Evaluation and refinement of an enhanced OCR process for mass digitisation

Dana Dannélls[1], Torsten Johansson[2], and Lars Björk[2]

[1] Språkbanken, University of Gothenburg, Sweden
dana.dannells@gu.se
[2] Kungliga biblioteket, Stockholm, Sweden
{lars.bjork/torsten.johansson}@kb.se

Abstract. Great expectations are placed on the capacity of heritage institutions to make their collections available in digital format. Data driven research is becoming a key concept within the humanities and social sciences. Kungliga biblioteket's (National Library of Sweden, KB) collections of digitised newspaper can thus be regarded as unique cultural data sets with information that rarely is conveyed in other media types. The digital format makes it possible to explore these resources in ways not feasible while in printed form. As texts are no longer only read but also subjected to computer based analysis the demand on the correct rendering of the original text increases. OCR technologies for converting images to machine-readable text play a fundamental part in making these resources available, but the effectiveness vary with the type of document being processed. This is evident in relation to the digitisation of newspapers where factors relating to their production, layout and paper quality often impair the OCR production. In order to improve the machine readable text, especially in relation to the digitisation of newspapers, KB initiated the development of an OCR-module where key parameters can be adjusted according to the characteristics of the material being processed. The purpose of this paper is to present the project goals and methods.

Keywords: Language Technology · OCR · Digitisation

1 Introduction

Great expectations are placed on the capacity of heritage institutions to make their collections available in digital format. Large collections of digitised newspaper are unique cultural data sets with information that rarely is conveyed in other media types. Newspaper collection which are available in digital format make it possible for researchers in the digital humanities to explore these resources in ways not feasible while in printed form. Consequently, data driven research is becoming a key concept within the humanities and social sciences.

As texts are no longer only read but also subjected to computer based analysis the demand on the correct rendering of the original text increases. OCR technologies for converting images to machine-readable text play a fundamental part in making these resources available, but the effectiveness vary with the type of document being processed. This is evident in relation to the digitisation of newspapers where factors relating to their production, layout and paper quality often impair the OCR results.

At the National Library of Sweden (Kungliga biblioteket, KB) and at the Swedish Language Bank (Språkbanken, SB) at the University of Gothenburg, we have recently embarked on a two year project to improve the machine readable text, especially in relation to the digitisation of newspapers. The purpose of this project is to carry out a formal evaluation of, and improve an OCR process through systematic text analyses, dictionaries and word lists with the aim of implementing it in the mass digitisation process. Another objective of this project is to further investigate the feasibility of establishing a model for the benchmarking of OCR processes.[3]

Possible results of such an undertaking could be a collection of reference material that can be fed through the digitisation production line in order to evaluate the OCR process and a framework for quality standards that can be used to classify products of various types of processes. At present there are no such benchmarking tools or standards available, neither nationally nor on an international level. Concepts such as trustworthiness and the degree of correct rendering of the original text have to be evaluated in ways that correspond with new modes of use. This project aims at addressing such needs by focussing on the quality and usability of digital text resources.

1.1 Newspaper digitisation at the KB

In 2014 KB replaced the microfilming of legal deposit newspapers with digitisation. As a result, and with the help of external funding, KB's collection now holds more than 20 million pages of digitised historical and contemporary newspapers, a collection that is estimated to reach 31 million pages within the coming four years.

As texts in digital format no longer only are read as images on a display but also subjected to computer based analysis, new requirements are placed on the data provided being true to the original text. OCR processing was initially introduced in the mass digitisation of newspaper in order to enable indexing and retrievability of relevant articles. The production of machine readable text had the purpose of facilitating manoeuvring in image-based representation of text-pages, it was not a means of extracting text resources for further analysis or reuse. The increased reliance on large amount of data within the humanities and social sciences has resulted in new demands on the resources produced by the digitisation of newspaper. The proposed project will address this need by improving and benchmarking the OCR process and hence the accessibility to reliable data resources for research purposes.

[3] The project is financed by a grant from the Swedish Tercentenary Foundation (Riksbankens Jubileumsfond, RJ) 2019-2020.

