How many Mountains are there in Switzerland? Explorations of the SwissTopo Name List

Martin Volk

Institute of Computational Linguistics, University of Zurich

Abstract. In the project "Text+Berg" we digitize all yearbooks of the Swiss Alpine Club from 1864 until today. The books comprise articles in German, French and Italian, a total of around 100,000 pages. This paper describes the corpus and the project phases towards its digitalization. We then focus on the classification of named entities, in particular geographic entities. We explore the usefulness of a large list of geographical names that is distributed by the Swiss Federal Office of Topography. A first experiment indicates that the recognition and classification of geographical names remains difficult despite the large gazetteer.

1 Introduction

In the project Text+Berg[1] we digitize the heritage of alpine literature from various European countries. A shorter version of this paper has been accepted for publication as [Volk et al., 2009].[2] The current paper adds a detailed study of the SwissTopo database.

Text+Berg is a joint project by the German Department and the Institute of Computational Linguistics at the University of Zurich. Currently our group digitizes all yearbooks of the Swiss Alpine Club (SAC) from 1864 until today. Each yearbook consists of 300 to 600 pages and contains reports on mountain expeditions, culture of mountain peoples, as well as the flora, fauna and geology of the mountains. The corpus is thus a valuable knowledge base to study the evolution in all these areas. It provides a time line of information on hot topics like climate change, sustainable use of alpine resources, and technological developments ranging from transportation to media, as well as communication and climbing equipment. But the corpus is also a resource to catch the spirit of Switzerland in cultural terms: What does language use in alpine texts show about the cultural identity of the country and its change over time?[3]

Computational challenges of the Text+Berg project include the automatic correction of OCR errors, text structure annotation, as well as information extraction based on named entity recognition.

[1] See `http://www.textberg.ch`.

[2] We gratefully acknowledge contributions by the coauthors of this paper.

[3] See [Bubenhofer, 2009] for the theoretical and methodological background of research in this domain.

This paper describes the corpus and the project phases towards its digitalization. We then describe the recognition and classification of named entities, in particular person names and geographic entities. We focus on the description of a large gazetteer of Swiss geographical names, the SwissTopo list with more than 150,000 entries.

2 The Text+Berg Corpus

The Swiss Alpine Club was founded in 1863 as a reaction to the foundation of the British Alpine Club a year before. The Swiss did not want to leave the exploration of "their" alps to the English. But contrary to the British Alpine Club they did not restrict membership to mountain climbers but opened it to "all those who love the mountains". Thus our corpus has a clear topical focus: conquering and understanding the mountains. But at the same time it covers a wide variety of text genres as for example expedition reports, (popular) scientific papers, book reviews, etc. The articles focus mostly on the Alps, but over the 144 years the books have probably covered any mountain region on the globe.

Fig. 1. Books from different periods of the Text+Berg Corpus

Some example headers from the 1911 yearbook may illustrate the diversity. There are the typical reports on mountain expeditions: "*Klettereien in der Gruppe der Engelhörner*" (English: Climbing in the Engelhörner group) or "*Aus den Hochregionen des Kaukasus*" (English: From the high regions of the Caucasus). But there are also articles on scientific topics such as topography, geology or glacierology. The 1911 book contains scientific articles on the formation of

caves ("*Über die Entstehung der Beaten- und Balmfluhhöhlen*") and on the periodic variations of the Swiss glaciers ("*Les variations périodiques des glaciers des Alpes suisses*").

The corpus is multilingual. Initially the yearbooks contained mostly German articles and few in French. Since 1957 the books appeared in parallel German and French versions (with some Italian articles) which allows for interesting cross-language comparisons. The parallel versions followed the same article sequence and page structure, but not all articles were translated. For example, the parallel 1958 German and French yearbooks contain 27 translated articles (German-French), plus 38 of the same articles in both yearbooks (DE:18, FR:17, IT:3).

3 Project Phases

We have already collected all books in two copies (as a result of a call for book donations by the Swiss Alpine Club). One copy was cut open so that the book can be scanned with automatic paper feed. The other copy remains as reference book.

