Open Source Tesseract in Re-OCR of Finnish Fraktur from 19th and Early 20th Century Newspapers and Journals – Collected Notes on Quality Improvement

Kimmo Kettunen [0000-0003-2747-1382] and Mika Koistinen

The National Library of Finland, DH projects Saimaankatu 6, 50 100 Mikkeli, Finland
Firstname.lastname@helsinki.fi

Abstract. This paper presents work that has been carried out in the National Library of Finland to improve optical character recognition (OCR) quality of a Finnish historical newspaper and journal collection 1771–1910. Work and results reported in the paper are based on a 500 000 word ground truth (GT) sample of the Finnish language part of the whole collection. The sample has three different parallel parts: a manually corrected ground truth version, original OCR with ABBYY FineReader v. 7 or v. 8, and an ABBYY FineReader v. 11 re-OCRed version. Based on this sample and its page image originals we have developed a re-OCRing process using the open source software package Tesseract[1] v. 3.04.01. Our methods in the re-OCR include image preprocessing techniques, usage of morphological analyzers and a set of weighting rules for resulting candidate words. Besides results based on the GT sample we present also results of re-OCR for a 29 year period of one newspaper of our collection, Uusi Suometar.

The paper describes the results of our re-OCR process including the latest results. We also state some of the main lessons learned during the development work.

Keywords: OCR; historical newspapers; Tesseract; Finnish

1 Introduction

The National Library of Finland has digitized historical newspapers and journals published in Finland between 1771 and 1929 and provides them online [1-2]. The last decade of the open collection, 1920–1929, was released in early 2018. The collection contains approximately 7.45 million freely available pages primarily in Finnish and Swedish. The total amount of pages on the web is over 14.5 million, and about half of them are in restricted use due to copyright restrictions. The National Library's Digital Collections are offered via the *digi.kansalliskirjasto.fi* web service, also known as *Digi*. An open data package of the collection's newspapers and journals from period 1771 to 1910 has been released in early 2017 [2].

[1] https://github.com/tesseract-ocr

When originally non-digital materials, e.g. old newspapers and books, are digitized, the process involves first scanning of the documents which results in image files. Out of the image files one needs to sort out texts and possible non-textual data, such as photographs and other pictorial representations. Texts are recognized from the scanned pages with Optical Character Recognition (OCR) software. OCRing for modern prints and font types is considered a resolved problem, that usually yields high quality results, but results of historical document OCRing are still far from that [3].

Newspapers of the 19th and early 20th century were mostly printed in the Gothic (Fraktur, blackletter) typeface in Europe. Fraktur is used heavily in our data, although also Antiqua is common and both fonts can be used in same publication in different parts. It is well known that the Fraktur typeface is especially difficult to recognize for OCR software. Other aspects that affect the quality of OCR recognition are the following [3–5]:

- quality of the original source and microfilm
- scanning resolution and file format
- layout of the page
- OCR engine training
- unknown fonts
- etc.

Due to these difficulties scanned and OCRed document collections have a varying amount of errors in their content. A quite typical example is *The 19th Century Newspaper Project* of the British Library [6]: based on a 1% double keyed sample of the whole collection Tanner et al. report that 78% of the words in the collection are correct. This quality is not good, but quite common to many comparable collections. The amount of errors depends heavily on the period and printing form of the original data. Older newspapers and magazines are more difficult for OCR; newspapers from the early 20th century are easier (cf. for example data of Niklas [7], that consists of a 200 year period of The Times of London from 1785 to 1985). There is no exact measure of the amount of errors that makes OCRed material useful or less useful for some purpose and the use purposes and research tasks of the users of digitized material vary hugely [8]. A linguist who is interested in the forms of words needs as errorless data as possible; a historian who interprets texts on a broader level may be satisfied with text data that has more errors. Anyhow, very high error rate of texts may cause serious discomfort and squeamishness for researchers as e.g. article of Jarlbrink and Snickars about quality of one OCRed Swedish newspaper, *Aftonbladet* 1830–1862, shows [9].

