet, and J. Sankaranarayanan (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th International Conference on Data Engineering (ICDE)*, pp. 201–212.

Lin, X. and Y. Ban (2008). On the framework and key techniques of modern GIR systems.

In *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Volume 37, pp. 269–276. ISPRS Congress Beijing.

Lingsoft Language Solutions (s.a.). A short introduction to GERTWOL. Overview. http://www2.lingsoft.fi/doc/gertwol/intro/overview.html. (Date accessed: 22.06.2011).

Linguateca (2006). HAREM - evaluation contest for named entity recognizers in Portuguese. http://www.linguateca.pt/HAREM/. (Date accessed: 03.08.2011).

Linguateca (2010). HAREM: NER for Portuguese. http://www.linguateca.pt/HAREM/. (Date accessed: 03.08.2011).

Liu, X., S. Zhang, F. Wei, and M. Zhou (2011). Recognizing named entities in tweets. To appear in ACL 2011.

MacLean, A. (1976). *The Golden Gate.* Greenwich Connecticut: Fawcett Publications.

Mani, I., J. Hitzeman, J. Richer, D. Harris, R. Quimby, and B. Wellner (2008). Spatialml: Annotation scheme, corpora, and tools. In *6th International Conference on Language Resources and Evaluation (LREC 2008.*

Manning, C. D., P. Raghavan, and H. Schütze (2008). *Introduction to Information Retrieval.* New York, NY, USA: Cambridge University Press.

Martin, H.-J. (2011). Geschichtlicher Abriß der Rechtschreibung. http://www.schriftdeutsch.de/orth-his.htm. (Date accessed: 13.08.2011).

McDonald, D. D. (1996). *Internal and external evidence in the identification and semantic categorization of proper names*, pp. 21–39. Cambridge, MA, USA: MIT Press.

McDonald, K. R. and L. Di (2003). Serving NASA EOS data to the GIS community through the OGC-standard based NWGISS system. In *Asia GIS Conference. Session 14. GIS Method and Technology (IV).* http://www0.hku.hk/dupad/asiagis/. (Date accessed: 10.08.2011).

Mikheev, A., M. Moens, and C. Grover (1999). Named entity recognition without gazetteers. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, EACL '99, Stroudsburg, PA, USA, pp. 1–8. Association for Computational Linguistics.

Miller, S., M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, R. Weischedel, and The Annotation Group (1998). Algorithms that learn to extract information BBN: Description of the sift system as used for MUC-7. In *In Proceedings of MUC-7*.

Montello, D. R., M. F. Goodchild, J. Gottsegen, and P. Fohl (2003). Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation 3*(2-3), 185–204.

MUC-6 Appendix (1995). Appendix C: Named entity task definition (v2.1). In *MUC6 '95: Proceedings of the 6th conference on message understanding*, Stroudsburg, PA, USA, pp. 317–332. Association for Computational Linguistics.

Müller, C. and M. Strube (2006). Multi-level annotation of linguistic data with MMAX2. In S. Braun, K. Kohn, and J. Mukherjee (Eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pp. 197–214. Frankfurt a.M., Germany: Peter Lang.

Naaman, M., Y. J. Song, A. Paepcke, and H. Garcia-Molina (2006). Assigning textual names to sets of geographic coordinates. *Computers, Environment, and Urban Systems 30*(4), 418–435.

Nadeau, D. and S. Sekine (2007). A survey of named entity recognition and classification. *Linguisticae Investigationes 30*(1), 3–26. Publisher: John Benjamins Publishing Company.

Overell, S. E. and S. Rüger (2006). Indentifying and grounding descriptions of places. In *SIGIR Workshop on Geographic Information Retrieval*, pp. 14–16.

Oxford University Press (s.a.). Oxford dictionaries. http://oxforddictionaries.com/definition/token?region=us&rskey=HNSPtj&result=14. (Date accessed: 22.06.2011).

Palmer, D. D. and D. S. Day (1997). A statistical profile of the named entity task. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP)*, Washington, D.C., pp. 190–193.

Pasley, R., P. Clough, R. S. Purves, and F. A. Twaroch (2008). Mapping geographic coverage of the web. In *Proceedings of the 16th ACM SIGSPATIAL international con-*

ference on Advances in geographic information systems, GIS '08, New York, NY, USA, pp. 19:1–19:9. ACM.

