Crowdsourcing an OCR Gold Standard for a German and French Heritage Corpus

Simon Clematide, Lenz Furrer, Martin Volk
Institute of Computational Linguistics
University of Zurich, Switzerland
{simon.clematide,furrer,volk}@cl.uzh.ch

Abstract

Crowdsourcing approaches for post-correction of OCR output (Optical Character Recognition) have been successfully applied to several historic text collections. We report on our crowd-correction platform Kokos, which we built to improve the OCR quality of the digitized yearbooks of the Swiss Alpine Club (SAC) from the 19th century. This multilingual heritage corpus consists of Alpine texts mainly written in German and French, all typeset in Antiqua font. Finding and engaging volunteers for correcting large amounts of pages into high quality text requires a carefully designed user interface, an easy-to-use workflow, and continuous efforts for keeping the participants motivated. More than 180,000 characters on about 21,000 pages were corrected by volunteers in about 7 months, achieving an OCR gold standard with a systematically evaluated accuracy of 99.7 % on the word level. The crowdsourced OCR gold standard and the corresponding original OCR recognition results from Abby FineReader 7 for each page are available as a resource. Additionally, the scanned images (300 dpi) of all pages are included in order to facilitate tests with other OCR software.

Keywords: Optical Character Recognition, Crowdsourcing, Cultural Heritage Corpora

1. Introduction

Crowdsourcing approaches for post-correction of OCR output (Optical Character Recognition) have been successfully applied to several historic text collections. We report on our crowd-correction platform Kokos, which we built to improve the OCR quality of the digitized yearbooks of the Swiss Alpine Club (SAC) from the 19th century. This multilingual heritage corpus consists of Alpine texts mainly written in German and French, all typeset in Antiqua font.

Finding and engaging volunteers for correcting large amounts of pages into high quality text requires a carefully designed user interface, an easy-to-use workflow, and continuous efforts for keeping the participants motivated.

The scanned images, the uncorrected output of a standard OCR software and the high-quality text corrected by our crowd build a valuable resource.[1] It can be used for extracting heritage lexicons covering 19th century German in particular, or for training as well as testing automatic OCR error correction systems.

In the following section, we introduce our multilingual corpus and describe the process of its digitization. We report on our efforts in building and maintaining a crowd-correction platform and compare them to other work in the field. In Section 3, we analyze and evaluate the corrections performed by the volunteer collaborators. The released resource is described in the last section.

2. Materials and Methods

2.1. Corpus Data

In the Text+Berg project[2] we digitized the yearbooks of the SAC from 1864 until today (henceforth SAC corpus) for building a multilingual heritage corpus of Alpine texts (Göhring and Volk, 2011).

In this paper we focus on the yearbooks of the 19th century. Those yearbooks from 1864 to 1899 amount to 21,247 pages (including tables of content and index pages) with around 304,000 sentences and 6.3 million tokens before correction. This is about 16 % of our complete SAC corpus.

Thematically, the corpus contains detailed mountaineering and travel reports (mostly from Switzerland, but also from abroad), historical and biological articles (flora and fauna of the Alps), geological and geographical studies (including frequent glacier observations), linguistic articles (e. g. on language boundaries in the Alps), and protocols of the annual club meetings. The text contains a huge number of proper names, geographical names, and Latin botanical names.

Our automatic sentence-based language identification assigns 5.5 million tokens to German and 0.74 million tokens to French. See Figure 3 for the distribution of these languages across yearbooks. Additionally, there are a few thousand tokens in English (mostly book and article titles), Italian, Swiss German, and Romansh.[3]

2.2. OCR

All the yearbooks from 1864 until 2000, we have collected in printed form. From 2001 until 2009 the SAC has provided us with PDF files, and since 2011 the SAC generates structured XML files directly out of their authoring system.

[1] http://pub.cl.uzh.ch/purl/OCR19thSAC
[2] http://textberg.ch

[3] Most of the 384 sentences (the vast majority) of the 19th century that were automatically classified as Romansh are in fact not Romansh. They are Latin or French or toponyms or OCR errors. Unfortunately, we cannot reliably detect Romansh sentences, the yearbook 1899, for instance, contains an article with a number of Romansh house inscriptions only few of which we identify correctly.

We obtained the first 10 yearbooks as leather-bound copies. Through collaboration with the Austrian Alpine Club (AAC) in Vienna, we were able to scan them without destroying them. All yearbooks from 1874 until 2000 were cut open so that we were able to use a normal scanner with paper feed. From 1957 onwards, the SAC has published parallel French and German versions of the yearbooks, both of which we processed in the same manner.

