A Two-OCR Engine Method for Digitized Swedish Newspapers

Dana Dannélls
Språkbanken Text, Dept. of Swedish
University of Gothenburg, Sweden
dana.dannells@gu.se

Lars Björk
Kungliga biblioteket
Stockholm, Sweden
lars.bjork@kb.se

Ove Dirdal
Zissor
Oslo, Norge
ove@zissor.com

Torsten Johansson
Kungliga biblioteket
Stockholm, Sweden
torsten.johansson@kb.se

Abstract

In this paper we present a two-OCR engine method that was developed at Kungliga biblioteket (KB), the National Library of Sweden, for improving the correctness of the OCR for mass digitization of Swedish newspapers. To evaluate the method a reference material spanning the years 1818–2018 was prepared and manually transcribed. A quantitative evaluation was then performed against the material. In this first evaluation we experimented with word lists for different time periods. The results show that even though there was no significant overall improvement of the OCR results, some combinations of word lists are successful for certain periods and should therefore be explored further.

1 Introduction

The process of converting images into digitized editable text is called Optical Character Recognition (OCR). OCR techniques have been applied since the late 90s and their performances have been improved significantly during the last decade with the advances of neural networks (Amrhein and Clematide, 2018; Nguyen et al., 2020). However, OCR processing of historical material, especially newspapers, remains a challenge because of low paper and print quality, variation in typography and orthography, mixture of languages and language conventions (Gregory et al., 2016; Chiron et al., 2017).

Kungliga biblioteket (KB), the National Library of Sweden, is the central source for digitized Swedish newspapers, offering access to more than 25 million pages via the web service "Svenska dagstidningar".[1] The accuracy of the OCR system is therefore an important factor in order to maximize the access and usability of the digitized collections. To address this, KB, in collaboration with the Norwegian software company Zissor,[2] has implemented a novel OCR technique for combining two OCR engines: Abbyy and Tesseract. The two-OCR engine method has so far only been used as an internal testbed, and was only evaluated manually, awaiting automatic evaluation and suggestions of possible improvements.

In 2019, KB embarked on an infrastructure project together with Språkbanken Text,[3] the Swedish Language bank at the University of Gothenburg, which is the coordinating Swedish CLARIN (Swe-CLARIN) node and CLARIN B Center, with the aim of evaluating and improving the results of the method (Dannélls et al., 2019).

In this paper we describe the two-OCR engine method, how we selected and prepared the ground truth material, spanning the years 1818-2018, and report the first quantitative evaluation. We further describe

[1] https://tidningar.kb.se/
[2] https://zissor.com/
[3] https://spraakbanken.gu.se/en

Dana Dannélls, Lars Björk, Torsten Johansson and Ove Dirdal 2021. A Two-OCR Engine Method for Digitized Swedish Newspapers. *Selected papers from the CLARIN Annual Conference 2020.* Linköping Electronic Conference Proceedings 180: 180 65–74.

our attempt to improve the OCR performance for this period by increasing the OCR engines build-in vocabulary. By experimenting with different word lists for different time periods we take diachronic aspects into consideration and thereby, hope to adhere to linguistic change in the course of time (Springmann and Lüdeling, 2017). To our knowledge, this is the first use case study of evaluating and improving the OCR accuracy of Swedish newspaper texts over a period of 200 years.

2 Related Work

Newspapers are challenging material because of their mixture of typeface, size and complex layout. Gregory et al. (2016) present some of the biggest challenges in working with digitization of the British Library's nineteenth century newspaper collection. They emphasise the importance of knowing the source material and looking into the original data. Their work provides indication of developing methods and solutions that are tailored to the original text segments. In this project we follow their recommendations, but instead of text segments we are focusing on smaller units, namely on paragraph levels.

Earlier work on historical English demonstrated the challenges with combining multiple OCR engines (Lund et al., 2011). They reported an improvement over the word accuracy rate using voting and dictionary features. Reul et al. (2018) applied a confidence voting scheme between OCR models that were trained on a single engine. Their evaluation on Latin books showed a relative improvement over the character accuracy rate.