Patullo, I. (2008). Improved document geocoding for geo-complex text. Master's thesis, School of Computer Science and Software Engineering, The University of Western Australia.

Pettersson, M., D. Randall, and B. Helgeson (2004). Ambiguities, awareness and economy: a study of emergency service work. *Computer Supported Cooperative Work (CSCW) 13*(2), 125–154.

Philip Smart, Florian Twaroch, C. J. (2009). TRIPOD. Deliverable 6.5: Final toponym ontology prototype. http://tripod.shef.ac.uk/outcomes/public_deliverables/Tripod_D6.5.pdf. (Date accessed: 14.07.2011).

Piotrowski, M., S. Läubli, and M. Volk (2010). Towards mapping of alpine route descriptions. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10, New York, NY, USA, pp. 2:1–2:2. ACM.

Piskorski, J., D. Saarbrucken, and G. Neumann (1999). An intelligent text extraction and navigation system. In *Proceedings of the RIAO-2000.*

Pouliquen, B., R. Steinberger, C. Ignat, and T. De Groeve (2004). Geographical information recognition and visualization in texts written in various languages. In *Proceedings of the 2004 ACM symposium on Applied computing*, SAC '04, New York, NY, USA, pp. 1051–1058. ACM.

Purves, R. and C. Jones (2006). Geographic information retrieval (GIR). *Computers, Environment and Urban Systems 30*(4), 375–377.

Rauch, E., M. Bukatin, and K. Baker (2003). A confidence-based framework for disambiguating geographic terms. In *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, HLT-NAACL-GEOREF '03, Stroudsburg, PA, USA, pp. 50–54. Association for Computational Linguistics.

Rössler, M. (2004). Corpus-based learning of lexical resources for German named entity recognition. In *Proceedings of LREC 2004*, Lisboa, Portugal.

Rössler, M. (2007). *Korpus-adaptive Eigennamenerkennung*. Ph. D. thesis, Universität Duisburg-Essen.

Santos, D., N. Seco, N. Cardoso, and R. Vilela (2006). Harem: An advanced ner evaluation contest for portuguese. In *Odjik and Daniel Tapias (eds.), Proceedings of LREC 2006 (LREC'2006) (Genoa)*, pp. 22–28.

Sarawagi, S. (2008). Information extraction. *Found. Trends databases 1*, 261–377.

Schoch, M. (2009). Flurnamen sorgen für rote Köpfe. Thurgauer Zeitung. http://www.thurgauerzeitung.ch/ostschweiz/thurgau/kantonthurgau/tz-tg/art123841,1359563. (Date accessed: 15.06.2011).

Schoch, M. (2010). "Habe das Thema unterschätzt". Thurgauer Zeitung. http://www.thurgauerzeitung.ch/ostschweiz/thurgau/kantonthurgau/tz-tg/art123841,1552014. (Date accessed: 15.06.2011).

Schorta, A. (1999). *Wie der Berg zu seinem Namen kam. Kleines Rätisches Namenbuch mit zweieinhalbtausend geographischen Namen Graubündens* (3. Auflage ed.). Chur: Terra Grischuna Verlag.

Schweizer Alpen-Club SAC (s.a.). Archiv "SAC Jahrbücher". http://www.sac-cas.ch/Zeitschrift-Die-Alpen.1504.0.html. (Date accessed: 04.07.2011).

Sekine, S. and H. Isahara (1999). IREX project overview. http://as.nyu.edu/object/satoshisekine.html. The IREX Workshop. Tokyo, Japan. (Date accessed: 20.08.2011).

Shannon, C. (1984). A mathematical theory of communication. *Bell System Technical Journal 27*, 379–423 and 623–656.

Siebenhaar, B. and A. Wyler (1997). *Dialekt und Hochsprache in der deutschsprachigen Schweiz*, Volume 5., vollständig überarbeitete Auflage. Zürich: Edition "Pro Helvetia". http://www.uni-leipzig.de/~siebenh/pdf/Siebenhaar_Wyler_97.pdf (Date accessed: 13.08.2011).

Singhal, A. (2001). Modern information retrieval: a brief overview. *Bulletin of the IEEE computer society technical committee on data engineering 24*, 2001.