3.1 Scanning and OCR

We use state-of-the-art OCR software to convert the images to text. This software comes with two lexicons for German (and one for French and Italian) which match the spelling after 1901 and the new orthography after the spelling reform of the late 1990s. For the German spelling of the 19th century (e.g. old *Nachtheil* and *passiren* instead of modern *Nachteil* and *passieren*) we have added special word lists to the OCR lexicon. In all other cases we have to rely on the quality of the character recognition without lexicon look-up. Fortunately the yearbooks were set in Antiqua font from the start in 1864. So we do not have to deal with old German Gothic font (Fraktur).

A group of student volunteers helps in the correction of the OCRed text. The idea is to get the text structure right and to eliminate the most obvious OCR errors. We are also experimenting with methods for automatic OCR-error correction, e.g. statistical approaches as described in [Reynaert, 2008].

3.2 Mark-up of the Text Structure

We first introduce a mark-up of the text structure. Specially developed programs annotate the text with TEI-conformant XML tags for the beginning and end of each article, the title and the author, for page numbers, footnotes and caption texts. Much of that information can be checked against the table of contents and table of figures in the front matter of the yearbooks. This increases the annotation accuracy.

3.3 Language Identification

Proper language identification is important for most of the subsequent steps of automatic text analysis, e.g. part-of-speech recognition and lemmatization. In addition, it should be possible to limit the search in the text corpus to a specific language. Of course it is also interesting to evaluate the change in language use in our corpus over the years (German vs. French vs. Italian vs. other languages).

Therefore we use a language identification program[4] to determine the language for each sentence. Such a fine granularity helps us to detect quotes and direct speech in languages different from the text language (as for example English sentences in German text).

With respect to other languages we are interested in learning more about the usage of the Swiss minority language Rhaeto-Romanic, a Romance language that is still spoken today by a few 10,000 people in the canton Graubünden. When is this language used in our corpus? And to what extent does the corpus contain texts, passages and quotes of direct speech in Swiss German? Finally, what is the role of English in these books given that British mountaineers and tourists were amongst the first and most active in the 19th century? For example, we were surprised to find that the 1903 yearbook contains a German article with the English statement that Switzerland has turned into "the playground of Europe".

3.4 Archiving, Access and Distribution

In the final phase the annotated corpus will be stored in a database which can be searched via the internet. Because of our detailed annotations the search options will be more powerful and lead to more precise search results than usual search engines. For example, it will be possible to find the answer to the question "List the names of all glaciers in Austria that were mentioned before 1900." We also annotate the captions of all photos and images so that they can be included in the search indexes.

In addition to the query module we will provide easy access to the texts and images through a variety of intuitive and appealing graphical user interfaces. We plan to have clickable geographic maps that lead to articles dealing with certain regions or places. This is one of the reasons why we work on the automatic classification of toponyms in the texts.

4 Named Entities in our Text+Berg Corpus

Named entity recognition is an important aspect for information extraction. But it has also been recognized as an important aspect for the access of heritage data. [Borin et al., 2007] argue for named entity recognition in 19th century Swedish literature, distinguishing between 8 name types and 57 subtypes.

We have investigated methods for named entity recognition in newspaper texts [Volk and Clematide, 2001], and in this paper we discuss how these methods work on our Text+Berg corpus.

[4] We use Michael Piotrowski's Lingua-Ident.

4.1 Person Names

In a previous project [Volk and Clematide, 2001] we had built three rule-based modules for the recognition of person names, geographical names and company names respectively in a corpus of a weekly business-oriented computer newspaper. The module for person name recognition relied on a long list of person first names and the fact that person names in newspapers are usually introduced by a first name followed by a last name. Thereafter the last name can be used alone within the same newspaper article.

If the name is not mentioned for a longer period of text, then it needs to be reintroduced. We had modelled this observation with an algorithm which we called learn-apply-forget. When the person's family name is introduced after a first name trigger, then the family name is saved with a certain priming level (this is the learning step). This priming level is decreased for each subsequent sentence that does not contain the name. If the name does occur, then this increases the priming number. In our current research we investigate whether the same algorithm also works for our Text+Berg corpus.

One striking difference between our computer newspaper corpus and our current corpus is that many person names in the alpine yearbooks are not introduced by first name plus last name but rather by a title or a function term followed by a last name. Address forms like English *Mr., Mrs.*, German *Herr*, academic titles (*Prof., Dr.*) but also *Ingenieur* (engineer), *Major*, are particularly common in the early 1900s.