After scanning all book pages with 300 dpi, we used the OCR software Abbyy FineReader Pro 7 to convert the images to text (selecting the recognition languages German, French and Italian). This led to mixed text recognition results. The text on some pages was recognized excellently whereas other pages contained a multitude of OCR errors.

Our initial idea was to manually correct these errors in the OCR application since it preserves the mapping between words recognized in the text and the corresponding position on the page. But we soon realized that manual correction is very time-consuming even when working on well-recognized yearbooks of the 20th century. It is prohibitively time-consuming for the yearbooks of the 19th century, where recognition accuracy is inferior because of (a) words that are unknown to the OCR system lexicon (foreign words, toponyms, alpinistic special terms, person names, old spellings, dialect words), (b) special characters (fraction glyphs, old symbols), (c) more stains on the paper and curved pages. Generally, the corpus contains many OCR challenges such as tables, mathematical formulae, spaced type, or words in images.

We investigated various means of improving the OCR quality and correcting OCR errors automatically (Volk et al., 2010). There are only few ways in which a commercial OCR system can be tuned. The most obvious way is to add "unknown" words to its lexicon. In order to extend the coverage of the built-in lexicon, we collected words with old German spelling patterns (e. g. *acceptiren, acceptieren, Mittheilung*) and also added the names of 4000 Swiss mountains and cities. This led to some improvements of the OCR quality but a multitude of seemingly random OCR errors persisted.

Then we experimented with two ways of automatic error correction. First we employed a second OCR system (OmniPage) and compared the output of the two systems (Volk et al., 2010). Wherever they disagreed we checked with a German morphological system (Gertwol, see Koskeniemmi and Haapalainen (1996)) whether both words are known German words. If so, then we chose the word that occurred more frequently in our corpus. If only one of the words was known, then this was the obvious choice. If none of the words was known, then we trusted Abbyy FineReader as the more reliable system. This method also led to a small reduction of errors.

Finally we experimented with automatic error correction based on character similarities of words. If an unknown word deviates only in one or two characters from another known word which frequently occurs in our corpus, then we automatically substitute the unknown word with the known word. This method is similar to grammar checking as used in popular text processing software, but needs to work with high precision since human intervention (i. e. manual choice of the correct option) is not possible given the large amounts of text. Therefore we applied this method only for words with more than 15 characters.

2.3. Crowd Correction

It became obvious that we can only achieve a clean corpus if we organize a large distributed effort for correcting OCR errors via a crowd of volunteers. Therefore, we built the collaborative web-based correction system *Kokos*.[4] Kokos is based on the wiki idea and is actually technically built on top of *PmWiki*.[5]

User Interface We modified the wiki such that it displays the OCRed text of a page and the scan image side by side (see Figure 1). The text is an HTML export from the OCR software, and the layout, paragraphs and font sizes resemble the facsimile.

In the recognized text, each word is a clickable and editable unit. While reading through the text, Kokos correctors can simply click on faulty words in order to open a small editing window (Figure 1). In this window they can modify the word and save the correction. Quick access buttons help to insert frequently mis-recognized special characters, e. g. æ, ß, ¼, or Greek letters. The corrected word becomes visible immediately in the text. In addition to corrections within a word, three frequent operations are needed. A delete button removes spurious tokens caused by dirt on the page. Another button joins incorrectly split tokens into the edit window, for instance in the case of spaced type. Third, inadvertently connected words can be split by inserting a blank character.

When the editing window is open or when the user hovers over a word, the corresponding rectangle in the facsimile is highlighted. This is an important and motivating feature that allows the user to quickly spot and doublecheck a dubious word in the image. The positions of each word were computed by the OCR system during recognition. These coordinates provide the alignment between each word in the text and the corresponding area in the image.

In order to draw the reader's attention to words where the OCR software had low recognition confidence (that is, potential OCR errors), a blue font color was used. Unfortunately, the confidence values of the software are not really reliable, and therefore not helpful for the correction task.

