Human Language Technologies – The Baltic Perspective
I. Skadiņa and R. Rozis (Eds.)

doi:10.3233/978-1-61499-701-6-122

Between Diachrony and Synchrony: Evaluation of Lexical Quality of a Digitized Historical Finnish Newspaper and Journal Collection with Morphological Analyzers

Kimmo KETTUNEN[a,1], Tuula PÄÄKKÖNEN[b] and Mika KOISTINEN[b]
[a] *National Library of Finland, Centre for Digitization and Conservation*
[b] *NLF, Centre for Digitization and Conservation*

Abstract. The National Library of Finland has digitized the historical newspapers and journals published in Finland between 1771 and 1910 [1,2]. The size of the whole collection up to 1910 is about 3.1 M pages. The newspaper collection contains approximately 1.961 million pages mostly in Finnish and Swedish. Finnish part of the collection consists of about 1 063 648 pages, and Swedish part of 892 101 pages. Additionally there are 11 548 pages in German and Russian. Finnish part of the collection has about 2.407 billion words. The National Library's Digital Collections are offered via the *digi.kansalliskirjasto.fi* web service, also known as *Digi*. An open data delivery package of the whole text material has been produced recently and it will be made publicly available later this year [3]. The quality of OCRed collections is an important topic in digital humanities, as it affects general usability, searchability and advanced processing, such as content mining, of collections [4, 5]. There is no single available method to assess the quality of large collections, but different methods can be used to approximate quality. This paper uses corpus analysis style methods to approximate overall lexical quality of the Finnish part of the Digi collection. Methods include usage of parallel samples and word error rates, usage of morphological analyzers, frequency analysis of words and comparisons to comparable edited lexical data of the same era. Our aim in the quality analysis is twofold: firstly to analyze the present state of the lexical data and secondly, to establish a set of methods that build up a compact procedure for quality assessment after e.g. re-OCRing or post-correction of the material.

Keywords. historical newspapers, Finnish, quality evaluation

1. Introduction

Newspapers of the 19th and early 20th century were mostly printed in the Gothic (Fraktur, blackletter) typeface in Europe. The typeface is almost notoriously difficult to recognize for OCR software [6, 7]. Other aspects that affect the quality of the OCR recognition are the following, among others: quality of the original source and

[1] Corresponding Author, National Library of Finland, Centre for Digitization and Conservation, Saimaankatu 6, FI-50100, Mikkeli, Finland; e-mail: kimmo.kettunen@helsinki.fi

microfilm, scanning resolution and file format, layout of the page, noisy typesetting process, OCR engine training, and unknown fonts.

As a result of these difficulties scanned and OCRed document collections have a varying number of errors in their content. The number of errors depends heavily on the period and printing form of the original data. Older newspapers and journals are more difficult for OCR; newspapers from the early 20th century are usually easier (cf. for example data of Niklas [8] that consists of a 200 year period of The Times of London from 1785 to 1985). Digital collections may be small, medium sized or large and different methods of quality assessment are useful or practical for different sizes of collections. Smallish and perhaps even medium sized collections may be assessed and corrected by human inspection [9]. When the size of the collection increases, human inspection becomes impossible, or human inspection can only be used to assess samples of the collection.

Thus quality assessment of OCRed collections is most of the times *sample-based*, as in the case of the British Library [10]. A representative part of the collection is assessed by using a gold standard collection, when such is available or can be produced cost effectively. Word and character level comparisons can then be made and error rates of the OCRed collections can be reported and compared. Another, fully automatic possibility to assess the quality of the collection is usage of digital dictionaries. Niklas [8], for example, uses dictionary look-up to check the overall word level quality of The Times of London collection from 1785 to 1985 in his OCR post-correction work. Same kind of approach is used by Alex and Burns [11]. This kind of approach gives a word accuracy approximation for the data [9]. Its strength is in easy implementation with available dictionaries, but naturally the procedure will also produce false recognitions and misrecognitions due to gaps in the dictionary data.

Usage of digital dictionaries suits only languages like English that have only a little inflection in words and thus the words in texts can be found in dictionaries as dictionary entries. A heavily inflected, morphologically complex language like Finnish needs other means, as the language has potentially thousands of grammatical word forms for noun, verb and adjective lexemes. Full morphological analysis of the material is needed for this type of language. If the analyzer can relate an input word after application of rules to a base form or forms in its lexicon, it has successfully recognized/analyzed the word. We shall employ and discuss this approach with our material.

