LIBER QUARTERLY
The Journal of the Association of European Research Libraries
Vol. 30 (2020) 1–20 | e-ISSN: 2213-056X

Ground Truth OCR Sample Data of Finnish Historical Newspapers and Journals in Data Improvement Validation of a re-OCRing Process

Kimmo Kettunen

University of Helsinki, National Library of Finland, Finland
kimmo.kettunen@helsinki.f , orcid.org/ 0000-0003-2747-1382

Mika Koistinen

University of Helsinki, National Library of Finland, Finland
j.m.o.koistinen@gmail.com, orcid.org/ 0000-0003-0471-314X

Jukka Kervinen

University of Helsinki, National Library of Finland, Finland
jukka.kervinen@helsinki.f

Abstract

The National Library of Finland (NLF) has digitized historical newspapers, journals and ephemera published in Finland since the late 1990s. The present collection consists of about 16.51 million pages mainly in Finnish and Swedish. Out of these about 7.64 million pages are freely available on the web site https:/ / digi.kansalliskirjasto.f / etusivu. The copyright restricted part of the collection can be used at six legal deposit libraries in different parts of Finland. The time period of the open collection is from 1771 to 1929. The last nine years, 1921–1929, were opened in January 2018.

This paper presents brief y the ground truth Optical Character Recognition data of about 500,000 words that has been compiled at the NLF for

Uopen Journals | http://liberquarterly.eu/ | DOI: 10.18352/lq.10322

development of an improved OCR process for the Finnish collection. We discuss compilation of the data generally and show results of the new OCR process in comparison to current OCR, using the ground truth data as an evaluation benchmark. We also show with real newspaper data of 30 years and 109 million words that the re-OCRing process is improving the quality of the OCRed data.

Keywords: OCR quality; ground truth data; evaluation; measurement; Finnish historical newspapers

1. Introduction

The National Library of Finland has digitized historical newspapers, journals and ephemera (small prints) published in Finland since the late 1990s. The digitized collection of NLF is part of a globally expanding network of historical data, produced by libraries that offers researchers and lay persons insight into the past. In 2012 it was estimated that there were about 129 million pages and 24,000 titles of digitized newspapers available in the web in Europe alone (Dunning, 2012). A very conservative estimation about the worldwide number of titles is 45,000 (The State of the Art, 2015). The number of currently available titles is probably much bigger, as the national libraries have been working steadily with digitization both in Europe, Northern America and the rest of the world.

Besides producing and publishing the digitized raw data all the time, the NLF has been involved in research and improvement of the digitized material during the last years. In September 2019 we ended a two year European Regional Development Fund (ERDF) project. NLF was also involved in the research consortium *Computational History and the Transformation of Public Discourse in Finland, 1640–1910* (COMHIS) that was funded by the Academy of Finland (2016–2019) and utilized the newspaper and journal data in its research of historical changes of publicity in Finland. We participate in and provide our data also for the EU project NewsEye[1] that started in May 2018.

One part of our data improvement effort has been the quality analysis of Finnish data. Out of this we have learned that about 70–75% of the words in the data are probably right and recognizable. In a collection of about 2.4 billion words[2] this means that 600–800 million word tokens are wrong

(Kettunen & Pääkkönen, 2016). This is a huge proportion of the words in the collection. The documents are shown to users as pdf files in the web presentation system, but also results of optical character recognition can be seen by the user in the user interface. We also provide the raw textual data as such for research use. OCR errors in the digitized newspapers and journals may have several harmful effects for users of the data. One of the most important effects of poor OCR quality – besides lower readability and comprehensibility — is worse on-line searchability of the documents in the collections. Also general usefulness and linguistic post-processing is harmed by OCR errors (Järvelin, Keskustalo, Sormunen, Saastamoinen, & Kettunen, 2016; Lopresti, 2009; Traub et al., 2016). Although users of the NLF collections have not complained about the quality much, its improvement is a natural first step in adding more value to the collection.[3]

In order to fulfill this mission, we started to consider re-OCRing of the data in 2015. The main reason for this was that the collection had been OCRed with a proprietary OCR engine, ABBYY FineReader (v.7 and v.8). Newer versions of the software exist, the latest being 15.0,[4] but the cost of the Fraktur font for OCR is too high a burden for re-OCRing the collection with ABBYY FineReader. We ended up using the open source OCR engine Tesseract v. 3.04.01 and started to train Fraktur font for it. This process and its results are described in detail in Koistinen, Kettunen, and Pääkkönen (2017), Koistinen, Kettunen, and Kervinen (2018) and in Kettunen and Koistinen (2019).

The rest of the paper is arranged as follows: section 2 introduces the data in the ground truth collection. Section 3 compares the results of the new OCR in the GT with the results of the current/old OCR using different measures and types of analysis. Finally, section 4 concludes the paper.

2. Data in the GT Collection

The main reason for setting up a re-OCRing procedure for a digitized text collection is usually bad or mediocre data quality of the collection. To properly evaluate the results of re-OCRing one needs to establish ground truth (GT) data[5] that can be used for comparing the old and the new OCRed data. For this purpose we chose manually a set of newspaper and journal pages that had Fraktur font, originating from different publications and decades. Our

budget for creation of the GT was minimal: we were able to pay for a subcontractor for the creation of the basic GT, but the budget was limited (about 4,000 €). This also limited the amount of data that could be used for the GT.