Workflow In order to attract correctors to work on the task it is important to make initial access as easy as possible. In Kokos we allow all interested persons to read through the text by browsing and searching. It is

[4] http://kokos.cl.uzh.ch

[5] http://www.pmwiki.org

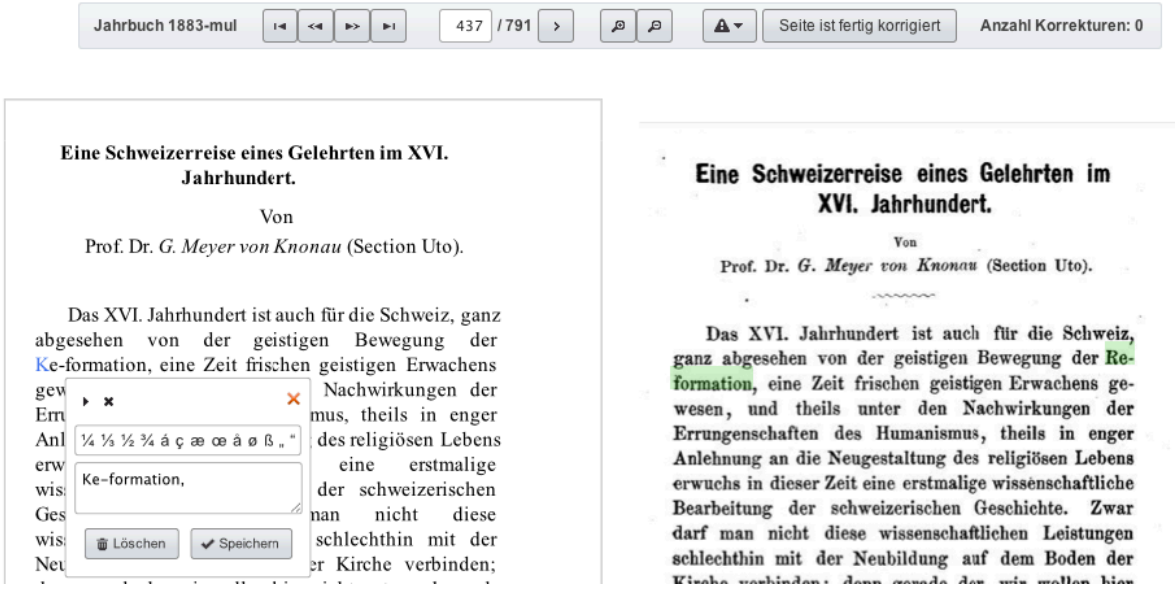

Figure 1: A book page in Kokos: synoptic view of the editable text on the left and the facsimile image on the right. Note the small edit window within text and the corresponding highlighted word in the facsimile.

then an easy step to register with user name, password and email address in order to sign up as volunteer corrector. The downside of this is that we know very little about our correctors. We kept our correction guidelines as short and concise as possible. It probably would have been a good idea to introduce the task and the correction guidelines with a short tutorial video.

Users can access the text through a table of contents sorted by yearbook, or a text search, or a "Quick Start" button that leads to a page without or only a few corrections, or an overview of finished and unfinished pages.[6] Our basic workflow is "correct errors while reading a text of interest". By clicking on a button users can mark a page as finished when they consider the text carefully corrected. This button will also automatically advance the view to the next page. Other users can still apply corrections to "finished" pages if need be.

Kokos also supports an orthogonal workflow via global search and replace, which includes a keyword-in-context view of the search results with facsimile image snippets of the search word (see Figure 2). This speeds up the correction of repeated recognition errors. In order to prevent users from introducing damage by accidental mass replacements, we limited the amount of global replacements to 15 hits per user interaction.

Crowd Management In January 2014, the SAC magazine (in all three language versions: French, German and Italian) published a call for volunteer helpers to correct the SAC heritage yearbooks. Dozens of users registered and started to work within days. After 7 months our active crowd had finished correcting all of the 21,000 pages. We observed a performance pattern which seems to be typical for crowd correction: there were not thousands of volunteers doing tiny bits of work (typical for paid micro-work crowdsourcing), but there was a small crowd of dedicated persons doing most of the work in collaborative and goal-oriented fashion.

In order to keep the top performers motivated and to give them feedback, a user ranking based on the number of corrections proved to be useful. In our opinion, this form of gamification is sufficient for volunteers who are inherently interested in a task. For community building, we regularly sent emails to the correctors, informing them about progress and system improvements.

In order to achieve consistent corrections, we provided concise guidelines and an FAQ section, which was regularly updated. Our correctors were very cooperative, and we never had to deal with vandalism. Our initial fears that we need to invest time to monitor the correction quality, or that a double correction of all pages would be necessary in order to achieve high quality turned out to be wrong.