Smith, D. A. and G. Crane (2001). Disambiguating geographic names in a historical digital library. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '01, London, UK, pp. 127–136. Springer-Verlag.

Srihari, R., C. Niu, and W. Li (2000). A hybrid approach for named entity and sub-type tagging. In *Proceedings of the sixth conference on Applied natural language processing*, ANLC '00, Stroudsburg, PA, USA, pp. 247–254. Association for Computational Linguistics.

Stevenson, M. and R. Gaizauskas (2000). Using corpus-derived name lists for named entity recognition. In *Proceedings of the sixth conference on Applied natural language processing*, ANLC '00, Stroudsburg, PA, USA, pp. 290–295. Association for Computational Linguistics.

Swiss Federal Office of Topography (2002). SwissNames. Die Namendatenbank der Schweiz. (Detailierte Produktinfo). http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/toponymy.html. (Date accessed: 23.06.2011).

Swiss Federal Office of Topography (2005). SwissNames. Die Geodaten der Schweiz des Budesamtes für Landestopografie für den professionellen Einsatz. (Produktflyer). http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/toponymy.html. (Date accessed: 23.06.2011).

Swiss Federal Office of Topography (2008). Swissnames 25.

Swiss Federal Office of Topography (2009). Geographic names. http://www.swisstopo.admin.ch/internet/swisstopo/en/home/topics/toponymie.html. (Date accessed: 23.06.2011).

Swiss Federal Office of Topography (2010). Kartenviewer. Geoportal Bund. http://map.geo.admin.ch/. (Date accessed: 23.06.2011).

Swiss Federal Office of Topography (s.a.). map.geo.admin.ch. http://map.geo.admin.ch/. (Date accessed: 23.06.2011).

Tague-Sutcliffe, J. (1992). An introduction to informetrics. *Informatioin Processing & Management 28*(1), 1–3.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. In *Proceedings of the 6th conference on Natural language learning*, Volume 20 of *COLING-02*, Stroudsburg, PA, USA, pp. 1–4. Association for Computational Linguistics.

Tjong Kim Sang, E. F. and F. De Meulder (2003). Introduction to the conll-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, Stroudsburg, PA, USA, pp. 142–147. Association for Computational Linguistics.

Twaroch, F. A., R. S. Purves, and C. B. Jones (2009). Stability of qualitative spatial relations between vernacular regions mined from web data. In *Workshop on Geographic Information on the Internet*, Toulouse , France.

Vallez, M. and R. Pedraza-Jimenez (2007). Natural language processing in textual information retrieval and related topics. *Hipertext.net 5*. http://www.hipertext.net (Date accessed: 03.06.2011).

Vestavik, Ø. (2004). Geographic information retrieval: An overview. Norwegian University of Technology and Science, Deptarment of Computer and Information Science. http://www.idi.ntnu.no/~oyvindve/article.pdf (Date accessed: 04.06.2011).

Volk, M. (2009). How many mountains are there in Switzerland? Explorations of the swisstopo name list. In S. Clematide, M. Klenner, and M. Volk (Eds.), *Searching Answers: Festschrift in Honour of Michael Hess on the Occasion of His 60th Birthday*, pp. 127–140. Münster, Germany: Monsenstein und Vannerdat.

Volk, M., N. Bubenhofer, A. Althaus, and M. Bangerter (2009). Classifying named entity in an alpine heritage corpus. *Künstliche Intelligenz (KI) 4*, 40–43.

Volk, M. and S. Clematide (2001). Learn - filter - apply - forget. Mixed approaches to named entity recognition. In *Proceedings of the 6th International Workshop on Applications of Natural Language to Information Systems*, pp. 153–163. GI.

Welt-Blick.de (s.a.). Die Staaten der Erde. www.welt-blick.de. (Date accessed: 04.07.2011).

Werlen, I. (2008). Die Grundwörter der Oberwalliser Gipfelnamen. In B. Huber, M. Volkart, and P. Widmer (Eds.), *Chomolangma, Demawend und Kasbek. Festschrift* *für Roland Bielmeier zu seinem 65. Geburtstag. Band II: Demawend und Kasbek*, pp. 577–614. Halle (Saale): International Institute for Tibetan and Buddhist Studies GmbH.