1.2 Språkbanken

SB offers access to historical and modern corpora and lexicons which are freely searchable through SB's corpora and lexical research infrastructures Korp and Karp [6, 5]. These infrastructures are interlinked through their annotations that are assigned by Korp's annotation pipeline. Annotations include linguistic information, i.e. part-of-speech, morphology, syntactic and semantic annotations. The rich linguistic information the historical and modern corpora are equipped with allows for searching and extracting valuable pieces of information from them. Currently, SB offers access to 13 billion words of modern texts, 3 billion words of historical texts, and nearly 1 million lexical entries. Among this data is a set of old newspaper texts from the Digidaily project at KB.[4] The amount of OCR errors is high in some parts of this collection, a limitation which leads to increasing annotation errors. This, in turn, has a huge impact on the users who wish to access and explore this material. One of the main contributions of this project is to improve the annotations of historical material, making it more reliable and searchable through SB's annotation pipeline.

2 Background

During 2017 a test platform for OCR processing (henceforth the OCR-module) was developed by Kungliga biblioteket in cooperation with the Norwegian software company Zissor.[5] The module is designed to enable adjustment and control of some key parameters of the post-capture stage of the OCR process, e.g. dictionaries and linguistic processing, to match typical features of the newspaper as a printed product, characteristics that in a historic perspective change over time, such as layout, typography, and language conventions. The approach taken in the design of the module (outlined in the following section), where the verification of the correctness is based on the evaluation of the results from two separate programs, will enable the development of reliable indicators of the quality of an OCR produced text.

The underlying principle is to utilise the individual differences in capacity of two commonly used OCR programs: ABBYY FineReader,[6] and Tesseract,[7] by processing the image-file with the two programs, comparing the results (on the word level) and choosing the output that has the highest validity according to a scoring system. The approach is discussed by e.g. [8], [20], and [15] but has not been tested in mass digitisation of newspaper.

[4] The corpus is now going under the name KubHist (Kungliga bibliotekets historiska tidningar) in Korp.

[5] www.zissor.com

[6] https://www.abbyy.com /en-eu/, (ver. 11)

[7] https://github.com/tesseract-ocr/, (ver. 3.05.01)

2.1 Development of the OCR-module – phase 1

The OCR-module was based on one of Zissor's products – Zissor Content System – an article-segmentation and OCR processing application. This application enables a high degree of control and automation of the segmentation and OCR conversion (including the use of dictionaries). Zissor Content System uses ABBYY FineReader for the OCR and the initial step was to integrate the second OCR program, Tesseract, in the application. A great challenge in this phase was to match the two OCR processes with respect to the production of the ALTO (Analyzed Layout and Text Object) XML-files, i.e. a standard for storing layout and content information of diversified digital objects.

A distinguishing feature of the approach taken in the development of the OCR-module is the production of an individual ALTO XML-file for each OCR program. The ALTO XML-file holds the positions of words and layout features in the digital image, thus enabling a direct comparison between the outputs – on word level – of the two programs. The importance of detailed documentation and the possibility to relate the results of the OCR production to the digital image are important features when evaluating the precision of the OCR process, as underlined in [19].

The first test phase focussed on a manual analysis of the variations between the two programs. 532 digitised pages (1831, 4p. 1840, 4p. 1945, 48p. 2002, 96 p. and 2010, 284 p.) were processed separately in ABBYY and Tesseract and the analysis consisted of visually comparing the results from the two OCR programs. The OCR-module was used for this purpose as it enables a comparison between the zoning and segmentation applied in the process as well as the words confirmed or rejected by the dictionaries. No ground-truth was available for this test, which means that the results are approximations. Among the observations it can be noted that ABBYY appears to perform better than Tesseract with respect to OCR and has a higher precision in its zoning. Tesseract appears to be better at handling variations in font type and sizes, although being more sensitive to background noise in the image. Most importantly, the coordinates for words and letters in the ALTO-XML-files from the two OCR programs were the same, thus verifying the possibility to automatise the process of comparing the two outputs. On the basis of this test it was decided that the OCR-module should use ABBYY for zoning as the Tesseract zoning regularly generated overlapping text fields.