The final GT data consists of 479 pages of both journals and newspapers from the time period of 1836–1918. Most of the data is from 1870 onwards, as the majority of publications in the collection is from 1870–1910 (Kettunen & Pääkkönen, 2016). When the pages were picked, only the year of publication, type of publication (journal/newspaper), font type and number of pages and

Fig. 1: Number of characters in newspaper GT data for each year.

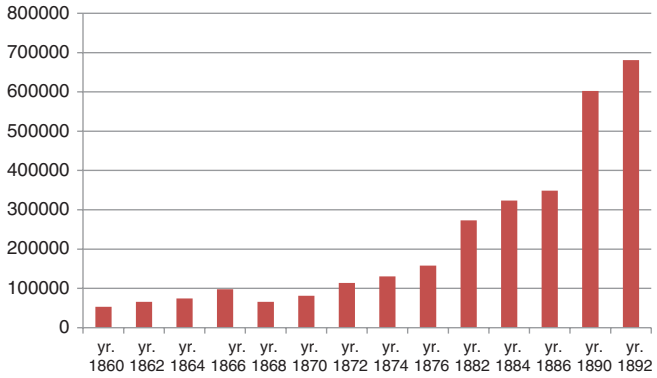


Fig. 2: Number of characters in journal GT data for each year.

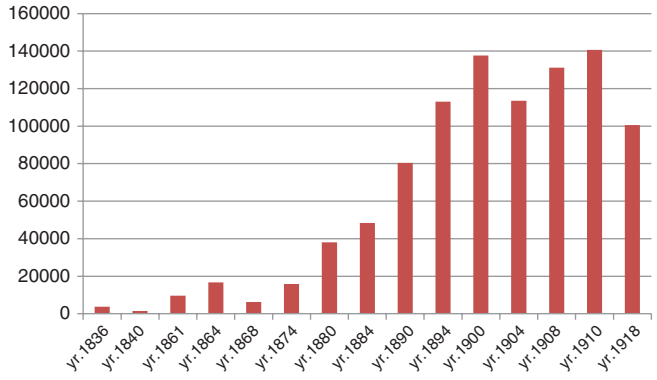


Fig. 3: Example of parallel GT data.

PUBTYPE	PUBYEAR	ISSN	PAGENAMI	SORTORDE	GT	TESSERACT	OLD	FR11
AIK	1884	fk00010	fk00010_1	3	taan	taan	taan	taan
AIK	1884	fk00010	fk00010_1	4	sekä	sekä	sekä	sekä
AIK	1884	fk00010	fk00010_1	5	ehdottoma	ehdottoma	ehdottoma	ehdottoma
AIK	1884	fk00010	fk00010_1	6	raittiita	raittiita	raittiita	raittiita
AIK	1884	fk00010	fk00010_1	7	että	että	että	että
AIK	1884	fk00010	fk00010_1	8	raittiuden	raittiuden	raittiuden	raittiuden
AIK	1884	fk00010	fk00010_1	9	harrastajia,	harrastajia,	harrastajia,	harrastajia,
AIK	1884	fk00010	fk00010_1	10	jotka	jotka	jotka	jotka
AIK	1884	fk00010	fk00010_1	11	tekewät	tekewät	tekemät	tekemät
AIK	1884	fk00010	fk00010_1	12	työtä	työtä	työtä	työtä
AIK	1884	fk00010	fk00010_1	13	raittiuden	raittiuden	raittiuden	raittiuden
AIK	1884	fk00010	fk00010_1	14	edestä,	edestä,	edestä,	edestä,
AIK	1884	fk00010	fk00010_1	15	katsoen	katsoen	katsoen	katsoen
AIK	1884	fk00010	fk00010_1	16	sitä	sitä	sitä	sitä
AIK	1884	fk00010	fk00010_1	17	tulewaisuu	tulewaisuu	tulemaisuu	tulemaisuu

characters was known of the data. In the final selection 56% of the pages are from journals, 44% from newspapers. Journal data has about 950 K of characters, newspaper data 3.06 M. Figures 1 and 2 show the amounts of characters in newspaper and journal GT data for different years.

Figure 3. shows an excerpt of the GT data. Information includes the type of the publication (AIK is a journal, SAN – not shown in the figure – is a newspaper), year of publication, ISSN of the publication and page information of the page image file. GT, Tesseract, Old (ABBYY FineReader 7/8) and FR11 (ABBYY FineReader 11[6]) are different OCR versions of the data.

The final ground truth text was corrected manually in two phases: the first correction was by a subcontractor from the output of ABBYY FineReader 11 and the final correction was performed in house at the National Library of Finland. The resulting GT is not errorless, but it is the best reference available. The final data used for this paper has 471,903 parallel lines[7] of words or character data. The words in the GT have 3,290,852 characters without spaces, including punctuation, and 4,234,658 characters with spaces. Medium length of the words is 6.97 characters.