4.2 Geographical Names

In our work on named entity recognition in newspaper texts [Volk and Clematide, 2001] we had only distinguished two types of geographical names: city names and country names. This was sufficient for texts that dealt mostly with facts like a company is located in a certain country or has started business in a certain city. But our Text+Berg corpus deals with much more fine-grained location information: mountains and valleys, glaciers and climbing routes, cabins and hotels, rivers and lakes. In fact the description of movements (e.g. in mountains) requires all kinds of intricate references to positions and directions in three dimensions.

Thus it is no surprise that geographers and computational linguistics alike are working on the problems of the semantic ordering of spatial expressions. There are numerous initiatives to build geographic ontologies (e.g. [Nudelman Hess et al., 2007]), and there are special workshops that deal with the analysis of geographic references in natural language text, for example the HLT-NAACL 2003 Workshop on Analysis of Geographic References. According to the organizers of this workshop the analysis of geographic references in text involves four distinct stages:

1. geographic entity reference detection (hypothesizing that the strings *Matterhorn, Reuss, Zurich* are referring to geographical entities, i.e. a mountain, a river and a city respectively; this step includes the grouping of multiword

names like *Mont Blanc, Col de Peuterey, Kleine Windgällen, Crans Montana, St. Moritz*)
2. contextual information gathering (classification and possible locations)
3. disambiguation (*Freiburg im Breisgau, Germany* vs. *Freiburg im Üechtland, Switzerland*; *Simon Bolivar* as a person name vs. as a mountain name; *Essen, Halle, Hof* as city names vs. as nouns)[5]
4. grounding (assignment of geographic coordinates; *Zurich* is on 47°22'N 8°33'E)

There have been a number of approaches on the identification of geographical references (e.g. [Rauch et al., 2003], [Leidner et al., 2003] and [Axelrod, 2003]). But they have focused mostly on newspaper texts. Our texts are much denser in terms of geographical references since mountain climbing is the central topic. For example, in the following paragraph we can identify the names of mountains (*Bocktschingel, Kleiner Ruchen, Hintere Kalkschyen*), of a glacier (*Hüfigletscher*), a cabin (*Hüfihütte*) and a snow formation on a mountain (*Bocktschingelfirn*).

> *In kurzer Zeit erreichten wir sodann auf öfter beschriebenem Wege über den* ***Bocktschingel*** *und den* ***Hüfigletscher*** *das südliche Ufer und die* ***Hüfihütte****. Ganz wider Erwarten brach am 16. August ein glanzvoller Tag an. Um 4 Uhr verließen wir die Hütte und gewannen auf dem Weg, auf dem wir hergekommen waren, den* ***Bocktschingelfirn*** *um 6 Uhr, dann weiter über den* ***Kleinen Ruchen*** *den Nordgipfel der* ***Hintern Kalkschyen*** *um 8 Uhr 30 Min. Ohne Aufenthalt stiegen wir über den Südgrat ab zur berüchtigten Scharte, deren Überwindung wir nach kritischer Musterung sogleich in Angriff nahmen.* (SAC-Jahrbuch 1910, p.298, bold face added)

In addition to the names there are other descriptive elements that provide for the textual coherence of the spatial description, many of those provide directions (*das südliche Ufer, den Nordgipfel, den Südgrat*).

In our previous project we had identified geographical names based on large gazetteers for city and country names. In addition to the listed base forms our program was able to recognize genitive forms (*Frankreichs, Münchens*) as well as adjectival forms (*Münchner, Bad Homburger*). In recognizing mountain names we also need to take care of occasional plural forms (*Fergenhorn - die drei Fergenhörner*).

Since it is inefficient to compute all inflected forms beforehand, we will have to use decompounding and lemmatization wherever possible. For example, the genitive form *Gornergrates* is split into *Gorner+grates* and then reduced to the base form *Gornergrat*. The corpus itself serves as dictionary source for verification of the computed lemmas. By splitting we also collect compounding elements like *Gorner* which occurs frequently in *Gornergletscher* but rarely alone. The

[5] We use the term "noun" in this article to denote a class of words which is often called "regular noun" in contrast to proper names. When we write "nouns", then this excludes proper names.

parallel French version *glacier de Gorner* helps to identify such compound elements.

In order to recognize the geographical names in our corpus we have acquired a large list of Swiss toponyms. In the next section we explore its contents.