Widmer, C. (2009). Mit Dialekt fürs Vaterland. Thurgauer Zeitung. http://www.thurgauerzeitung.ch/thurgau-alt/ostschweiz/thurgau/kantonthurgau/tz-tg/art131331,1706794. (Date accessed: 15.06.2011).

Wikipedia (2009). Liste der Erstbesteigungen. http://de.wikipedia.org/wiki/Liste_der_Erstbesteigungen. (Date accessed: 04.07.2011).

Wikipedia (2011a). Höchster Berg. http://de.wikipedia.org/wiki/Höchster_Berg. (Date accessed: 04.07.2011).

Wikipedia (2011b). Liste der grössten Gebirge der Erde. http://de.wikipedia.org/wiki/Liste_der_grö\T1\ssten_Gebirge_der_Erde. (Date accessed: 04.07.2011).

Wikipedia (2011c). Liste der Viertausender in den Alpen. http://de.wikipedia.org/wiki/Liste_der_Viertausender_in_den_Alpen. (Date accessed: 04.07.2011).

Wilke, G. (2009). Approximate geometric reasoning with extended geographic objects. In *ISPRS-COST Workshop on quality, scale and analysis aspects of city models*, Lund, Sweden. Extended Abstract.

Winkler, E. (s.a.). Die höchsten Berge der Welt. http://www.eckart-winkler.de/reise/specials/berge.htm. (Date accessed: 04.07.2011).

Zubizarreta, Á., P. de la Fuente, J. Cantera, M. Arias, J. Cabrero, G. García, C. Llamas, and J. Vegas (2009). Extracting geographic context from the Web: Georeferencing in MyMoSe. In M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy (Eds.), *Advances in Information Retrieval*, Volume 5478 of *Lecture Notes in Computer Science*, pp. 554–561. Springer Berlin / Heidelberg.

A Lists and Code

Some of the lists used in this thesis are included in this appendix. Since most of the lists and programmes are long documents, the majority has been saved on a CD which is part of every hard copy of this thesis.

A.1 Lists

The following lists are included in this appendix:

- Characteristic toponym endings (subsection A.1.1)
- Characteristic toponym beginnings (subsection A.1.2)
- An excerpt of the extracted toponyms with characteristic endings (subsection A.1.3)
- An excerpt of the extracted toponyms with characteristic beginnings (subsection A.1.4)
- An excerpt of the list of toponyms and toponym parts excluded from the SwissNames look-up list because they were classified as geo-non geo ambiguous (subsection A.1.5)

These lists are available on the CD accompanying each hard copy of this thesis:

- List of words recognised by GERTWOL (*gertwol.xls*)
- List of foreign toponyms (*foreign_toponyms.xls*)
- List of entries in SwissNames and toponyms parts thereof (*swissnames.xls*)
- Complete list of toponyms and toponym parts excluded from the SwissNames look-up list because they were classified as geo-non geo ambiguous (*excluded_toponyms.xls*)
- Characteristic toponym beginnings and endings (*beginnings_endings.xls*)
- Complete list of NER results (one-word candidate toponyms) after the last data postprocessing step (*one-word_candidate_toponyms.xls*)

- Complete list of NER results (compound candidate toponyms) after the last data postprocessing step (*compound_candidate_toponyms.xls*)

A.1.1 Characteristic Toponym Endings

-gletscher
-glätscher
-stock
-stöckli
-stöcke
-firn
-horn
-hörner
-hörnli
-hütte
-hütten
-nordwand
-südwand
-westwand
-ostwand
-tobel
-pass
-passhöhe
-paß
-paßhöhe
-massiv
-grat
-grätli
-lücke
-joch
-furka
-fluh
-flue
-first
-matt
-matten
-stein
-gruppe
-tal
-tälchen
-thal
-thälchen
-berge
-berg
-talstrasse
-see
-seen
-seelein
-bach
-brücke
-kette
-couloir
-alp
-schlucht
-gipfel
-sattel
-spitz
-spitze
-spitzen
-gletschers
-stocks
-stöcklis
-firns
-horns
-hörnlis
-tobels
-massivs
-grätlis
-jochs
-firsts
-tals
-tälchens
-thals
-thälchens
-bergs
-sees
-seeleins
-bachs
-couloirs
-gipfels
-sattels
-gletschern
-stöcken
-hörnern
-bergen
-steins
-stockes
-firnes
-hornes
-passes
-paßes
-massives
-grates
-joches
-tales
-thales
-berges
-baches
-spitzes

