Challenges in Building a Multilingual Alpine Heritage Corpus

Martin Volk, Noah Bubenhofer, Adrian Althaus, Maya Bangerter, Lenz Furrer, Beni Ruef

Institute of Computational Linguistics, University of Zurich
volk@cl.uzh.ch

Abstract

This paper describes our efforts to build a multilingual heritage corpus of alpine texts. Currently we digitize the yearbooks of the Swiss Alpine Club which contain articles in French, German, Italian and Romansch. Articles comprise mountaineering reports from all corners of the earth, but also scientific topics such as topography, geology or glacierology as well as occasional poetry and lyrics. We have already scanned close to 70,000 pages which has resulted in a corpus of 25 million words, 10% of which is a parallel French-German corpus. We have solved a number of challenges in automatic language identification and text structure recognition. Our next goal is to identify the great variety of toponyms (e.g. names of mountains and valleys, glaciers and rivers, trails and cabins) in this corpus, and we sketch how a large gazetteer of Swiss topographical names can be exploited for this purpose. Despite the size of the resource, exact matching leads to a low recall because of spelling variations, language mixtures and partial repetitions.

1. Introduction

In the project Text+Berg[1] we digitize the heritage of alpine literature from various European countries. 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.

Some examples 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 the 1911 book also contains scientific articles on the development 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 thus a valuable knowledge base to study the changes in all these areas. 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?[2]

This paper describes the corpus (part of which is parallel French-German) and the project phases from digitalization through annotation to publication. We focus on the language technology challenges in improving optical character recognition (OCR), language identification, and the classification of named entities, in particular geographic entities.

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. Thus our corpus has a clear topical focus: conquering and understanding the mountains. The articles focus mostly on the Alps, but over the 145 years the books have probably covered any mountain region on the globe.

[1] See www.textberg.ch.

[2] See (Bubenhofer, 2009) for our research in this area.

Figure 1: Books from different periods of the Corpus

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 and may serve as training material for Statistical Machine Translation systems.

3. Project Phases

We have 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, one for the spelling after 1901 and one for the new orthography following the spelling reform of the late 1990s. The system does not have a lexicon for the German spelling of the 19th century (e.g. old *Nachtheil*, *passiren* and *successive* instead of modern *Nachteil*, *passieren* and *sukzessive*).

We have therefore added 19th century word lists to the system. We have manually corrected one book from 1890, and subsequently extracted all words from that book that displayed old-style character sequences (such as 'th', 'iren', and 'cc').

The 20th century books follow the Swiss variant of German spelling. In particular, the Swiss spelling has abandoned the special character 'ß' in favor of 'ss'. For example, the word 'ließ' (English: let) is spelled 'liess' in Switzerland. The OCR lexicons list only the spelling from Germany. We have therefore compiled special word lists with Swiss spelling variants taken from the GNU Aspell program.

Names that are not in the system's lexicon pose another problem to character recognition. Our books contain a multitude of geographical names many of which are unknown to the OCR system. We have therefore purchased a large list of geographical names from the Swiss Federal Office of Topography (see section 4.1.) and extracted the names of the major Swiss cities, mountains, valleys, rivers, lakes, hiking passes and mountain cabins. In total we added 14,800 toponyms to the OCR system.

These steps have helped us to further improve the OCR quality which was good from the very start of the project, thanks to the fact that the yearbooks were set in Antiqua font from the first publishing year 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 automatically recognized text. The idea is to get the text structure right and to eliminate the most obvious OCR errors. Additionally, we post-correct errors caused by graphemic similarities which have been missed by the OCR engine. This automatic correction happens after tokenisation on a word by word level, using a rule-based approach. For example, a word-initial 'R' is often misinterpreted as 'K', resulting in e.g. *Kedaktion* instead of *Redaktion* (English: editorial office). To minimize false positives, our rules fall in one of three categories: First, strict rule application: The tentative substitute must occur in the corpus and its frequency must be at least 2.5 times as large as the frequency of the assumingly mistyped word, and the suspect must not occur in the German newspaper corpus TIGER. Second, normal rule application: The tentative substitute must occur in the corpus. Third, unconditional substitution. The above *K→R* example falls in the strict category; substituting 'ii' by either 'n', 'u', 'ü', 'li' or 'il' (of the five tentative substitutes the word with the highest frequency is selected; e.g. *iiberein* → *überein*, English: in agreement) falls in the normal category; and substituting *Thai* with *Thal* (the 19th century spelling of *Tal*, English: valley) is an example of the unconditional category.