The size of the data seems relatively small in comparison with the overall size of the collection which was 1,063,648 pages of Finnish newspapers and

journals at the time of creation. With regards to limited means, however, the size can be considered adequate for our purposes. It is far from the one per cent of the original data that Tanner, Muñoz and Ros (2009) used for error rate counting with 19th century British newspapers, but it is also much larger than typical OCR research paper evaluation data sets. Berg-Kirkpatrick and Klein (2014) use 300–600 lines of text, Drobac, Kauppinen, and Lindén (2017) 9,000–27,000 lines of text in their re-OCRing trials as evaluation data. Silfverberg, Kauppinen, and Lindén (2016) use 40,000 word pairs in postcorrection evaluation and Kettunen (2016) uses 3,800–12,000 word pairs. Dashti (2018) uses about 300,000 word tokens for evaluation of a real-word error correction algorithm. The ICDAR Post-OCR Text Correction 2017 competition uses a dataset of more than 12 million characters of English and French.[8] In comparison to current usage in the field, our 471,903 words and 3,290,852 characters can be considered a medium sized data set.

3. Comparison of New OCR to GT and Old OCR

We have described the components of the re-OCRing process and its evaluation thoroughly in Koistinen et al. (2017, 2018) and Kettunen and Koistinen (2019). Here we discuss only the evaluation results of the re-OCR process using the GT data.

Basic statistics of the data show that 85.4% of the words in Tesseract's output are identical to words of the ground truth. In the old OCR this figure is 73.1% and in ABBYY FineReader v.11 79%.

We have performed different analyses for the data and have found that the new Tesseract OCR is clearly better than the old ABBYY Finereader v.7/8 OCR in all respects. Tesseract OCR is also better than ABBYY FineReader v. 11 OCR for the same data (Koistinen et al., 2017, 2018). Table 1 shows recognition results of the data with two automatic morphological analyzers, Omorfi[9] and a version of Omorfi[10] that has some enhanced capability to recognize 19th century Finnish. We call this version HisOmorfi. We have earlier used morphological analyzers to get an overall picture of the word level correctness of the data in Kettunen and Pääkkönen (2016) and Kettunen, Pääkkönen, and Koistinen (2016) without available ground truth. Although the method is prone to estimation errors, it gives a good enough analysis of the data and it is easy to use.

Table 1. Recognition rates for different comparable data: 471,903 words.

	Ground truth version	Tesseract OCR version	Current (old) OCR version	FR11 OCR version
Omorfi 0.3	88,413 unrecognized words	102,507 unrecognized words	107,838 unrecognized words	69,461 unrecognized words
	81.26% recognition rate	78.27% recognition rate	77.14% recognition rate	85.28% recognition rate
HisOmorfi	24,054 unrecognized words	47,747 unrecognized words	89,800 unrecognized words	65,984 unrecognized words
	94.9% recognition rate	89.88% recognition rate	80.97% recognition rate	86.01 recognition rate

Plain Omorfi recognizes Tesseract words slightly better than the words of current OCR, the difference being 1.13% units. The seemingly small difference is caused by the fact that HisOmorfi was used in the re-OCRing process to choose words from output of Tesseract and it favors *w* to *v*;[11] thus more words with *w* than *v* are produced in the process. The old OCR words have 27,127 w's, Tesseract OCR words 64,180, GT 74,046 and FR11 only 3,732. Plain Omorfi does not recognize most of the words that include *w,* but HisOmorfi is able to recognize them, which is shown in the high recognition percentage in Tesseract's and GT's HisOmorfi result column. The words OCRed with Tesseract achieve almost a 9% unit improvement in recognition with HisOmorfi compared to current OCR.

3.1. Precision, Recall and F-score

The GT data allows the usage of other evaluation measures, too. We can use for example standard measures of recall and precision and their combination, F-score (Manning & Schütze, 1999, pp. 267–270; Märgner & El Abed, 2014), to get an overall picture of the results. These measures that originate from information retrieval evaluation have been used in both postcorrection and re-OCRing evaluations. Other similar measures exist, too, but many of them, as for example correction rate (CR) used in Silfverberg et al. (2016), are closely

related to P/R scores and based on the same basic ideas. Recall and precision measures are useful also in the sense that they allow more detailed analysis of the results.

The re-OCRed data consists of four different types of words: 1) true positives (TP) are originally wrongly OCRed words that are corrected in the re-OCRing; 2) false positives (FP) are correct words that are changed to a misspelling in the re-OCRing; 3) false negatives (FN) are wrongly spelled words that are still wrong after the re-OCRing; 4) true negatives (TN) are correct words that are correct after the re-OCRing.

Out of these we define Recall, R, as *TP / (TP+FN)*, Precision, P, as *TP / (TP+FP)* and F-score, F, as *2*R*P / (R + P)* (Manning & Schütze, 1999, pp. 268–269). Correction rate, a novel and slightly modified metric, used in Silfverberg et al. (2016), is defined as (TP-FP)/(TP+FN).

Table 2 shows P/R results and F-scores of the re-OCRed data and the correction rate for the data. We show two results: one in the left column is a comparison of the data without cleaning. The result in the right column shows results with punctuation and all other non-alphabetic and non-number characters removed from the lines. Removed character set is: .,;\':\"\'_!@#%&*()+=<>[]{}?\\/—~ | ^\"„ ¦ «©»®°¡.

