Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means

Kimmo Kettunen, Tuula Pääkkönen
National Library of Finland, The Centre for Preservation and Digitisation
Saimaankatu 6, FI-50100 Mikkeli, Finland
kimmo.kettunen@helsinki.fi, tuula.paakkonen@helsinki.fi

Abstract

The National Library of Finland has digitized a large proportion of the historical newspapers published in Finland between 1771 and 1910 (Bremer-Laamanen 2001). This collection contains approximately 1.95 million pages in Finnish and Swedish. Finnish part of the collection consists of about 2.39 billion words. The National Library's Digital Collections are offered via the *digi.kansalliskirjasto.fi* web service, also known as *Digi*. Part of this material is also available freely downloadable in The Language Bank of Finland provided by the Fin-CLARIN consortium. The collection can also be accessed through the Korp environment that has been developed by Språkbanken at the University of Gothenburg and extended by FIN-CLARIN team at the University of Helsinki to provide concordances of text resources. A Cranfield-style information retrieval test collection has been produced out of a small part of the Digi newspaper material at the University of Tampere (Järvelin et al., 2015). The quality of the OCRed collections is an important topic in digital humanities, as it affects general usability and searchability of collections. There is no single available method to assess the quality of large collections, but different methods can be used to approximate the quality. This paper discusses different corpus analysis style ways to approximate the overall lexical quality of the Finnish part of the Digi collection.

Keywords: OCR evaluation, historical newspaper collections, Finnish

1. Introduction

Digitization of both hand-written and printed historical text material during the last 10–15 years has been an ongoing academic and non-academic industry. Most probably this activity will only increase in the current Digital Humanities era. As a result of the past and current work we have lots of digital historical document collections available and will have more of them in the future.

This paper discusses different corpus analysis style ways to approximate the overall lexical quality of the Finnish part of the Digi collection. Methods include usage of parallel samples and word error rates, usage of morphological analysers, frequency analysis of words and comparisons to comparable lexical data. Our aim in the quality analysis is to establish a set of available simple assessment methods that also build up a compact procedure for quality assessment after e.g. re-OCRing or post-correction of the material. In the conclusion part of the paper we shall synthesise results of our different analyses.

2. Problems of Optical Character Recognition

Newspapers of the 19th and early 20th century were mostly printed in the Gothic (Fraktur, blackletter) typeface in Europe. It is well known that the typeface is difficult to recognize for OCR software (Holley, 2008; Furrer & Volk, 2011; Volk et al., 2011). Other aspects that affect the quality of the OCR recognition are the following, among others (cf. Holley, 2008; Klijn, 2008, for a more detailed list):

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

As a result of these difficulties scanned and OCRed document collections have a varying amount of errors in their content. 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 (2010) that consists of a 200 year period of The Times of London from 1785 to 1985). There is no clear measure of the amount of errors that makes the material useful or less useful for some purpose, and the use purposes and research tasks of the users of the digitized material vary hugely (Traub et al., 2015). A linguist who is interested in the forms of the words needs as errorless data as possible; a historian who interprets the texts on a more general level may be satisfied with text data that has more errors. In any case, the quality of the OCRed word data is of crucial importance.

2.1. Analyzing OCRed Text

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 intellectually, by human inspection (cf. Strange et al., 2014). When the size of the collection increases, human inspection becomes impossible, or human inspection can only be used to assess samples of the collection. In our case, the size of the collection makes comprehensive human inspection impossible: almost 2 million newspaper *pages* of varying quality cannot be assessed by human labour.

Thus quality assessment of OCRed collections is most of

the times *sample-based*, as in the case of the British Library (Tanner et. al., 2009).[1] A representative part of the collection is assessed e.g. 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 quality of the collection is usage of digital dictionaries. Niklas (2010), 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 (2014). This kind of approach gives a word accuracy approximation for the data (Strange et al., 2014).

Unfortunately usage of dictionaries suits only languages like English that have only little inflection in words and thus the words in texts can be found in dictionaries as dictionary entries. A heavily inflected 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. We shall discuss this approach with our material next.