2.2 Development of the OCR-module – phase 2

A seminar was held at KB in the spring of 2017 with representatives from KB, Zissor, SB, and MKC (Media Conversion Center – the mass digitisation facility run by the National Archives, where KB's newspapers are digitised) in order to discuss how these initial findings could be utilised and automatised with the aim of improving the output of the OCR process.

Based on the outcome of that seminar a second phase was initiated. Two goals were set for this phase: (1) implement the parallel processing of the image file in the two OCR programs, and (2) automatise the handling of dictionaries. Based on a rule set for the selection of words in the OCR process, should ABBYY's and Tesseract's suggestions differ.

2.3 The OCR-module – present version

A scoring model is implemented, based on the dictionaries of the two OCR programs. The comparison and process of selection between the two outputs of the OCR conversion is now fully automated. Each individual word is either verified (if confirmed by both dictionaries) or falsified (if rejected by one or both dictionaries) and subjected to further comparison, according to the rule set in the scoring model.[8] Figure 1 demonstrates the workflow in the OCR-module.

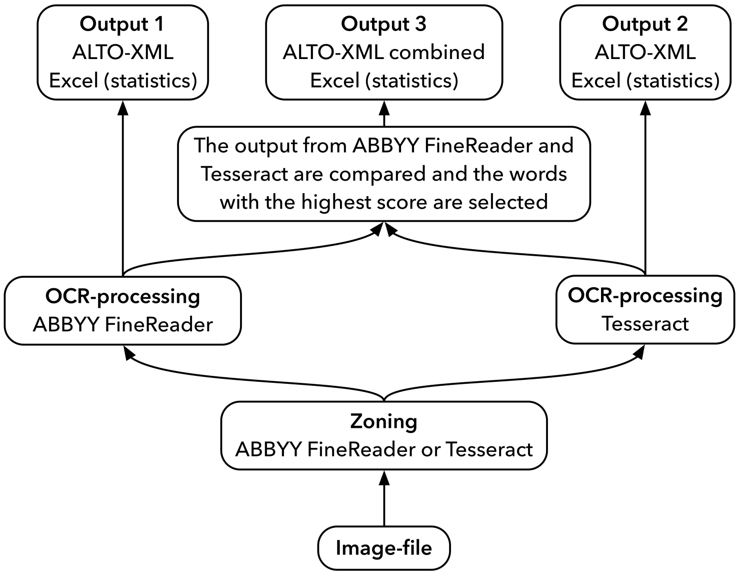

Fig. 1. The workflow in the OCR-module.

The OCR-module is designed in such a way that the results are fully traceable by maintaining a link between the individual words, the OCR program and dictionary that was involved, and the position of the word in the actual image file. Three ALTO XML-files are generated for each page, one for each OCR program and a third with the combined results from the scoring process. Statistics as to errors and verifications regarding the three ALTO XML-files is also generated as an Excel-file for each page. In this way the consequences of modifying the parameters in the OCR process can be closely monitored.

[8] The user documentation is available from KB on request (Zissor Content System 2.0, User Documentation, January 2018).

3 Material and method

3.1 Reference material

The first part of our project will be devoted to creating a reference material, a selected sample of already digitised newspapers, to which the results from the OCR-module will be compared for evaluations. We are aiming at a set of approximately 400 pages covering the time period 1818–2018 with individual articles, advertisements and other types of text components managed as separate xml-files in the transcription. The selected sample of the reference material will be carefully chosen to reflect typical variations in layout, typography, and language over time. We will use both statistical and manual selection procedures to ensure that the material is broad and representative.

The reference material will be transcribed (double keyed) by human annotators in order to provide a ground truth. The transcription will be made externally by a company who specialises in double keying. We expect the transcription process, including preparation of source material, to take three months. For consistent annotations of the material, we will define the guidelines for the transcription. These will be based on our inspections of the selected material; they will contain instructions for typeface, size and location changes. We aim to make the complete reference material along with its manual transcriptions and the guidelines for producing them freely available (as far as possible, within the legal restrictions caused by the present copyright legislation) under the open Creative Commons Attribution-Share Alike licence (CC BY-SA).

