Strategies for reducing and correcting OCR errors

Volk, Martin ; Furrer, Lenz ; Sennrich, Rico

Abstract: In this paper we describe our efforts in reducing and correcting OCR errors in the context of building a large multilingual heritage corpus of Alpine texts which is based on digitizing the publications of various Alpine clubs. We have already digitized the yearbooks of the Swiss Alpine Club from its start in 1864 until 1995 with more than 75,000 pages resulting in 29 million running words. Since these books have come out continuously, they represent a unique basis for historical, cultural and linguistic research. We used commercial OCR systems for the conversion from the scanned images to searchable text. This poses several challenges. For example, the built-in lexicons of the OCR systems do not cover the 19th century German spelling, the Swiss German spelling variants and the plethora of toponyms that are characteristic of our text genre. We also realized that different OCR systems make different recognition errors. We therefore run two OCR systems over all our scanned pages and merge the output. Merging is especially tricky at spots where both systems result in partially correct word groups. We describe our strategies for reducing OCR errors by enlarging the systems' lexicons and by two post-correction methods namely, merging the output of two OCR systems and auto- correction based on additional lexical resources.

DOI: https://doi.org/10.1007/978-3-642-20227-8_1

Strategies for Reducing and Correcting OCR Errors

Martin Volk, Lenz Furrer and Rico Sennrich

Abstract In this paper we describe our efforts in reducing and correcting OCR errors in the context of building a large multilingual heritage corpus of Alpine texts which is based on digitizing the publications of various Alpine clubs. We have already digitized the yearbooks of the Swiss Alpine Club from its start in 1864 until 1995 with more than 75,000 pages resulting in 29 million running words. Since these books have come out continuously, they represent a unique basis for historical, cultural and linguistic research. We used commercial OCR systems for the conversion from the scanned images to searchable text. This poses several challenges. For example, the built-in lexicons of the OCR systems do not cover the 19th century German spelling, the Swiss German spelling variants and the plethora of toponyms that are characteristic of our text genre. We also realized that different OCR systems make different recognition errors. We therefore run two OCR systems over all our scanned pages and merge the output. Merging is especially tricky at spots where both systems result in partially correct word groups. We describe our strategies for reducing OCR errors by enlarging the systems' lexicons and by two post-correction methods namely merging the output of two OCR systems and auto-correction based on additional lexical resources.

Martin Volk
Institute of Computational Linguistics, University of Zurich, e-mail: volk@cl.uzh.ch

Lenz Furrer
Institute of Computational Linguistics, University of Zurich, e-mail: lenz.furrer@access.uzh.ch

Rico Sennrich
Institute of Computational Linguistics, University of Zurich, e-mail: sennrich@cl.uzh.ch

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.

Digitization of this corpus requires a large-scale scanning effort followed by converting the images to text, a procedure known as optical character recognition (OCR). There are a few commercial OCR products (Abbyy FineReader, Nuance OmniPage) and one open-source product (previously named Tesseract, now called OCRopus[2]). Initial experiments indicated that the open-source tool's recognition quality is far lower than the commercial products, therefore we decided early-on to focus on Abbyy FineReader, the alleged market leader. Of course, the commercial OCR systems only deliver error-free text under ideal conditions like modern font, evenly printed on spotless white paper. For most of our input texts, these conditions are not given.

Some of the books of our corpus have yellowed pages, sometimes even grey spots from the paper, the printing or the handling. The books were typeset in Antiqua from the start but the letters are not always evenly printed. In addition the books contain special symbols (e.g. the clock time 15:45h is often written as 15 $^3/_4$ where the fraction is one symbol). Moreover, they comprise 19th century spelling variants, a complex layout with a mix of images and text, and text in multiple languages. Challenges for the OCR systems abound. We have noticed that this results in a recognition accuracy of several errors per page. Our aim was to reduce the error rate.

In principle there are three ways to improve the text accuracy resulting from OCR. We can improve the input to the OCR system, the OCR system itself, or the output of the OCR system. On the input side we can try to improve the image so that the contrasts are ideal and the image is clean. We have experimented with greyscale vs. black-and-white scanning and tried various contrast settings. The OCR results did not differ much, so we decided in favor of 300 dpi greyscale images, not least because Abbyy claims that their OCR system is optimized for these. We have not tried cleaning the images with despeckle programs. Experiments reported in [5] indicate that this does not improve the accuracy noticeably.

As a second option, we can try to improve the OCR system. Abbyy FineReader allows three ways of tuning the system. The user can train certain characters (or variants thereof) so that they are added to the recognition alphabet. This is cumbersome and time-consuming. The user can select additional characters from a list so that they are added to the recognition alphabet of the selected language. For example, if the German alphabet is chosen, the user may add Icelandic diacritic characters when processing a text about an expedition to Iceland that contains geographical

[1] See www.textberg.ch.

[2] See code.google.com/p/ocropus/

names with these special characters. Finally, the user can enlarge the systems' built-in lexicon to add domain-specific vocabulary. This can be expected to increase the recognition accuracy since the OCR system decides on word hypotheses based on language-specific word lists. We have experimented with lexicon enlargement and report on our results in section 4.2.

Thirdly, we can improve the text accuracy by repairing the OCR system output. This amounts to automatic spelling and grammar correction and thus can be tackled in a large number of different ways. We present two methods in this paper. One method is based on merging the output of two different OCR systems, the other is based on producing spelling variants for "unknown" words and predicting which variant is most likely correct.

This paper first describes the Text+Berg project with its multitude of challenges for OCR. We then describe our experiments with a number of strategies for reducing and correcting OCR errors.