2.2. Analyzing the Digi Newspaper Collection

There has been on-going work on the assessment of the quality of the Digi since 2014. Part of this work has been described in Kettunen et al. (2014) and Kettunen (2015). These publications describe mainly post-correction trials of the newspaper material. To that effort we set up semi-automatically seven smallish parallel corpora (circa 212 000 words) upon which post correction trials were done. The results of the evaluation showed that the quality of the evaluation data varied from about 60 % word accuracy at worst to about 90 % accuracy at best, the mean being circa 75 % word accuracy. The evaluation samples, however, were small, and on the basis of the parallel corpora it is hard to estimate what the overall quality of the data is. Scarce availability of clean 19th century parallel newspaper material makes this approach also hard to continue any further and there are no resources to set up larger parallel data for evaluation purposes by ourselves. Thus another type of approach is needed.

Since the first trials we have done further work on lexical level with our data. In winter 2015 we extracted the words of the newspaper index from the page texts of the index dump. It consists of material from about 320 Finnish newspapers from the era 1770–1910. We got two different word lists: the first and smaller one consists of all the Finnish newspaper word material up to year 1850. It has about 22 million word form tokens, which is less than 1 % of the whole data. The second and more interesting word list consists of the Finnish words during the period 1851–1910 and it contains about 2.39 billion word form tokens. As the main volume of the lexical data of the collection is in the 1851–1910 section of the corpus, we shall concentrate mainly on the analysis of this part of the corpus, but will show also some results of the time period of 1771–1850.

As far as we know there is no single method or IT system available that could be used for analyzing the quality of the word data in a very large newspaper collection. Thus we ended up in using a few simple ways to approximate the quality. Firstly we analyzed all the words of the index with two modern Finnish morphological analyzers, commercial FINTWOL[2] and open source Omorfi[3]. As there is no morphological analyzer of historical Finnish available, this is the only possible way to do morphological analysis for the data[4]. We ran our data through the analyzers and counted how many of the words were known or unknown for the analyzers. Obviously the number of unknown words contains both historically unknown words for the modern Finnish analyzers (out-of-vocabulary, OOV) **and** OCR errors. Also a positive recognition does not guarantee that the word was what it was in the original text. However, when the figures of our data are compared to analyses of existing edited dictionary and other data of the same period, we can at least approximate, what amount of our data could be OCR errors.

Secondly, we made frequency calculations of the word data and took different samples out of that data for further analysis with the morphological analyzers. These analyses show a more detailed picture of the data.

Table 1 shows the recognition rates of all the word tokens and types in the Digi with the two morphological analyzers.

Collection	Number of words	Rec. by Omorfi	Rec. by FINTWOL	Type of data
Digi up to 1850 *tokens*	22.8 M	65.6 %	65.2 %	OCRed index words
Digi 1851–1910 *tokens*	2.385 G	69.3 %	N/A	OCRed index words
Digi up to 1850 *types*	3.24 M	15.6 %	14.9 %	OCRed index words
Digi 1851–1910 *types*	177.3 M	3.8 %	3.5 %	OCRed index words

Table 1. Recognition rates for word types and tokens of Digi

[1] "To discover the actual OCR accuracy of the newspaper digitization program at the BL we sampled a significant proportion (roughly 1%) of the total 2 million plus pages..." This kind of approach, where a clean parallel data for the OCRed sample is produced in house or by a contractor is also beyond our means.

[2] http://www2.lingsoft.fi/doc/fintwol/; We use FINTWOL's version 1999/12/20.

[3] https://github.com/flammie/omorfi; We use omorfi-analyse version 0.1, dated 2012.

[4] After writing the paper in autumn 2015 we were made aware of a version of Omorfi that is capable of handling some of the historical variations in the 19th century Finnish. cf. https://github.com/jiemakel/omorfi

At this stage we need also comparable recognition rates for lexical data of the same period. For comparison purposes we used material from the Institute for the Languages of Finland. From their web page[5] we collected two different word corpuses from two different historical periods of Finnish and four different dictionaries from the 19th century. Figures of this *edited* data are shown in Table 2. Sizes of dictionaries refer to dictionary entries extracted from the data, not to all of the words in the material. Unless otherwise mentioned, the data consists of word types.

