How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine

Mika Koistinen, Kimmo Kettunen and Jukka Kervinen

The National Library of Finland, DH projects, Saimaankatu 6, FI-50100, Mikkeli
firstname.lastname@helsinki.fi

Abstract

The current paper presents work that has been carried out in the National Library of Finland (NLF) to improve optical character recognition (OCR) quality of the historical Finnish newspaper collection 1771–1910. Results reported in the paper are based on a 500 000 word 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. Using this sample and its page image originals we have developed a re-OCRing procedure using the open source software package Tesseract v. 3.04.01. Our method achieves 27.48% improvement vs. ABBYY FineReader 7 or 8 and 9.16% improvement vs. ABBYY FineReader 11 on document level. On word level our method achieves 36.25% improvement vs. ABBYY FineReader 7 or 8 and 20.14% improvement vs. ABBYY FineReader 11. Precision and recall results on word level show that both recall and precision of the re-OCRing process are on the level of 0.69-0.71 compared to old OCR. Other measures, such as recognizability of words with a morphological analyzer and character accuracy rate, show also clear improvement after re-OCRing.

Keywords: Optical Character Recognition, historical newspaper collections, evaluation

1. Introduction

The National Library of Finland has digitized historical newspapers and journals published in Finland between 1771 and 1920 and provides them online (Kettunen et al. 2014; Kettunen et al., 2016). The last decade of the open collection, 1911–1920, was released recently in February 2017. This collection contains approximately 5.11 million freely available pages primarily in Finnish and Swedish. The total amount of pages on the web is over 11 million, slightly over half (54%) 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 from period 1771 to 1910 has been released in early 2017 (Pääkkönen et al., 2016). The digitized collection has about 100 000 users and in 2016 it had about 18 million page downloads.

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 (Piotrowski, 2012).

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 (Holley 2009; Piotrowski, 2012; Springman and Lüdeling, 2017). Other aspects that affect the quality of OCR recognition are the following (cf. Holley 2009; Piotrowski, 2012, for a more detailed list):

- 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 (Tanner et al. 2009): 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.

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. Also all kind of post processing of the textual data is harmed by bad quality. Thus improvement of OCR quality of digitized historical collections is an important step in improving overall usability of the collections.

This paper reports results of re-OCR for a historical Finnish newspaper collection. The re-OCR process consists of combination of different image pre-processing techniques, and a new Finnish Fraktur model for Tesseract OCR enhanced with morphological recognition and some simple rules to weight the result words.

2. How to Improve OCR Quality

Ways to improve quality of OCRed texts are few, if total rescanning is out of question, as it usually is due to labour costs. Improvement can be achieved with three principal

methods: manual correction with different aids (e.g. editing software, Clematide et al., 2017), re-OCRing (Piotrowski, 2012) or algorithmic post-correction (Reynaert, 2008). These methods can also be mixed. One popular method to realise manual correction has been crowdsourcing. Although this method can be useful, if there is enough population to carry it out (cf. Holley 2010), the method does not suit to large collections of languages that don't have enough people to carry out massive correction. Kettunen and Pääkkönen (2016) have approximated earlier, that about 25–30% out of 2.4. billion Finnish words in the data of 1771–1910 are wrong. This means about 600-800 million word tokens and a few hundred million word types. Effective manual correction of this amount of data is impossible. An earlier crowdsourcing effort resulted in correction of only about 65 000 words (Crohns and Sundell 2011), which shows clearly the futility of this approach with a large heavily erroneous collection of a small language.[1]

Algorithmic post-correction can improve quality of texts, but its capabilities are still limited with low quality original data (Reynaert, 2008). Thus we chose re-OCRing with open source OCR engine Tesseract v. 3.04.01 as our primary method for improving the quality of the texts. Post-correction can be tried later or it can be attached to the process as there are now available tools for doing post-correction of historical Finnish (cf. Silfverberg et al. 2016; Drobac et al., 2017).

2.1. Our Re-OCR Process

OCRing of historical Finnish documents is difficult mainly because of the varying quality newspaper images and lack of model(s) for Finnish Fraktur. However, the character set of Finnish is very similar to other common Fraktur fonts: Finnish has *ä*, *ö* and *å* letters, but no *ü*, and *ß* like German Fraktur. Thus some existing fonts can be used in producing a new Fraktur font for Finnish.