We are also experimenting with other methods for automatic OCR-error correction, e.g. statistical approaches as described in (Reynaert, 2008) and an ensemble approach based on processing our scanned images with two different OCR systems and merging their output.

3.2. Mark-up of the Text Structure

We then add 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, its title and author, subheaders and paragraphs, page breaks, footnotes and caption texts. For example, footnotes are recognized by their bottom position on the page, their smaller font size and their starting with any character followed by a closing parenthesis character.

Some of the text structure information can be checked against the table of contents and table of figures in the front matter of the yearbooks. We manually correct these tables as the basis for a clean database of all articles in the corpus. Matching entries from the table of contents to the articles in the books is still not trivial. It requires that the article title, the author name(s) and the page number in the book are correctly recognized. Therefore, we use fuzzy matching to allow for OCR errors and small variations between table of content entries and the actual article header in the book.

However, not all data from the OCR output is kept in the XML-based mark-up. We discard redundant elements: On the yearbook level we discard the front matter (table of contents and table of figures) as well as the index at the end of the book. On the page level header and footer lines containing page numbers are discarded, including binding signatures.

Another challenge is the changing layout of the books over time. In the initial years the books were pocket-size (11×18 cm), the page numbers on the top of the page together with alternating header lines containing either the article author or the title (often an abbreviation of the title). Picture pages were not numbered and were not even counted in the page numbering. Footer lines contained binding signatures and binding numbers on regular intervals (e.g. every 16th page). These footer lines can easily be mistaken for footnotes if one considers only the font size. At the end of the 19th century the book format was enlarged to account for the increased interest in photos. It was enlarged again in 1910 (to 18×25 cm, cf. figure 1) which triggered a period of ornaments including ornamented initial letters for each article (which are a nightmare for OCR).

From 1925 the ornaments were dropped and gave way for an austere layout with article headers in sans-serif fonts and articles no longer starting on a new page. With the advent of the parallel French-German versions in 1957 the layout changed radically with page numbers now at the bottom of the page and no more header lines. Since 1970 the yearbooks have been set in two-column layout.

3.3. Language Identification

Proper language identification is important for most of our subsequent steps of automatic text analysis, e.g. part-of-speech tagging, lemmatization and named entity classification.

Therefore we use a character-n-gram-based language identification program[3] to determine the language for each sentence. The hope is that a granularity on the sentence level helps us to detect quotes and direct speech in languages different from the text language (as for example English quotes in German text). Language identification is unreliable for short sentences. Therefore we work with the heuristic that sentences with less than 75 characters are not

[3]We use Michael Piotrowski's language identifier Lingua-Ident from search.cpan.org/dist/Lingua-Ident/.

analyzed by the language identifier but rather they are assigned the language of the article.
With respect to other languages we are interested in screening our corpus for the Swiss minority language Romansch, a Rhaeto-Romance language that is still spoken today by a few 10,000 people in the canton Graubünden.
We are also interested in investigating to what extent the corpus contains texts, passages and quotes of direct speech in Swiss German dialects. There is no standardized orthography for the dialects but they are still sometimes used in our books. For example a 1962 article on humor in Switzerland contains a number of jokes with direct speech like "Ihr nemmeds aber au chaibe gmüetli mit Schaffe" (p.46) which translates into German as "Ihr nehmt es aber auch ziemlich gemütlich mit dem Arbeiten" (English: You take it rather easy with the work).
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 a German article with the English statement that Switzerland has turned into "the playground of Europe" as early as 1903. We later learned that this expression dates back to the title of a 1870 book by Leslie Stephen.
It should be noted that tokenisation and language identification lead to a circularity in sequencing. Tokenisation and in particular sentence boundary recognition need to precede language identification so that we are able to feed sentence by sentence to the language identifier. But high quality tokenisation relies heavily on language-specific abbreviation lists and conventions. We therefore do a rough tokenisation and sentence boundary recognition before language identification. Afterwards we do another round of tokenisation in order to correct possible tokenisation errors.

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.
(Witte et al., 2008) emphasize that advanced access methods are crucial for Cultural Heritage Data. They distinguish different user groups having different requirements (Historians, Practitioners, Laypersons, Computational Linguists). 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.
As of March 2010 we have scanned and OCR-converted 142 books from 1864 to 1982. This is the black-and-white period. Although the yearbooks contained occasional color photos from the beginning, these books of the first 118 years contain mostly black-and-white images. In 1983 this changed to a new layout and mostly color photos.
We have 90 books from 1864 to 1956. In 1870, 1915 and 1924 no yearbooks were published. From 1957 to 1982 we have parallel French and German versions of the yearbooks. Overall we have scanned nearly 70,000 pages. This resulted in 6101 articles in German, 2659 in French, 155 in Italian, 12 in Romansch, and 4 in Swiss-German. This count includes duplicates from the parallel French and German yearbooks. 458 of the articles in these parallel books are not translated but reprinted in the same language. This means we currently have a corpus of 8931 articles in one of the languages. Our parallel corpus currently contains 701 articles amounting to 2.6 million tokens in French and 2.3 million tokens in German. Table 1 gives an overview of the token frequencies per language. Work on scanning and converting the yearbooks from 1983 is ongoing and will be finished later this year.