Collection	Number of words	Rec. by Omorfi [6]	Rec. by FINTWOL	Type of data
VKS frequency corpus	285 K	15 %	16.6 %	15–18th century material
VKS frequency corpus tokens	3.43 M	49 %	50.3 %	15–18th century material
VNS frequency corpus [7]	530 K	55. 9 %	58. 1%	19th century material
VNS frequency corpus tokens [8]	4.86 M	86.1 %	86.5 %	edited 19th century material
Ahlman dictionary 1865	91.4 K	73 %	71.5 %	dictionary material
Europaeus dictionary 1853	43.2 K	76 %	69 %	dictionary material
Helenius dictionary 1838	25.8 K	49 %	50 %	dictionary material
Renvall dictionary 1826	25.8 K	43 %	45.5 %	dictionary material
Four dictionaries combined	132.5 K	62 %	61 %	dictionary material

Table 2. Historical edited word data from Institute for the Languages of Finland

We can summarize the recognition rates of the Digi and comparable same period materials as a graph shown in Figure 1. Dictionary data is on the left side of the figure, data from Digi and other clean corpora than dictionaries on the right side.

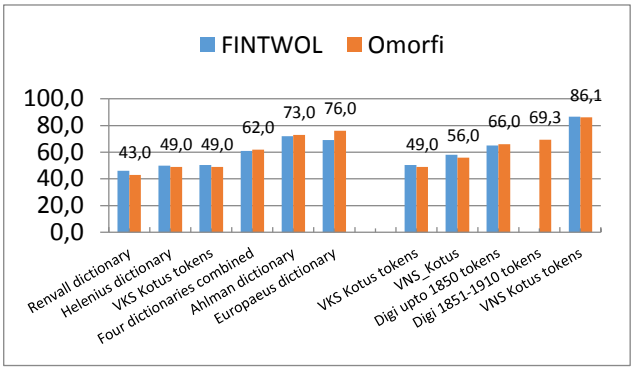

Figure 1. Recognition rates of OCRed and hand edited data. Token level data marked separately, percentages for Omorfi's recognition shown in numbers.

As can be seen from Table 1 and Figure 1, type level recognition rates of the Digi data are very low compared to clean hand edited comparable material of the same period. The main reason for this is the high number of hapax legomena words, once occurring words, most of which are OCR errors (Kettunen & Pääkkönen, 2016; cf. also Ringlstetter et al., 2006). When token level of the Digi data is looked at, recognition rates are quite reasonable, 66–69 %. There is a 17–20 % unit difference to the clean comparable data on token level. The quite high recognition rate of the VNS corpus is partly due to the fact that 1000 most frequent types in the corpus consist already 44.6 % of the whole corpus on token level and among these 2.17 M tokens recognition rate is 99.2 % (cf. also footnote 11 in section 2.3).

Out of this data we can approximate, that 56–76 % of the words on type level from 19th century data can be recognized by modern language morphological analyzers. On token level the recognition rate can be up to 86.5 %. If there is older material in the data, recognition will drop, and the drop can be quite large. Interestingly, the small earlier part of the Digi corpus has a better recognition rate on type level than the large Digi data. This might hint that there are less OCR errors in this part. This might be due to different printing style of the older material: font sizes were bigger and many times only one column was used for printing, which helps optical character recognition.

Figure 2 shows recognition rates by decade (data from 1780 to 1820 not included here, it contains only magazines).

5 http://www.kotus.fi/aineistot/tietoa_aineistoista/sahkoiset_aineistot_kootusti

6 Recognition rates of Omorfi are slightly harmed by the fact that all the data is lower cased and Omorfi recognizes proper nouns from upper case initial letter only

7 This material contains the dictionary materials

8 This material contains the dictionary materials

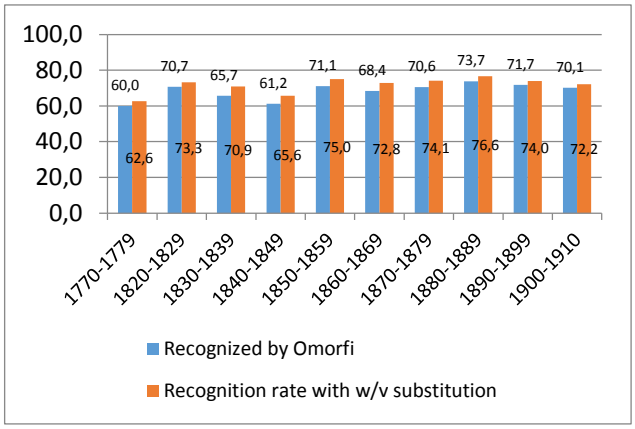

Figure 2. Recognition rates of OCRed data by decade

Figure 2. shows that recognition level decade-by-decade keeps quite even, ranging from 60 per cent to 74. If w's are substituted with v, recognition rates are slightly higher, as will be discussed in section 2.3.