OCR errors are often classified into two groups: non-word errors and real-word errors (Mei et al., 2016; Nguyen et al., 2019). Real-word errors are more challenging because they require human inspections. Correcting real-word errors by increasing the engine's vocabularies has been studied by several authors who have proven the usefulness of lexicons, among other successful strategies for improving the OCR accuracy (Kissos and Dershowitz, 2016; Schulz and Kuhn, 2017; Nguyen et al., 2018). Authors have shown that a lexicon-based approach is competitive if it is adapted to certain domains or time periods. Our assumption here is that curated word lists are suitable to experiment with for Swedish material that spans over hundred years since a great number of changes in orthography and morphology occurred during this time, in particular around 1906. Real-word errors can also be caused by insertion, replacement or deletion of characters, for example *föreda* 'provide' became *breda* 'sprea' because "fö" was merged to "b". Several successful approaches have been proposed to detect these types of errors, ranging from Levenshtein distance (Samanta and Chaudhuri, 2013) to Finite-State (Silfverberg et al., 2016) and n-gram (Eger et al., 2016) methods. We have so far not explored any of these, but have plans to incorporate Levenshtein distance in the next phase of our work.

The work presented by Koistinen et al. (2017) aims to calculate the accuracy of the OCR system by comparing the Abbyy average confidence score that is assigned to each processed documents automatically. In this work we do not take this score into consideration because it was proven unreliable in an internal study which we conducted after the method was implemented.

Clematide and Ströbel (2018) discuss how to improve the quality of newspapers texts. An important outlook from their work is to understand how the performance of the OCR system varies in relation to studying how often mixture between Antiqua and Blackletter occurs. One conclusion was that high number of Blackletter articles often results in low OCR accuracy, therefore it is important to check the distribution of Antiqua and Blackletter. Something will address when we evaluate the complete material.

The results we report here are almost as accurate as the results reported for Finnish and Swedish newspaper texts (Drobac et al., 2019). However, they are not directly comparable because Drobac et al. experiments were done on a smaller selected set from 1771 until 1874, which we at the time of writing do not have any access to.

3 Two-OCR Engine Method

The two-OCR engine method was developed in 2017 in cooperation between KB and Zissor. The method was designed to enable adjustment and control of some key parameters of the post-capture stage of the OCR process, including 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 working principle of the method is based on the evaluation and comparison between the results from two separate OCR engines: the proprietary system Abbyy FineReader, version 11.1.16,[4] and the open source system Tesseract, version 4,[5] developed by Google and is based on Long Short Term Memory (LSTM) Recurrent Neural Networks (Smith, 2007). The pre-trained models provided by each system were used to OCR the Swedish newspaper material. Abbyy and Tesseract have been integrated in such a way that allows to choose between one of them or run both in parallel. Figure 1 provides an overview of the two-OCR engine architecture.

When Abbyy and Tesseract run in parallel the results from the two engines are analysed and the best results on the word level are prioritised to create a new combined ALTO XML file. This process is accomplished by comparing the two ALTO XML files, word by word. The comparison and process of selecting between the results of the OCR systems relies on a scoring model that is based on the internal dictionaries of each system (see Section 3.1).