Another problem is quality of page images of OCRed data. Scanned historical document images have many times different types of noise, such as scratches, tears, ink spreading, low contrast, low brightness, and skewing etc. (Piotrowski, 2012). Smitha et al. (2016) present that document image quality can be improved by binarization, noise removal, deskewing, and foreground detection. We use a set of different image preprocessing techniques in our process to improve the original page images. The image processing methods used in our process are explained in detail in Koistinen et al (2017). It suffices to mention here, that use of different image processing methods and their combinations has been essential to achieve improvement in re-OCRing of our data.

Our re-OCRing process consists of four parts: 1) image preprocessing, 2) Tesseract OCR, 3) choosing of the best candidate from Tesseract's output and 4) transformation of Tesseract's output to ALTO format. The process is shown in Figure 1.

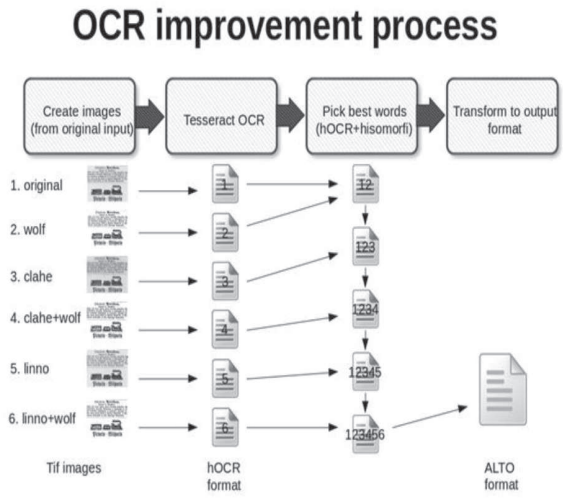

Fig. 1: Re-OCR process

The process uses five different image pre-processing techniques before sending the page images to Tesseract for OCRing. Different combinations of image preprocessing are tried and best combinations are chosen based on the hOCR confidence values and results of morphological recognition of output words in phase three. After that the results are transformed to ALTO format.

After image preprocessing documents are OCRed using Tesseract OCR with new font models *fin* and *fi_frak_mk41* that have been developed for the process. Our Finnish Fraktur model was developed using an existing German Fraktur model[2] as a starting point. The Fraktur model was iteratively improved. The characters that had most errors were improved in training data boxes (single letters and two letter combinations). Then Tesseract was run 1 to N times with the developed Finnish Fraktur model and already existing Finnish Antiqua model[3] in dual model mode, where best alternative from Fraktur and Antiqua results is chosen.

The third phase of the process**, pick the best words,** selects the best word candidates**.** Tesseract uses hOCR format[4] as output. hOCR is an open standard for presenting OCR results and it has confidence value for each word produced by the used OCR tool (Breuel, 2007). Best words are selected by using hOCR word confidence values and a morphological analysis software Omorfi[5] to check recognizability of the words. If candidate word is recognized by Omorfi, the hOCR confidence value of the word gets +10 points and if it is not recognized by Omorfi, it gets -2 points (on a scale of 0–100). If the word is a number, +10 extra points are not given, since there were multiple long number series errors among the first selected results if extra points were given.

[1] A typical success story in crowdsourcing is described e.g. in Clematide et al. (2017), where 180 000 characters on about 21 000 pages were corrected in about 7 months.

[2] https://github.com/paalberti/tesseract-dan-fraktur

[3] https://github.com/tesseract-ocr/langdata/tree/master/fin

[4] https://kba.github.io/hocr-spec/1.2/

[5] https://github.com/jiemakel/omorfi. We call this version HisOmorfi.

Frequency of characters in Finnish is taken into consideration in the process, too. Rarely used characters like *c* and *f* are given -3 points for each occurrence in the word. Thus word candidate *kokonkfcsfa,* for example, would get -9 points, and *kokoukscssa* would get -3 points. This seems like a good rule for Finnish, but would not work for Swedish, the second major language of our collection, as Swedish texts contain lots of correct *f* and *c* characters. Similarly special characters *' ! ; : _&"* are given minus points in the results. Also other special characters like *[ ] ( ) / {} % # ? " &* etc. should be considered to be given minus points in future.