2 The Text+Berg Project

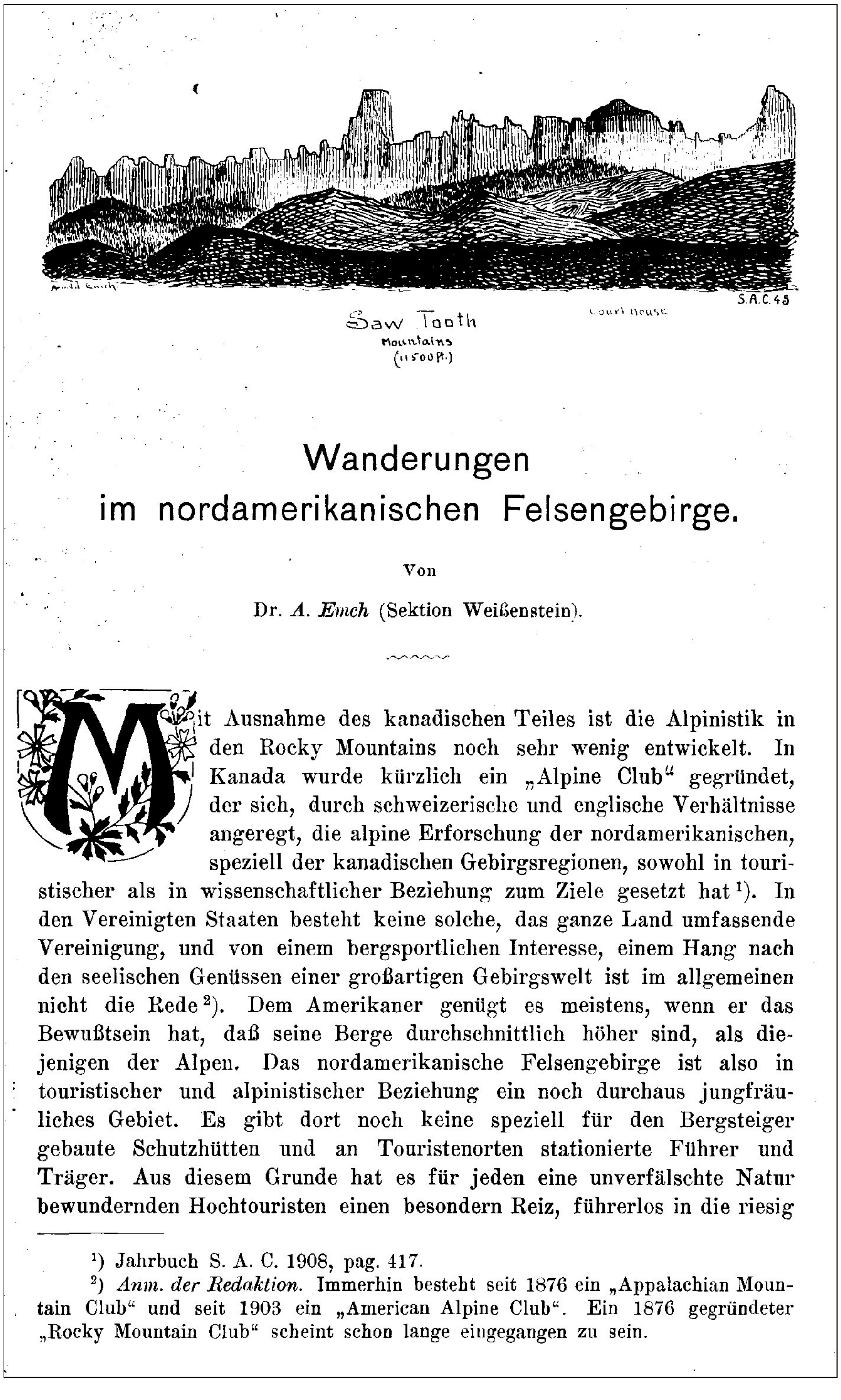
The Text+Berg project at the University of Zurich aims at building a large corpus of Alpine texts. As a first step we digitized the yearbooks of the Swiss Alpine Club (SAC). The SAC was founded in 1863 as a reaction to the founding of the British Alpine Club a year before. From the very start it produced a sizable yearbook documenting its mountaineering activities. 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 every mountain region on the globe.

Some examples from the 1911 yearbook may illustrate the diversity. It has 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? See [3] for our research in this area.

Let us briefly describe how we processed the books. Initially 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.

Then all books were scanned and processed by the OCR systems. The main challenges for OCR which we encountered were the multilingual nature of the text, dia-



Arnold Emch
S.A.C.45
Saw Tooth Mountains (11500 ft.)
Court house

Wanderungen im nordamerikanischen Felsengebirge.

Von

Dr. *A. Emch* (Sektion Weißenstein).

Mit Ausnahme des kanadischen Teiles ist die Alpinistik in den Rocky Mountains noch sehr wenig entwickelt. In Kanada wurde kürzlich ein „Alpine Club“ gegründet, der sich, durch schweizerische und englische Verhältnisse angeregt, die alpine Erforschung der nordamerikanischen, speziell der kanadischen Gebirgsregionen, sowohl in touristischer als in wissenschaftlicher Beziehung zum Ziele gesetzt hat[1]. In den Vereinigten Staaten besteht keine solche, das ganze Land umfassende Vereinigung, und von einem bergsportlichen Interesse, einem Hang nach den seelischen Genüssen einer großartigen Gebirgswelt ist im allgemeinen nicht die Rede[2]. Dem Amerikaner genügt es meistens, wenn er das Bewußtsein hat, daß seine Berge durchschnittlich höher sind, als diejenigen der Alpen. Das nordamerikanische Felsengebirge ist also in touristischer und alpinistischer Beziehung ein noch durchaus jungfräuliches Gebiet. Es gibt dort noch keine speziell für den Bergsteiger gebaute Schutzhütten und an Touristenorten stationierte Führer und Träger. Aus diesem Grunde hat es für jeden eine unverfälschte Natur bewundernden Hochtouristen einen besondern Reiz, führerlos in die riesig

[1] Jahrbuch S. A. C. 1908, pag. 417.

