Assessing lexical quality of a digitized historical Finnish newspaper collection with modern language technology tools

Kimmo Kettunen and Tuula Pääkkönen
National Library of Finland, Centre for Digitization and Conservation
kimmo.kettunen@helsinki.fi, tuula.paakkonen@helsinki.fi

Digitization by means of scanning and optical character recognition (OCR) of both hand-written and printed historical 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 ongoing Digital Humanities era. As a result of past and current work we have lots of digital historical document collections available and will have more of them in the future.

The National Library of Finland has digitized a large proportion of the historical newspapers published in Finland between 1771 and 1910 (Bremer-Laamanen 2014; Kettunen et al. 2014). 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 the newspaper material (years 1771–1874) is also available freely downloadable in The Language Bank of Finland provided by the FinCLARIN consortium[1]. The collection can also be accessed through the Korp[2] 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 also been produced out of a small part of the Digi newspaper material at the University of Tampere (Järvelin et al. 2015).

The web service *digi.kansalliskirjasto.fi* contains different material besides newspapers, including journals, and ephemera (different small prints). Recently a new service was created: it enables marking of clips and storing of them to a personal scrapbook. The web service is used, for example, by genealogists, heritage societies, researchers, and history enthusiast laymen. There is also an increasing desire to offer the material more widely for educational use. In 2014 the service had over 10 million page loads. User statistics show that about 88.5 % of the usage of the Digi comes from Finland, but a 11.5 % share of use is coming outside of Finland.

Quality of OCRed collections is an important topic in digital humanities, as it affects general usability and searchability of collections (Holley, 2008, Tanner et al., 2009). There is no single available method to assess quality of large collections, but different methods can be used to approximate quality. This paper discusses different 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 analysers, frequency analysis of words and comparisons to comparable edited lexical data. Our aim in the quality analysis is twofold: firstly to analyse the present state of the lexical data and secondly, to establish a set of assessment methods that build up a compact procedure for overall quality assessment after e.g. re-OCRing or post-correction of the material. In the discussion part of the paper we shall synthesise results of our different analyses.

Our results show, that about 69 % of all the word tokens of the Digi can be recognized with a modern Finnish morphological analyser. If orthographical variation of *v/w* in the 19th century Finnish is taken into account and number of out-of-vocabulary words (OOVs) is estimated, the recognition rate increases to 74–75 %. The rest, about 625 M words, is estimated to consist mostly of OCR errors, at least half of them being hard ones. 1 M most frequent word types in the data make

[1] https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/KielipankkiAineistotDigilibPub

[2] https://korp.csc.fi/

2.043 billion tokens, out of which 79.1 % can be recognized. If words that occur only once in the data (hapax legomena) are analysed, 98 % of them are unrecognized by morphological software.

The lexical quality approximation process we have set up is relatively straightforward and does not need complicated tools. It is based on frequency calculations and usage of off-the-shelf modern Finnish morphological analyzers. Even though we have done the estimation now in a partially automatized way, it is possible to automatize the operation completely. It is also 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.

Acknowledgements

This work is supported by European Regional Development Fund (ERDF), Leverage from the EU, 2014–2020.

References

Baayen, H. (2001). *Word Frequency Distributions*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bremer-Laamanen, Maj-Lis, 2014. In the spotlight for crowdsourcing. Scandinavian Librarian Quarterly 1: 18–21.

Holley, Rose. 2008. 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, Anni, Keskustalo, Heikki, Sormunen, Eero, Saastamoinen, Miamaria and Kettunen, Kimmo, 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

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

Kettunen, Kimmo, Honkela, Timo, Lindén, Krister, Kauppinen, Pekka, Pääkkönen , Tuula and Kervinen, Jukka, 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

Tanner, Simon, Muñoz, Trevor and Ros, Pich Hemy, 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