Table A.1: Characteristic toponym endings

A.1.2 Characteristic Toponym Beginnings

Ober
Val
Unter
Piz
Alpe
Bois
Alp
Vorder
Hinter
Plan
Sur
Under
Pra
Forêt
Col
Ruine
God
Combe
Gross
Fuorcla
Pizzo
Schloss
Chli
Valle
Mont
Grand
Cima
Hinder
Ruisseau

Plaun
Pian
Côte
Passo
Bocchetta
Petit
Bosco
Mittler
Oberi
Pointe
Sous
Château
Glacier
Prés
Piano
Costa
Torrent
Usser
Bosch
Vers
Monti
Monte
Underi
Creux
Fin
Undere
Chalet
Uaul
Corte

Lac
Vordere
Crêt
Tête
Inner
Crap
St.
Obers
Gros
Lago
Lai
Derrière
Munt
Pass
Grande
Schulhaus
Sasso
Ova
Motto
Pont
Motta
Refuge
Vadret
Ufem
Hindere
Bim
Untere
Pro
Petite

Unders
Vorderi
Moulin
Clos
Mot
Cabane
Sex
Halte
Hinderi
Grosse
Pâturage
Punta
Aua
Las
Capanna
Laviner
Ils
Grands
Marais
Fil
Dent
Rià
Hintere
Prau
Punt
Métairie
Ruina
San
Stavel

Foppa
Riale
Montagne
Pianca
Poncione
Cascina
Pas
Plang
Roche
Sass
Lej
Obri
Ponte
Neu
Kloster
Grange
Ganne
Burgstelle
Blais
Crest
Crête
Pâquier
Lui
Cresta
Chamanna
Glatscher
Vorders
Unterer
Rochers

Grossi
Croix
Prise
Got
Plaunca
Six
Rote
Mayens
Côtes
Muot
Hinders
Höch
Undri
Grandes
Aiguille
Ual
Bad
Ava
Moille
Flugfeld
Cuolm
Hoch
Essert
Wiss
Zum
Roc
Klein
Grosses
Sogn

Table A.2: Characteristic toponym beginnings

Lang	Petits	Mött	Suot	Droit	Revers
Planche	Underem	Nider	Ronco	Fontana	Lengi
Entre	Roti	Bleis	Esserts	Vanil	Burg
Bisse	Mittleri	Blaisch	Dos	Strandbad	Ussers
Canal	Lagh	Monts	Hinterer	Mittlist	Chlys
Sut	Crasta	Jeur	Mittel	Run	Kleine
Châble	Etang	Petites	Gane	Belle	Bella
Innere	Planches	Chlei	Envers	Inder	Pascul
Ban	Ferme	Plauncas	Grosser	Zen	Grasso
Zer	Schwarz	Corn	Case	Gana	Remointse
Sot	Lag	Madonna	Ehemalige	Castello	Cappella
Gianda	Sotto	Gorges	Schwarze	Ciernes	Ovel
Corona	Rifugio	Cuolms	Vallon	Chant	Muotta
Cras	Vieille	Usseri	Tegia	Üsseri	Chlein
Faura	Devant	Valletta	Vallone	Becca	Kleines
Comba	Nant	Cul	Gîte	Port	Plam
Acla	Ghiacciaio	Rière	Sunnig	Aiguilles	Rocher
Ganna	Murtel	Oberes	Ussere	Rots	Fond
Löita	Platta	Ehemaliges	Muletg	Cierne	Mayen
Bir	Mittlere	Unteri	Contour	Funtana	Croce
Alti	Bec	Vadrec	Dosso	Plans	Gualdo
Unners	Pointes	Ovi	Maison	Obem	Communs
Inneri	Casa	Schattig	Plain	Üsser	Unteres
Spi	Bot	Rein	Prada	Obre	Vorderer
Inners	Cerneux	Dzeu	Stabbio	Foura	Igls
Vallun	Laghetto	Tecc	Cugnolo	Pianche	