2. Analyzing the Data

Most of the data, 82.7 %, in the Digi collection, is from the last two decades of the period, 1890–1910. 92.3 % of the data is from the last four decades, 1870–1910. Proportions of data of newspapers in words in in different decades are shown in Figure 1 without material of 1780–1819, as data between 1780 and 1819 contains only Swedish.

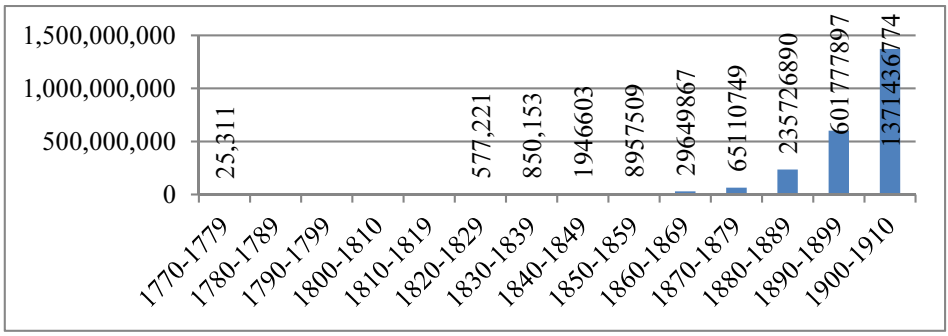

Figure 1: Proportions of data in words in newspapers

Figure 2 shows the number of pages in Finnish and Swedish newspaper data during the publication period of 1771–1910. Time spans in the figure are based on the zip packages of the forthcoming open data delivery [3]. As can be seen from the figure, Swedish was the dominant language in newspaper and journal printing in Finland up to 1890, but since that Finnish has been the prevalent language. A very small number of pages were written in Russian (8 997) and German (2 551) during this time. Language is the primary language of the title as listed in our newspaper database.

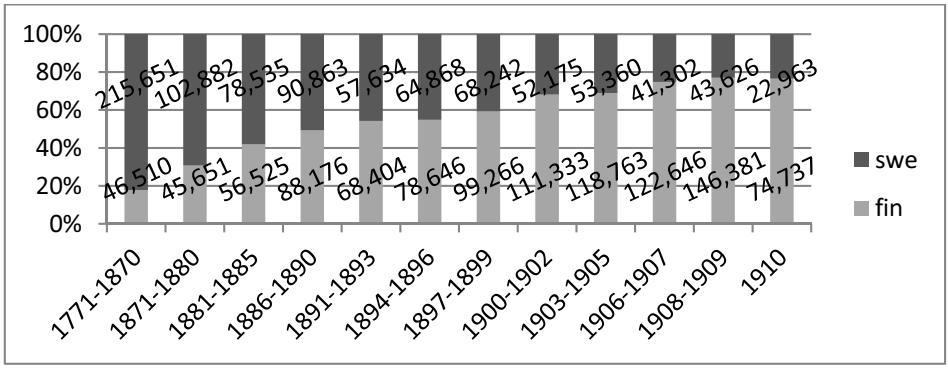

Figure 2: Number of pages in Finnish and Swedish newspapers in different time spans of 1771–1910. Total number of Finnish pages is 1 063 648, and total number of Swedish pages 892 101.

Our first word quality analysis results were published in Kettunen and Pääkkönen [12], and the results were achieved with an older version of Omorfi and FINTWOL (version 1999/12/20). Our first results with version 0.1 of Omorfi (dated 2012) showed that about 69 % of the words of the Digi can be recognized with a modern Finnish morphological analyzer. If the most salient orthographical difference in the 19th century Finnish, *v/w* variation, was taken into account and number of out-of-vocabulary words (OOVs) was estimated, the recognition rate increased to 74–75 %. The rest, about 625 M words, was estimated to consist mostly of OCR errors, at least half of them being hard ones.

After initial analysis of the data with version 0.1 of Omorfi, we have used two newer versions of the software. The other one is version 0.2, (dated 2014) and another version, that we shall call HisOmorfi. This version is based on v. 0.2 of Omorfi and

modified to deal with many historical features of 19th century Finnish[2]. Table 1 shows analysis results with these analyzers. Omorfi 0.2 does not recognize words much better than version 0.1, but HisOmorfi achieves improved recognition of 3 % units with the main part of the data. There is improvement in recognition with HisOmorfi for every type of data, although for word types improvement is small. This and our earlier analyses [12] show that there are lots of once occurring strings that are mostly errors.