3.5. Corpus Linguistic Research

Quantitative corpus linguistics has proofed to be a valuable technique in many domains of philological, sociological and historical research. The digitised and linguistically annotated corpus is therefore an interesting source for studies in many fields and facilitates the investigation of changing patterns of language use, and how these reflect underlying cultural shifts.
In linguistics and cultural studies, the change of language use over time, special terminology and cultural shifts are of interest. The "speaking" about mountains is characterised by cultural, historical and social factors; therefore, language use can be viewed as a mirror of these factors. The extra-linguistic world, the essence of a culture, can be reconstructed through analyzing language use within alpine literature in terms of temporal and local specifics that emerged from this typical use of language (Bubenhofer, 2009). For instance, frequent use of personal pronouns and specific intensifiers in texts between 1930 and 1950 can be interpreted as a shift to a more subjective, personal role that mountaineering played in society. In contrary, between 1880 and 1900, the language surface shows less emotionality which probably is a mirror of a period when the mountain pioneers claimed more seriousness (Bubenhofer and Schröter, 2010).
But also in literature studies, the corpus can help to understand the parallels and differences of narrative structures in literary and alpine texts. Historical, ethnological or economic studies will also profit from keyword analyses and easy access to structured alpine texts.

4. Geographical Names 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.
In a previous project we have investigated methods for named entity recognition in newspaper texts (Volk and Clematide, 2001). In that work we had only distinguished two types of geographical names: city names and country names. This was sufficient for texts that dealt mostly with

	German	French	Italian	English	**Total**
all tokens in corpus	17,253,000	8,126,000	329,000	44,000	25,753,700
tokens in parallel corpus	2,295,000	2,604,000			

Table 1: Token counts (rounded) in the Text+Berg corpus

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.

Therefore geographers and computational linguists alike are working on the problems of structuring the semantics 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 Geographic Information Retrieval (e.g. the GIR workshop series) and 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 a regular nouns))

4. grounding (assignment of geographic coordinates; *Zurich* is on 47°22'N 8°33'E)

Approaches to the identification of geographical references include (Rauch et al., 2003), (Leidner et al., 2003) and (Axelrod, 2003). 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 (*südliche Ufer, Nordgipfel, 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 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 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.

4.1. The SwissTopo Name List

The Swiss Federal Office of Topography[4] 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. 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.

[4] www.swisstopo.ch

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 the different occurrences refer to 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* (English: bath, swimming pool) 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.

4.1.1. 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. This is particularly troublesome in German since all nouns are spelled with an initial upper case letter. 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 a total of 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 noun lemmas 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). Of course, it is debatable whether all entries in the SwissTopo list would pass a linguistic test that distinguishes proper names from other nouns. When a public swimming-pool is marked as *Bad*, the word denotes an instance of a class of objects and therefore should fall into the category noun rather than proper name. If however a mountain is called *Kamm* (English: comb, ridge), we tend to accept this as proper name. Since it is difficult to draw this line automatically (and often even intellectually), we assume all entries in the SwissTopo list to be names for the sake of this discussion.

More than 50% of all SwissTopo sports facility names (Sportanlage) are homographs with nouns from the TIGER corpus (e.g. *Bad, Rodelbahn, Stadion*). And more than 25% of all SwissTopo words denoting public buildings are TIGER nouns (e.g. *Schulhaus, Spital, Zoo*). Furthermore there are 270 field name types (Flurnamen, e.g. *Matte, Rebberg, Winkel*) that are nouns in our TIGER reference corpus. Therefore we will have to treat these three categories (sport facilities, public buildings, and field names) 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 inflected noun form is also a candidate. Therefore we also 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* (English: to bath) is probably the best example of a verb form that serves as a name.

Obviously our method for checking 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 Romansch 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".

4.1.2. 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 with and 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*)

4.1.3. 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 and cantons are listed with German names (e.g. *Basel* but not *Bâle*, *Aargau* but not *Argovie*). For few bilingual cities which are located at a language border both names occur in one string separated by a slash (e.g. *Biel/Bienne, Disentis/Mustér, Celerina/Schlarigna*).