We concentrate on the analysis of the left column results in more detail from now on. The number of erroneous words in the data is 126,758 (and errorless thus 345,145). Re-OCRing corrects 90,877 of errors (true positives, 71.7% of errors) and leaves 35,881 uncorrected (false negatives, 28.3% of errors). It also

Table 2. P/R results of re-OCRing in comparison to old OCR.

Basic results with parallel columns	Results without non-alphabet data
Recall = 0.72	Recall = 0.74
Precision = 0.73	Precision = 0.77
F-score = 0.73	F-score = 0.75
Correction rate = 0.46	Correction rate = 0.51

produces 32,953 new errors to the data (false positives). In general it seems thus, that the recall of the re-OCRed data with regards to erroneous words is satisfactory, but precision is low, as the process produces quite a lot of new errors. This harms the overall result.

In comparison, a simple Levenshtein distance based postcorrection algorithm used in Kettunen (2016) for small data samples of 3,850 – 12,000 word pairs had usually a high precision of 0.85–0.95, but much lower recall than our re-OCRing process. With the current data set the postcorrection algorithm achieves recall of 0.47, precision of 0.42 and F-score of 0.44. If non-alphabetic data is pruned from the data, the F-score is 0.57. The postcorrection algorithm handles only lower case characters, which affects its results. If case distinction is omitted in words and non-alphabet data pruned, postcorrection algorithm's best F-score is 0.63.

3.1.1. False and True Positives

Recall and precisions figures give an overall picture of the improvements in the re-OCRing process. In order to get a more detailed view of the process, one needs to examine the set of false and true positives more closely: what are the most frequent errors, what kind of errors are corrected, what new errors generated. In our case part of the false positives of the re-OCRed data is due to the recurring trouble with quote marking or division of the word on two lines when it ends with a hyphen. These data, when re-OCRed, miss a quote or two in the result word, or it contains the HTML code *&quote;* instead of the quote itself. Many words are also incorrectly divided on the line. The same applies to false negatives, too. The number of all faulty word divisions in the data of false and true positives together is about 10,000, which makes this error type one of the most common. Missing punctuation or extra punctuation also causes errors. This can be seen in the right column of Table 2 where results with cleaned output are shown.

When true positives are examined, one can see that about 54% of the errors corrected are one character corrections and about 89% are 1–3 character corrections. But re-OCR corrects also truly hard errors, where more than three characters are corrected. Even errors with a Levenshtein distance

Table 3: Corrections of Levenshtein distance of 11.

Original OCR	Tesseract 3.04.01
eiifuroauffellt»	esikuwauksellisesti
KarjlltijoloSluSyhbiStytsen	Karjanjalostusyhdistyksen
ttfcnfäMtämifeSfä,	itsensäkieltämisessä,
liiannfiljtccvillc	maansihteerille

Table 4: Corrections of Levenshtein distance of 5.

Original OCR	Tesseract 3.04.01
fofoufsessct,	kokouksessa
silmciyfsert	silmäyksen
ncihbessciän	nähdessään
roäliHä	wälillä.
yfsincicin.	yksinään
tylyybestcicin	tylyydestään
fitsattbestaan,	kitsaudestaan.
Iywäzlyllln	Jywäskylän
pairoana	päiwänä

(Levenshtein, 1966)[12] (LD) over 10 are corrected, a few examples being the word pairs of edit distance of 11 in Table 3.

Another example of corrected hard errors are 2,376 words that have a Levenshtein edit distance of five. When the error count is this high, words are becoming unintelligible. Some examples of corrections with five errors are shown in Table 4.

The bigger the error count is, the harder the error would be to correct for a postcorrection software, and here lies the strength of re-OCRing at its best. Reynaert (2016), e.g., states that his postcorrection system of Dutch, TICCL, corrects best errors of LD 1–2. It can be run with LD 3, "but this has a high processing cost and most probably results in lower precision." Error correction for LD 4 and higher values he considers too ambitious for the time being. This is also one of the conclusions in Choudhury, Thomas, Mukherjee, Basu, and Ganguly (2007).[13]

The number of corrected words with edit distances of 1–10 in true positives of our re-OCR process can be seen in Table 5.

Table 5: Number of corrected words with edit distances of 1–10: 99.2% of all the true positives.

Edit distance	Number of corrections
LD 1	47,783
LD 2	22,713
LD 3	9,182
LD 4	4,375
LD 5	2,376
LD 6	1,519
LD 7	920
LD 8	629
LD 9	423
LD 10	315
	SUM = 90,235 (of 90,877 total true positives)

3.2. Further Analysis of Results

Overall, the sum of character errors in the data decreased from old OCR's 293,364 to 220,254 in Tesseract OCR, which is about a 25% decrease. Tesseract produces significantly more errorless words than the old OCR (403,069 vs. 345,145), but it produces also more character errors per erroneous word. The old OCRing has about 2.32 errors per erroneous word, Tesseract OCR 3.2. This is a mixed blessing: erroneous words are encountered more seldom in Tesseract's output, but they may be harder to read and understand when they occur.

Mean length of the word tokens – including punctuation – in different versions of OCR does not vary much: in the current OCR it is 6.94 characters, in GT 6.97 and in Tesseract OCR 6.99 characters. The length of words does not bring great variance to improvement of OCR. Words that are up to seven characters long (total of 286,066) in the current OCR get F score of 0.72 and correction rate of 0.44. Words that are longer than seven characters (total of 185,387) get F score of 0.73 and correction rate of 0.47.