The SwissTopo Name List The Swiss Federal Office of Topography (`www.swisstopo.ch`) maintains a database of all names that appear on its topographical maps. We have obtained a copy of their database called "SwissNames25" (since it contains all names from its maps with a 1:25,000 resolution) and investigated its usefulness for our purposes.

The SwissTopo database contains 156,755 names in 61 categories. Categories include settlements (10 categories ranging from large cities to single houses), bodies of water (13 categories from major rivers to ponds and wells), mountains (7 categories from mountain ranges to small hills), valleys, mountain passes, streets and man-made facilities (like e.g. bridges and tunnels), and single objects like hotels, mountain cabins, monuments etc. We include the complete overview of the categories with frequency counts in the appendix. It is striking that the bulk of the names refers to field names (Flurnamen: 54,980) and single houses (Einzelhaus: 42,456). Furthermore we notice that some objects are subclassified according to size. For example cities are subdivided into main, large, middle and small cities according to their number of inhabitants.

It should also be noted that a geographical entity might be included several times in the list. It appears in the list as often as its name appears on a topographical map. For example, the river name *Rhein* appears 28 times in the list (associated with different communities in different Swiss cantons). Unfortunately there is no way to tell that this is the same river. It could be that there are two (or more) rivers with the same name in different parts of the country. We need to use other information sources to derive a name list without duplicates.

The problem of multiple name occurrences referring to the same object is naturally more eminent with large rivers than with small creeks. And it is more eminent with longish entities like rivers than with mountains or cities. Even the large cities Basel, Bern and Zürich occur only once in the list.

Of course, more generic names occur more often and refer to different entities. The name *Bad* (bath) occurs 269 times in 6 different categories distributed over 22 cantons. In 231 cases it is classified as a sports facility (Sportanlage) and 24 times as a single house (Einzelhaus). But it is also listed as the name of communities of different size and also as the name of a castle (in canton Zurich). These counts refer only to the name *Bad* as a stand-alone name. There are another 26 occurrences of *Bad* as a name prefix. These fall into 6 categories, including the names of 8 different towns, with *Bad Ragaz* (SG) and *Bad Zurzach* (AG) being the largest and best known. Lastly, there are 10 occurrences of *Bad* as a name suffix (e.g. *Alvaneu Bad, Luthern Bad, Schwarzsee Bad*).

Every name is listed in the SwissTopo database with its coordinates, its altitude (if applicable and if available), the source document (almost all names

stem from topographical maps 1:25,000), the administrative unit to which it belongs (usually the name of a nearby town), and the canton.

The altitude is specified for 23,802 names (15% of all name entries). This information is distributed over 54 name categories. Fortunately, the altitude coverage for mountains (from mountain ranges to small peaks) is almost complete (99.5%). Also for the city entries the percentage of altitude information is high; around 95% of all city and town names come with figures for their elevation. There are two more categories with good altitude coverage, road passes (97%) and artificial lakes (Stausee, 82%). For all other categories the altitude coverage is more coincidental. Bodies of running water ("Fluss, Bach, Gletscher") never have altitude information.

Name-Noun Ambiguities Potentially every one of the SwissTopo names may occur in a mountaineering report. On the other hand each name that is also a noun is a potential source of ambiguity. Therefore we need to know how many names are homographs with nouns and which of the SwissTopo name categories introduce the most ambiguities.

In order to check this we compiled a list of nouns from the TIGER corpus, which contains 888,299 tokens from a German newspaper with manually checked Part-of-Speech tags and lemmas. This corpus contains 184,000 noun tokens (tagged as NN) which result in a list of 37,846 unique noun lemmas. When we compare the SwissTopo name list with this list of TIGER nouns, we find that of the 105,000 name types in the SwissTopo list only 495 name types are ambiguous with nouns in the TIGER corpus. These 495 name types account for 4024 entries in the name list. The most frequent ones are *Bad* (269), *Berg* (192), *Brand* (153), *Loch* (140) and *Feld* (136). More than 50% of all SwissTopo sports facility names (Sportanlage) are nouns (e.g. *Bad, Rodelbahn, Stadion*). And more than 25% of all SwissTopo words denoting public buildings are nouns (e.g. *Schulhaus, Spital, Zoo*). Furthermore there are 270 field name types (Flurnamen, e.g. *Matte, Rebberg, Winkel*) that are nouns in the TIGER corpus. Therefore we will have to treat these three categories with special care or even exclude them from automatic name matching.