On each page, the correctors were reminded to preserve the spelling of the printed text, even if it deviated from modern German orthography. Additionally, we asked them to perform *dehyphenation*, i. e. recomposing words that were hyphenated at a line break.

Via social media buttons the users could promote interesting pages to common social media channels. However, this feature was not used a lot by our volunteer correctors.

2.4. Related Work

The Recaptcha system (von Ahn et al., 2008) has earned fame for hiding crowdsourcing effort in OCR correction behind an access system to websites. Users are shown two image snippets where one is known to the system and used for verifying access to a website. The

[6] Especially in the last phase of correction, this view guided our volunteers to correct yearbooks completely.

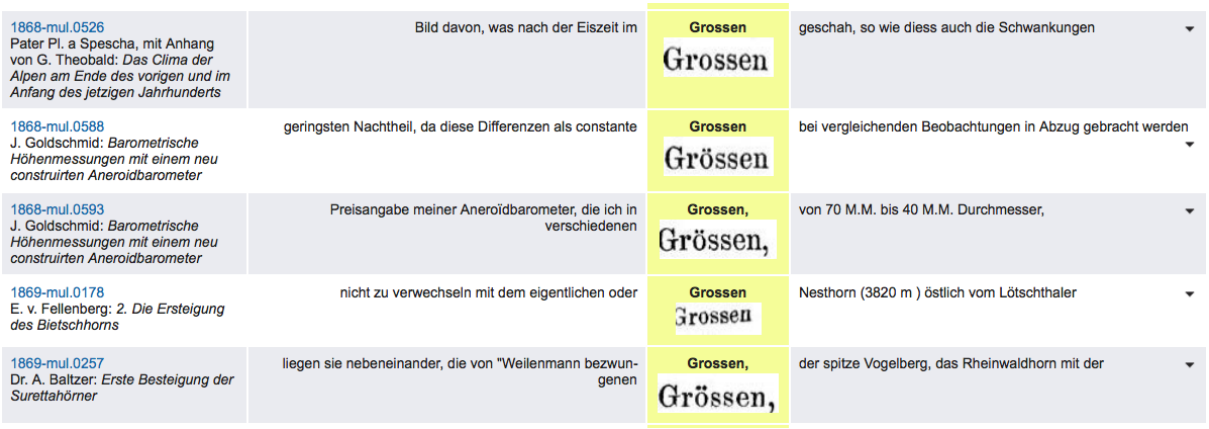

Figure 2: Search result in KWIC view with facsimile snippets for each hit for quick verification.

other is unknown and its text content will be determined by majority vote of many contributors. Of course users do not know which word is known and which is unknown.

The National Library of Australia has set up *trove*,[7] a system for crowd correction in historical newspapers (Holley, 2009). Meanwhile, user correction of digitized historical newspapers is successfully integrated in commercial platforms for digitization and document collection management such as *Veridian*,[8] a system which is used by several large libraries around the world.[9] See also Rose Holley's blog,[10] which lists five US historical newspapers that use crowdsourcing for OCR corrections, as well as one Australian, Finnish, Russian and Vietnamese newspaper.

Chrons and Sundell (2011) present Digitalkoot, a gamification-based system for correcting OCR errors in old Finnish newspapers typeset in gothic font. The words are taken out of context and inserted into simple games. The authors monitored the activities for the initial two months, in which 4800 persons tried out the games and completed 2.5 million micro tasks. This was the result of heavy media coverage with more than 30 newspaper articles and some TV programs reporting on the project. The authors remark that a small percentage of users provided one third of the work. The quality of the crowd corrections was very high and improved the text from 85 % word accuracy to over 99 %.

Yet another approach for involving the crowd into OCR correction reports Wang et al. (2013) for ancient Chinese books. They first extract graphically similar Chinese characters and present them to the users in a row for quick verification. This reduces the correction task to the question whether all logograms in a row are the same.

In the course of the IMPACT project (Tumulla, 2008), a collection of ground-truth texts were created from digitizing historical printed texts. The resource is advertised on the *Impact Centre of Competence*'s website;[11] however, it is only accessible to members.

3. Results

We investigated the crowdsourced corrections in two ways: with a quantitative analysis of the modifications, which is detailed in Section 3.1., and by evaluating the quality of the corrected texts in a representative sample that was checked separately by two persons (Section 3.2.).