[2] *Anm. der Redaktion.* Immerhin besteht seit 1876 ein „Appalachian Mountain Club“ und seit 1903 ein „American Alpine Club“. Ein 1876 gegründeter „Rocky Mountain Club“ scheint schon lange eingegangen zu sein.

Fig. 1 Example page from the SAC yearbook 1909-10, with decorated initial letter and footnotes.

chronic changes in spelling and typesetting, and the wide range of proper nouns. In section 3, we will give a detailed account on our efforts to improve OCR quality.

After text recognition we added a mark-up of the text structure. Specially developed programs annotated the text with 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. Figure 1 shows the start page of an article from the 1910 yearbook with title, author and two footnotes.

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 corrected 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.

2.1 Language Identification

Proper language identification is important for most steps of automatic text analysis, e.g. part-of-speech tagging, lemmatization and named entity classification. The SAC yearbooks are multilingual, with most articles written in German and French, but also some in Italian, Romansch and Swiss German[3]. We use a character-n-gram-based language identification program[4] to determine the language for each sentence.

While language identification may help improve automatic text analysis, the dependency is circular. OCR, tokenization and sentence boundary recognition need to precede language identification so that we are able to feed individual sentences to the language identifier. But high quality tokenization relies heavily on language-specific abbreviation lists and conventions. We therefore perform an initial tokenization and sentence boundary recognition before language identification. Afterwards, we retokenize the text in order to correct possible tokenization errors.

OCR is performed without prior language identification. We configured the OCR systems to use the dictionaries for the following four languages: German, French, Italian and English.

[3] See section 2.3 for information on the amount of text in each language.

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

2.2 Further Annotation

Apart from structural mark-up and language identification, the corpus is automatically tagged with Part-of-Speech information. We also aim to provide a fine-grained annotation of named entities.

Named entity recognition is an important aspect of information extraction. But it has also been recognized as important for the access to heritage data. For example, Borin et al. [2] argue for named entity recognition in 19th century Swedish literature, distinguishing between 8 name types and 57 subtypes.

In the latest release of our corpus (all yearbooks from 1864 to 1995), we have annotated all mountain names that we could identify unambiguously through exact matching. We have obtained a large gazetteer with 156,000 toponyms from the Swiss Federal Office of Topography. It contains geographical names in 61 categories. We have extracted the SwissTopo mountain names from the 4 highest mountain classes plus the names classified as ridges (*Grat*). This resulted in 6227 names from which we have manually excluded 50 noun homographs. For example *Ofen* (English: *oven*) is a Swiss mountain name, but in order to avoid false hits we eliminated it from the list. The resulting gazetteer triggered the identification of 95,400 mountain names in our corpus.

2.3 Aims and Current Status

In the final processing phase, the 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 via the usual search engines. For example, it will be possible to find the answer to the query "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.

[15] 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 December 2010, we have scanned, OCR-converted and annotated 168 books from 1864 to 1995 (cf. [4]).

We have 90 books from 1864 to 1956. (In 1870, 1914 and 1924 no yearbooks were published.) From 1957 to 1995 we have parallel French and German versions of the yearbooks. Overall we have scanned around 75,000 pages. The corpus is made up of around 6000 articles in German, 2700 in French, 140 in Italian, 11 in Romansch, and 3 in Swiss-German. Our parallel corpus currently contains 950 articles amounting to 3.3 million tokens in French and 2.9 million tokens in German.

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

	German	French	Italian	Other	**Total**
tokens in entire corpus	18,700,000	9,770,000	320,000	100,000	28,890,000
tokens in parallel subcorpus	2,910,000	3,300,000			

Table 1 gives an overview of the token frequencies per language. Work on scanning and converting the yearbooks from 1996 is ongoing and will be finished soon. More details on the project phases can be found in [14].

3 Scanning and OCR

Let us return to the OCR step. After scanning the pages in greyscale with 300 dpi we embarked on converting all pages to text. We started by using Abbyy-FineReader 7. We have initially evaluated Abbyy-FineReader version 7 to version 9, but found the older version more stable and of equal OCR quality.

Even though the performance of OCR applications is satisfactory for most purposes, we are faced with thousands of OCR errors in large text collections. Since we aim to digitize the data as cleanly as possible, we wish to minimize the number of errors. Additionally, OCR errors can be especially damaging for some applications. The numerous named entities, i.e. names of mountains, streams and Alpine cabins are especially prone to OCR errors, especially because many of them do not occur in the dictionaries used by OCR tools. At the same time, these named entities are highly relevant for our goal of building a searchable database. In our project, OCR is complicated by the fact that we are digitizing a multilingual and diachronic corpus with texts spanning from 1864–1995.

3.1 Enlarging the OCR Lexicon

The OCR 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’). In this way we added 1500 word forms to the OCR lexicon.

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 and added around 5000 entries to the OCR lexicon.

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 (www.swisstopo.ch) 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. In section 4.2 we present the results of this lexicon enlargement.

3.2 Post-correcting OCR Errors

Following the OCR we have experimented with three post-correction methods. Here we introduce the methods, while section 4.2 has the comparative evaluation results.

3.2.1 Pattern-based Corrections

When the annotation process has completed tokenisation, a polisher module is invoked. Here, some heuristics are applied to catch and correct common OCR errors, such as misrecognised *Htttte* for correct *Hütte* (English: *cabin*) or erroneous *Eichtung* for correct *Richtung* (English: *direction*). Using regular expressions, a closed set of substitution patterns is applied to the corpus.