Noun lemmas are the most likely candidates for name homographs, but in principle every other noun form is also a candidate. Therefore we searched the TIGER corpus for noun forms that are not lemmas. The TIGER corpus contains 17,513 such noun forms. Most of them are plural forms. To our surprise we found that they account for 594 SwissTopo name ambiguities (150 name types). Examples are city names like *Meilen, Seen, Wangen*, mountain names like *Läden, Scharten, Zwillinge*, and then again a multitude of field names (e.g. *Betten, Öfen, Sagen*), and names for single houses (e.g. *Dellen, Gründen, Heulen*). The latter example indicates that not only noun forms but also other word classes can be ambiguous with names. The city name *Baden* is probably the best example of a verb form that serves as a name.

This comparison also tells us that 7 main mountains have names that are also nouns (*Dom, Esel, Jungfrau, Krone, Mönch, Ochse, Speer*). Furthermore there

are 26 name types of minor mountains that are nouns (the most frequent ones are *Stock* (10), *Horn* (6) and *Stand* (4)). Some of the name-noun ambiguities are curious words like *Bär, Löffel, Prosecco* (field names in AG, FL, and TI respectively), or *Ast, Greuel, Gummi*, the names of small villages or hamlets.

Obviously our method for finding name-noun homographs has its limitations. The TIGER corpus is large, but it does not cover all German nouns. A larger corpus or a good dictionary will lead to more homographs. Moreover we searched only for German homographs. Naturally, there will also be French, Italian and Rhaeto-Romanic noun homographs among the SwissTopo names which need to be considered when we are processing texts in these languages. Occasionally there will even be cross-language homographs. For example the French word *Plage* (meaning "beach") occurs in the SwissTopo list as sports facility. This word is also a homograph with a German noun which means "menace, trouble".

Name Complexity Names in the SwissTopo list range from complex phrases to simple two-letter words. The list contains 1898 names with more than 20 characters. The longest names are the two airport names *Aérodrome régional de Lausanne la Blécherette* (45 characters) and *Aérodrome de La Chaux-de-Fonds-Les Eplatures* (44 chars) followed by an entry of category church *Frauenkloster Sankt Joseph der Clarissinnen* (43 chars). However, there are only 13 names with 35 or more characters.

Some of the long French names also account for those names with the most blank-separated tokens. The list is headed by two French cabin names: *Bivouac du Col de la Dent Blanche CAS* (8 tokens) and *Refuge de la Vue du Mont Blanc* (7) followed by the longest airport name *Aérodrome régional de Lausanne la Blécherette* (6). All the names with 6 or more tokens are French or Italian due to their use of articles and prepositions instead of German compounds and genitives. These counts indicate the required complexity of our matching routines in order to be able to match the names from the SwissTopo list.

At the other end of the spectrum there are 384 words with only 3 characters (e.g. *Elm, Inn, Vex*) and 38 words with only 2 characters. *Au* is probably the best known. But SwissTopo also lists small villages with the names *Gy, Lü* and *Oh*.

Acronyms as parts of SwissTopo names pose a special problem for recognition. For example, the cabins of the Swiss Alpine Club and of other clubs are listed with the respective club acronym (e.g. *Monte Rosahütte SAC, Bivouac du Dolent CAS, Mischabelhütten AAC Zürich*). Moreover some city names have the canton specified as part of the name (e.g. *Carouge (GE), St-Martin (VS)*). This occurs mostly with town names that refer to different towns in different cantons (e.g. *Kilchberg (BL) and Kilchberg (ZH)*). Still it is redundant information since the canton is always specified in a separate column. We cannot expect that these acronyms will be used in the corpus texts. The names need to be recognized even without these acronyms. Spelling variations that include other abbreviations also need to be taken into account (e.g *St-Martin* vs. *Saint-Martin*, *S. Nazzaro* vs. *San Nazzaro*)

Name Variants The multilingual nature of our corpus poses the question whether the SwissTopo name list includes name variants in French, German and Italian. For example, the name of the city of Zurich will appear as *Zürich* in German, *Zurich* in French and *Zurigo* in an Italian text. Unfortunately, SwissTopo lists only the name as it appears on the respective map. This means that French cities are listed with French names (e.g. *Genève* but not *Genf*, *Neuchâtel* but not *Neuenburg*) and German cities are listed with German names (e.g. *Basel* but not *Bâle*, *Aargau* but not *Argovie*). For few bilingual cities both names occur in one string separated by a slash (e.g. *Biel/Bienne, Disentis/Mustér, Celerina/Schlarigna*).