Ways to improve quality of OCRed texts are few, if total rescanning is out of question, as it usually is due to labor costs. Improvement can be achieved with three principal methods: manual correction with different aids (e.g. editing software), re-OCRing or algorithmic post-correction [3]. These methods can also be mixed. We don’t believe that manual correction e.g. with crowd sourcing is suitable for a large collection of a small language with small population: there just is not enough people to perform crowdsourcing. Also post correction’s capabilities are limited: errors of one to two characters can be corrected, but errors in historical OCR data do not limit to these. It seems that harder errors are still beyond performance of post correction algorithms [10-11].

Due to amount of data we have chosen re-OCRing with Tesseract v. 3.04.01 as our main method for improving the quality of our collection. In the rest of the paper we describe the results we have achieved so far and discuss lessons learned. In section two we describe our initial results, in section three improvements made in the re-OCR process and in section four the latest re-OCR results. Section five concludes the paper with some lessons that we have learned during the process.

2 Results – Part I

Our re-OCR process has been described thoroughly in [12–13]. As its main parts are unchanged, we describe it only briefly here. The re-OCRing process consists of four parts: 1) image preprocessing of page images using five different techniques: this yields better quality images for the OCR, 2) Tesseract OCR 3.04.01, 3) choosing of the best candidate from Tesseract's output and old ABBYY FineReader data and 4) transformation of Tesseract's output to ALTO format. We have developed a new Finnish Fraktur model for Tesseract using an existing German Fraktur model as a starting point.

We have evaluated the results of the re-OCR along the development process with different measures using our ground truth data of about 500 000 words [14]. This parallel data consists of proof read version of the data, current ABBYY FineReader OCR v.7/8, Tesseract 3.04.01 OCR and ABBYY FineReader v.11 OCR.

2.1 Precision and Recall

Measurement of OCR improvement does not have any real standard measure, and for this reason we have used several measures to be able to evaluate improvement of the process. Precision and recall are standard measures used in information retrieval, and they can also be applied to analysis of re-OCR results [10]. When we applied recall, precision and F-score to the data, we got recall of 0.72, precision of 0.73 and F-score of 0.73. Combined optimal OCR results of Tesseract and ABBYY FineReader v. 11 would give recall of 0.81, precision of 0.95, and F-score of 0.88**.** The latter figures show that possibility of using several OCR engines would benefit re-OCRing, as has been stated in research literature [15]. Unfortunately we do not have access to several new OCR engines in our final re-OCR.

Precision, recall and their combination, F-score, are useful figures, but it also benefits to take a closer look at the numbers behind the scores. As we analyzed the output of the P/R analysis further we noticed the following. Number of erroneous words in the data of was 126 758 and errorless 345 145. Re-OCR corrected 90 877 of errors (true positives, 71.7% of errors) and left 35 881 uncorrected (false negatives, 28.3% of errors). The OCR process also produced 32 953 new errors to the data (false positives). In general it seems, that the recall of the re-OCR 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. On the other hand, many of the errors were only errors in punctuation: if these were discarded, the results were slightly better. Although every character counts for algorithms that perform evaluation, not every difference in character is of equal importance for human understanding of the output results. Assuming that form *Porvoo* would be the right result*,* the three versions *Bor-*

woo/Porwo/Worwoo that are two characters away from it are not on equal status of intelligibility: the last one would probably be the hardest to understand even in context.

2.2 Character Error and Word Error Rate

Two other commonly used evaluation measures for OCR output are character error rate, CER, and word error rate, WER [16]. CER is defined as

$$\mathrm{CER} = \frac{i+s+d}{n}$$

and it employs the total number *n* of characters and the minimal number of character insertions *i*, substitutions *s* and deletions *d* required to transform the reference text into the OCR output.

Word error rate WER is defined as

$$WER = \frac{i_w + s_w + d_w}{n_w}$$

where n_w is the total number of words in reference text, i_w is the minimal number of insertions, s_w is number of substitutions and d_w number of deletions on word level to obtain the reference text. Smaller WER and CER values mean better quality. Our initial CER and WER results for the OCR process are shown in Table 1. These results have been analyzed with the OCR evaluation tool[2] described in Carrasco [16]. As can be seen from the figures, CER and WER values of the re-OCR are clearly better than those of the current OCR. Especially clear the difference is in word error rate which drops to about a half.

Table 1. Character and word error rates for the DIGI test set

	Re-OCR	**Current OCR**
CER	5.84	7.81
WER	13.65	27.3
WER (order independent)	11.88	25.25

[2] http://impact.dlsi.ua.es/ocrevaluation/. A similar software is PRImA Research's Text Evaluation tool that is available from http://www.primaresearch.org/tools/PerformanceEvaluation.