In this way we post-correct errors caused by graphemic similarities which have been missed by the OCR engine. This automatic correction happens after tokenization with heuristics that check each word. 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 three categories: First, strict rule application: The tentative substitute must occur in the corpus and its frequency must be at least 2 times as large as the frequency of the presumably mistyped word. The above *K*→*R* example falls in this category. Second, normal rule application: The tentative substitute must occur in the corpus. Substituting 'ii' by either 'n', 'u', 'ü', 'li' or 'il' (of the five tentative substitutes the word with the highest frequency is selected; *iiberein* → *überein*, English: in agreement) falls in the normal category. Third, unconditional substitution. For example, substituting *Thai* with *Thal* (the 19th century spelling of *Tal*, English: *valley*) is an example of the unconditional rule category.

3.2.2 OCR Merging

In an attempt to automatically detect and correct the OCR errors, we exploit the fact that different OCR systems make different errors. Ideally, we can eliminate all OCR errors that are only made by one of two systems. We have created an algorithm that compares the output of two OCR systems (Abbyy FineReader 7 and OmniPage 17) and performs a disambiguation, returning the top-ranking alternative wherever the systems produce different results.

3.2.3 The Merging Algorithm

For our task, we can avoid potential complexity problems since we do not have to compute a global alignment between the two OCR systems. Three factors help us keep the search space small: Firstly, we can extract differences page-by-page. Secondly, we ignore any differences that cross paragraph boundaries, defaulting to our primary system FineReader if such a large discrepancy should occur. Thirdly, the output of the two systems is similar enough that differences typically only span one or two words.

For each page, the algorithm traverses the two OCR-generated texts linearly until a difference is encountered. This point is then used as starting point for a longest common subsequence search in a 40-character-window. We extract as difference everything up to the start of the longest subsequence, and continue the algorithm from its end.

For selecting the best alternative, we consider the differences on a word level. If there are several differences within a short distance, all combinations of them are considered possible alternatives. As a consequence, we not only consider the output of FineReader (*Recensione-»,*) and OmniPage (*Rccensionen*), but also the combinations *Rccensione-»,* and *Recensionen*. In this way, the correct word form *Recensionen* can be constructed from two wrong alternatives.

Our decision procedure is based on a unigram language model trained on the latest release of the Text+Berg corpus. The choice to bootstrap the decision procedure with noisy data generated by Abbyy FineReader bears the potential risk of skewing the selection in Abbyy FineReader's favor. However, the language model is large (25.7 mio words), which means that possible misreadings of a word are far outnumbered by the correct reading. For instance, *Bergbauer* (English: *mountain farmer*) is twice misrecognized as *ßergbauer* by Abbyy FineReader. Still, *Bergbauer* is more than 20 times as frequent as *ßergbauer* in the corpus (47 vs. 2 occurrences), which lets the language model make a felicitous judgment.

It is worth noting that OCR merging is performed before language identification, and that we do not use one model per language, but a language model trained on the whole corpus, irrespective of language.

Words containing non-alphabetical characters have been removed from the language model, with the exception of hyphenated words. Punctuation marks and other special characters are thus penalized in our decision module, which we found to be an improvement.

A language model approach is problematic for cases in which the alternatives are tokenized differently. Generally, alternatives with fewer tokens obtain a higher probability. We try to counter this bias with a second score that prefers alternatives with a high ratio of known words. This means that *in Göschenen* is preferred over *inGöschenen*, even if we assume that both *Göschenen* (the name of a village) and *inGöschenen* are unknown words in our language model[5].

The alternatives are ranked first by the ratio of known words, second by their language model probability. If there are several candidates with identical scores, the alternative produced by Abbyy FineReader is selected.

3.2.4 First Evaluation of the OCR Merging

We have performed a manual evaluation of the merged algorithm based on all instances where the merged system produces a different output than Abbyy FineReader. The cases where Abbyy's system wins are not as interesting since we regard them as the baseline result. Out of the 1800 differences identified between the two systems[6] in the 1899 yearbook, the FineReader output is selected in 1350 cases (75%); in 410 (23%), the OmniPage reading is preferred; in 40 (2%), the final output is a combination of both systems. We manually evaluated all instances where the final selection differs from the output of Abbyy FineReader, which is our baseline and the default choice in the merging procedure.

Table 2 Examples where OmniPage is preferred over FineReader by our merging procedure.

Abbyy FineReader	OmniPage	correct alternative in context	jugdment
Wunseh,	Wunsch,	entstand in unserem Herzen der **Wunsch**,	better
East	Rast	durch die **Rast** neu gestärkt	better
Übergangspunkt,. das	Übergangspunktr das	ist Hochkrumbach ein äußerst lohnender Übergangspunkt, das	equal
großen. Freude	großen, Freude	zu meiner **großen Freude**	equal
halten	hatten	Wir **halten** es nicht mehr aus	worse
là	la	c'est **là** le rôle principal qu'elle joue	worse

Table 2 shows some examples and our judgment. We see clear improvements where non-words produced by Abbyy FineReader (e.g. *Wunseh*) are replaced with a known word produced by OmniPage (*Wunsch*, English *wish*). On the other hand, there are cases where a correctly recognized Abbyy word (e.g. *halten*, English: *hold*)

[5] Unknown words are assigned a constant probability > 0.

[6] Note that one difference, as defined by our merging algorithm, may span several words. Also, frequent differences that would be resolved in later processing steps (i.e. differences in tokenization or hyphenation) are ignored by the merging algorithm.

is overwritten by the OmniPage candidate (*hatten*, English: *had*) because the latter is more frequent in our corpus. As a third possibility, there are neutral changes where the Abbyy output is as wrong as the OmniPage output, as in the two examples judged as "equal", where the systems suggest different punctuation symbols where none is intended in the text.