Table A.3: Characteristic toponym beginnings (continued)

A.1.3 Extracted Toponyms with Characteristic Endings

Columbépass
Rheinwaldtal
Gellihorn
Draggaberg
Vermuntthal
Chelenalp
Kirstein
Zessettagletscher
Zberg
Lonzabrücke
Zippraspitz
Verzascagruppe
Sunnigstöcke
Oberaarjoehhütte
Medelsergletscher
Ledifluh
Gletscherseelein
Gufernstock
Gornernalp
Barrhörner
Teselalp
Gisigpass
Hohmadgletscher
Fassajoch
Krinnefirn
Solhorn
Golzerberg
Ciprianspitz
Mominggletscher
Göschenenthal
Aelplistock
Ennsthal
Ferpècletal
Rienthals
Fridolinshütte
Falzaregopass
Varaitatal
Valserhorn
Pillonpasses
Gonerlital
Twingischlucht
Stammtal
Fermeltal
Scheibepaß
Puntaiglas-Gletscher
Masone-Gletscher
Zindelspitze
Jjolligletscher
Dalathal
Medelser-Hütte
Cacciabellapaß
Rungspitzen
Centralgruppe
Mulixthal
Gurnigelkette
Rötihorn
Emmenthales
Steilertälchen
Gneissgrat
Zweigthal
Betempshütte
Kärpfgipfel
Gornerlibach
Juferalp
Frisalthal
Ebihorn
Gonschirolatobel
Valdrausgletscher
Biferten-Nordwand
Brandthal
Ducanpaß
Gliemslücke
Rojental
Agassizcouloir
Valsertal
Tilisunahütte
Saleinaz-Clubhütte
Niederental
Lemansee
Paßsattel
Grialetschpaß
Mühlebachtobel
Tösstal
Nufenenpass
Crozlinagletscher
Casatihütte
Etzlithal

Table A.4: An excerpt of the extracted toponyms with characteristic endings

A.1.4 Extracted Toponyms with Characteristic Beginnings

Burg Rinkenberg
Petit Arcellin
Punta Piodä
Alp Schweiben
Alp Monte Urmera
Monte Belìo
Piz Brascheng
Monte Agner
Pointe Ceresole
Alp Ulix
Piz Giendusas
Piz Lavinér
Mont Suc
Alp Galkerne
San Lucano
Grand Sassière
Col Rodella
Piz Ciumbraida
Alpe Fermunt
Pizzo Scalino
Alpe Porcheiro
Alp Zatélet-Prâ
Alp Kratzeren
Grande Crivola
Piz Bacoae
Mont Chauffée
Alp Verva
Petit Raim
Pizzo Cervendone

Kloster Roussano
Piz Gravasalvas
Mont Trelod
Alp Pozata
Grand Golliaz
Monte Camicia
Alp Luzeney
Punta Giapin
Piz Forbisch
Vadret Lischanna
Piz Mortaro
Mont Mallet
Alp Tscheng
Alp Fillar
Monte Sarera
Cima Busazza
Val Spadlatsch
Rifugio Citta
Piz Basodan
Alp Hornfeli
Alp Tusagn
Monte Cradiccioli
Alp Vergalden
Piz Dartjer
Monti Orsëra
Piz Musch
Ponte Zaglia
Val Prate
Valle Brembana

Val Storierà
Mont Coupeline
Piz Bacone
Cima Lago Spalmo
Rifugio Fratelli Longo
Cima Bianc
Monte Canusio
Piz Tranzera
Col Loson
Punta Moraschini
Val Schischenader
Gross Lohners
Monte Orsiera
Piz Tumbif
Cima Eötvös
Alp Alogna
Pizzo Filaut
Gross Sattelistock
Alp Stieren-Iffigen
Tête Rœse
Piz Giendusas
Alp Surpalix
Grosse Viescherhorn
Piz Corandini
Pointe Helbronner
Alp Bödmern
Cima Fràm-pola
Grand Lauzon
Stl Maria