Evaluation of OCR results can be done experimentally either with or without ground truth. After initial development and evaluation of the re-OCR process with the GT data, we started testing of the re-OCR process with realistic newspaper data, i.e. without GT to avoid overfitting of the data by using GT only in evaluation. We chose for testing *Uusi Suometar*, newspaper which appeared in 1869–1918 and has 86 068 pages. Table 2. shows results of a 10 years' re-OCR of Uusi Suometar with our first re-OCR process. We show here results of morphological recognition with (His)Omorfi that has been enhanced to process better historical Finnish. These results give merely an estimation of improvement in the word quality [1].

Table 2. Recognition rates of current and new OCR words of Uusi Suometar with morphological analyzer HisOmorfi (total of 7 937 pages)

Year	**Words**	**Current OCR**	**Tesseract 3.04.01**	**Gain in % units**
1869	658 685	69.6%	86.7%	17.1
1870	655 772	66.9%	84.9%	18.0
1871	909 555	73%	87%	14.0
1872	930 493	76%	88.7%	12.7
1873	889 725	75.4%	87.3%	11.9
1874	920 307	72.9%	85.9%	13.0
1875	1 070 806	71.5%	86%	14.5
1876	1 223 455	72.8%	86.7%	13.9
1877	1 815 635	73.9%	86%	12.1
1878	2 135 411	72%	85.4%	13.4
1879	2 238 412	74.7%	87%	12.3
ALL	13 448 256	73%	86.5%	13.5

Re-OCR is improving the recognition rates considerably and consistently. Minimum improvement is 11.9% units, maximum 18% units. In average the improvement is 13.5% units.

As can be seen, all our initial results show clear improvement in the quality of the OCR. The improvement could be characterized as noticeable, but not perhaps good enough.

2.3 Examination of the data: false and true positives

In a closer look part of the false positives of the re-OCR are due to recurring trouble with quote marking or division of the word on two lines when the word ends with a hyphen. The re-OCR misses a quote or two in the result word or it produces the HTML code *&quote;* instead of quote itself. Many words are also wrongly divided on the line. The same applies to false negatives, too. Number of all wrong word divisions in the data of false and true positives together is about 10 000, which makes the error type one of the most common. Also missing punctuation or extra punctuation causes errors.

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

Table 3. Corrections of Levenshtein distance of 11.

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

Another example of corrected hard errors are 2 376 words that have 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.

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

The bigger the error count is, the harder the error would be to correct for post correction software, and here lies the strength of re-OCR at its best. Reynaert (2016), e.g., states that his post correction 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 et al. (2007).[4]

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

[3] **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. https://en.wikipedia.org/wiki/Levenshtein_distance

[4] “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).”

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 (total of 90 877 true positives)

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. Old OCR has about 2.32 errors per erroneous word, Tesseract OCR 3.2. This can be seen as 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.

3 Improvements for the re-OCR Process

The results we achieved with our initial re-OCR process were at least promising. They showed clear improvement of the quality in the GT collection and also out of it with realistic newspaper data shown in Table 2. Slightly better OCR results were achieved by Drobac et al. [17] with Ocropy machine learning OCR system using character accuracy rate (CAR) as measure. Post-correction results of Silfverberg et al. [18], however, were worse than our re-OCR results.[5]

The main drawback of our re-OCR system is that it is relatively slow. Image processing and combining of images takes time, if it is performed to every page image as it is currently done. Execution time of the word level system was initially about 6 750 word tokens per hour when using a CPU with 8 cores in a standard Linux environment. With increase of cores to 28 the speed improved to 29 628 word tokens per hour. The speed of the process was still not very satisfying.

[5] Silfverberg et al. have evaluated algorithmic post correction results of *hfst-ospell* software with part of the historical data, 40 000 word pairs. They have used *correction rate* as their measure, and their best result is 35.09 ± 2.08 (confidence value). Correction rate of our initial re-OCR process data is 0.47, clearly better than post-correction results of Silfverberg et al. Our result is also achieved with almost a twelvefold amount of word pairs.