Frequency analysis of characters in different versions of the OCR does not show significant differences in alphabetical characters between GT and Tesseract. Tesseract seems to produce too many zeros and ones out of numbers and in other characters dash and backslash are over generated.

The number of different word types (unique words) in the current OCR data is 176,625. In GT data it is 135,433 and in Tesseract OCR data it is 156,459. The

Table 6: Combined P/R and corrections rate results of Tesseract and ABBYY FineReader 11.

Basic results: Tesseract only	Results with Tesseract + FR11
Recall = 0.72	Recall = 0.81
Precision = 0.73	Precision = 0.95
F measure= 0.73	F measure= 0.88
Correction rate = 0.46	Correction rate = 0.77

number of hapax legomena, that is words occurring only once, is 97,330 in GT, 120,878 in Tesseract OCR, and 140,802 in current OCR. The bigger number of unique words is one clear sign of more errors in the word data (Ghosh, Chakrabortya, Parui, & Majumder, 2016).

3.3. Combined OCR Results

Usage of combined results of several OCR software has proven fruitful in many evaluations (e.g. Klein & Kopel, 2002; Volk, Furrer, & Sennrich, 2011). As we have in our GT data results of another OCR software, ABBYY FineReader v.11, we can also evaluate the combined optimal results of Tesseract and ABBYY FineReader v.11. Recall of the optimal result of two combined OCR engines is 0.81, precision 0.95, F score 0.88 and correction rate 0.77 as shown in Table 6 in comparison to Tesseract's results only. Unfortunately we do not have available the other OCR engine for final re-OCRing, therefore we can only show upper limits for the results with these two engines.

3.4. Upper and Lower Case Characters

Upper and lower casing is a basic distinction in the Latin alphabet writing systems, and OCRing should maintain the distinction. We analyzed the effect of word initial capitalization on the results. If capitalization is neutralized from the data, results are almost the same. Thus it seems that the re-OCRing process is recognizing upper and lower case letters well.

Figures 4 and 5 show the distribution of upper and lower case letters of the Finnish basic alphabet in the ground truth and Tesseract data. Rare characters like *ü*, *á* etc. that occur only in foreign words are left out of the figures.

Fig. 4: Upper case letters of ground truth and Tesseract.

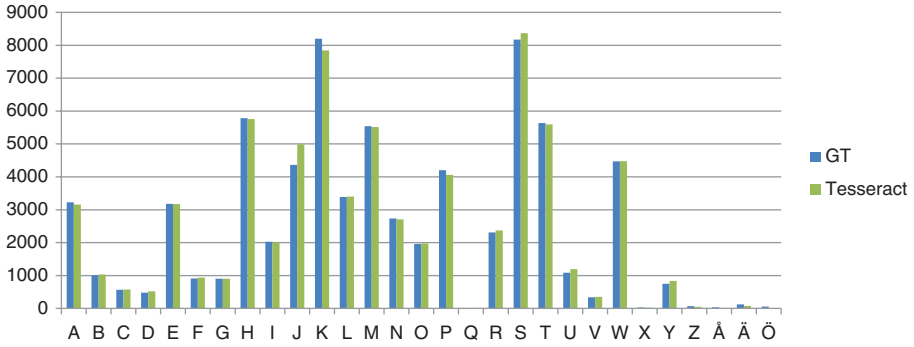
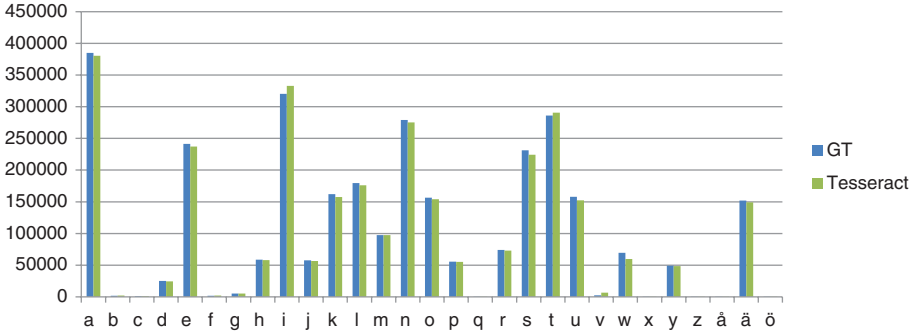


Fig. 5: Lower case letters of ground truth and Tesseract.



As can be seen from the figures, there are no enormous drops or spikes in recognition of any letter. Thus the OCRing process seems to handle the main characters of the Finnish alphabet quite consistently.

3.5. Stepping Outside of the Sandbox

Usage of a GT collection in OCR improvement is of vital importance. It can, however, have some drawbacks. Firstly, the collection may not be as representative as it should. Secondly, usage of the GT collection during development and evaluation may lead to an over-fit of data. To circumvent these

Table 7: Recognition rates of current and new OCR words of Uusi Suometar with morphological analyzer HisOmorfi.