3.1. Amount of Corrections

For assessing the amount of corrections, we compared two snapshots of the texts, taken at the start and at the end of the correction phase.

Identifiers hidden in the HTML export allowed us to easily align both versions at the paragraph level. In order to measure the effects of typical corrections in running text, we removed pages that are prone to errors originating from an early stage in OCR (particulary: Layout Recognition). Corrections of these high-level errors cannot be appropriately reflected by tools that focus on local changes (word/character level). We removed table-of-content pages (which had been manually corrected in the initial digitization phase already), pages with large tables or page-size images, and pages containing more than one language.[12] Furthermore, we discarded pages written in one of the sparsely-represented languages, i. e. languages other than French or German, and we removed individual paragraphs that were missing in one of the snapshots (which means that they were deleted or inserted in the correction phase). After this filtering, we were left with a set of just above 19,000 pages with a total of 111,000 paragraphs and 33.8 million characters.

We determined the amount of corrections by means of the modification rate between the two versions, which

[7] `http://trove.nla.gov.au`

[8] `http://www.veridiansoftware.com`

[9] For instance, the California Digital Newspaper Collection `http://cdnc.ucr.edu/cgi-bin/cdnc`

[10] `http://rose-holley.blogspot.ch/2013/04/crowdsourcing-text-correction-and.html`

[11] `http://www.digitisation.eu/tools-resources/image-and-ground-truth-resources/`

[12] As identified by our downstream processing pipeline.

is the character edit distance (Levenshtein, 1966) divided by the length of the corrected text. We computed the modification rate using the ISRI frontiers toolkit (Rice, 1996), which is a tool for analyzing modifications in OCR text. In total, the modifications in the French and German sentences affect 180,000 characters, which equals to an overall modification rate of 0.54 % (micro-average). For French the modification rate is 0.73 %, which is considerably higher than for German (0.52 %). The modification rate per paragraph (macro-average) shows an even more substantial difference between German (2.55 %, SD 18.0 %) and French (4.36 %, SD 23.3 %).

The difference between micro- and macro-average as well as the large variance indicate that a small number of paragraphs have an exceptionally high modification rate. Inspecting such cases revealed that the correctors occasionally had reorganized entire text regions or moved them from one paragraph to another, addressing failures of the OCR software where faulty layout recognition had disturbed the reading order (e. g. in tables that were not detected by our initial filtering method). Since our ID-based paragraph alignment did not expect text regions to be shifted, these corrections were counted as many character edit operations (character deletions in one place and insertions in another), even if the shifted text stayed the same.

As these macroscopic text editions have a distorting effect on the modification rate, we decided to exclude them from the evaluation. In an additional filtering step, we removed all paragraphs which showed a change in length of 10 % or more. This reduced the corpus size by less than 0.3 %, but the overall modification rate decreased by a fifth to 0.42 %, and the gap between German and French became smaller (0.42/0.46 % respectively in micro-average). As a side effect, this filtering also removed spurious paragraphs caused by spots or dirt.

Figure 3 shows the modification rates across all yearbooks, plotted against the text size. We found no correlation between the size of a yearbook and its modification rate, nor did we observe a clear tendency over time (correlation age–modification rate). Often, the modification rates for French and German develop in parallel. French tends to have a stronger amplitude, showing very low values for volumes with a low total rate and even much higher values for highly modified volumes. At least partially, the increased variability might be due to the relatively small amount of French texts, which gives more weight to individual outliers. Some of the slumps in the modification rate (e. g. 1890, 1899) can be attributed to correction efforts early in the digitization process, which were carried out using the user interface of the OCR software. This means that, occasionally, the text quality was already considerably improved before exporting into the online correction system, leaving less work to do for the crowd correctors.

A selection of frequent corrections is given in Table 1. All examples are misrecognized tokens that were corrected multiple times in different places. In many cases, the affected word posed increased challenges to the OCR system, in that it is not expected to be found in a dictionary that covers the general vocabulary of contemporary German or French. Often this is due to orthographic and linguistic variation, such as regional (Examples 13–17) and historical spelling (16–19) as well as outdated morphology (20–22), or because the word belongs to an open class, such as toponyms (23–27), and person names (28–31). Also, non-alphanumeric characters (10–12) and certain letters (e. g. upper-case *R*, see 7–9) are badly recognized.