Table 1: Recognition rates with Omorfi 0.2 and HisOmorfi for word types and tokens of Digi

Collection	Number of words	Recognized by Omorfi 0.2	Recognized by HisOmorfi	Type of data
Digi up to 1850 *tokens*	22.8 M	66.3 %	70.8 %	OCRed index words
Digi 1851–1910 *tokens*	2.385 G	69.7 %	72.7 %	OCRed index words
Digi up to 1850 *types*	3.24 M	16.0 %	19.4 %	OCRed index words
Digi 1851–1910 *types*	177.3 M	3.9 %	4.9 %	OCRed index words

We analyzed recognition rates of words in the data also decade by decade without data of 1780–1819 as it consists of Swedish only. Recognition rates are mainly between 65 and 77 per cent. Data of 1770–1779 and 1840–1849 are recognized slightly worse than other data Interestingly, there is no big variation in the recognition rates of earlier and late 19th century, although it would be expectable that older data contains more old vocabulary that is not recognized. One reason for quite good recognition of older data may be simpler column structures and larger fonts in older publications, which could have decreased OCR errors. Towards the end of the 19th century number of columns in newspapers increased and also fonts got smaller. Even if Finnish of the late 19th century as such should be easier to recognize for morphological analyzers, it may have more OCR errors due to printing format. We believe these two phenomena have a contrary overall effect on the recognition rate. Also the amount of data may have an effect.

To be able to approximate the level of unrecognition caused by historical Finnish and OCR errors we needed also recognition rates for clean lexical data of the same period as our OCRed data. In Table 2 recognition of Digi data is compared to comparable hand edited data of the 19th century. This data has been gathered from the web pages of The Institute for the Languages of Finland. The data consists of two wordlists, VKS and VNS, and four different dictionaries from different decades of the 19th century. Size of these corpora ranges from 28.5 K words to about 5 M words. Figures show that data of the earlier period (especially VKS and Renvall and Helenius dictionaries) are recognized clearly worse than data of later period.

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

Table 2: Recognition rates of four different morphological analyzers for the Digi data compared to hand-edited wordlist and dictionary data of The Institute for the Languages of Finland. Table legend: VKS = Corpus of Old Finnish (time span of 1543–1809) VNS = Corpus of Early Modern Finnish (time span of 1809–1899)

Collection	Rec. by Omorfi 0.1	Rec. by Omorfi 0.2	Rec. by HisOmorfi	Rec. by FINTWOL	Number of words
VKS frequency corpus types	15.00 %	16.00 %	28.70 %	16.60 %	285 K
VKS frequency corpus tokens	49.00 %	49.90 %	60.40 %	50.30 %	3.43 M
VNS frequency corpus types	55.90 %	58.00 %	62.20 %	58.10 %	530 K
VNS frequency corpus tokens	86.10 %	87.30 %	88.80 %	86.50 %	4.86 M
Renvall dictionary 1826	43.00 %	43.40 %	45.60 %	45.50 %	25.8 K
Helenius dictionary 1838	49.00 %	50.40 %	55.80 %	50.00 %	28.7 K
Europaeus dictionary 1853	76.00 %	77.00 %	79.00 %	69.00 %	43.2 K
Ahlman dictionary 1865	73.00 %	73.40 %	75.10 %	71.50 %	91.4 K
Four dictionaries combined	62.00 %	63.30 %	65.90 %	61.00 %	135.3 K
Digi index tokens 1851-1910	69.30 %	69.70 %	79.00 %	N/A	2.385 G
Post-corrected Digi index	78.10 %	78.19 %	N/A	N/A	2.385 G

Combining the results of analyses so far together, we suggest, that quality of the 19th century Finnish newspaper data from a digitized source can be estimated reasonably well with morphological analyzers of modern Finnish. Comparable edited data of the same period in Table 2 show that 50–88 % of the later 19th century clean word data (VNS, Europaeus dictionary, Ahlman dictionary) is recognized. Out of the older data (VKS, Renvall dictionary and Helenius dictionary) about 50–60 % of words are recognized. Thus the period of the data can be seen in the recognition rates to some extent. 69–79 per cent of our index data is recognized depending on the version of the recognizer. This is in the same range as recognition of clean data. With modern Finnish data Omorfi is able to analyze 92–97 % of the input words on the token level [13].