Year	Words	Current OCR	Re-OCR	Gain in % units
1869	658,685	69.42%	86.59%	17.18%
1870	655,772	66.98%	85.54%	18.57%
1871	910,707	72.81%	87.27%	14.46%
1872	930,493	75.09%	88.25%	13.16%
1873	892,745	74.61%	87.04%	12.43%
1874	921,603	72.70%	86.09%	13.39%
1875	1,075,339	70.62%	85.52%	14.90%
1876	1,223,455	71.50%	85.51%	14.01%
1877	1,818,803	72.09%	84.79%	12.70%
1878	2,193,869	70.78%	84.70%	13.91%
1879	2,238,412	73.52%	86.09%	12.57%
1880	2,135,334	70.11%	85.85%	15.74%
1881	2,617,533	67.98%	84.26%	16.28%
1882	2,736,109	62.41%	82.94%	20.53%
1883	3,182,853	70.19%	82.17%	11.98%
1884	3,365,356	69.60%	81.67%	12.07%
1885	3,965,632	68.11%	82.53%	14.42%
1886	4,247,173	68.21%	82.12%	13.92%
1887	4,393,615	65.25%	82.16%	16.91%
1888	5,030,160	70.27%	82.52%	12.25%
1889	5,152,628	65.71%	81.41%	15.70%
1890	5,676,613	64.69%	80.71%	16.02%
1891	6,275,418	65.16%	81.21%	16.05%
1892	6,372,156	62.01%	80.92%	18.91%
1893	6,331,905	60.62%	80.16%	19.54%
1894	6,618,095	66.63%	82.10%	15.47%
1895	6,485,491	67.96%	82.10%	14.14%
1896	6,802,715	64.29%	81.43%	17.14%
1897	7,366,360	61.69%	80.14%	18.45%
1898	7,113,723	63.87%	80.50%	16.63%
Total	109,388,752			Average 15.3%

possible effects, we show also quality improvement outside the GT data. After initial development and evaluation of the re-OCRing process with the GT data, we started final testing of the re-OCRing with newspaper data. We chose for testing *Uusi Suometar*, a newspaper which appeared in 1869–1918 and has 86,068 pages. Table 7 shows results of a 30 years' re-OCRing of this newspaper. Word level recognition rates using morphological analyzer are given for the old and the new OCR.

Re-OCRing is improving the quality of the newspaper clearly and consistently. The average improvement for the whole period of 30 years is 15.3% units. The largest improvement is 20.5% units, and the smallest 12% units. Although the usage of morphological recognition is no guarantee of the rightness of the result, these big improvements in recognition rate are a clear indication of quality improvement.

4. Conclusion

We have described in this paper generally our Optical Character Recognition GT sample for Finnish historical newspapers and journals. The data consists of 479 pages and 471,903 parallel words. It has been used in development and evaluation of a new OCRing process for our collection's Finnish Fraktur font part using Tesseract's open source OCR engine v. 3.04.01. According to our evaluation results, we can achieve a clear improvement on the OCR quality with Tesseract in the 500K GT data (Koistinen et al., 2017, 2018). All our analyses show that the re-OCR procedure works relatively well: it does not shorten or lengthen words significantly and it reduces the number of word types in Tesseract OCR in comparison to current OCR. Recognition of the produced words by morphological analyzers is improved with 9% units and P/R figures of the correction effect of the re-OCR are satisfactory. 89% of the corrections made to the words are corrections of 1–3 characters.

The GT data has been created as a tool for quality control of the re-OCRing process. We have published the word lists, ALTO XML and image files of the data on our web site *digi.kansalliskirjasto.fi/opendata* as open data. We have earlier published the text files of the collection's 1771–1910 part (Pääkkönen, Kervinen, Nivala, Kettunen, & Mäkelä, 2016) with metadata, ALTO XML and plain text. Publication of the GT data benefits those, who work on OCRing historical Finnish or who develop postcorrection algorithms for OCRing. Also development work of general OCR tools such as Transkribus[14] may benefit from the data. Earlier we have given the GT data for research use on demand, and it has been used in training of Ocropy OCR engine for the historical data (Drobac et al., 2017).

The old saying in computational linguistics is that *more data is better data*, and that applies in the case of OCR data too. It would have been nice to have an

even larger OCR GT data set, but with regards to resources at use, we are contented with the now available data. The data adds a useful resource for repertoire of somehow under-resourced collections of 19th century Finnish. We hope the data has use also outside of OCR and postcorrection field for those who work in the digital humanities.

Acknowledgements

This work was supported by the Academy of Finland as part of the project Computational History and Transformation of Public Discourse in Finland, 1640–1910, decision number 293341.

References

Berg-Kirkpatrick, T., & Klein, D. (2014). Improved typesetting models for historical OCR. In K. Toutanova & H. Wu (Eds.), *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics* (pp. 118–123). https://doi.org/10.3115/v1/P14-2020.

Carrasco, R. C. (2014). An open-source OCR evaluation tool. In A. Antonacopoulos & K. U. Schulz (Eds.), *Proceedings of the first international conference on digital access to textual cultural heritage (DATeCH '14)* (pp. 179–184). New York: ACM. https://doi.org/10.1145/2595188.2595221.