We have been able to improve the processing speed of re-OCR considerably during the latest modifications. We have especially improved the string replacements performed during the process, as they took almost as much time as the image processing. String replacements take now only a fraction of the time they took earlier, but image processing cannot be sped up easily. The new processing takes about half of the time it used to take with the GT data. We are now able to process about 201 800 word tokens an hour in a 28 core system.

We improved also the process for the word candidate selection after re-OCR. We have been using two morphological analyzers (Omorfi[6] and Voikko[7]), character trigrams and other character level data to be able to weight the suggestions given by the OCR process. We checked especially the trigram list and removed the least frequent ones from it.

4 Results – Part II

After improvements made to the re-OCR process we have been able to achieve also better results. The latest results are shown in Tables 6 and 7. Table 6 shows precision, recall and correction rate results and Table 7 shows results of CER, WER and CAR analyses using the ground truth data.

Table 6. Precision and recall of the re-OCR after improvements: GT data

Words without errors	374 299
Words with errors	131 008
Errorless not corrected	366 043
Sum (lines 1 and 2)	505 307
True positives	99 071
False negatives	31 937
False positives	8 256
Recall	0.76
Precision	0.92
F-score	0.83
Correction rate	0.69

[6] https://github.com/jiemakel/omorfi

[7] https://voikko.puimula.org/

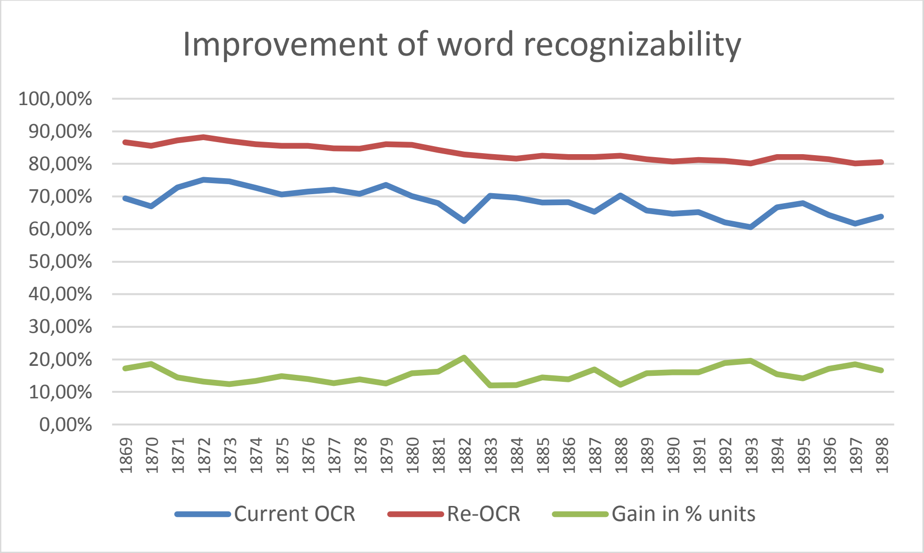
Table 7. CER, WER and CAR of the re-OCR after improvements: GT data

	Re-OCR	**Current OCR**[8]
CER	2.05	6.47
WER	6.56	25.30
WER (order independent)	5.51	23.41
CAR	97.64	92.62

Results in Table 6 and 7 show that the re-OCR process has improved clearly from the initial performance shown in Section 2. Precision of the process has improved considerably, and although recall is still slightly low, F-score is now 0.83 (earlier 0.73). CER and WER have improved also clearly. Our CAR is now also slightly better than Drobac's best value without post correction (ours 97.6 vs. Drobac's 97.3 [17].

Recognition results of the latest re-OCR of Uusi Suometar are shown in Figure 1. The data consists of years 1869–1898 of the newspaper with about 115 930 415 words and 33 000 pages.

Fig. 1. Latest recognition rates of Uusi Suometar 1869-1898 with HisOmorfi



[8] These figures differ slightly from figures of current OCR in Table 1 due to the fact that the improved re-OCR process finds now more matching word pairs in the image data.

Re-OCR is improving the quality of the newspaper clearly and consistently and the overall results are slightly better than in Table 2. The average improvement for the whole period of 30 years is 15.3% units. The largest improvement is 20.5% units, and smallest 12% units.

5 Conclusion

We have described in this paper results of a re-OCR process for a historical Finnish newspaper and journal collection. The developed re-OCR process consists of combination of five different image pre-processing techniques, a new Finnish Fraktur model for Tesseract 3.04.01 OCR enhanced with morphological recognition and character level rules to weight the resulting candidate words. Out of the results we create new OCRed data in METS and ALTO XML format that can be used in our docWorks document presentation system.