The phase of combining the OCRed documents is run in steps. First documents 1 and 2 are combined, and then the combination of 1 and 2 is combined with document 3 and so on. The last phase**, Transform to output format,** transfers the documents into ALTO XML format. ALTO is the format used by our production system docWorks, and the presentation system Digi.

3. Results

3.1. First results

Koistinen et al. (2017) reported **page level** evaluation results of re-OCR process with the 500 000 word sample comparing ABBYY FineReader v.7 and/or 8 (current OCR of the collection), ABBYY FineReader v.11 and Tesseract re-OCR with different image processing methods and by using page level confidence as a measure.

The best Tesseract OCR result on page level was achieved by combining four image pre-processing methods: Linear Normalization + WolfJolion, Contrast Limited Adaptive Histogram Equalization + WolfJolion, original image and WolfJolion. Page level system improves the word level quality of OCR by 1.91 percentage points (9.16%) against the best result of ABBYY FineReader 11 and by 7.21 percentage points (27.48%) against ABBYY FineReader 7 and 8. Thus our method could correct at best about 84.6 million words in the 1771–1910 1.06 million Finnish newspaper page collection (consisting of Finnish language) of the current OCR with ABBYY FineReader v. 7/8.

The method could still be improved. The method is 2.08 percentage points from the optimal Oracle result, which is 16.94% word error rate. Oracle result is the result when the truly best document is always selected, instead of choosing the result based on the hOCR confidence value.

The character accuracy results for Fraktur model show that characters *u, m* and *w* have less than 80 percent correctness even after re-OCRing. These letters are confused with partly overlapping letters such as *n* and *i*. It seems, however, that if accuracy for one of them is increased, accuracy of others will decrease. Also recognition of letter *ä* could possibly be improved, though it overlaps with letters *a* and *å*. From 20 most frequent errors in the character data only five characters are under 80% correct.

3.2. Further results

In the second **word level** evaluation document confidence was changed to select best single words from different images to make the method more accurate. In this method original image was changed into five different images using WolfJolion, Linear Normalization, Contrast Limited Adaptive Linear Normalization (CLAHE), Linear Normalization + WolfJolion, CLAHE + WolfJolion. Tesseract OCR was run on these six images and the best words were selected by the hOCR word accuracy value with Omorfi and rules *c-f* and special character detection to add/reduce points. Final result after the process is an ALTO format document for combined OCR results that contains the most accurate content and alternative blocks for less accurate content. On word level our method achieves 9.43% unit improvement vs. ABBYY FineReader 7 or 8 and 4.18% units improvement vs. ABBYY FineReader 11.

For further analysis of results we used a parallel version of the 500K collection with ground truth, old OCR and Tesseract OCR, and performed a detailed quality analysis for the results using different ways of evaluation. Kettunen and Pääkkönen (2016) have earlier estimated the quality of the whole historical collection with morphological analysis. We applied this method now with two morphological analyzers: original Omorfi v. 0.3[6] and HisOmorfi. Results of analyses are shown in Table 1.

	Ground truth	**Tesseract OCR**	**Current OCR**
Omorfi 0.3	81.2%	76.1%	76.9%
HisOmorfi	94.0%	87.4%	80.7%

Table 1. Word recognition rates with two morphological analyzers

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

As further evaluation measures we use standard measures of recall and precision and their combination, F-score (Manning and Schütze, 1999). These measures have been widely used in both post-correction and re-OCRing evaluations (Reynaert, 2008). Other measures exist, too, but most of them, as for example correction rate used in Silfverberg et al. (2016), are calculated only slightly differently than P/R figures.

As the data is not wholly parallel with number of words varying from 459 942 to 500 604 in different versions of the data, we based our calculations on lines where there was character data in every column of the table consisting of GT, CurrOCR, and TesseractOCR words. Number of these lines was 459 930.

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

Table 3 shows basic P/R results and F-scores of the data and also correction rate. We show two results: one on the left column is achieved by comparing all the data without cleaning. The result on the right column shows the results with punctuation and all other non-alphabet and non-number characters removed from the lines. Removed character set is: ,;\':\"\'_!@#%&*()+=<>[]{}?\\/—~|^\“„¦«©»®°¡. Variation of w/v is also neutralized.