Choudhury, M., Thomas, M., Mukherjee, A., Basu, A., & Ganguly, N. (2007). How difficult is it to develop a perfect spell-checker? A cross-linguistic analysis through complex network approach. In C. Biemann, I. Matveeva, R. Mihalcea, & D. Radev (Eds.), *TextGraphs-2: Graph-based algorithms for natural language processing – Proceedings of the workshop* (HTL-NAACL 2007) (pp. 81–88). New Brunswick, NL: Association for Computational Linguistics. https://arxiv.org/pdf/physics/0703198.pdf.

Dashti, S. M. (2018). Real-word error correction with trigrams: Correcting multiple errors in a sentence. *Language Resources and Evaluation, 52*, 485–502. https://doi.org/10.1007/s10579-017-9397-4.

Drobac, S., Kauppinen, P., & Lindén, K. (2017). OCR and post-correction of historical Finnish texts. In J. Tiedermann (Ed.), *NoDaLiDa, Proceedings of the 21th Nordic conference on computational linguistics* (pp. 70–76). Linköping: Linköping University Electronic Press. Retrieved January 23, 2020, from https://www.aclweb.org/anthology/W17-0209.pdf.

Dunning, A. (2012). *European newspaper survey report*. Retrieved January 23, 2020, from http://www.europeana-newspapers.eu/wp-content/uploads/2012/04/D4.1-Europeana-newspapers-survey-report.pdf.

Ghosh, K., Chakrabortya, A., Parui, S. K., & Majumder, P. (2016). Improving information retrieval performance on OCRed text in the absence of clean text ground truth. *Information Processing and Management*, *52*(5), 873–884. https://doi.org/10.1016/j.ipm.2016.03.006.

Järvelin, A., Keskustalo, H., Sormunen, E., Saastamoinen, M., & Kettunen, K. (2016). Information retrieval from historical newspaper collections in highly inflectional languages: A query expansion approach. *Journal of the Association for Information Science and Technology*, *67*(12), 2928–2946. https://doi.org/10.1002/asi.23379.

Kettunen K. (2016). Keep, change or delete? Setting up a low resource OCR post-correction framework for a digitized old Finnish newspaper collection. In D. Calvanese, D. De Nart, & C. nTasso (Eds.), *Digital libraries on the move (IRCDL 2015)* (Vol. 612, pp. 95–103). Communications in Computer and Information Science. Cham, CH: Springer. https://doi.org/10.1007/978-3-319-41938-1_11.

Kettunen, K., & Pääkkönen, T. (2016). Measuring lexical quality of a historical Finnish newspaper collection – Analysis of garbled OCR data with basic language technology tools and means. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, …, S. Piperidis (Eds.), *Proceedings of the tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 956–961). Retrieved January 23, 2020, from http://www.lrec-conf.org/proceedings/lrec2016/pdf/17_Paper.pdf.

Kettunen, K., Pääkkönen, T., & Koistinen, M. (2016). Between diachrony and synchrony: Evaluation of lexical quality of a digitized historical Finnish newspaper and journal collection with morphological analyzers. In I. Skadiņa & R. Rozis (Eds.), *Human language technologies – The Baltic perspective. Proceedings of the seventh international conference* (Baltic HLT 2016) (pp. 122–129). Amsterdam: IOS Press. Retrieved January 23, 2020, from http://ebooks.iospress.nl/volume/human-language-technologies-the-baltic-perspective-proceedings-of-the-seventh-international-conference-baltic-hlt-2016.

Kettunen, K., & Koistinen, M. (2019). Open Source Tesseract in Re-OCR of Finnish Fraktur from 19th and early 20th century newspapers and journals – Collected notes on quality improvement. In C. Navarretta, M. Agirrezabal, & B. Maegaard (Eds.), *Proceedings of the Digital Humanities in the Nordic countries 4th conference (DHN2019)* (pp. 270–282). Retrieved January 23, 2020, from http://ceur-ws.org/Vol-2364/25_paper.pdf.

Klein, S. T., & Kopel, M. (2002). A voting system for automatic OCR correction. In J. Callan, P. Kantor, & D. Grossmann (Eds.), *Proceedings of the SIGIR 2002 Workshop on information retrieval and OCR: From converting content to grasping meaning* (n.p.). Retrieved January 23, 2020, from http://boston.lti.cs.cmu.edu/callan/Workshops/IR-OCR-02/tklein.pdf.

Koistinen, M., Kettunen, K., & Pääkkönen, T. (2017). Improving optical character recognition of finnish historical newspapers with a combination of Fraktur & Antiqua models and image preprocessing. In J. Tiedermann (Ed.), *NoDaLiDa, Proceedings of the 21th Nordic conference on computational linguistics* (pp. 277–283). Linköping: Linköping University Electronic Press. Retrieved January 23, 2020, from http://www.ep.liu.se/ecp/131/038/ecp17131038.pdf.

Koistinen, M., Kettunen, K., & Kervinen, J. (2018). Bad OCR has a nasty character – re-OCRing historical Finnish newspaper material 1771–1910. Submitted to *International Journal of Document Recognition and Analysis.*

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Lopresti, D. (2009). Optical character recognition errors and their effects on natural language processing. *International Journal on Document Analysis and Recognition, 12*, 141–151. https://doi.org/10.1007/s10032-009-0094-8.

Mäkelä, Eetu. (2016). LAS: An integrated language analysis tool for multiple languages. *The Journal of Open Source Software, 1*(6):35, 1–2. https://doi.org/10.21105/joss.00035.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: The MIT Press.