Next we proceed to frequency analysis of parts of the data. Table 3. shows results of morphological analyses of 1000–1 M of the most frequent word types of the 1851–1910 part of the Digi word data on token level. For simplicity we analyzed the words on token level with FINTWOL only.

N of word types in the sample	N of word tokens	Unknown word tokens for FINTWOL	
1K	790 710 542	61 170 210	7.7 %
10K	1 317 532 256	152 388 093	11.6 %
100K	1 782 767 935	287 109 856	16.1 %
500K	1 983 275 749	387 237 305	19.5 %
1 M	2 043 976 151	427 214 868	20.9 %

Table 3. Number of unrecognized tokens for FINTWOL in 1000–1 M of the most frequent word types

Data in Table 3 shows that the 1M of the most frequent words is of quite good quality. On token level only about 21 % of them are unrecognized by FINTWOL, on type level the percentage is about 58 %. The bottom line of Table 3 is that 1.62 G of the tokens of 1 M of the most frequent word types is recognized. Out of the whole data this is 67.6 %. Thus for the rest circa 765 M of tokens the recognition rate is very low, only 30 M of them are recognized.

In Kettunen and Pääkkönen (2016) we analyzed also the least occurring word types (N=1–10) in the Digi 1851—1910. Due to space restrictions we only quote the main result of the analysis here; 85–98 % of these types are unrecognized by the recognizers, which means that most of the least occurring words are probably OCR errors. In sum they make about 309 M word tokens.

2.3 Other considerations

Orthography of Finnish was already reasonably stable in the mid-19th century, although there were phenomena that differ from modern language (cf. table 1. in Järvelin et. al, 2015). Also dialectical word forms were more common in newspapers of the 19th century. The biggest and most visible difference between modern Finnish and 19th century Finnish is variation of *w/v*, which does not exist in modern language. Thus words that have *w* and are not proper names like *Wien* (*Vienna* in Finnish) are not usually recognized by modern morphological analyzers. To approximate effect of this, we counted the occurrences of *w* in the 1 M of the most frequent words of the data. The data contains 92 749 word types with w, which makes 78 438 010 tokens (3.3 % of all the tokens). Out of the types 91 886 (99.1 %) are unrecognized by FINTWOL. This is 76 450 673 words (97.5 %) on token level. If we replace *w*'s with *v*'s, 54 049 types (58.3 %) are unrecognized by FINTWOL on type level and 24 016 996 (30.6 %) on token level.[9] Out of the whole 2.044 G of word tokens of the 1 M of most common types this makes 2.2 %. So effect of the w/v variation among the unrecognized words is significant although the number of the words in all the data is not very high.

To get an approximation of relation between OOV words of the analyzers and OCR errors proper we browsed through the 1 000 most common word types and their 120 unrecognized word types to FINTWOL. Out of these about 85 (70. 8 %) can be considered to be OCR errors, the rest being OOV's. Demarcation out of textual context is not always clear, but we can take the 70 % OCR error figure as a low estimate, and as OCR errors tend to increase with less frequent word types, OCR error percentage could be about 70–90 %.

2.4 Summing up

We have now reached a reasonably comprehensive result out of the quality assessment of our data. We have three different parameters that affect the results: number of OOV vocabulary, number of OCR errors proper and effect of *w/v* variation in the data. The effect of the OOV factor in the clean VNS_Kotus data is on token level about 14 % and in the VKS_Kotus about 50 %. Their mean is 32 %, but a fair approximation could be 14–20 % in edited material of the latter part of the 19th century word data. We believe that in the Digi data OCR errors tend to override vastly the OOV words. The variation of w/v has an effect of about 12 % among the unrecognized words.

The initial analysis without considering the w/v variation and effect of OOV's is shown concisely in Figure 3. We have 1.65 G of recognized words and 733 M of unrecognized words (cf. Table 1.). The 69 % recognition rate can be called *raw recognition rate* of the data.