In our manual evaluation, we found 277 cases where OCR quality was improved, 82 cases where OCR quality was decreased, and 89 cases where combining two systems neither improved nor hurt OCR quality.

We noticed that performance is worse for non-German text. Most notably, OmniPage tends to misrecognize the accented character *à*, which is common in French, as *A* or *a*, or to delete it. The misrecognition is a problem for words which exist in both variants, especially if the variant without accent is more common. This is the case for the French article *la* (English: *the*) and the adverb *là* (English: *there*), and leads to a miscorrection in the example shown in table 2. We are lucky that in our language model, the French preposition *à* (English: *to*) is slightly more probable than the French verb *a* (English: *has*); otherwise, we would encounter dozens of additional miscorrections.[7] Word deletions are relatively rare in the evaluation set, but pose a yet unsolved problem to our merging algorithm. In 8 cases, *à* is regrettably deleted by OmniPage. These alternatives always obtain a higher probability than the sequences with *à*[8], and are thus selected by our merging procedure, even though the deletion is incorrect in all 8 instances.

Considering that we are working with a strong baseline, we find it encouraging that using the output of OmniPage, which is considerably worse than that of Abbyy FineReader, allows us to further improve OCR performance.

3.2.5 Corrections based on External Resources

Our third approach for cleaning up the corpus in post-correction is based on external lexical resources. The aforementioned methods base their decision about "wrong" and "correct" word forms on frequencies of the words in the corpus itself. This entails the risk of categorizing frequent errors as "good" words, which is not an unlikely scenario since OCR systems misrecognize unknown words quite often; e. g. the Alpine toponym *Schneehorn* is rendered thrice as misspelt *Sehneehorn* or its genitive form *Sehneehorns*.

In order to reduce the danger of propagating OCR errors in post-correction, the following approach makes use of external resources for the correctness categorization. It is partly comparable to the pattern-based approach of 3.2.1 as it is all about substituting potentially misspelt words by close orthographical variants which are assumed to be correct. The main differences are in the decision routine, which is not based on corpus frequencies but on lexicon data, and in a search space not restricted to predefined regular expressions.

[7] Of course, one could devise rules to disallow particular corrections.

[8] Since every word has a probability < 1, each additional token decreases the total probability of an alternative.

We use two resources as categorizers to divide all word types of our corpus into *known words* and *unknown words*. For the general German vocabulary we use Gertwol, which takes care of the unlimited number of compounds in German. A subcorpus of the SwissTopo list mentioned in 2.2 is then used to recognize toponyms unknown to Gertwol. Subsequently, unknown words are re-classified as known words if they show to be ancient-spelling variants of known words, or if they can be analyzed as compounds with a toponym as the head and any known word as the tail. All words shorter than a predefined length threshold are excluded. The remaining unknown words are now considered potential OCR-errors and they are exposed to a correcting algorithm.

The procedure of searching correction candidates is done in three steps of ascending complexity. In the first and the second step, a small set of character substitutions found in a high number of OCR-errors is used to derive hypothetical spelling variants. For example, it is common to find mistaken *u* for correct *n* in the output of OCR systems, as in recognized *Küustlergesellschaft* for correct *Künstlergesellschaft* (English: *artist association*). For every unknown word, this substitution is applied to every occurrence of *u*, producing correction hypotheses. For example, from unknown *Tourenverzeichuissen* the variants *Tonrenverzeichuissen* and *Tourenverzeichnissen* are derived, the latter being the correct form meaning *tables of tours*. Of course, all of the substitution pairs (such as the inverse: $n \rightarrow u$) are applied to all words, as well as combinations of them up to a predefined recursion depth, rendering a large number of variants.

In the first step, the known words of the initial categorization serve as a lexicon. All hypothetical spelling variants derived from unknown words are looked up in this corpus-specific lexicon and, if present, are considered a spelling correction of the underlying unknown word. With this method, we can correct misrecognised words if their correct form is also present in the corpus (and has been judged correct by the categorizing mechanism) and if the differences of the known- / unknown-word pair can be described by means of the predefined substitution set. We found this to be safe (i. e. has a high precision), but only very few corrections are done overall (low recall).

In the second step, the hypothetical variants of unknown words that have not been corrected in the first step are analysed by Gertwol to find corrections beyond the corpus-derived lexicon. Word forms appearing only a few times throughout the corpus are likely to have no correct version in the corpus, if they happen to be misrecognised, so they cannot be caught in the first step. Therefore, the hypothetical variants are sent through the categorizing process as described above to find correct words among the bulk of non-words. The words found are taken as a correction.

In pre-evaluation during building, this method turned out to be error-prone in terms of categorizing non-words as correct words and hence introducing bad substitutions to the correction table. The reason for these erroneous jugdements lies in the “creativity” of Gertwol in analyzing words as compounds. This shall be illustrated by an example: The name of a famous gorge at the old route across the Gotthard pass, *Schöllenenschlucht*, had been correctly recognised by the OCR systems, but it was tagged ‘unknown’ by our categorizing method. As a consequence, the word is passed to the correction tool which produces, among others, the hypothetical variant *Sehölleuenschlucht* by substituting one *c* by *e* and one *n* by *u*. Unfortunately, Gertwol claims to know this word, analyzing it as `Seh#öl#leu\en#schlucht`, which can only be interpreted as a fantastic word creation, to be approximately translated as *gorge of the seeing-oil lions*. In order to suppress the sometimes amusing, but undesired analyses of this kind, we apply a filter to the Gertwol output, stopping compound segmentations with tiny elements like *-seh-*.

The processing configuration of the post-correction system including the first and the second step of searching for corrections, is referred to as the 'basic' configuration in the evaluation. To keep the number of searching variants small, only a limited number of character substitutions is followed in the basic configuration. Of course, real-life OCR-errors are not limited to a closed set of character substitutions, but rather show a wide range of deviations from their original word. The aim of the third step is to account for a broad variety of character operations to find correction candidates for unknown words.