3.2 Analysis and characterisation of the material

Once the ground truth is produced the image files that correspond to the transcribed pages will be processed by the OCR-module. The resulting text will be compared with the ground truth to define errors on character and word levels and to evaluate the accuracy of our OCR-module. Two general approaches will be applied: quantitative analysis, performed automatically by machines, and qualitative analysis, performed manually by human annotators. Previous approaches for improving the accuracy of the OCR results have considered quantitative analysis [13]. The advantage of quantitative analysis is that it accurately and very quickly produces generalisation and summarization of the input data. However, the disadvantage is that the analysis uses formal procedures that basically only provide statistical information about the data. Qualitative analysis, on the other hand, provide in-depth explanatory data. The disadvantage is that it requires human expertise and is time-consuming. Both approaches are however important to achieve this project objectives. Quantitative analysis is necessary for extracting word and character errors as the first step before applying post-processing methods. Qualitative analysis is necessary for learning about the peculiar typology exhibited in the data, which in itself would constitute a valuable contribution from the analysis undertaken in this project.

Since we have a fair amount of transcribed data with anticipated 300K transcribed words that we will produce at the first phase of the project, we will start out by statistical analysis. For this we will consider the OCR toolkit for measuring the performance of OCR systems developed by the Information Science Research Institute at the University of Nevada [2]. This group has developed several matrices that are attributed to variants of Levenshtein Edit Distance, i.e. mathematical methods for measuring the distance between two strings, in our case two sequences of characters and alphanumeric symbols, by calculating the minimal number of symbol insertions, deletions or substitutions between the sequences [14]. Statistical analysis of the data will provide us with performance measures including character and word accuracy, and frequencies of edit operations. The results from the statistical analysis will help us to identify the most frequent edit operations causing OCR errors, and propose suggestions for correcting wrongly recognised words. A similar approach was presented by [9] who corrected OCR errors in a German and French corpus. This approach will aid the development of a dictionary and context based post-correction method which we will experiment with in order to increase the accuracy of the OCR-module.

A challenging aspect of dictionary and context based method is the lack of access to dictionaries with high coverage. Språkbanken provides access to a large set of modern and historical computational lexical resources for Swedish, all interlinked on the lexical sense level. However, these lexical resources are by far insufficient to cover the whole time period we are aiming at. Two major lexical resources are in particular interesting to explore in this project: Dalin and SALDO, which comprise nearly 200K entries including a full morphological description of the vocabulary, covering late-modern Swedish (1831–1905), and modern Swedish (1906–2017) [4]. These dictionaries have been explored in the project "A free cloud service for OCR" [3] that have shown their potential to improve OCR errors when combined with n-gram lists complied from corpora. Here we follow the same lines as in the previous project. Similarly, in addition to dictionaries, we will explore the use of wordlists which we will compute automatically from 38 historical and modern corpora of 980 million annotated tokens from the period of 1800 until today.

3.3 Specific issues to be addressed

Some pilot studies exploring OCR errors of historical Swedish material [3, 7] found three major types of OCR errors that require systematic and consistent post-correction method: (1) graphemic similarities (e.g. "a" is confused with "c", "å" is recognised as "ä", or one character is misplaced with several, e.g. "cl" instead of "d"), (2) wrong recognition of white space and incorrect word split (e.g. "allamedhstärsta" instead of "alla medh största"), and (3) errors in name entities recognition (e.g. "KarinsSundh?" instead of "Karins Sundh"). Each of these error types will be addressed in this project to improve the performance of the OCR-module.

Many of the graphemic similarity errors, also similar to those that have been reported in [13] and [20], are common OCR problems which occur both in historical and modern texts. Recent research [11] has shown that context-sensitive replacement rules are an efficient method for correction OCR errors. We will apply a rule-based method to capture some frequent and some less frequent character replacements. It might be impossible to cover all miss-recognised character variations but since character replacements often reoccur throughout the document, it is known to be a reliable and efficient method, which leads to error reduction [20]. This method will be combined with the Levenshtein distance approach that assigns probabilities to string candidates from large corpora [1]. Here we will start with a few historical and modern corpora and increase the data incrementally. An important outcome from this analysis is a resource of orthographic replacement rules that is typical for different time periods.

