Creating and using ground truth OCR sample data for Finnish historical newspapers and journals

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

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

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 12.9 million pages mainly in Finnish and Swedish. Out of these about 7.36 million pages are freely available on the web site digi.kansalliskirjasto.fi. 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 years 1920–1929 were opened in January 2018.

This paper presents the ground truth Optical Character Recognition data of about 500 000 Finnish words that has been compiled at the NLF for development of a new OCR process for the collection. We discuss compilation of the data and show basic results of the new OCR process in comparison to current OCR using the ground truth data.

Keywords: historical newspaper collections, Optical Character Recognition, ground truth, Finnish

1 Introduction

The National Library of Finland has digitized historical newspapers, journals and ephemera published in Finland since the late 1990s. Besides producing and publishing the digitized raw data all the time NLF has been involved in research and improvement of the digitized material during the last years. We ended a two year European Regional Development Fund (ERDF) project in July 2017 and started another two year ERDF project in August 2017. NLF is also involved in research consortium *Computational History and the Transformation of Public Discourse in Finland, 1640-1910* (COMHIS) that is funded by the Academy of Finland (2016–2019). COMHIS utilizes the newspaper and journal data in its research of historical changes of publicity in Finland.

One part of our data improvement effort has been quality analysis of Finnish data. Out of this we have learned that about 70–75% of the words in the data of 1771–1910 are recognizable and probably right. In a collection of about 2.4 billion words this means that 600–800 million word tokens are wrong [1]. This is a huge proportion of the words in the collection, and the erroneous words harm both online search of the

collection and general usefulness of the textual data [2]. 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 worse readability and comprehensibility - is worse on-line searchability of the documents in the collections. Although users of the NLF collections have not complained much about the quality, improvement of the quality is a natural first step in improvement of the collection.[1]

In order to improve the quality of the collection, 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 14.0, but the cost of the Fraktur font is too high a burden for re-OCRing the collection with ABBYY FineReader. We ended up using open source OCR engine Tesseract v. 3.04.01[2] and started to train Fraktur font for it. This process and its results are described in detail in Koistinen et al. [3–4].

2 Data in the GT collection

To be able to evaluate the re-OCR results properly, we needed to establish ground truth (GT) data[3] that could be used for comparison of the re-OCR results. For this purpose we chose manually a set of newspaper and journal pages that were printed in Fraktur font and were from different publications and decades. Our budget for creation of the GT was minimal: we were able to pay for a subcontractor for creation of the basic GT, but the budget was scarce – about 4 000 €. This 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 time period 1836–1918. Most of the data is from 1870 onwards, as the majority of publications in the collection is from 1870–1910 [1]. When the pages were picked, only year of publication, type of publication (journal/newspaper), font type and number of pages and characters was known of the data. In the final selection 56% of the pages are

[1] About half of the collection is in Swedish, the second official language of Finland and until about year 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.

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

[3] '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/

from journals, 44% from newspapers. Journal data has about 950 K of characters, newspaper data 3.06 M. Figures 1 and 2 show character amounts in newspaper and journal GT data for each year.