It should, of course, be kept in mind, that recognizability of words is not the same as correctness in the original text. A word may be wrongly OCRed, but still recognizable as a form of some other word. Nonexistent compounds may be recognized, if their composite parts are in the lexicon of the analyzer. As Omorfi has a very large lexicon (424 259 lexemes according to Pirinen [13]), this may cause lots of false recognitions of compounds. Many words in the Digi's database are split wrongly to parts due to hyphenation in the original text, which may cause both false positive recognition and false negative recognition. Compounds were also written differently in the 19th century Finnish. OOVs, words that are not in the lexicon of the analyzer, bring complexity of their own to results. Amount and effect of these kinds of phenomena are hard to estimate, but it is clear that all these phenomena cause uncertainty in the results and make an estimation of error margins in the analysis hard to establish.

3. Improving the Data

One of our aims is to improve word correctness of our data and our recognition procedure is important also in this respect. There are two practically possible routes to improve the data: re-OCRing of the newspaper and journal data and post-correction of the data. Due to proprietary software license of ABBYY FineReader, re-OCRing with the original OCR engine is too expensive, and we are in the process of configuring Tesseract's open OCR software to recognize our page images.

Recently we have had collaboration with the FIN-CLARIN consortium in the Department of Modern Languages at the University of Helsinki. Usage of their post correction software has yielded word recognition improvement of about 9 % units with the whole index[3] so far (cf. Table 2, last row). In Tables 3 and 4 we show recognition results with a 500 000 word sample of the data with our old OCRed material, a vendor provided ground truth out of it, a manually corrected ground truth out of the vendor GT and an ABBYY FineReader v. 11 re-OCRing of the data.

Table 3: Word recognition rates for data of the 500 000 word samples. Table legend: GT = ground truth, GT_proof = proof read ground truth, FR11= ABBYY FineReader v. 11 OCR, Old OCR = original OCR data

	Omorfi 0.1 recognition	Recognition with w/v substitution	**Omorfi 0.2 recognition**	Recognition with w/v substitution	**HisOmorfi recognition**	Recognition with w/v substitution
Old OCR	76.6 %	79.9 %	77.1 %	80.4 %	80.9 %	80.9 %
GT	80 %	92.6 %	80.6 %	93.2 %	93.5 %	93.5 %
GT_proof	80.6 %	93.8 %	81.1 %	94.4 %	94.6 %	94.7 %
FR11	84.7 %	85.1 %	85.3 %	85.6 %	86 %	86 %

Table 4: Word recognition rates for post-corrected data of 500 000 words and Digi's index

	Omorfi 0.1 recognition	**Omorfi 0.2 recognition**	**HisOmorfi recognition**
Old OCR	76.4 %	77.1 %	86.8 %
GT	79.5	80.2 %	94.4 %
GT_proof	80.0 %	80.6 %	95.1 %
FR11	81.1 %	81.2 %	91.3 %

Figures in table 3 show that there is a clear improvement in the re-OCRed data, which is recognized even better than the GT with Omorfi 01. and 0.2 With HisOmorfi and w/v substitution the situation changes: GT data is recognized best. This implies that re-OCRed data has problems with recognition of *v, w* and *m*, which was confirmed in a detailed character comparison with ISRI's OCR evaluation software. It can also be seen that differences between the recognition of Omorfi 0.1 and 0.2 are small. HisOmorfi's recognition rate with all the data is much higher. This is mainly due to the fact that HisOmorfi handles the frequently occurring *w* in the words. If w's are substituted with v's also Omorfi 0.1 and 0.2 recognize words almost as well as HisMorfi as can be seen in Table 3.

So far our work for with configuring Tesseract's open source OCR engine[4] to read our data is a work in progress. Word Error Rate (WER) for the 500 K of the old OCRed data in comparison to the ground truth is 26.10, and for the best Tesseract model 27.26. We have tried various models for teaching the Fraktur font. Our model has 203 156

[3] Correction results shown here are provided by Mr. Pekka Kauppinen, Department of Modern Languages, University of Helsinki.

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

pre-classified characters as teaching data. The model needs still a lot of improvement so that the quality gets better than the old OCR engine's level. Problems on character level include e.g. breaking letters in the newspaper images, for example *iii-m-n-u, v-w, w-m, f-s, I-J, I-1-]-i*. Broken *m,* for example, might be recognized as *ni 1n 1u iii* etc. and broken *n* as *u* or *ii*. On word level one of the problems is that documents contain multiple different Fraktur/Antiqua fonts, which appear mostly in titles and advertisements.