This last task is done by a module that partly implements Martin Reynaert's TICCL algorithm (see [10]). As the algorithm is already well described there, only the basic idea is given here: With the anagram hashing technique, Reynaert found a way to treat words as "bags of letters" to reduce searching complexity. The words are stored by a numerical hash key and character operations (such as deletion, insertion, substitution) can be modelled with arithmetic operations. By addition and subtraction one can easily get from one word to all other words having a majority of characters in common.

4 Evaluation

The aim of our evaluation was to measure the influence of several methods for improving OCR accuracy in the Text+Berg project. At different stages of the digitization process, various attempts were made to reduce the rate of OCR-errors. First we briefly specify the evaluation method. The results are then presented in section 4.2.

4.1 Evaluation Setup

The evaluation is based on the Text+Berg *Release 131* [4]. The test corpus was compiled from four volumes of different periods, namely the SAC yearbooks from 1890, 1899, 1912 and 1950. Based on the automatically recognized text, these four books had been manually corrected. The corrected books seem to be of high quality and serve as a gold standard for this evaluation. However, it cannot be ruled out that a certain amount of OCR-errors remained undetected in the manual correction process.

In order to reduce the complexity of measuring and to increase the reliability of the results while working with a multilingual and diachronic corpus, only the German parts of the books were used. Tokens not containing at least one of the basic Latin alphabet's letters were rejected (i.e. numbers and punctuation were left out, as well as noisy tokens). The text was extracted from the XML files as found in the *Release 131* and compared using the ISRI OCR-Evaluation Frontiers Toolkit [12] for aligning and measuring word accuracy.

4.2 Evaluation Results

4.2.1 Standard Processing Modules

To measure the influence of the dictionaries and the merging and polishing module, which are already standard parts of our corpus building process, the whole processing pipeline has been run four times with different configurations, each having one of the modules switched off. One pipeline run was done with standard settings including all modules. For every output, word accuracy was measured by determining agreement with the gold standard. Since the modules evaluated are expected to have an improving effect on the data, their switching off should lead to a lower score than the standard settings. This assumption is shown to hold in most cases.

Table 3 Word accuracy of the standard modules

	standard settings		without ancient spelling dict		without merging module		without polisher module	
1890	119008	Words						
	94.38	% Accuracy	94.36	% Accuracy	94.17	% Accuracy	94.32	% Accuracy
	6689	Errors	6718	Errors	6944	Errors	6765	Errors
			+29	(0.43 %)	+255	(3.67 %)	+76	(1.12 %)
1899	111967	Words						
	99.49	% Accuracy	99.53	% Accuracy	99.35	% Accuracy	99.50	% Accuracy
	575	Errors	527	Errors	727	Errors	557	Errors
			-48	(9.11 %)	+152	(20.91 %)	-18	(3.23 %)
1912	93750	Words						
	99.23	% Accuracy			99.11	% Accuracy	99.17	% Accuracy
	720	Errors			835	Errors	776	Errors
					+115	(13.77 %)	+56	(7.22 %)
1950	135844	Words						
	99.49	% Accuracy			99.42	% Accuracy	99.49	% Accuracy
	691	Errors			785	Errors	699	Errors
					+94	(11.97 %)	+8	(1.14 %)

The results are shown in table 3. As can be seen from the accuracy values, overall quality is high (> 99 %) as compared to the gold standard, with the oldest yearbook being an exception (approx. 94 %). The deviations are not exceedingly high when seen in relation to the book size, but they are remarkable with respect to the number of misrecognized words. The merging module achieves the best results. In all books of the test corpus, turning it off leads to a serious increase of misrecognized words; e.g. in yearbook 1912 a total of 727 words differ from what they should (according to the gold standard) if the merging is not done, but with the standard settings, where the merging is included, this number is reduced by a fifth to 575 differing words. Note that we found a higher number of word error corrections in the manual evaluation of the 1899 yearbook than in the automatic one. We attribute this discrepancy to errors in the gold standard, since we found several corrections being counted as new errors.

The results of the pattern-based correction module are less convincing, but it seems to be helping in the most cases, too. As for the yearbooks 1890, 1912 and 1950 a low to moderate improvement can be achieved by applying some regular-expression patterns during postprocessing; but as for 1899 the module had better not been used since its switching off leads to a higher agreement with the gold standard. The reasons for this unexpected result have not been analyzed yet, but see also 4.2.2 for concerns about the gold standard accuracy.

The results are bad for the semiautomatically built (and very incomplete) dictionary for the 19th-century spelling. The word list was based on a manually corrected list of old spelling variants found in the yearbook 1890. Nevertheless, adding this list to the OCR system has almost no effect on the recognition of that very yearbook (a plus of 29 words out of 6718 are correctly recognized), although the dictionary was overfitted for this data. One might conclude, that the OCR system does not trust the dictionaries provided by the user. This finding corresponds with the observations reported in [5]. The score of the other 19th-century yearbook, 1899, even shows worse results when adding the dictionary. Since this impairment cannot be explained by non-use of the dictionary by the OCR system (if the dictionary was simply ignored, the result would be the same as with the standard setting), the dictionary must be either misleading the system or we are dealing with gaps in the gold standard (cf. 4.2.2).

4.2.2 New Post-correction Module

Our post-correction tool using Gertwol and the SwissTopo list as categorizers (introduced in 3.2.5) has not yet been integrated into the processing pipeline. Therefore the new tool was evaluated against the existing standard settings to see if it leads to any improvement. Seven configurations have been evaluated, four of which will be presented here. The configurations differ in the minimal-length threshold and the combinations of the different correction steps as described in 3.2.5.