The "monolingual" SwissTopo name list requires us to complement it with other information sources to cover language variants. One option is to use name correspondence lists from Wikipedia. For example, the Wikipedia page titled "Liste deutscher Bezeichnungen Schweizer Orte" contains 400 French, Italian and Rhaeto-Romanic city names with their German correspondences. Other possible resources are the list of Swiss postal codes (distributed via `match.postmail.ch`) or the geographic information packages from the US National Geospatial-Intelligence Agency (`www.nga.mil`).[6]

The multilinguality of our corpus also requires us to cater for German-French combinations. French names sometimes contain the name category as part of the name (e.g. *Glacier du Petit Mont Collon, Ruine du Château de Martinet*). It is likely that these names will also occur in German texts as *der Gletscher des/von Petit Mont Collon* or *die Ruine Château de Martinet* (without the French preposition).

A First Experiment: Finding Mountain Names We selected an article from the SAC yearbook of 1900 to check the precision and recall of automatically identifying mountain names based on the SwissTopo name list. The article is titled "Bergfahrten im Clubgebiet (von Dr. A. Walker)". It is an article in German with a wealth of French mountain names since the author reports about his hikes in the French speaking part of Switzerland. We took the article after OCR without any further manual correction. After our tokenisation (incl. the splitting of punctuation symbols) it consisted of 9380 tokens. We first used the SwissTopo mountain names classified as Massiv, HGipfel and GGipfel, i.e. the 3 highest mountain classes. They consist of 1174 mountain names. Our matching routine searched the article for exact matches in the mountain name list and identified 29 mountain names (10 different mountain names), all of which are correct matches (*Aiguille d'Argentière, Bietschhorn, Pointe d'Orny* etc.).

In a second step we enlarged the gazetteer to include the 4th mountain class (KGipfel) which resulted in a list of 5588 mountain names. This increased the recall to 54 mountain names (20 different mountain names) but at the expense of erroneously marking 6 nouns *Gendarm, Haupt, Kamm, Stand, Stein, Turm* as mountain names. This indicates that it is a good idea to exclude nouns from

[6] We would like to thank Simon Clematide and Michael Piotrowski for pointing us to these resources.

the gazetteer when precision is important. These nouns should only be marked as mountain names when the context clearly favors this interpretation.

How many mountain names have we missed to identify? A manual inspection showed that there are another 92 mountain names (35 different mountain names) missing. So recall of the naive exact matching is still below 40% despite the large gazetteer. We have identified the following reasons for missed names.

1. Some mountains in the article are simply not in Switzerland and therefore not in the SwissTopo list (e.g. *Mont Blanc*).
2. Sometimes there is an upper case vs. lower case distinction between the SwissTopo list and the corpus (e.g. *Tour Noir* vs. *Tour noir* both of which occur in our text). This is easily remedied.
3. Some mountain names appear in hyphenated compounds and need to be separated for identification (e.g. *Monte Rosa-Massiv*).
4. Spelling variations necessitate fuzzy matching (e.g. the book mentions the mountains *Tour Salière* and *Aiguille de la Neuva* but the SwissTopo list writes *Tour Sallière* and *Aiguille de l'A Neuve*).
5. OCR errors prohibit some exact matches (e.g. *Aiguille duTour* with a missing blank).
6. Substitutions of French name parts by German translations lead to missed matches. For example, the program correctly marked *Grand Combin*, but failed to identify *Großen Combin*.
7. Partial repetitions (of the type *Grand Combin* vs. *Combin* or *Le Catogne* vs. *Catogne*) lead to missed matches and need to be handled in a coreference resolution module. These partial references come in a great variety which makes it a hard problem.