4. Discussion

We have analyzed in this article recognizability of 19th century OCRed Finnish newspaper and journal text in order to estimate the word level correctness of the collection. For this purpose we have used modern Finnish morphological analyzers and one analyzer, that has been modified to analyze also some historical phenomena of Finnish. Our best analysis result is 72.7 % recognized words in the 1851–1910 part of the index of the collection, which contains 99.1 % of the words of the whole collection. Out of all the analyses we can estimate, that about 69–73 % of the words in the collections are recognizable. The rest consists of OOV's, OCR errors and possible misinterpretations. All in all 20–30 % of the words in the collection are susceptible, and would need correction. In smaller samples recognition rates are slightly higher.

In Tables 3 and 4 we showed preliminary results of post-correction and re-OCRing of the data. Both post correction of the whole index and re-OCRing of a 500 K sample out of it show a clear improvement in the recognition rate. The recognition rate of the whole index improved by 8.8 % units and recognition rate of the 500 K sample with HisOmorfi with 0.5–6.5 % units. It seems thus plausible, that the overall recognition rate of the index's words could be pushed to round 80 %, but probably not much above it. Already this would improve the quality of the data significantly and further processing and use of the material would benefit. We do not believe that the worst part of the data can be corrected by these means. Perhaps only a total rescanning from original newspapers and journals (vs. microfilms) would decrease significantly the amount of wholly unintelligible data. But this is not possible or cost-effective.

Users of the Digi collection have complained about the poor OCR of the collection relatively little, but some of them have reported curious search results and been annoyed by the OCR quality [14]. Basing on the empirical search results with the evaluation collection derived from a small subset of the whole Digi material [15], it is evident that search results in the Digi collection itself are not optimal, and better OCR quality would probably improve them. Thus any improving of the word level quality is important for the collection.

Acknowledgements

This work is funded by the EU Commission through its European Regional Development Fund and the program Leverage from the EU 2014–2020.

References

[1] M.-L. Bremer-Laamanen, A Nordic Digital Newspaper Library, *International Preservation News* **26** (2001), 18–20.

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

[3] T. Pääkkönen, J. Kervinen, A. Nivala, K. Kettunen, E. Mäkelä, Exporting Finnish Digitized Historical Newspaper Contents for Offline Use, *D-Lib Magazine* **22**, (2016), http://www.dlib.org/dlib/july16/paakkonen/07paakkonen.html

[4] J. Evershed, K. Fitch, Correcting Noisy OCR: Context beats Confusion. DATeCH, 45–51, (2014), http://dl.acm.org/citation.cfm?doid=2595188.2595200

[5] D. Lopresti, Optical character recognition errors and their effects on natural language processing. *International Journal on Document Analysis and Recognition,* **12** (2009), 141–151.

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

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

[8] K. Niklas, Unsupervised Post-Correction of OCR Errors. Diploma Thesis, Leibniz Universität, Hannover, (2010), www.l3s.de/~tahmasebi/Diplomarbeit_Niklas.pdf

[9] C. Strange, J. Wodak, I. Wood, Mining for the Meanings of a Murder: The Impact of OCR Quality on the Use of Digitized Historical Newspapers, *Digital Humanitites Quarterly **8** (2014),* http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html

[10] S. Tanner, T. Muñoz, 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* July/August (2009), http://www.dlib.org/dlib/july09/munoz/07munoz.html

[11] B. Alex, J. Burns, Estimating and rating the quality of optically character recognized text, *DATeCH '14 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage,* (2014), 97–10,. http://dl.acm.org/citation.cfm?id=2595214

[12] K. Kettunen, T. Pääkkönen, Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means, LREC 2016, http://www.lrec-conf.org/proceedings/lrec2016/pdf/17_Paper.pdf

[13] T. Pirinen, Development and Use of Computational Morphology of Finnish in the Open Source and Open Science Era: Notes on Experiences with Omorfi Development. SKY Journal of Linguistics, **28**, (2015), 381–393, http://www.linguistics.fi/julkaisut/SKY2015/SKYJoL28_Pirinen.pdf

[14] T. Hölttä, Digitoitujen kulttuuriperintöaineistojen tutkimuskäyttö ja tutkijat. M. Sc. thesis (in Finnish), University of Tampere, School of Information Sciences, Degree Programme in Information Studies and Interactive Media (2016), https://tampub.uta.fi/handle/10024/98714

[15] A. Järvelin, H. Keskustalo, E. Sormunen, M. Saastamoinen, K. Kettunen, Information retrieval from historical newspaper collections in highly inflectional languages: A query expansion approach, *Journal of the Association for Information Science and Technology* (2015). http://onlinelibrary.wiley.com/doi/10.1002/asi.23379/abstract