The post-correction tool was run with the corpus of the Text+Berg *Release 131*. Each configuration run led to a correction file, listing OCR-errors with a single correction each. Every correction list was applied to a copy of the pipeline output for each of the four yearbooks of the test corpus, substituting every occurrence of an OCR-error by the proposed correction. The post-corrected books were then each compared with their corresponding gold standard book.

Table 4 sums up the results. The leftmost column shows the original standard-settings pipeline output with no further post-correcting and is identical with the first column of the evaluation table 3. The basic configuration includes steps 1 and 2, i. e. correction candidates were only searched by applying a small set of character substitutions. The same holds for the second configuration, where the minimal-length threshold was reduced from 15 to 10 characters, consequently leading to a larger correction list. In the third configuration, a minimal version was tried, which performed only step 1 (corrections were only sought within the known words of the corpus). The last configuration was done using all of three correction steps, including the TICCL module, which has no limitations on character substitutions.

Table 4 Word accuracy of the new post-correction module

	standard settings		basic min. length = 15		corpus-derived lexicon only		with TICCL	
1890	119008	Words						
	94.38	% Accuracy	94.37	% Accuracy	94.38	% Accuracy	94.22	% Accuracy
	6689	Errors	6698	Errors	6689	Errors	6884	Errors
			+9	(0.13 %)	+0		+195	(2.83 %)
1899	111967	Words						
	99.49	% Accuracy	99.46	% Accuracy	99.49	% Accuracy	99.41	% Accuracy
	575	Errors	601	Errors	575	Errors	665	Errors
			+26	(4.32 %)	+0		+90	(13.53 %)
1912	93750	Words						
	99.23	% Accuracy	99.21	% Accuracy	99.23	% Accuracy	99.14	% Accuracy
	720	Errors	736	Errors	722	Errors	802	Errors
			+16	(2.17 %)	+2	(0.28 %)	+82	(10.22 %)
1950	135844	Words						
	99.49	% Accuracy	99.48	% Accuracy	99.49	% Accuracy	99.41	% Accuracy
	691	Errors	703	Errors	691	Errors	803	Errors
			+12	(1.71 %)	+0		+112	(13.94 %)

Unlike the previous evaluation setup, where the modules were switched off to observe their effect, the post-correction tools here are added to existing settings. Improvements could therefore be seen in a decrease of the misrecognized words. Disturbingly, not a single instance of a reduced word error rate can be found in the evaluation table. In the reduced configuration doing only one correction step, hardly any substitutions can be seen at all. The other configurations introduce heaps of words judged as additional misrecognized ones.

Since observations and partial evaluations during building did not indicate such disastrous results, manual re-checking was done. All word replacements that have actually been done while applying the correction lists to the test corpus had been saved to logfiles and could easily be reconstructed. It became clear that the tool measuring word accuracy ignored such substitutions as the deletion of noisy characters in word tokens, which makes us lose a good part of the good substitutions in the evaluation score. Then, the gold standard's reliability has to be questioned when measuring word accuracy in ranges above 99 %. For example, in the reduced configuration ('corpus-derived lexicon only'), the – truly – misrecognized word *Verkanfsmagazine* is replaced by correct *Verkaufsmagazine* (English: *vending magazine*). Unfortunately, the gold standard has the misspelt version here, so the good correction is falsely judged a bad replacement. Although the rate of OCR-errors still present in the gold-standard books is probably low, the few remaining errors have a great chance to be found by the post-correction module, which gives them considerable weight in the automatic evaluation.

Two random samples of each 50 replacements have been manually checked thereafter. In the TICCL-configuration, 18 replacements were found to be good, 29 lead to an impairment and 3 were judged neutral (the text is neither improved nor impaired by those substitutions). Among the undesired replacements, mostly correct words became another known word. A major problem is the haphazardness of correct words to be known or not known by the resources which we used. For example, the compound participle *bestgemachten* (English: *best made*) cannot be analyzed by Gertwol, but *selbstgemachten* (a compound participle as well, meaning *self made*) is known, and since the two words happen to be within a Levenshtein distance of 3, a replacement is triggered. The same holds for the two mountain names *Kienthaleralpen* and *Simmenthaleralpen* (with 19th-century spelling, but both correctly recognized), where the first is known, but not the latter.

While the full configuration of the post-correction tool – after a closer look to the data – still seems not to be performing too well as of now, the basic configuration turns out to be far better than at first sight. Out of the random sample of 50 replacements, 44 have shown to be perfectly well, 5 had better not happened and 1 could not be categorized. So the overall result is positive after all. Again, the two circumstances mentioned above lead to the poor evaluation scores: The – desirable – removing of non-alphanumeric characters is not reported in the accuracy values, and uncorrected OCR errors in the gold standard can be critical. Having the latter in mind, one might suppose that the evaluation results of the first evaluation (see 4.2.1) could be expected to be higher, if the gold-standard books were further improved.

5 Related Work

Holley [5] provides an excellent overview of the issues involved in improving the text quality when converting historic newspaper corpora with OCR. She concludes that the most promising way is collaborative correction and describes their approach of having users correct the Australian Newspaper collection in [6]. A project to digitize a large collection of Romansch texts also aims at collaborative user input [9].

Other methods for automatic OCR-error correction include e.g. statistical approaches as described in [10] and [7], as well as lexical approaches as in [13]. As mentioned before, some of our experiments were inspired by Reynaert [10] who worked on cleaning a digitized collection of historical Dutch newspapers. He has developed an efficient way of mapping an unknown word to its most similar known words which can then be used as substitute in the text. In contrast Kolak et al. [7] present a finite-state character sequence transformation system. They are interested in improving OCR in new languages and show in their experiments that an OCR system for English can produce good results for French when combined with GIZA-style word alignment. Strohmaier [13] has collected domain-specific lexicons by crawling the web. He has shown that these lexicons help to improve the OCR accuracy (for the OCR systems of his time, Abbyy FineReader version 5 and OmniPage version 10). But he has also demonstrated that the combination of two OCR systems leads to improved accuracy which is well in line with our results.