We have shown that the re-OCRing process yields clearly better results than commercial OCR engine ABBYY FineReader v. 7/8, which is our current OCR engine. We have also shown that a 29 year time span of newspaper Uusi Suometar (33 000 pages and ca. 115.9 million words) gets significantly and consistently improved word recognition rates for Tesseract output in comparison to current OCR. We have also shown that our results are either equal or slightly better than results of a machine learning OCR system Ocropy in Drobac et al. [17]. Our results outperform clearly post correction results of Silfverberg et al. [18].

Let us now turn to lessons learned during the re-OCR process so far. Our development cycle for a new re-OCR process has been relatively long and taken more time than we were able to estimate in advance. We started the process by first creating the GT collection for Finnish [14]. The end result of the process was a ca. 525 000 word collection of different quality OCR data with ground truth. The size of the collection could be larger, but with regards to limited means it seems sufficient. In comparison to GT data used in OCR or post correction literature, it fares also well, being a mid-sized collection. The GT collection has been the cornerstone of our quality improvement process: effects of the changes in the re-OCR process have been measured with it.

The second time consuming part in the process was creation of a new Fraktur font model for Finnish. Even if the font was based on an existing German font model, it needed lots of manual effort in picking letter images from different newspapers and finding suitable Fraktur fonts for creating synthesized texts. This was, however, crucial for the process, and could not be bypassed.

A third lesson in our process was choice of the actual OCR engine. Most of the OCR engines that are used in research papers are different versions of latest machine learning algorithms. They may show nice results in the narrowly chosen evaluation data, but the software are usually not really production quality products that could be used in an industrial OCR process that processes 1–2 million page images in a year. Thus our slightly conservative choice of open source Tesseract that has been around for more than 20 years is justifiable.

Another, slightly unforeseen problem have been modifications needed to the existing ALTO XML output of the whole process. As ALTO XML[9] is a standard approved

[9] https://www.loc.gov/standards/alto/

by the ALTO board, changes to it are not made easily. An easy way to circumvent this is to use two different ALTOs in the database of docWorks: one conforming to the existing standard and another one that includes the necessary changes after re-OCR. We have chosen this route by including some of the word candidates of the re-OCR in the database as variants.

We shall continue the re-OCR process by re-OCRing first the whole history of *Uusi Suometar*. Its 86 000 pages should give us enough experience so that after that we can move over to re-OCRing the whole Finnish collection. As there are hundreds of publications to be re-OCRed, usage data of the collections are informative in planning of the re-OCR: the most used newspapers and journals need to be re-OCRed first.

We have also created a Swedish language GT collection to be able to start re-OCRing our Swedish language part of the collection. The size of the Swedish GT collection will be about 250 K of words from Swedish language newspapers and journals published in Finland in 1771–1775 and 1798–1919. We should be able to start quickly re-OCR trials with the Swedish data with our so far developed re-OCR process. There should be no need for new font model generation for Swedish Fraktur, as such a font is already available.

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 worse readability and comprehensibility – is worse on-line searchability of the documents in the collections [19–20]. Although information retrieval is quite robust even with corrupted data IR works best with longer documents and long queries, especially when the data is of bad quality. Empirical results of Järvelin et al. [21] with a Finnish historical newspaper search collection, for example, show that even impractically heavy usage of fuzzy matching in order to circumvent effects of OCR errors will help only to a limited degree in search of a low quality OCRed newspaper collection, when short queries and their query expansions are used.

Weaker searchability of the OCRed collections is one dimension of poor OCR quality. Other effects of poor OCR quality may show in the more detailed processing of the documents, such as sentence boundary detection, tokenization and part-of-speech-tagging, which are important in higher-level natural language processing tasks [22]. Part of the problems may be local, but part will cumulate in the whole pipeline of natural language processing causing errors. Thus quality of the OCRed texts is the cornerstone for any kind of further usage of the material and improvements in OCR quality are welcome. And last but not least, user dissatisfaction with the quality of the OCR, as testified e.g. in Jarlbrink and Snickars [9], is of great importance. Digitized historical newspaper and journal collections are meant for users, both researchers and lay person. If they are not satisfied with the quality of the content, improvements need to be made.