To correct the errors caused by word boundary recognition we will experiment with frequency word lists including compounds and multi-word units analysis that is available in SALDO. There are a number of possible methods to address word boundary recognition errors using morphological analysis [10, 17]. We will follow the approach applied for Icelandic [10] and use dictionaries and morphological description to find word boundaries and identify compounds. Compound analysis for modern corpora has been recently improved in the Språkbanken's annotation pipeline. However, there is no computational compound analysis available for SB's historical dictionaries. The comprehensive word boundary analysis that we will carry out in this project may lead to an improvement of the compound analysis at SB in general and Dalin in particular.

Identifying person and place names is an extremely difficult problem. In prints around 1800 we find references to place names with large spelling variation or to places which no longer exist. One way to address this is by compiling gazetteer lists about name entities. Our major sources for compiling these lists will be corpora annotated with name entities and historical maps. Lantmäteriet (the Swedish mapping, cadastral, and land registration authority) has recently released historical maps with place names from the period of 1628–1880. Historical name entities, in particular folklore in different areas of Sweden is a major focus of research of the Swedish Institute for Language and Folklore, we will explore their name, organisation and place name lists in our approach. In addition, we will exploit other types of authorities that are managed by KB and the National Archives. The gain of compiling gazetteer lists for the proposed time period is twofold: improve the name entity annotations at Språkbanken and contribute with a valuable resource for historians.

Quality assessment of OCR plays an essential role in the project and will be carried out throughout the project period. Since OCR processes produce significant amount of OCR errors it is desirable to have quality measures for the entire material, as well as for parts of the material. In this relation it is also important to have access to: (1) tools for finding and extracting the correct lexical items directly from the source, i.e. the image, and (2) dictionaries and word-lists for ranking the corrected lexical items along with their possible interpretations. At this stage the impact of layout, (e.g. irregular columns), typography (e.g.

mixture of small fonts and oversized headings) and language conventions on the overall result is documented as well as the occurrence of repetitive errors.

The tools we will explore for annotating and manually correcting OCR errors are freely available open source interactive tools developed within the IMPACT project.[9] These will facilitate and reduce the annotation effort and time of the annotators and at the same time constitute a valuable process of estimating errors and adjusting parameters. This will have to be performed several time for each batch of samples, in order to arrive at an optimal and stable result. Quality assurance and consistency checking of the manual annotation will be done automatically.

4 Evaluation of the OCR-module

4.1 Case 1: Evaluation based on comparison with a ground truth

A selection of image files will be OCR processed and the result is compared with the ground truth produced by the same material. On the basis of this evaluation the functionality and the application of dictionaries and word-lists of the OCR-module will be subject to further development and refinement with the aim at achieving optimal correctness in the process.

4.2 Case 2: Evaluation based on comparison with previously processed material

A selection of already OCR processed image files will be re-processed in the OCR-module. The resulting textfiles will be compared with previously produced textfiles and the ground truth in order to evaluate the differences in capacity between the OCR-module and the present newspaper production line.

The method for evaluating the OCR-module in Case 1 and 2 will follow the same procedure. In both cases we will apply quantitative and qualitative analyses. In the quantitative analysis we will measure the performance of the module by comparing the output results with the ground truth. We will use the OCR Frontiers toolkit developed by [2] to compute accuracies of the percentage of correct characters and words based on their frequencies in the OCR processed textfiles. This evaluation method will help us improve our module incrementally. Qualitative analysis will be carried out using human evaluators who will perform manual inspections of the textfiles. The purpose of the qualitative analysis is twofold. In Case 1 it is applied to form a typology from the material and learn more about how to enhance the resources for improving the OCR process. In Case 2 it is applied to evaluate the usefulness of the final module that will be evaluated against some “new” material that is not included in the ground truth and thus is lacking transcription. Figure 2 demonstrates the stages in the evaluation process.