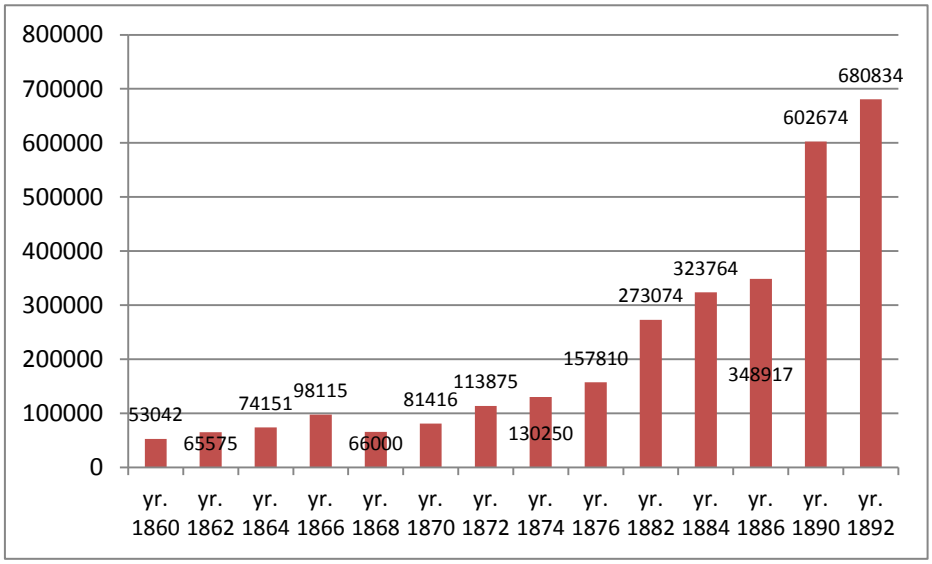

Fig. 1. Number of characters in newspaper GT data for each year

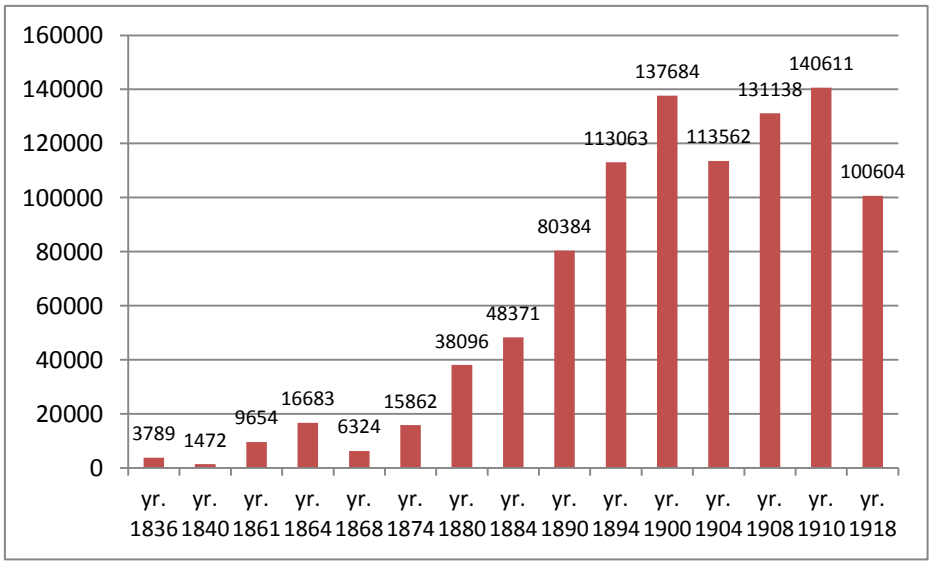

Fig. 2. Number of characters in journal GT data for each year

Figure 3. shows an excerpt of the GT data as an Excel table view. Information includes 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) are different OCR versions of the data.

PUBTYPE	PUBYEAR	ISSN	PAGENAMI	SORTORDE	GT	TESSERACT	OLD	FR11
AIK	1884	fk00010	fk00010_18	3	taan	taan	taan	taan
AIK	1884	fk00010	fk00010_18	4	sekä	sekä	sekä	sekä
AIK	1884	fk00010	fk00010_18	5	ehdottoma	ehdottoma	ehdottoma	ehdottoma
AIK	1884	fk00010	fk00010_18	6	raittiita	raittiita	raittiita	raittiita
AIK	1884	fk00010	fk00010_18	7	että	että	että	että
AIK	1884	fk00010	fk00010_18	8	raittiuden	raittiuden	raittiuden	raittiuden
AIK	1884	fk00010	fk00010_18	9	harrastajia,	harrastajia,	harrastajia,	harrastajia,
AIK	1884	fk00010	fk00010_18	10	jotka	jotka	jotka	jotka
AIK	1884	fk00010	fk00010_18	11	tekewät	tekewät	tekemät	tekemät
AIK	1884	fk00010	fk00010_18	12	työtä	työtä	työtä	työtä
AIK	1884	fk00010	fk00010_18	13	raittiuden	raittiuden	raittiuden	raittiuden
AIK	1884	fk00010	fk00010_18	14	edestä,	edestä,	edestä,	edestä,
AIK	1884	fk00010	fk00010_18	15	katsoen	katsoen	katsoen	katsoen
AIK	1884	fk00010	fk00010_18	16	sitä	sitä	sitä	sitä
AIK	1884	fk00010	fk00010_18	17	tulewaisuu	tulewaisuu	tulemaisuu	tulemaisuu

Fig. 3. An example of parallel GT data

The final ground truth text was corrected manually in two phases: first correction was performed by a subcontractor from output of ABBYY FineReader 11, and 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 set has 471 903 parallel lines of words or character data. The words in the GT have 3 290 852 characters without spaces, punctuation included, and 4 234 658 characters with spaces. Medium length of the words is 6.97 characters.

Size of the data seems relatively small in comparison to the overall size of the collection which was at the time of creation 1 063 648 pages of Finnish newspapers and journals. 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 et al. [5] used for error rate counting with 19th century British newspapers, but it is also quite a lot larger than a typical OCR research paper evaluation data set. Berg-Kirckpatrick and Klein [6] use 300–600 lines of text, Drobac et al. [7] 9 000 – 27 000 lines of text in their re-OCR trial as evaluation. Silfverberg et al. [8] use 40 000 word pairs in post correction evaluation and Kettunen [9] uses 3 800–12 000 word pairs. Dashti [10] 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[4]. In comparison to

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

current usage in the field our 471 903 words and 3 290 852 characters can be considered a medium sized data set.