Acknowledgment

This work is funded by the European Regional Development Fund and the program Leverage from the EU 2014-2020.

References

1. Kettunen, K., Pääkkönen, T.: Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means," Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).
2. Pääkkönen, T., Kervinen, J., Nivala, A., Kettunen, K., Mäkelä, E.: Exporting Finnish Digitized Historical Newspaper Contents for Offline Use. D-Lib Magazine, July/August (2016).
3. Piotrowski, M.: Natural Language Processing for Historical Texts. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers (2012).
4. Holley, R.: How good can it get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. D-Lib Magazine, 15(3/4) (2009).
5. Doermann, D., Tombre, K. (Eds.): Handbook of Document Image Processing and Recognition. Springer (2014).
6. Tanner, S., Muñoz, T., Ros, P.H.: 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) (2009).
7. Niklas, K.: Unsupervised Post-Correction of OCR Errors. Diploma Thesis, Leibniz Universität, Hannover. www.l3s.de/~tahmasebi/Diplomarbeit_Niklas.pdf (2010).
8. Traub, M. C., Ossenbruggen, J. van, Hardman, L.: Impact Analysis of OCR Quality on Research Tasks in Digital Archives. In: Kapidakis, S., Mazurek, C., Werla, M. (eds.), Research and Advanced Technology for Libraries. Lecture Notes in Computer Science, vol. 9316, pp. 252-263 (2015).
9. Jarlbrink, J., Snickars, P.: Cultural heritage as digital noise: nineteenth century newspapers in the digital archive. Journal of Documentation, https://doi.org/10.1108/JD-09-2016-0106 (2017).
10. Reynaert, M.: OCR Post-Correction Evaluation of Early Dutch Books Online – Revisited. In Proceedings of LREC, pp. 967–974 (2016)
11. Choudhury, M. Thomas, M., Mukherjee, A., Basu, A., Ganguly, N.: How difficult is it to develop a perfect spell-checker? A cross-linguistic analysis through complex network approach. In Proceedings of the second workshop on TextGraphs: Graph-based algorithms for natural language processing, pp. 81–88, (2007).
12. Koistinen, M., Kettunen, K., Kervinen, J.: How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine. Proc. of LTC 2017, Nov. 2017, pp. 279–283 (2017).
13. Koistinen, M., Kettunen, K., Pääkkönen, T.: Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing. Proc. of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, May 2017, pp. 277–283 (2017).
14. Kettunen, K., Kervinen, J., Koistinen, M.: Creating and using ground truth OCR sample data for Finnish historical newspapers and journals. In DHN2018, Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference, 162-169. http://ceur-ws.org/Vol-2084/ (2018).
15. Volk, M., Furrer, L., Sennrich, R.: Strategies for reducing and correcting OCR errors. In C. Sporleder, A. van den Bosch, and K. Zervanou, Eds. Language Technology for Cultural Heritage, 2011, 3–22 (2011).

16. Carrasco, R.C.: An open-source OCR evaluation tool. In: Proceeding DATeCH '14 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, 179-184 (2014)
17. Drobac, S., Kauppinen, P., Lindén, K.: OCR and post-correction of historical Finnish texts. In: Tiedemann, J. (ed.) Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden, 70-76 (2017)
18. Silfverberg, M., Kauppinen, P., Linden, K.: Data-Driven Spelling Correction Using Weighted Finite-State Method. In: Proceedings of the ACL Workshop on Statistical NLP and Weighted Automata, 51–59, https://aclweb.org/anthology/W/W16/W16-2406.pdf (2016)
19. Taghva, K., Borsack, J., Condit, A.: Evaluation of Model-Based Retrieval Effectiveness with OCR Text. ACM Transactions on Information Systems, 14(1), 64–93 (1996)
20. Kantor, P. B., Voorhees, E. M.: The TREC-5 Confusion Track: Comparing Retrieval Methods for Scanned Texts. Information Retrieval, 2, 165–176 (2000)
21. Järvelin, A., Keskustalo, H., Sormunen, E., Saastamoinen, M. and Kettunen, K.: 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 (2016)
22. Lopresti, D.: Optical character recognition errors and their effects on natural language processing. International Journal on Document Analysis and Recognition, 12: 141–151 (2009)