[9] When same analysis is carried out with the 4.86 M tokens of the VNS_Kotus, the original recognition rate of 86.5 improves only to 86.7 %. VNS_Kotus data contains considerably less w's due to the editing policy of the data, cf. http://kaino.kotus.fi/korpus/1800/viite/1800-luvun_korpuksen_koodauksesta.php

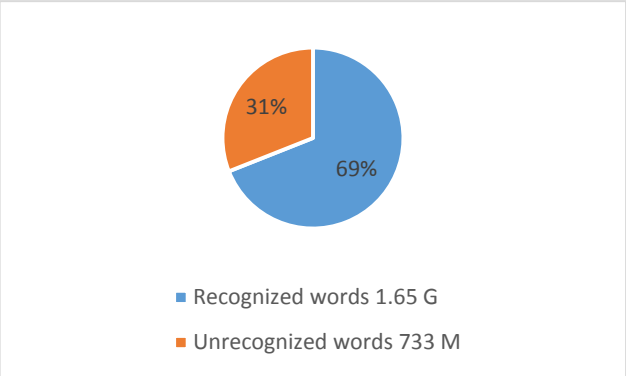

Figure 3. Raw recognition rates for the data

We can now proceed to a more detailed analysis of the 733 M unrecognized words. It is safest to assume, that *w/v* variation and OOV's have most of their effect in the 427 M part of the unrecognized words (cf. Table 3), because they belong to the most frequent words and constitute 85.6 % of the whole word data. If the number of words recognized when *w*'s are replaced with *v* (52 M) is taken into consideration, the share of recognized words goes up to 71.3 % and share of unrecognized words drops to 28.7 %, absolute number of unrecognized remaining words being circa 375 M. The approximate share of OOV words among the still unrecognized 375 M of words could be somewhere between 50–75 M. Thus the real number of OCR errors in the whole data is round 600–625 M, approximately about 25–26 % of the whole. The *approximated recognition rate* of the words in the data could thus be 74–75 %.

3. Conclusion

Out of all analyses presented, we can make the following conclusions:

- Main part, over 99 % of the data in the Digi is between the years 1851–1910. This implies that the effect of OOV words in the data should not be prohibitively large for analysis with modern language morphological analyzers (Tables 1. and 2.)
- Vocabulary of this time period can be recognized reasonably well with modern language morphological analyzers; about 56–76 % of the word types in edited data are recognized, on token level the recognition rate can be as high as 86.5 % (Table 2.)
- In the Digi data, raw token level recognition rate is round 65–69 % (Table 1.) When effect of v/w variation and OOV words is taken into account, the approximated recognition rate is about 74–75 %; This is near the 75 % mean word correctness of our earlier sample-based analysis in Kettunen et al., 2014
- modern Finnish morphological analyzers lack about 20-50 % of the vocabulary of the mainly 19th century edited data on type level; on token level the figure is at minimum about 14 % (Table 2.)
- In up to almost 75 % of the most frequent data (almost 1.8 G of words), raw recognition rate of the word form tokens is about 68.5–70 %. After this the recognition rate drops heavily (Table 3.)

The main result that our analysis gives is that the collection has a relatively good quality part, about 69–75 %, and a very bad quality part, about 9-12 %. The set of about 13 % of the words that are not recognized is harder to estimate. As they belong to the most frequent part of the data, they could be at least partly easier OCR errors and OOVs. Anyhow, about a 25-30 % share of the collection needs further processing so that the overall quality of the data would improve.

It is apparent that we need to be cautious in conclusions, as different data are of different sizes which may cause errors in estimations (Baayen, 2001; Kilgariff, 2001). However, we believe that our analyses have shed considerable light into quality of the Digi collection and our procedure can be used for quality approximation also after possible improvements in the data.

At this stage we can also reflect usefulness of the analysis procedure from point of view of improvement of the OCR quality of the Digi collection. If correction of the data is performed it should be focused on the 24–25 % unrecognized part of the data. Out of this the ca. 300 M possibly easier part could perhaps be improved by post-correction of the material with algorithmic correction software. We have tried post-correction with a sample (Kettunen, 2015), but the results were not good enough for realistic post-correction. If post-correction would be focused to only the easier part of the Digi's erroneous data, it could work quite well. General experience from algorithmic post-correction of OCR errors shows, that good quality word material can be corrected relatively well (e.g. Niklas, 2010**;** Reynaert, 2008). This may also apply to the medium quality word data. But the worst 9–12 % part of the Digi data cannot be corrected with post-processing; only re-OCRing could help with it, as there is so much of it.[10]

Taken that some action had been taken to improve the quality of the Digi data, we have to consider, whether our procedure would be useful in showing quality improvement, if such had been achieved. We suggest that improvement of the lexical quality could be shown e.g. with following analyses:

- clear improvement in overall recognition rate of the data: at least 3–5 % units in both type and token level analyses
- recognition rate in the top 1 M of the most frequent word types should improve significantly, especially in the 100 K–1 M range, that is now beyond mean recognition rate of edited data
- a very large drop in the number of unrecognized hapax legomena and other rare word types; in practice this would mean tens of millions of word forms to be become recognizable.