4.3 Future Work

Coreference Resolution As a step towards extracting spatial descriptions we will identify coreference chains throughout each text. Coreference means that two words refer to the same object. Detecting coreference involves the mapping of different words to the same object, but it also requires the disambiguation of the "same" name referring to different objects. Here are some examples of coreference variants:

- name abbreviations (*Bad Homburg - Homburg, Eiger - Vordereiger*)
- definite noun phrases (*Tambohorn - die elegante Pyramide*)
- partial references (*Eiger - die Nordwand*)
- name variants (historical, dialectal, across languages: *Mons Egere - Eiger, Weisshorn - Wysshorn*)[7]

We intend to solve the coreference problem with a combination of rule-based and statistical methods.

[7] [Werlen, 2008] discusses dialectal variants and the etymology of Swiss mountain names and comments on the SwissTopo name list.

Mark-up of Semantic Relations Eventually we intend to go beyond the recognition of geographical entities towards the recognition of spatial descriptions. This means that we will automatically identify in which spatial relation two physical objects are. For example, we want to determine that "a cabin is located on a mountain" or that "a town is between a mountain and a river".

In order to reach this goal we need to analyze the texts in more detail. In other words we need to apply an automatic parser for syntax analysis. Our research group has recently developed robust and fast dependency parsers for English and German [Schneider, 2008]. This work has shown that the parser can be efficiently ported to other languages [Sennrich et al., 2009]. The parsers have a rule-based backbone, but they resolve ambiguities with statistical means. The statistical disambiguation, which is extracted from a treebank, allows the parser to prune substantially during parsing and to return likely analyses that are licensed by the grammar, ranked by their probability. Our parser has been shown to achieve state-of-the-art speed and accuracy.

5 Conclusion

We are working on the digitalization and annotation of alpine texts. In a first step we compile a corpus of German 145 yearbooks and 52 French yearbooks from the Swiss Alpine Club. The Club has agreed to host the complete archive on its web server. Users will be able to inspect the search results as PDF documents so that they can enjoy the original look and feel (including all the pictures). In addition we will distribute our corpus as an XML-tagged textual resource for researchers.

As part of the XML annotation we are working on the automatic classification of person names and geographical names. Large resources like the SwissTopo list will serve as the backbone of toponym classification. However, as we have shown in this paper, exact matching is not enough. In order to reach a satisfactory recall a lot of fine-tuning and linguistic processing is necessary.

If we can attract sufficient funding, we hope to have a preliminary version of the SAC yearbooks on the web for full-text search and article-wise inspection in early 2010. We would like to have a preliminary version of the annotated corpus (as XML files) available for distribution by the summer of 2010.

6 Acknowledgments

We would like to thank the many student helpers who have contributed their time to this project. We also thank Michael Hess for his input in defining the Text+Berg project. But most of all we would like to thank Michael for creating an open and inspiring work atmosphere that allows us to follow our research ideas as wild as they may seem at times. Congratulations to Michael. We are looking forward to many more years of collaboration.

References

[Axelrod, 2003] Axelrod, A. (2003). On building a high performance gazetteer database. In *Proceedings of Workshop on the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Budapest.

[Borin et al., 2007] Borin, L., Kokkinakis, D., and Olsson, L.-J. (2007). Naming the past: Named entity and animacy recognition in 19th century Swedish literature. In *Proceedings of The ACL 2007 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, Prague.

[Bubenhofer, 2009] Bubenhofer, N. (2009). *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse.* Number 4 in Sprache und Wissen. de Gruyter, Berlin, New York.

[Leidner et al., 2003] Leidner, J. L., Sinclair, G., and Webber, B. (2003). Grounding spatial named entities for information extraction and question answering. In *Proceedings of Workshop on the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Budapest.

[Nudelman Hess et al., 2007] Nudelman Hess, G., Iochpe, C., Ferrara, A., and Castano, S. (2007). Towards effective geographic ontology matching. In *Proceedings of the Second International Conference on GeoSpatial Semantics (Mexico City)*, Lecture Notes in Computer Science, pages 51–65. Springer.

[Rauch et al., 2003] Rauch, E., Bukatin, M., and Baker, K. (2003). A confidence-based framework for disambiguating geographic terms. In *Proceedings of Workshop on the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, Budapest.

[Reynaert, 2008] Reynaert, M. (2008). Non-interactive OCR post-correction for giga-scale digitization projects. In Gelbukh, A., editor, *Proceedings of the Computational Linguistics and Intelligent Text Processing 9th International Conference, CICLing 2008*, Lecture Notes in Computer Science, pages 617–630, Berlin. Springer.