OCR	corr.	
nnd	und	(1)
zn	zu	(2)
sieh	sich	(3)
Ton	von	(4)
Über	über	(5)
lieber	Ueber	(6)
Eichtung	Richtung	(7)
Kichtung	Richtung	(8)
Bedaktion	Redaktion	(9)
-f-	+	(10)
°/o	%	(11)
Va	½	(12)
Hessen	liessen	(13)
massig	mässig	(14)
Händen	Handen	(15)
Centralcomite	Centralcomité	(16)
Bureau	Büreau	(17)
Thaies	Thales	(18)
grossenteils	grossentheils	(19)
altern	ältern	(20)
Schütze	Schutze	(21)
Schlüsse	Schlusse	(22)
Eimer	Elmer	(23)
Gesehenen	Geschenen	(24)
Mordes	Morcles	(25)
Unterwaiden	Unterwalden	(26)
Bergeil	Bergell	(27)
Bubi	Dübi	(28)
Imfeid	Imfeld	(29)
111.	Ill.	(30)
Franche	Francke	(31)
Ueber-gang	Uebergang	(32)
all-mälig	allmälig	(33)
Schnee-und	Schnee- und	(34)

Table 1: Frequent word corrections.

From an natural language processing point of view, it is worthwhile to look at cases that are particularly hard to tackle in automated post-correction. As such, many corrections deal with *real-word* errors (3–6, 13–17, 19–25), i. e. tokens that match an existing word, which means that their erroneous nature can only be revealed through their context or by comparison with the facsimile. A similarly tricky issue is dehyphenation, which cannot be performed mechanically in a linguistically unaware fashion (see Example 32 vs. 34).

Table 2 shows the most frequent edit operations. Most

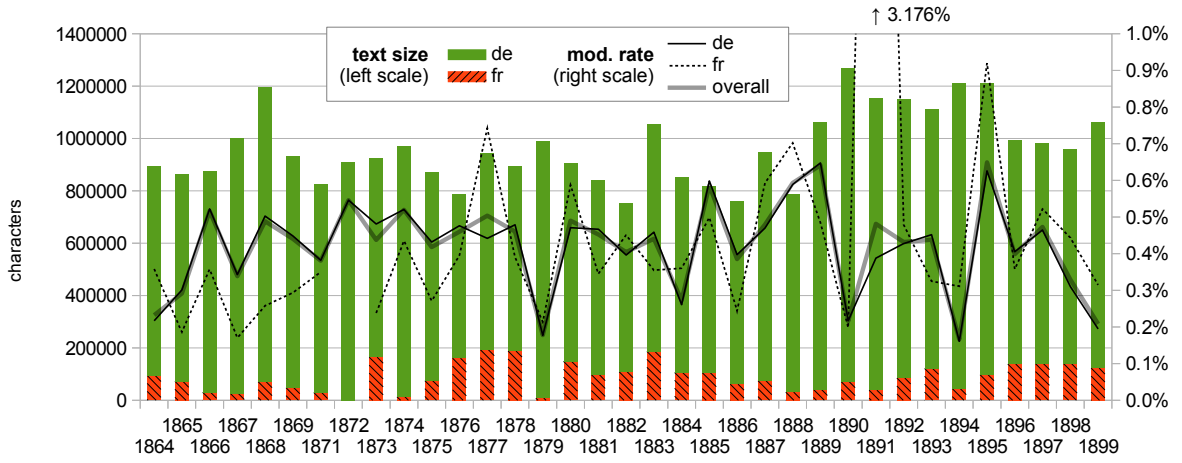

Figure 3: Text size and modification rate (micro-average) in the filtered data.

of the top modifications are concerned with fixing word boundaries through insertion and deletion of spaces and hyphens. Some operations had been carried out mechanically, such as the global replacement of quotation marks or the removal of squares and bullets. 35.9 % of the modifications are deletions of one or more characters (mostly punctuation and whitespace characters). Many corrections are related to letters with diacritic marks, which appear to be particularly challenging for OCR in a multilingual corpus. 7.6 % of the observed edit operations differ only by diacritics (e. g. a→ä or vice versa).

3.2. Quality of Corrections

In order to assess the quality of the corrected pages, we decided to carefully validate them using a sufficiently large sample of text snippets. We divided the corpus into snippets of approximately 100 characters (excluding whitespace), respecting word and page boundaries. We randomly sampled according to yearbook and language into 200 stratified folds with a size of approximately 1440 snippets each.