Basic results	Results with cleaned data
Recall = 68.4 Precision = 70.1 F measure= 69.3 Correction rate = 39.3	Recall = 71.0 Precision = 71.0 F measure= 71.0 Correction rate = 43.0

Table 2. P/R results for Tesseract OCR vs. current OCR

The results achieved are clearly better than previous post-correction trial results in Kettunen (2016), where F-scores of about 55-60 at best were reached with small test samples. As current results are also achieved with a more realistic sample of the data, they seem promising. It seems that our re-OCR has a satisfying recall of the errors, but it is not very precise. This is mainly due to new erroneous words introduced by the re-OCR.

We can additionally compare our re-OCRing results to some other correction results of data that originates from our newspaper data but where the data sample is only a part of our sample. Silfverberg et al (2016) have evaluated post-correction results of *hfst-ospell* software with the historical data using about 40 000 word pairs. They used *correction rate* as their measure, and their best result is 35.09 ± 2.08 (confidence value). Correction rate of our re-OCR process data in Table 2. is 39.3, which is slightly better than result of post-correction in Silfverberg et al. (2016). Besides, our result is achieved with a tenfold amount of word pairs.

Drobac et al. (2017) have used neural network based software Ocropy to re-OCR a sample of historical Finnish newspaper material. They have used two differently trained models, which they call DIGI and NATLIB. Besides these OCR models they use also post-correction with hfst-ospell. Drobac et al. use character accuracy (CAR) as their evaluation measure. Results reported in Drobac et al. (2017) and comparative results using CAR for our re-OCR data are shown in Table 3.

	Drobac et al. (2017)	NLF re-OCR
Ocropy OCR	93.0	N/A
DIGI model+post corr.	93.3	N/A
NATLIB model+post corr.	95.2	N/A
NLF ReOCR	N/A	93.2
NLF FR11	N/A	94.5
NLF current OCR	N/A	90.9

Table 3. Results of Drobac et al. (2017) compared to results of NLF's re-OCR results using character accuracy

Figures show, that plain Ocropy OCR is on the same level of performance as our re-OCR method. Post-correction brings some gain for the character accuracy with the NATLIB model, but not with the DIGI model. Version 11 of ABBYY FineReader performs slightly better than Ocropy, but is slightly beyond performance of NATLIB model and post-correction.

3. Discussion

We have described in this paper a re-OCRing process for a historical Finnish newspaper collection. The process consists of combination of different image pre-processing techniques, a new Finnish Fraktur model for Tesseract OCR enhanced with morphological recognition and some simple 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 better results than commercial OCR engine ABBYY FineReader. Compared to older versions of ABBYY FineReader (7 and 8, available for us), the increase on page level correctness of words is 7.21% units. Compared to ABBYY FineReader v. 11, the improvement is 1.91% units. On word level our method achieves 9.43% unit improvement vs. ABBYY FineReader 7 or 8 and 4.18% unit improvement vs. ABBYY FineReader 11.

On word level we achieve word recognition improvement of 6.7% units in comparison to old OCR using morphological recognition. F-score of our re-OCR is 69.3. Character accuracy of our results is on the same level or slightly below results of Drobac et al. (2017) who use Ocropy OCR engine and post-correction. Thus the developed process is competitive in its results in comparison to other existing re-OCR systems for historical Finnish and slightly better than the post-correction system reported in Silfverberg et al. (2016).

The results are promising initially, but probably they could be improved. First of all, some improvements could be considered for the re-OCR process. Post-correction of the re-OCR using Finnish hfst-ospell model could be beneficial, as shown in Drobac et al. (2017). As the image quality of the documents is one of the most important factors in the recognition accuracy, further research with image processing algorithms could also be performed. In addition to utilizing the confidence measure value, methods to determine noise level in the image could possibly be utilized to choose only bad quality images for further pre-processing.

The OCR process could also benefit from general profiling of the data to pinpoint parts of data that have the lowest quality. A readily available OCR document error profiler is described in Reffle and Ringlstetter (2013) and Fink et al. (2017). The method described in Reffle and Ringlstetter computes data's statistical profile that provides an estimate of error classes with associated frequencies and points to conjectured errors and suspicious tokens. The system combines lexica, pattern sets and advanced matching techniques in a specialized Expectation Maximization (EM) profile (Reffle and Ringlstetter, 2013). We plan to investigate embedding of the system within our OCR process.