[9] http://impacteurope.eu/

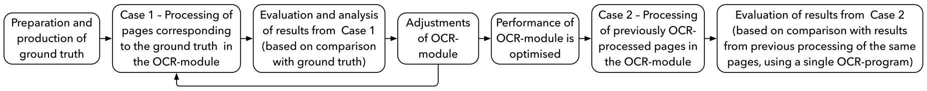
Preparation and production of ground truth
Case 1 – Processing of pages corresponding to the ground truth in the OCR-module
Evaluation and analysis of results from Case 1 (based on comparison with ground truth)
Adjustments of OCR-module
Performance of OCR-module is optimised
Case 2 – Processing of previously OCR-processed pages in the OCR-module
Evaluation of results from Case 2 (based on comparison with results from previous processing of the same pages, using a single OCR-program)

Fig. 2. The stages in the evaluation process.

In our material, the amount of digitised newspapers available for the time period before 2010 is less than half, compared to the amount available after 2010. There is consequently a risk that the reference material and the training material in case 2 will not be representative. This will require careful manual inspection of the selected material in both cases. We will try to balance the material according to two time periods, including late modern Swedish, i.e. 1818–2000, and the material for modern data, i.e. 2000–2018. This will be done by first applying statistical based selection of the material and then manually select additional material for increasing the scope of the late modern Swedish data.

5 Related work

The question of quality in relation to the output of OCR-processes is complex. [19] observes that most studies addressing the topic of OCR quality tend to be carried out with the perspective of the producer, rather than the user. For example [18] discusses OCR quality in relation to newspaper digitisation at the British Library. Here the central focus of OCR conversion is to enable indexing and information retrieval as opposed to "full text representation". Words with low significance are in this approach treated as noise and subsequently not included in the evaluation of the capacity of the OCR-process. The article concludes that an accuracy of 80% on the word level is sufficient for these purposes.

[19] takes a different approach by focussing on the intention of the user. Four specific types of usage of digitised text material were identified based on the types of tasks of a group of researchers, e.g. finding the first mention of a specific concept, identifying articles of relevance to a specific interest, analysis of word frequencies, and qualitative content analysis. Based on such a typology [19] sets to task to define acceptable error levels that could be used to indicate the suitability of a given text for a specific purpose. This approach to the problem of quality was in many ways hampered by the absence of metadata relating to the OCR-process, making it impossible to relate components in the process to the specific outcomes – the individual words. Their findings however indicate a promising way forward and has provided the base for the initiative taken by KB to address the question of quality declaration of OCR-produced text resources.

Computational linguistic studies have shown that the quality of the material varies depending on the software that performs the evaluations. [13] present the evaluation results of four different off-the-shelf tools on the same material. They show there is a variation of around 3-10% between the results of the different tools. Therefore an exact measure of OCR accuracy is not a sufficient indicator for determining what is an acceptable accuracy rate. In particular for historic newspapers where accuracy rates usually are lower because of the condition of the original material. [12] emphasizes that the most reliable way to determine the quality of the OCR-process is through proofreading the output documents. Identifying character and word error accuracy at the document level is a good indicator that could give a reliable confidence level about the accuracy level of the OCR-process, but since this is not feasible in large scale historical newspaper digitalisations, the method should be combined with automatic methods.

The computational resources that are available for a particular time period also are relevant for the quality assessment. Because of the lack of comprehensive morphological description and lexica describing the orthography and spelling variation of a particular language, in our case Swedish, false negatives, viz-á-viz falsely identified errors, are common phenomenon for historic material. To cater for this deficiency computational linguistic methods should be combined [16].

6 Summary

In this paper we present an ongoing project initiated by KB and SB. In the project we intend to carry out a formal evaluation of, and improve the OCR-module through systematic text analyses, dictionaries and word lists with the aim of implementing it in a mass digitisation process.

Acknowledgements

This work has been partially supported by Swe-Clarin, a Swedish consortium in Common Language Resources and Technology Infrastructure (CLARIN) Swedish CLARIN (grant agreement 821-2013-2003).