Table A.5: An excerpt of the extracted toponyms with characteristic beginnings

A.1.5 Excluded from the SwissNames Look-Up List

Aa
Abc
Abi
Abrahams
Absatz
Abschlag
Abschlagen
Abschwung
Absturz
AC
Acht
Acker
Ackersand
Adam
Adams
Adler
Adrian
AG
Agen
Ahorn
Ahorne
Ahornen
Ahornwald
Air
Airport
Aktien
Al
Alb
Albert
Albin
Ale
Alexander
Algier
Alle
Allerheiligen
Alp
Alpe
Alpen
Alpenrose
Alpes
Alpina
Alps
Alpweg
Alt
Altar
Alte
Alten
Alter
Alters
Altersheim
Altes
Am
Amerika
Amphitheater
An
André
Andrea
Andreas
Angst
Ängsten
Anker
Anna
Anne
Ans
Anstalt
Antoine
Anton
Antonio
Antonius
Aquädukt
AR
ARA
Ara
Arche
Aren
Arena
Ari
Arme
Armes
Arni
Ars
Arsch
Arten
Arven
As
Äschen
Ass
Asses
Ast
At
Atlas
Au
Aua
Auditorium
Auf
Aufstieg
Aufzug
Auge
Auges
Äugst
Aula
Ausbildungszentrum
Ausser
Aussicht
Autofähre
Ava
Ave
Ba
Bach
Bäche
Bächen
Bächle
Bachs
Bacon
Bad
Badanstalt
Baden
Badeplatz
Bader
Bäder
Badstube
Bahnhof
Bahntunnel
Bald
Bali
Balis
Balkans
Ballone
Bambi
Ban
Band
Bande
Bänder
Bandes
Bank
Bann
Banne
Baptiste
Bar
Bär
Baracken
Barbara
Bären
Bargen
Barmasse

Table A.6: An excerpt of excluded geo-non geo ambiguous toponyms and toponym parts

A.2 Code

The following programmes are available on the CD accompanying each hard copy of this thesis:

Data Preprocessing

- Extracting German words recognised by GERTWOL (*gertwol.pl*)
- Extracting toponyms and toponym parts registered in SwissNames (*swissnames.pl*)
- Comparing Swiss toponyms and toponym parts to German words recognised by GERTWOL (*sn-in-gertwol.pl*)
- Sorting Swiss toponyms and toponym parts, removing the geo-non geo ambiguous tokens (*sn-sort.pl*)

NER

- NER for one-word tokens (*nouns.pl*)
- NER for token sequences (*nouns_mul.pl*)

Data Postprocessing

- Checking for antiquated and different orthography in one-word tokens (*spelling.pl*)
- Checking for antiquated and different orthography in multi-word tokens (*spelling_mul.pl*)
- Checking for separators in one-word tokens (*slash-hyphen.pl*)
- Checking for separators in multi-word tokens (*slash-hyphen_mul.pl*)
- Checking for joining elements in one-word tokens (*hyphen2.pl*)
- Checking for joining elements in multi-word tokens (*hyphen2_mul.pl*)
- Extracting one-word candidate toponyms with characteristic endings (*endings.pl*)
- Extracting multi-word candidate toponyms with characteristic beginnings (*beginnings.pl*)

Result Analysis

- Determining year of last mention of each one-word candidate toponym (*ages.pl*)
- Determining year of last mention of each compound candidate toponym (*ages_mul.pl*)
- Identifying one-word candidate toponyms which are similar to registered SwissNames toponyms (*levenshtein.pl*)
- Identifying compound candidate toponyms which are similar to registered SwissNames toponyms (*levenshtein_mul.pl*)
- Randomly selecting 10% of the one-word candidate toponyms (*random.pl*)
- Randomly selecting 10% of the compound candidate toponyms (*random_mul.pl*)
- Identifying the toponyms (names of mountains, glaciers and cabins) recognised by the corpus NER system and are not in SwissNames (*nerfiles.pl*)
- Identifying the toponyms (names of mountains, glaciers and cabins) recognised by the corpus NER system, are not in SwissNames and are in a German article written about Switzerland (*nerfiles-swissarticles.pl*)
- Identifying which toponyms (names of mountains, glaciers and cabins) were detected by both the corpus and the thesis NER systems (*mount-cab-glac.pl*)
- Extracting toponym coordinates from the SwissNames gazetteer (*sn-coordinates.pl*)

Personal Declaration

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

Zurich, 30^{th} August 2011

Linda Ettlin