As for the combination of multiple OCR systems, research has identified two main questions: how to efficiently align the output of multiple OCR systems (e.g. [8]), and how to select the optimal word among different candidates. The question of output alignment arises because multiple OCR systems will result in different tokenisations. The word selection step has been performed using voting algorithms [11], dictionaries [8], or human post-editing [1].

6 Conclusion

We are working on the digitization and annotation of Alpine texts. Currently we compile a corpus of German and French yearbooks from the Swiss Alpine Club that span 145 years. 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 SAC collection. We also have an agreement with the British Alpine Club to include their texts in our corpus.

In this paper we have presented various methods aimed at reducing OCR errors. We discussed the enlargement of the OCR system lexicon and various post-correction methods. The lexicon enlargement had surprisingly little impact on the results. It seems that the OCR system does not make good use of the additional lexical material and, annoyingly, the OCR company leaves the user in the dark as to when and how additional lexicons will improve the accuracy rate.

In addition, we have implemented three different post-correction methods. Our post-correction heuristics based on regular expression substitutions aim at obvious OCR errors and are thus a reliable correction method with low recall. The other two methods are more flexible, but only the merging of the output of two OCR systems leads to clearly improved text accuracy. Our final attempt, employing external resources, has not resulted in clear improvements, although our manual inspections showed a number of interesting automatic corrections.

An obvious extension of our merging approach is the inclusion of further OCR systems. For this, Tesseract is an attractive candidate since it is open-source and can thus be tuned to handle those characters well where we observe special weaknesses in the commercial OCR systems.

Our merging procedure also triggered further ideas for combining other textual sources. Our parallel French and German books since the 1950s contain many identical texts. These books are only partially translated, and they partially contain the same article in both books. We have already found out that even the same OCR system (Abbyy FineReader) makes different errors in the recognition of the two versions of the (same) text (e.g. *in der Gipfelfaüinie* vs. *inj der Gipfelfallinie*). This gives us more variants of the same text which we can merge.

We are also wondering whether the same text scanned under different scanner settings, e.g. different contrasts or different resolution, will lead to different OCR results which could be merged towards improved results. For instance, a certain scanner setting (or a certain image post-correction) might suppress dirt spots on the page which may lead to improved OCR quality.

Finally we would also like to explore whether translated texts can help in OCR error correction. Automatic word alignment might indicate implausible translation correspondences which could be corrected via orthographically similar, but more frequent aligned words.

Acknowledgements The authors would like to thank Torsten Marek for his contributions to the OCR merging. Many thanks also to the many student helpers who have contributed to the Text+Berg project. We are especially grateful for the support by the Swiss Alpine Club and by Hanno Biber and his team from the Austrian Academy Corpus. Part of this research has been funded by the Swiss National Science Foundation in the project “Domain-specific Statistical Machine Translation”.

References

1. Ahmad Abdulkader and Matthew R. Casey. Low cost correction of OCR errors using learning in a multi-engine environment. In *Proceedings of the 10th International Conference on Document Analysis and Recognition*, 2009.
2. Lars Borin, Dimitrios Kokkinakis, and Leif-Jöran Olsson. 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, 2007.
3. Noah Bubenhofer. *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Number 4 in Sprache und Wissen. de Gruyter, Berlin, New York, 2009.
4. Noah Bubenhofer, Martin Volk, Adrian Althaus, Maya Bangerter, Torsten Marek, and Beni Ruef. Text+Berg-Korpus (Release 131). XML-Format, 2010. Digitale Edition des Jahrbuch des SAC 1864-1923 und Die Alpen 1925-1995.
5. Rose Holley. How good can it get? Analysing and improving the OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine*, 15(3/4), March/April 2009.

6. Rose Holley. Many hands make light work: Public collaborative OCR text correction in australian historic newspapers. Technical report, National Library of Australia, March 2009.
7. Okan Kolak, William Byrne, and Philip Resnik. A generative probabilistic OCR model for NLP applications. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 55–62, Morristown, NJ, USA, 2003. Association for Computational Linguistics.
8. William B. Lund and Eric K. Ringger. Improving optical character recognition through efficient multiple system alignment. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JDLC09)*, pages 231–240, Austin, TX, 2009.
9. Claes Neuefeind and Fabian Steeg. Digitale rätoromanische Chrestomathie - Werkzeuge und Verfahren für die kollaborative Volltexterschließung digitaler Sammlungen. In *Poster bei der DGfS Jahrestagung*, Göttingen, Februar 2011.
10. Martin Reynaert. 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, 2008. Springer.
11. S. V. Rice, J. Kanai, and T. A. Nartker. A report on the accuracy of OCR devices. Technical report, University of Nevada, 1992. Technical Report.
12. Stephen V. Rice. *Measuring the Accuracy of Page-Reading Systems*. PhD thesis, University of Nevada, 1996.
13. Christian M. Strohmaier. *Methoden der Lexikalischen Nachkorrektur OCR-Erfasster Dokumente*. PhD thesis, Ludwig-Maximilians-Universität, München, 2004.
14. Martin Volk, Noah Bubenhofer, Adrian Althaus, Maya Bangerter, Lenz Furrer, and Beni Ruef. Challenges in building a multilingual alpine heritage corpus. In *Proceedings of LREC*, Malta, 2010.
15. René Witte, Thomas Gitzinger, Thomas Kappler, and Ralf Krestel. 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, 2008.