References

1. Ahlberg, M., Bouma, G. A best-first anagram hashing filter for approximate string matching with generalized edit distance. In: Proc. 24th international conference on computational linguistics COLING, pp. 13–22 (2012)
2. Bagdanov, A.D., Rice, Stephen V., Nartker, T.A.. The OCR Frontiers Toolkit. Version 1.0. Information Science Research Institute (1999)
3. Borin, L., Bouma, G., Dannélls, D. A free cloud service for OCR / En fri molntjänst för OCR. Project report (GU-ISS 2016-01), Department of Swedish, University of Gothenburg (2016)
4. Borin, L., Forsberg, M. A Diachronic Computational Lexical Resource for 800 Years of Swedish. Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series, Theory and Applications of Natural Language Processing, C. Sporleder et al. (eds.). Springer:Berlin (2011)
5. Borin, L., Forsberg, M., Olsson, LJ., Olsson, O., Uppström, J. The lexical editing system of Karp. In: Proceedings of the eLex conference, Tallin (2013)
6. Borin, L., Forsberg, M., Roxendal J. Korp the corpus infrastructure of Språkbanken. In: Proceedings of LREC 2012. Istanbul: ELRA (2012)

7. Borin, L., Kokkinakis, D., Olsson, L.-J: Naming the Past: Named Entity and Animacy Recognition in 19th Century Swedish Literature. In: Proceedings of the Workshop on Language Technology for Cultural Heritage Data (2007)
8. Cecotti, H., Belaïd, A.: Hybrid OCR combination approach complemented by a specialized ICR applied on ancient documents. In: Document Analysis and Recognition. Eighth International Conference on, pp. 1045–1049, IEEE (2005)
9. Clematide, S., Furrer, L., Volk, M.: Crowdsourcing an OCR Gold Standard for a German and French Heritage Corpus. Proceedings of the Tenth International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA) (2016)
10. Daðason, J.F., Bjarnadóttir, K., Rúnarsson, K.: The Journal Fjölnir for Everyone: The post-processing of historical OCR texts. In: Language Resources and Technologies for Processing and Linking Historical Documents and Archives-Deploying Linked Open Data in Cultural Heritage–LRT4HDA Workshop. ELRA (2014)
11. Drobac, S., Kauppinen, P., Lindén, K. OCR and post-correction of historical Finnish texts. In: Proceedings of the 21st Nordic Conference of Computational Linguistics, pp. 70–76 (2017)
12. Holley, R.. How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs. In: D-Lib Magazine, 15(3/4) (2009)
13. Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T., Kervinen, J. 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 Proceedings: 80th IFLA General Conference and Assembly (2014)
14. Levenshtein, Vladimir I.. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, 10:707 (1966)
15. Lund, W. B.: Ensemble Methods for Historical Machine-Printed Document Recognition (Doctoral Dissertation). Brigham Young University, Provo, UT (2014)
16. Mihov, S., Koeva, S., Ringlstetter, C., Schulz, K.U., Strohmaier, C.M. Precise and Efficient Text Correction using Levenshtein Automata, Dynamic Web Dictionaries and Optimized Correction Models. In: Proceedings of Workshop on International Proofing Tools and Language Technologies (2004)
17. Silfverberg, M., Rueter, J.: Can morphological analyzers improve the quality of optical character recognition? In Septentrio Conference Series, number 2, pp. 45–56 (2015)
18. Tanner, S., Munoz, 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. In: D-Lib Magazine, 15(7/8) (2009)
19. Traub, M., Van Ossenbruggen, J., Hardman, L.: Impact Analysis of OCR Quality on Research Tasks in Digital Archives. In: Proceedings of Research and Advanced Technology for Digital Libraries – 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, pp. 1–12, Poznan (2015)
20. Volk, M., Furrer, L., Sennrich, R.: Strategies for Reducing and Correcting OCR Errors. I C. Sporleder, A. van den Bosch, K. Zervanou (Red.), Language Technology for Cultural Heritage, pp. 3–22, Berlin: Springer (2011)