In order to determine the minimal required size of a sample for estimating the accuracy of the complete texts, we regarded the problem as an application of empirical probability of a character being misrecognized. Preliminary investigation suggested that the character error rate did not exceed 0.1 %. This confirmed that one fold (approx. 170,000 characters) constitutes a sample of sufficient size within a confidence level of ± 0.02 % ($p<0.01$).[13]

One fold was then separately proofread by two German native speakers with good knowledge of French. They were asked to correct the snippets according to the guidelines of the correctors. For each snippet, a cropped facsimile image was provided for collation.

5 % of the snippets were modified by at least one proofreader. Well over half of the modifications were done by both correctors in agreement, the rest was contributed by either of them in similar parts. All corrections were identical when detected by both, i. e. they never disagreed on how to correct an error, but only on its mere presence. It is most likely that the disagreements arose from varying attentiveness, rather than disparate judgment.

If corrections are very rare as in our case, the modifications can be modeled as a binary classification task (namely error *detection*) on the word level. Measuring the corrector agreement by Fleiss' κ (Fleiss, 1971) yielded a value of 0.67. Discussions between the proofreaders revealed that the guidelines were not detailed enough concerning whitespace (e. g. spaces separating the integer and fractional part in decimal numbers). By applying appropriate adjudication to these cases, κ raised to 0.73.

We then merged the proofread snippets into a single gold standard. Judging from the κ score, a few errors may have remained undetected, but we expect them not to be more than a handful, as the number of errors found by only one of the proofreaders and missed by the other was relatively small.

By comparing to the gold standard, we measured the spelling quality of the sample as corrected by the online collaborators. The crowd-corrected texts achieved a high accuracy of 99.71 % on the level of words and 99.93 % on the level of characters. Qualitatively, most of the remaining errors were hard-to-spot details, such as missing commas or diacritics (e. g. *avance*/*avancé*) or substitutions of similarly looking letters (e. g. *Clnbhütte*/*Clubhütte*, *Generalyersammlung*/*Generalversammlung*).

Based on the accuracy figures, we estimated the proportion of errors removed. The modification rate of the filtered corpus tells us that 0.42 % of the characters were edited. Under the assumption that every modification by the correctors contributed to an error correction, this is also the percentage by which accuracy was improved through the crowd corrections. Thus, the original character accuracy prior to the correction phase can be extrapolated to approximately 99.5 % (0.9993–0.0042). This means that an estimated initial error of

[13] $\frac{2.58^2}{0.0002^2} \times 0.001 \times (1 - 0.001) = 165706.54$

German			French		
freq.	OCR	corr.	freq.	OCR	corr.
13970	‹ ›	‹›	2024	‹ ›	‹›
10006	‹-›	‹›	701	‹e›	‹é›
7175	‹"›	‹“›	526	‹-›	‹›
3669	‹›	‹ ›	417	‹›	‹ ›
2644	‹i›	‹l›	372	‹"›	‹“›
1942	‹e›	‹c›	183	‹.›	‹›
1354	‹.›	‹›	141	‹-›	‹ ›
1147	‹K›	‹R›	128	‹ä›	‹à›
1146	‹E›	‹R›	126	‹,›	‹›
1079	‹u›	‹ü›	124	‹e›	‹è›
1070	‹"›	‹›	114	‹n›	‹„›
912	‹,›	‹›	113	‹i›	‹l›
847	‹B›	‹R›	112	‹e›	‹c›
824	‹li›	‹h›	98	‹é›	‹e›
783	‹-›	‹ ›	98	‹■›	‹›
782	‹U›	‹ü›	95	‹•›	‹›
766	‹n›	‹u›	94	‹'›	‹›
741	‹ö›	‹o›	91	‹E›	‹R›
722	‹ü›	‹u›	85	‹—›	‹-›
713	‹'›	‹›	82	‹ii›	‹ü›
655	‹■›	‹›	82	‹K›	‹R›
625	‹ä›	‹a›	80	‹'›	‹1›
604	‹'›	‹1›	76	‹™›	‹ m›
573	‹e›	‹é›	73	‹n›	‹u›
533	‹m›	‹rn›	50	‹"›	‹›
528	‹a›	‹ä›	49	‹»›	‹s›
496	‹Y›	‹V›	45	‹B›	‹R›
486	‹é›	‹e›	44	‹œ›	‹æ›
461	‹•›	‹›	42	‹›	‹.›
411	‹u›	‹n›	41	‹*›	‹1›
407	‹ii›	‹ü›	39	‹O›	‹0›
383	‹o›	‹ö›	38	‹V›	‹l'›
380	‹ii›	‹n›	38	‹*›	‹t›
358	‹ti›	‹ü›	37	‹- ›	‹›
353	‹™›	‹ m›	36	‹,›	‹.›
337	‹0›	‹O›	35	‹. ›	‹›
332	‹,›	‹.›	35	‹I›	‹l›
294	‹›	‹.›	34	‹se›	‹æ›
275	‹*›	‹1›	34	‹-›	‹—›
269	‹a›	‹u›	33	‹'›	‹›
268	‹—›	‹-›	33	‹è›	‹é›
236	‹«›	‹e›	32	‹li›	‹h›
229	‹ii›	‹u›	32	‹0›	‹O›
226	‹- ›	‹›	32	‹Y›	‹V›
225	‹B›	‹D›	32	‹'›	‹t›
221	‹tt›	‹ü›	32	‹e›	‹ê›
218	‹*›	‹›	31	‹.›	‹ ›
214	‹a›	‹n›	31	‹a›	‹s›
213	‹i›	‹›	27	‹k ›	‹›
202	‹»›	‹s›	27	‹'›	‹ ›