3 Results of re-OCR

We have described the re-OCR process and its evaluation thoroughly in Koistinen et al. [3–4]. Here we mention only the basic 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 equal to words of the ground truth. In the old OCR this figure is 73.1% and in ABBYY FineReader v.11 79%. When the words of the GT are analyzed with morphological analyzer Omorfi[5], the words OCRed with Tesseract achieve about 9% unit improvement in recognition compared to current OCR (90% recognition vs. 81%). In the GT data the recognition rate is 94.9%.

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 was published in 1869–1918 and has 86 068 pages. Table 1. shows results of a 10 years' re-OCR of Uusi Suometar.

Table 1. Recognition rates of current and new OCR words of Uusi Suometar with morphological analyzer Omorfi (total of 7 937 pages). Number of columns in the newspaper increased from three to five during this period.

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. This shows clearly that the new OCR process yields good results also outside the GT sample.

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

4 Discussion

We have described in this paper our Optical Character Recognition GT sample for Finnish historical newspapers and journals. The parallel data with hand corrected GT and two versions of OCR results consists of 479 pages and 471 903 words and it has been used in development of a new OCR 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 [3–4]. The new OCR process shows also clear improvement with data of Uusi Suometar 1869–1879: in average word recognition is improved with 13.5% units.

The GT data has been created as a tool for quality control of the re-OCR process [11]. The data package is published on our web site *https://digi.kansalliskirjasto.fi/opendata* as open data. The package contains the preservation quality image file (.tif) and the ALTO XML file of each page of the data. We have earlier published the text files of the collection's 1771–1910 part with metadata, ALTO XML and plain text [12]. Publication of the GT data benefits those, who work on OCR of historical Finnish or develop post correction algorithms for OCR. Also development work of general OCR tools such as Transkribus[6] may benefit from the data. So far 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 newspaper and journal data of th4 NLF [7].

The old saying in computational linguistics claims 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. We believe it offers a useful tool both for our re-OCR process development and post correction algorithm development for OCRed historical Finnish. It adds a useful resource for somehow under resourced historical late 19th century Finnish. We hope it has use also outside of OCR and post correction field for those who work in the digital humanities. In November 2017 we started creation of GT data for our collection's historical Swedish language prints that use Fraktur font.

Acknowledgements

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. In: Calzolari, N et al. (Eds.), Proceedings of the Tenth International Confer-

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

ence on Language Resources and Evaluation (LREC 2016) http://www.lrec-conf.org/proceedings/lrec2016/pdf/17_Paper.pdf (2016).
2. Järvelin, A., Keskustalo, H., Sormunen, E. Saastamoinen, M., 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).
3. Koistinen, M., Kettunen, K., Kervinen, J.: Bad OCR has a nasty character - re-OCRing historical Finnish newspaper material 1771–1910. International Journal of Document Recognition and Analysis (IJDAR). Submitted (2017).
4. Koistinen, M., Kettunen, K., Pääkkönen, T.: Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing. In: Proceedings of Nodalida 2017 http://www.ep.liu.se/ecp/131/038/ecp17131038.pdf (2017).
5. 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) http://www.dlib.org/dlib/july09/munoz/07munoz.html (2009).
6. Berg-Kirkpatrick, T., Klein, D.: Improved Typesetting Models for Historical OCR. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 118-123. www.aclweb.org/anthology/P14-2020 (2014).
7. 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, pp. 70-76 (2017).
8. 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, pp. 51–59, https://aclweb.org/anthology/W/W16/W16-2406.pdf (2016).
9. Kettunen K.: 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 (2016).
10. Dashti, S.M.: Real-word error correction with trigrams: correcting multiple errors in a sentence. Language Resources and Evaluation. DOI 10.1007/s10579-017-9397-4 (2017).
11. Märgner, V., El Abed, H. Tools and Metrics for Document Analysis System Evaluation. In: Doermann, D. and Tombre, K. (Eds.), Handbook of Document Image Processing and Recognition, pp. 1011–1036. Springer (2014).
12. Pääkkönen, T., Kervinen, J., Nivala, A., Kettunen, K., Mäkelä, E.: Exporting Finnish Digitized Historical Newspaper Contents for Offline Use. D-Lib Magazine, July/August. http://www.dlib.org/dlib/july16/paakkonen/07paakkonen.html (2016).