[Schneider, 2008] Schneider, G. (2008). *Hybrid Long-Distance Functional Dependency Parsing.* Phd thesis, University of Zurich, Institute of Computational Linguistics.

[Sennrich et al., 2009] Sennrich, R., Schneider, G., and Volk, M. (2009). A new hybrid dependency parser for German. In *Proceedings of GSCL-Conference*, Potsdam.

[Volk et al., 2009] Volk, M., Bubenhofer, N., Althaus, A., and Bangerter, M. (2009). Classifying named entities in an alpine heritage corpus. *Künstliche Intelligenz (Sonderheft: Kulturerbe und KI)*, (4). To appear in the fall 2009.

[Volk and Clematide, 2001] Volk, M. and Clematide, S. (2001). Learn-filter-apply-forget. Mixed approaches to named entity recognition. In Moreno, A. M. and van de Riet, R. P., editors, *Applications of Natural Language for Information Systems. Proc. of 6th International Workshop NLDB'01*, volume P-3 of *Lecture Notes in Informatics (LNI) - Proceedings*, pages 153–163, Madrid.

[Werlen, 2008] Werlen, I. (2008). Die Grundwörter der Oberwalliser Gipfelnamen. In *Chomolangma, Demawend und Kasbek, Festschrift Für Roland Bielmeier*, pages 577–614.

Appendix: Overview of the SwissTopo Categories

The explanations in this overview are based on a brochure by the Swiss Federal Office of Topography: "SwissNames. Die Namendatenbank der Schweiz. La banque de données toponymiques de la Suisse" (not dated). See `www.swisstopo.ch`.

SwissTopo Code	***English explanation***	***Freq***
Siedlungen	**Settlements**	
HGemeinde	city > 50,000 inhabitants	10
GGemeinde	city 10,000 - 50,000 inhabitants	117
MGemeinde	town 2000 - 10,000 inhabitants	678
KGemeinde	village < 2000 inhabitants	1993
GOrtschaft	large settlement > 2000 inhabitants	112
MOrtschaft	middle settlement < 2000 inhabitants	1987
KOrtschaft	small settlement 50 - 100 inhabitants	2849
Weiler	hamlet < 50 inhabitants	10697
Streusiedl	dispersed settlement	1424
Einzelhaus	single house	42456
	Total	**62323**
Täler	**Valleys**	
Haupttal	main valley	178
Nebental	subordinate valley	2046
Graben	rift	2628
	Total	**4852**
Gebiete	**Areas**	
Gebiet	area	87
Flurname	field name	54980
Wald	forest	7093
	Total	**62160**
Gewässer und Seen	**Bodies of Water**	
Fluss	river	399
Bach	brook, creek	3960
KBach	small creek	1004
GSee	large lake	53
KSee	small lake	817
Quelle	source	69
Stausee	artificial lake	83
Wasserfall	waterfall	52
Sumpf	swamp	191
Weiher	pond	101
Brunnen	well, spring	86
Staumauer	dam	5
Gletscher	glacier	730
	Total	**7550**

SwissTopo Code	*English explanation*	*Freq*
Berge	**Mountains**	
Massiv	important mountain range	143
HGipfel	main Alpine peak	165
GGipfel	minor Alpine peak	866
KGipfel	small peak	4414
Grat	ridge	1440
Fels	rock	1998
Huegel	hill	2543
	Total	**11569**
Pässe	**Passes**	**Freq**
Strassenpass	road pass	102
Fusspass	foot pass	1898
	Total	**2000**
Strassen und Anlagen	**Streets and Facilities**	
Strasse	street	44
Tunnel	tunnel	60
Weg	route, way	197
Park	park	39
Bahnhof	train station	464
Bruecke	bridge	309
Flugplatz	airport	68
Hafen	harbour	24
	Total	**1205**
Einzelobjekte	**Single Objects**	
Denkmal	monument	59
ErrBlock	erratic block	185
Friedhof	cemetry	19
HistOrt	archeological place	226
Hoehle	cave	80
Hotel	hotel	61
Huette	cabin	771
Industrie	industry facility	1018
Kirche	church	459
OeffGeb	public building	648
Ruine	ruin	603
Schloss	castle	401
Sportanl	sports facility	488
Turm	tower	24
Zoll	customs	54
	Total	**5096**