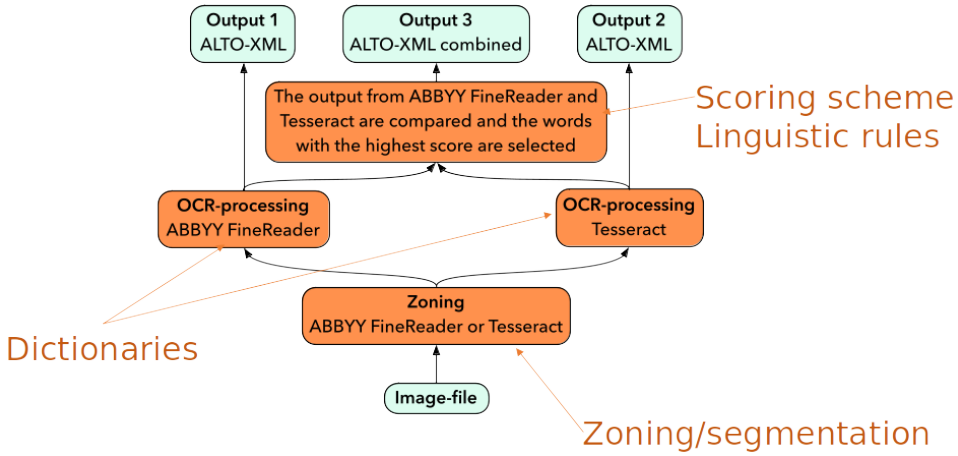

Figure 1: The design of the two-OCR engine method

3.1 Scoring Scheme

Abbyy and Tesseract apply different OCR techniques for analysis, recognition and segmentation of the scanned image. Their internal dictionaries also vary with respect to size and how confidence scores are calculated for each OCRed word. In a series of experiments Zissor analysed the OCR results generated by Abbyy and Tesseract and examined the differences between them in more details (Zissor, 2017). The aim of the experiments was to determine how to combine the OCR results from both systems to yield the most accurate results. The results of the experiments showed Tesseract has more overlapping zones compared to Abbyy – a feature which makes article segmentation difficult. Tesseract is also more sensitive to image noise, something that affects both zoning and the OCR quality. On the other hand, Tesseract seems to handle larger text fonts better, and in general greater variance in font compared to Abbyy. Each of the systems generates a confidence score for the OCRed words, but because this score is calculated differently for Abbyy and Tesseract, a comparison based on the OCR confidence scores is not straightforward. Character and word coordinates are the same for Abbyy and Tesseract which implies the same word can be taken from both systems for in-depth character and word comparisons.

Following the results of the experiments a rule-driven scoring model was implemented to determine automatically how to prioritise between the systems when disagreement on word level occurs. 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. Figure 2 demonstrates the scoring scheme in its present version. The results from the automatic comparison is a combined ALTO XML file containing words from each system.

[4] `http://finereader.abbyy.com/`, accessed via Server Web Services API
[5] `https://github.com/tesseract-ocr/`

Step	Rule	Sequence/consequence and choice of word
1.	Both words are equal = ABBYY	If ABBYY´s suggestion and Tesseract's are equal: ABBYY´s word is selected. **[IF NOT: STEP 2]**
2.	ABBYY's suggestion is blank = no word is given	If ABBYY's suggestion is blank no word is given. **[IF NOT: STEP 3]**
3.	Both words are equal after the removal of "noise characters" = ABBYY	If ABBYY's suggestion and Tesseract's suggestion are equal after the removal of one "noise character" in the beginning and end of the given word: ABBYY's word is selected. **[IF NOT: STEP 4]**
4.	The word is found in ABBYY's dictionary = ABBYY	If ABBYY's suggestion is found in ABBYY's dictionary or in a customised dictionary: ABBYY's word is selected. **[IF NOT: STEP 5]**
5.	ABBYY's suggestion is a numeral = ABBYY	If ABBYY's suggestion is a numeral that numeral is selected. **[IF NOT: STEP 6]**
6.	Tesseract's suggestion is blank = ABBYY	If Tesseract's suggestion is blank: ABBYY's word is selected. **[IF NOT: STEP 7]**
7.	If Tesseract's suggestion is a single character = ABBYY	If Tesseract's suggestion is a single character: ABBYY's word is selected. **[IF NOT: STEP 8]**
8.	If Tesseract's suggestion is found in Tesseract's dictionary or in a customised dictionary = Tesseract	If Tesseract's suggestion is found in Tesseract's dictionary or in a customised dictionary: Tesseract's word is selected. **[IF NOT: STEP 9]**
9.	If Tesseract's suggestion is a numeral = Tesseract	If Tesseract's suggestion is a numeral: Tesseract's numeral is selected. **[IF NOT: STEP 10]**
10.	ABBYY word = ABBYY	If none of the previous steps has rendered a word: ABBYY's word is selected.

Figure 2: The scoring schema underlying the voting principles between Abbyy and Tesseract

3.2 Segmentation Tool

In addition to the ALTO XML files produced by the OCR process, statistics about the word errors and applied corrections are generated as Excel-files for each system. In this way, the process can be manually monitored and adjusted using Zissor's article segmentation tool as illustrated in Figure 3.

3.3 External Swedish Word Lists

As mentioned above, Abbyy and Tesseract have their own internal dictionaries incorporated in the systems for improving the OCR word accuracy. The size of Abbyy's internal dictionary is unknown and the dictionary in Tesseract is rather small, containing around one million entries that have been complied from unknown time periods. One possible way to improve the accuracy of the systems is therefore by increasing the internal vocabulary with external word lists for specific periods in the post-processing step. For this purpose we compiled four word lists from dictionaries and lexical resources from different time periods:

- Dalin, a full form lexicon for the 19th century, covering the morphology of late modern Swedish (Borin and Forsberg, 2011), containing 509,924 entries;
- Saldo, a full form modern lexicon (Borin et al., 2013), containing 1,704,718 entries;[6]
- Saol-hist,[7] a subset of the Swedish Academy historical lexicon, containing only base forms, amounting to 128,720 entries;
- Fem, a word list over name entities that was compiled from five lexical sources at KB, containing in total 311,481 entries.