Table 2: Most frequent edit operations.

0.5 % was reduced to 0.07 % by our correctors, which is a reduction rate of 85 %.

4. Conclusion

We have shown that interested volunteers can effectively solve annoying OCR quality problems for the scientific community. We identified several success factors in our project. In our case we could recruit volunteers from an inherently interested community that additionally has a long tradition of citizen science. However, feedback about the correction progress and high-score rankings are also needed in order to keep motivation up.

While correcting our OCR errors in a heritage corpus of German and French texts from the Alpine domain, we created a large OCR gold standard which can be used as OCR training material, or as a resource for lexicon extraction. The estimated word-level accuracy of 99.7 % provides a good basis for evaluating systems that either process the scanned images of the pages or try to improve upon the output of a standard OCR system. We provide both resources in combination with the crowd-corrected gold standard.

We provide the textual portion of our OCR19thSAC corpus in three versions:

1. Complete corpus, without filtering, page-wise aligned with the scan images; best suited for training an OCR engine.

2. Corpus with page-wise filtering (as described in Section 3.1.), paragraph-wise aligned across snapshots; suitable for training a post-correction system.

3. Like 2, but with additional paragraph-wise filtering (based on change in length across snapshots), also suitable for post-correction training.

All versions are provided for both snapshots (OCR quality and crowd-correction quality), as UTF-8 encoded plain text under a Creative Commons Attribution 4.0 International License.[14]

Acknowledgements

We would like to thank our volunteer correctors who invested a lot of time and engagement into our crowd-correction project. We also thank Adrian Althaus and Matthias Fluor for implementing the Kokos web platform.

Bibliographical References

Chrons, O. and Sundell, S. (2011). Digitalkoot: making old archives accessible using crowdsourcing. In *Proceedings of the 2011 AAAI Workshop on Human Computation*, page 20–26.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Göhring, A. and Volk, M. (2011). The Text+Berg corpus: An alpine French-German parallel resource. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN 2011)*, Montpellier.

[14] See `http://pub.cl.uzh.ch/purl/OCR19thSAC` for download

Holley, R. (2009). How good can it get? analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine*, 15(3/4).

Koskeniemmi, K. and Haapalainen, M. (1996). GERTWOL – Lingsoft Oy. In Hausser, R., editor, *Linguistische Verifikation : Dokumentation zur Ersten Morpholympics 1994*, number 34 in Sprache und Information, pages 121–140. Niemeyer, Tübingen.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8).

Rice, S. V. (1996). *Measuring the Accuracy of Page-Reading Systems*. PhD thesis, University of Nevada, Las Vegas.

Tumulla, M. (2008). IMPACT: Improving access to text. *Dialog mit Bibliotheken*, 20(2):39–41. (German article).

Volk, M., Marek, T., and Sennrich, R. (2010). Reducing OCR errors by combining two OCR systems. In *ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH 2010, page 61–65.

von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M. (2008). reCAPTCHA: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468.

Wang, S., Wang, M., and Chen, K. (2013). Boosting OCR accuracy using crowdsourcing. In *Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts, An Adjunct to the Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing, November 7-9, 2013, Palm Springs, CA, USA*, volume WS-13-18 of *AAAI Workshops*. AAAI.