Märgner, V., & El Abed, H. (2014). Tools and metrics for document analysis system evaluation. In D. Doermann & K. Tombre (Eds.), *Handbook of document image processing and recognition* (pp. 1011–1036). London: Springer Verlag.

Pääkkönen, T., Kervinen, J., Nivala, A., Kettunen, K., & Mäkelä, E. (2016). Exporting Finnish digitized historical newspaper contents for offline use. *D-Lib Magazine*, *22*(7/8), n.p. https://doi.org/10.1045/july2016-paakkonen.

Pletschacher, S., Clausner, C., & Antonacopoulos, A. (2015). Europeana newspapers OCR workflow evaluation. In B. Coüasnon, V. Märgner, V. Frinken, & B. Barrett (Eds.), *HIP '15, Proceedings of the 3rd International workshop on historical document imaging and processing* (pp. 39–46). New York: ACM Digital Library. https://doi.org/10.1145/2809544.2809554.

Reynaert, M. (2016). OCR Post-correction evaluation of early Dutch books online – Revisited. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, …, S. Piperidis (Eds.), *Proceedings of the tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 967–974). Retrieved January 23, 2020, from https://pure.uvt.nl/ws/portalfiles/portal/14518959/LREC2016.EDBOeval.FinalSubmittedVersion.redownloaded20160318.pdf.

Silfverberg, M., Kauppinen, P., & Lindén, K. (2016). Data-driven spelling correction using weighted finite-state method. In B. Jurish, A. Maletti, U. Springmann, & K.-M. Würzner, *Proceedings of the ACL Workshop on Statistical NLP and Weighted Automata*

(pp. 51–59). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved January 23, 2020, from https://aclweb.org/anthology/W/W16/W16-2406.pdf.

The "State of the Art": A Comparative analysis of newspaper digitization to date (2015). Retrieved January 23, 2020, from https://www.crl.edu/sites/default/files/d6/attachments/events/ICON_Report-State_of_Digitization_final.pdf.

Tanner, S., Muñoz, T., & Ros, P. H. (2009). Measuring mass text digitization quality and usefulness. Lessons learned from assessing the OCR accuracy of the British Library's 19th Century Online Newspaper Archive. *D-Lib Magazine, 15*(8), n.p. https://doi.org/10.1045/july2009-munoz.

Traub, M. C., Samar, T., Ossenbruggen, J. van, He, J., Vries, A. de, & Hardman, L. (2016). Querylog-based assessment of retrievability bias in a large newspaper corpus. In J. S. Downie & R. H. McDonald (Eds.), *JCDL '13, 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 7–16). New York: ACM. https://doi.org/10.1145/2910896.2910907.

Volk, M., Furrer, L., & Sennrich, R. (2011). Strategies for reducing and correcting OCR error. In C. Sporleder, A. van den Bosch, & K. Zervanou (Eds.), *Language technology for cultural heritage – Selected papers from the LaTeCH workshop series* (pp. 3–22). Berlin: Springer. https://doi.org/10.1007/978-3-642-20227-8_1.

Notes

[1] https://www.newseye.eu/.

[2] This estimation is based on the period 1771–1910. Pletschacher, Clausner and Antonacopoulos (2015) report a word accuracy of 67.5% for the Finnish part of the Europeana newspaper collection. This estimation is based on a selection of about 132 000 pages included in the Europeana data set.

[3] About half of the collection is in Swedish, the second official language of Finland and up till about 1890 the main publication language of newspapers and journals. We have not estimated the quality of the Swedish data as thoroughly as quality of the Finnish data, but it seems that quality of the Swedish data is worse than quality of the Finnish data.

[4] https://www.abbyy.com/en-eu/finereader/.

[5] "In digital imaging and OCR, ground truth is the objective verification of the particular properties of a digital image, used to test the accuracy of automated image analysis processes. The ground truth of an image's text content, for instance, is the

complete and accurate record of every character and word in the image. This can be compared to the output of an OCR engine and used to assess the engine's accuracy, and how important any deviation from ground truth is in that instance." https://www.digitisation.eu/tools-resources/image-and-ground-truth-resources/. Cf. also Märgner and El Abed (2014) and Carrasco, (2014).

[6] This version was produced by the subcontractor when the GT data was formed.

[7] The original data has 500 640 words. Parallelization of the different OCR versions of the data has proven hard, and we use the results of 471K of data that has content for every different OCR version.

[8] https://sites.google.com/view/icdar2017-postcorrectionocr/dataset.

[9] https://github.com/flammie/omorfi.

[10] https://github.com/jiemakel/omorfi, Mäkelä (2016).

[11] Variation of *w* and *v* is one of the main differences between 19th century and modern Finnish spelling. *W* was used much more in 19th century, in modern Finnish it is used mostly in foreign names (e.g. *Wagner*).

[12] Levenshtein distance is a string metric for measuring the difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. It is named after Vladimir Levenshtein, who considered this distance in 1965. See https://en.wikipedia.org/wiki/Levenshtein_distance.

[13] "It is impossible to correct very noisy texts, where the nature of the noise is random and words are distorted by a large edit distance (say 3 or more)."

[14] https://transkribus.eu/Transkribus/.