4. Acknowledgements

First author is funded by the EU Commission through its

[10] We have ongoing trials and evaluations with both post-correction and re-OCRing.

European Regional Development Fund and the program *Leverage from the EU 2014–2020.*

5. Bibliographical References

Alex, B., Burns, J. (2014). Estimating and rating the quality of optically character recognized text. In *DATeCH '14 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage,* pp. 97–102. http://dl.acm.org/citation.cfm?id=2595214

Baayen, H. (2001). *Word Frequency Distributions*. Kluwer Academic Publishers.

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

Furrer, L., Volk, M. (2011). Reducing OCR Errors in Gothic-Script Documents. In *Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop*, pp. 97–103, Hissar, Bulgaria, pp. 97–103. http://www.aclweb.org/anthology/W11-4115

Holley, R. (2009). 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

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

Kettunen, K. (2015). Keep, Change or Delete? Setting up a Low Resource OCR Post-correction Framework for a Digitized Old Finnish Newspaper Collection. *11th Italian Research Conference on Digital Libraries - IRCDL 2015*, Bozen-Bolzano, Italy, 29–30 January, 2015. http://ircdl2015.unibz.it/papers/paper-01.pdf.

Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T., Kervinen, J. (2014). 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. http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-honkela-en.pdf

Kettunen, K., Pääkkönen, T. (2016). How to do lexical quality estimation of a large OCRed historical Finnish newspaper collection with scarce resources. Submitted.

Kilgariff, A., (2001). Comparing Corpora. *International Journal of Corpus Linguistics* 6:1, 97-133.

Klijn, E. (2008). The Current State-of-art in Newspaper Digitization. A Market Perspective. *D-Lib Magazin* January/February. http://www.dlib.org/dlib/january08/klijn/01klijn.html

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

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*. Lecture Notes in Computer Science vol. 4919, pp. 617–630.

Ringlstetter, C., Schulz, K., Mihanov, S. (2006). Orthographic Errors in Web Pages: Toward Cleaner Web Corpora. *Computational Linguistics* 32(3), pp. 295–340.

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

Tanner, S., Muñoz, T., 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. *D-Lib Magazine* July/August http://www.dlib.org/dlib/july09/munoz/07munoz.html

Traub, M. C., Ossenbruggen, J. van, Hardman, L. (2015). Impact Analysis of OCR Quality on Research Tasks in Digital Archives. In *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science vol. 9316, pp. 252–263.

Volk, M., Furrer, L., Sennrich, R. (2011). Strategies for reducing and correcting OCR errors. In C. Sporleder, A. Bosch, & K. Zervanou (Eds.), *Language Technology for Cultural Heritage,* Springer-Verlag, Berlin/Heidelberg, pp. 3–22.

6. Language Resource References

Ahlman dictionary. Helsinki: Kotimaisten kielten tutkimuskeskus. Available at: http://kaino.kotus.fi/korpus/1800/meta/ahlman_sanastot/ahlman_sanastot_coll_rdf.xml

Europaeus dictionary. Helsinki: Kotimaisten kielten tutkimuskeskus. Available at: http://kaino.kotus.fi/korpus/1800/meta/europaeus_sanastot/europaeus_sanastot_coll_rdf.xml

Helenius dictionary. Helsinki: Kotimaisten kielten tutkimuskeskus. Available at: http://kaino.kotus.fi/korpus/1800/meta/helenius/helenius_coll_rdf.xml

Renvall dictionary. Helsinki: Kotimaisten kielten tutkimuskeskus. Available at: http://kaino.kotus.fi/korpus/1800/meta/renvall/renvall_coll_rdf.xml

VKS frequency corpus (Corpus of Old Finnish). Helsinki: Kotimaisten kielten tutkimuskeskus. Available at: http://kaino.kotus.fi/sanat/taajuuslista/vks.php

VNS frequency corpus. (Corpus of Early Modern Finnish) Helsinki: Kotimaisten kielten tutkimuskeskus. Available at: http://kaino.kotus.fi/sanat/taajuuslista/vns.php.