[6]Both Dalin and Saldo are part of CLARIN lexical resources.

[7]`http://spraakdata.gu.se/saolhist/`

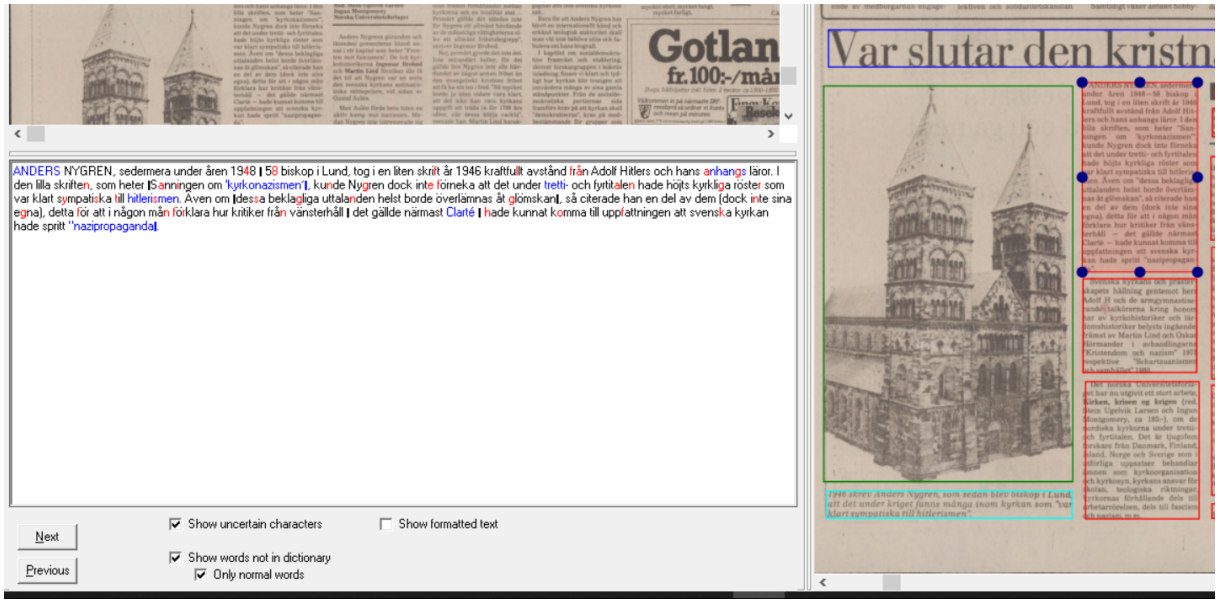

Figure 3: Zissor's article segmentation and OCR tool with Abbyy. Selected Zone; Abbyy OCR text. Blue words are words that are not included in the Abbyy standard dictionary or in additional imported Swedish dictionaries used in the OCR process. Red letters are letters that Abbyy is unsure of during the OCR process.

The first three word lists (Dalin, Saldo and Saol-hist) were extracted from the original sources. Fem was extracted from a selected set of corpora. While Saol-hist and Fem only contain plain vocabulary lists, Dalin and Saldo contain morphological gazetteers.

To get an estimate of the coverage of each word list, we counted the number of tokens in our ground truth material (see Section 4) and compared each token with the words in each word list. For each word list we also kept a count of the tokens without any match. We summarise the results periodically, grouped by frequency in Figure 4. As Figure 4 shows the total amount of tokens in the whole material is 1,112,996 of which 521,816 tokens appear in newspapers between 1818-1906, 512,182 appear in newspapers between 1907-1996 and 78,998 tokens in 1997-2018. We can observe that for each period about 50% of the words are covered in Saldo and Dalin, and about 20% are covered in Saol-hist and Fem. When we inspected the list of words that were not found in any of the word lists (category "None" in Figure 4) we found that the majority of these words are numbers, place names, and proper names.

4 Reference Material and Processing

Our reference material consists of 400 pages, selected from 200 newspapers spanning the years from 1818 until 2018. Pages were carefully chosen to reflect typical variations in layout and typography. Two pages were selected from each newspaper; the second and the fourth. The underlying assumption for this decision was that there are generally less advertisements and pictures on the second and fourth pages.[8]

Each page in the reference material was segmented down to paragraph level, and each paragraph was marked with an ID number. This segmentation scheme is kept as a matrix that can be reused for the comparison between the reference material and the corresponding section in the OCR processed material. The resulting ground truth material amounts to 43823 IDs.[9]

[8] We acknowledge that advertisements and pictures are important mediums to study our cultural heritage, but for the purpose of this study where text is the primary focus, we chose to minimize these.