A crucial condition for the OCR algorithm is speed of execution, when one needs to OCR a collection containing millions of documents. Current execution time

of our word level system is about 6 750 word tokens per hour when using a CPU with 8 cores in a standard Linux environment. With 56 cores the speed improved to 29 628 word tokens per hour. Thus a realistic scenario for re-OCRing of our material would be to first start with one popular newspaper and re-OCR its whole history. A suitable candidate for this would be for example *Uusi Suometar*, which appeared in 1869–1919 and has 86 068 pages. Out of the Finnish language newspapers it is the most used in the collection according to our usage statistics. Gaining experience of re-OCRing a whole newspaper would give invaluable experience of the re-OCR process. If re-OCRing could be directed with profiling to only those documents or document parts that have most errors, the process could become faster.

Acknowledgements

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

References

Breuel, T. (2007) The hOCR Microformat for OCR Workflow and Results. Document Analysis and Recognition, 2007. In: *ICDAR 2007. Ninth International Conference on* http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4377078

Chrons, O., Sundell, S. (2011). Digitalkoot: making old archives accessible using crowdsourcing. In*: Human Computation, Papers from the 2011 AAAI Workshop* http://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/view/3813/4246

Clematide, S., Furrer, L. and Volk, M. (2017). Crowdsourcing an OCR Gold Standard for a German and French Heritage Corpus. In: *Language Resources and Evaluation* (to appear*)*.

Drobac, S., Kauppinen, P. and Lindén, K. (2017). 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, pp. 70–76.

Fink, Florian, Schulz, Klaus U. and Springmann, Uwe (2017). Profiling of OCR'ed Historical Texts Revisited. In: *DaTeCH2017*, pp. 61-66.

Holley, R. (2009). How good can it get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. In: *D-Lib Magazine*, 15(3/4). http://www.dlib.org/dlib/march09/holley/03holley.html

Holley, R. (2010). Crowdsourcing: How and Why Should Libraries Do It? In: *D-Lib Magazine,* 16(3/4). http://www.dlib.org/dlib/march10/holley/03holley.html

Kettunen K. (2016) Keep, Change or Delete? Setting up a Low Resource OCR Post-correction Framework for a Digitized Old Finnish Newspaper Collection. In: Calvanese D., De Nart D., Tasso C. (eds.) *Digital Libraries on the Move. IRCDL 2015*. Communications in Computer and Information Science, vol. 612. Springer, Cham, pp. 95–103.

Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T., and Kervinen, J. (2014). Analyzing and Improving the Quality of a Historical News Collection using Language Technology and Statistical Machine Learning Methods. In: *IFLA World Library and Information Congress*, Lyon. http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-honkela-en.pdf.

Kettunen, K. and 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: Calzolari, N. et al. (Eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* http://www.lrec-conf.org/proceedings/lrec2016/pdf/17_Paper.pdf

Koistinen, M., Kettunen, K. and Pääkkönen, T. (2017). Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing. In: Tiedemann, J. (ed.) Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden, pp. 277-283

Manning, C.D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Reffle, U. and Ringlstetter, C. (2013). Unsupervised Profiling of OCRed historical documents. In: *Pattern Recognition* 46, pp. 1346–1357.

Reynaert, M. (2008). Non-interactive OCR post-correction for giga-scale digitization projects. In: *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing, CICLing'08*, pp. 617–630.

Silfverberg, M., Kauppinen, P., and Linden, K. (2016). Data-Driven Spelling Correction Using Weighted Finite-State Methods. In: *Proceedings of the ACL Workshop on Statistical NLP and Weighted Automata,* pp. 51–59. https://aclweb.org/anthology/W/W16/W16-2406.pdf

Smitha, M.L, Antony, P.J. and Sachin, D.J. (2016). Document Image Analysis Using Imagemagick and Tesseract-ocr. In: *International Advanced Research Journal in Science, Engineering and Technology*, 3(5), pp. 108–112.

Springmann, U. and Lüdeling, A. (2017). OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus. In: *Digital Humanities Quarterly* 11(2), http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html

Tanner, S., Muñoz, T., and 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. In: *D-Lib Magazine*, (15/8) http://www.dlib.org/dlib/july09/munoz/07munoz.html.