Re-OCR in Action – Using Tesseract to Re-OCR Finnish Fraktur from 19th and Early 20th Century Newspapers and Journals

Kimmo Kettunen
The National Library of Finland
DH projects
Mikkeli, Finland
firstname.lastname@helsinki.fi

Mika Koistinen
The National Library of Finland
DH projects
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 procedure using the open source software package Tesseract[1] v. 3.04.01. Our methods in the re-OCR include image preprocessing techniques, usage of a morphological analyzer and a set of weighting rules for resulting words. Besides results based on the GT sample we present also results of re-OCR for a 10 year period of one newspaper of our collection, Uusi Suometar.**

OCR; historical newspapers; Tesseract; Finnish

I. INTRODUCTION

The National Library of Finland has digitized historical newspapers and journals published in Finland between 1771 and 1920 and provides them online [1-2]. The last decade of the open collection, 1911–1920, was released in February 2017. The collection contains approximately 5.11 million freely available pages primarily in Finnish and Swedish. The total amount of pages on the web is over 12.5 million, over half of them being in restricted use due to copyright reasons. 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, years 1911–1920 will be released later [2].

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 yields high quality results, but results of historical document OCRing are still far from that [3].

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

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 [3–4]:

- 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 [5]: 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 realistic.

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.

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.

II. RESULTS

Our re-OCR process has been described more thoroughly in [6–7]. Here we describe it only briefly. The re-OCRing process consists of four parts: 1) image preprocessing of page images using five different techniques, 2) Tesseract OCR 3.04.01, 3) choosing of the best candidate from Tesseract's output 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 so far with different methods using our ground truth data of about 500 000 words. This parallel data consists of proof read version of the data, current OCR, Tesseract 3.04.01 OCR and ABBYY FineReader v.11 OCR. We performed detailed quality analyses for the results using different

ways of evaluation. Kettunen and Pääkkönen [1] have earlier estimated the quality of the whole historical collection of Finnish with automatic morphological analysis. We applied this quality approximation method now with two morphological analyzers: Omorfi v. 0.3[2] and HisOmorfi, a modified version of Omorfi. Results of analyses are shown in Table 1.

TABLE I. RECOGNITION RATES FOR DIFFERENT COMPARABLE DATA: 471 903 WORDS

	GT	Tesseract 3.04.01	Current OCR	ABBYY FineReader v.11
Omorfi 0.3	81.3%	78.3%	77.1%	85.3 %
HisOmorfi	94.9%	89.9%	81%	86%

Figures show that the manually edited ground truth version is recognized clearly best, as it should be. Plain Omorfi recognizes words of the current OCR version slightly better than Tesseract words, the difference being 0.8% units. This is caused by the fact that HisOmorfi is used in the re-OCRing process and it favors *w* to *v*. 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 percentage of Tesseract's HisOmorfi result column.

When we applied standard measures of 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 [3, 8]. Unfortunately we do not have access to several OCR engines in our final re-OCR.

After initial development and evaluation of the re-OCR process with the GT data, we have started final testing of the re-OCR process with realistic newspaper data. 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.

TABLE II. 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

As can be seen from the figures, 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.

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

III. CONCLUSION

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

We have shown that the re-OCRing process yields clearly better results than commercial OCR engine ABBYY FineReader v. 7/8 and v. 11 with our GT data. We have also shown that a 10 year time span of newspaper Uusi Suometar (7937 pages and ca. 13.45 M words) gets significantly and consistently improved word recognition rates for Tesseract output in comparison to current OCR.

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. We have also started creation of a Swedish language GT collection to be able to start re-OCRing our Swedish language part of the collection.

ACKNOWLEDGMENT

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

REFERENCES

[1] K. Kettunen and T. Pääkkönen, "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] T. Pääkkönen, J. Kervinen, A. Nivala, K. Kettunen, and E. Mäkelä "Exporting Finnish Digitized Historical Newspaper Contents for Offline Use," D-Lib Magazine, July/August 2016.

[3] M. Piotrowski, Natural Language Processing for Historical Texts. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers 2012.

[4] R. Holley, "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] S. Tanner, T. Muñoz, and P.H. Ros, "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.

[6] M. Koistinen, K. Kettunen, and J. Kervinen, "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.

[7] M. Koistinen, K. Kettunen, and T. Pääkkönen, "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.

[8] M. Volk, L. Furrer, and R. Sennrich, "Strategies for reducing and correcting OCR errors," in C. Sporleder, A. van den Bosch, and K. Zervanou, Eds. Language Technology for Cultural Heritage, 2011, pp.3–22.