The "monolingual" SwissTopo name strategy 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 Romansch 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).[5]

4.2. 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)" (English: Mountain Tours in the Club Region). 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 used the SwissTopo mountain names classified as Massiv, HGipfel, GGipfel, and KGipfel, i.e. the 4 highest mountain classes. They consist of 5588 mountain names. This leads to a recall of 54 mountain names (20 different mountain names) at the expense of erroneously marking 6 nouns *Gendarm, Haupt, Kamm, Stand, Stein, Turm* as mountain names.

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 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. Some mountain names appear in hyphenated compounds and need to be separated for identification (e.g. *Monte Rosa-Massiv*).
3. 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*).
4. OCR errors prohibit some exact matches (e.g. *Aiguille duTour* with a missing blank).
5. 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*.
6. 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

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

module. These partial references come in a great variety which makes it a hard problem.

Precision obviously suffers from name-noun and name-name ambiguities. For example, the SwissTopo list contains 6 different mountains called *Breithorn*. Often the disambiguation of toponyms is based on geographic proximity. For instance, such map-based methods (Buscaldi and Rosso, 2008) are used for disambiguating city names on a national or world-wide scale (Athens, Greece vs. Athens, Georgia, USA). It is unclear how well they work on the narrow relations in mountaineering reports.

5. Conclusion

We are working on the digitalization and annotation of alpine texts. Currently we compile a corpus of 145 German yearbooks and 52 French yearbooks from the Swiss Alpine Club. In the next step we will digitize the French yearbooks *L'Echo des Alpes* that were published in Switzerland from 1871 until 1924 to counterbalance the German language dominance in the yearbooks of the Swiss Alpine Club.

Language identification works reliably for French, German, Italian and English for longer sentences. We need to adapt the identifier for Romansch, Swiss German dialects and occasional Latin quotes.

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, exact matching is not enough. In order to reach a satisfactory recall and precision a lot of fine-tuning and linguistic processing (incl. coreference resolution) is necessary.

Mountaineering accounts primarily focus on routes, i.e. they are not about a peak (as a location), but about *how* that peak was reached. Thus, routes and their descriptions are of particular interest for research. We therefore aim to automatically extract and map route descriptions from mountaineering reports. This requires both the geo-coding of the toponyms and the recognition of all temporal cues in order to identify the sequence of events. One of the challenges in geo-coding mountaineering reports is the distinction between route points and general panorama descriptions. In (Piotrowski et al., 2010) we discuss some preliminary ideas on how to distinguish these.

6. Acknowledgments

We would like to thank the many student helpers who have contributed their time to this project and Torsten Marek for his programming work on the XML annotation. We are also grateful for the support by the Swiss Alpine Club and by Hanno Biber and his team from the Austrian Academy Corpus.

7. References

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

Lars Borin, Dimitrios Kokkinakis, and Leif-Jöran Olsson. 2007. Naming the past: Named entity and animacy recognition in 19th century Swedish literature. In *Proceedings of The ACL Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, Prague.

Noah Bubenhofer and Juliane Schröter. 2010. Die Alpen. Sprachgebrauchsgeschichte – Korpuslinguistik – Kulturanalye. In *Wohin steuert die historische Sprachwissenschaft? Tagung vom September 2009*, Debrecen, Ungarn.

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

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

Jochen L. Leidner, Gail Sinclair, and Bonnie Webber. 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.

Guillermo Nudelman Hess, Cirano Iochpe, Alfio Ferrara, and Silvana Castano. 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, November.

Michael Piotrowski, Samuel Läubli, and Martin Volk. 2010. Towards mapping of alpine route descriptions. In R. Purves, P. Clough, and C. Jones, editors, *Proceedings of the 6th Workshop on Geographic Information Retrieval (GIR'10)*, pages 15–16, Zurich.

Erik Rauch, Michael Bukatin, and Kenneth Baker. 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.

Martin Reynaert. 2008. Non-interactive OCR post-correction for giga-scale digitization projects. In A. Gelbukh, 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.

Martin Volk and Simon Clematide. 2001. Learn-filter-apply-forget. Mixed approaches to named entity recognition. In Ana M. Moreno and Reind P. van de Riet, 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.

René Witte, Thomas Gitzinger, Thomas Kappler, and Ralf Krestel. 2008. A Semantic Wiki Approach to Cultural Heritage Data Management. In *Proceedings of LREC Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, Marrakech, Morocco.