[9] A selection of the material is freely available under open source license from Språkbanken Text `https://spraakbanken.gu.se/en/resources#refdata`.

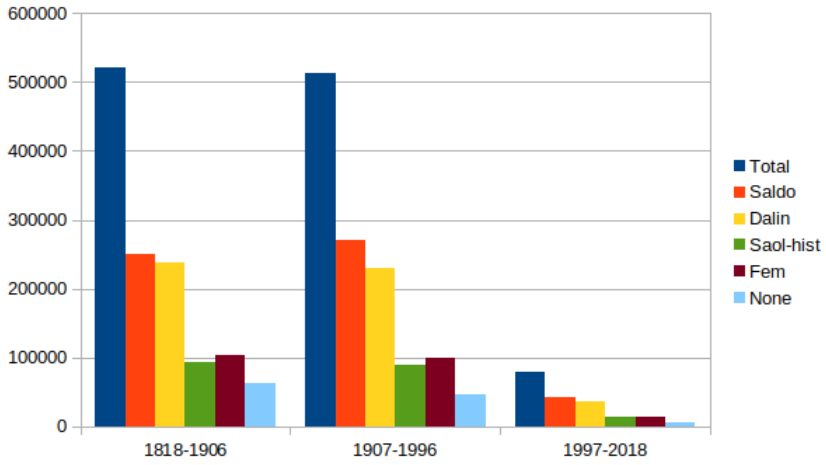

Figure 4: The frequency of tokens for three time periods, their total amount in the ground truth material, their coverage across four word lists (Saldo, Dalin, Saol-hist and Fem) and the number of tokens that does not appear in any of the word lists (marked with the "None" category).

The selected reference material was sent to Grepect, a transcription company who specialises in manual transcription of older material.[10] Based on our inspections of the material we defined the guidelines for the transcription which contained instructions for typeface, size and location changes.[11]

To produce our baseline we run the material through the two-OCR engine system using Abbyy's and Tesseract's internal dictionaries exclusively. Next, we run the material in the two-OCR engine system four times, each run with a different word list (see Section 3.3). For each run the system delivers three results: one result for Abbyy, one for Tesseract and one calculated/verified Abbyy-Tesseract.

5 Experiments and Evaluation

For evaluation we used the OCR frontier toolkit (Carrasco, 2014). The method calculates the results of the OCR errors by measuring character accuracy rate (CAR) and word accuracy rate (WAR).

First we run the evaluation against the prepared ground truth material of the 400 newspaper pages, once without external word lists (our baseline) and once with external word lists. Table 1 shows the evaluation results of these runs. At first glance we note that the best performing system on the character level is achieved with the combined method (91.64%), but the improvement is minor compared to Abbyy and Tessearct respectively. On the word level we observe a similar tendency. Surprisingly, the accuracy of the Abbyy improves with the external word lists, as opposed to the other systems whose results actually decrease when run with the external word lists.

	Without word lists		With four word lists	
OCR engine	CAR (%)	WAR (%)	CAR (%)	WAR (%)
Abbyy	91.54	83.39	91.45	83.6
Tesseract	91.39	87.37	91.39	85.53
Abbyy-Tesseract	91.64	84.21	91.5	83.9

Table 1: Evaluation results of 400 pages for the whole time period 1818-2018, run with three types of OCR engine setups. Without word lists (our baseline) on the left and with all word lists on the right.

[10] `http://www.grepect.de/`

[11] The resources that are developed in this project, covering the years up to 1909 will be made freely available for download through `https://vlo.clarin.eu`

Second, we calculated the CAR and WAR results for each run on the same material, this time separately with each word list and divided into three time periods. Table 2 shows the evaluation results of our runs without external word lists (baseline) and with the external word lists both separately and all combined.[12]

	1818-1906		1907-1996		1997-2018	
OCR engine	CAR (%)	WAR (%)	CAR (%)	WAR (%)	CAR (%)	WAR (%)
Abbyy baseline	87.38	71.47	94.22	92.11	93.34	92.58
Tesseract baseline	88.14	81.26	92.48	89.93	92.84	91.17
Abbyy-Tesseract baseline	87.46	72.66	93.36	90.05	93.43	90.11
Abbyy Dalin	87.31	**71.87**	94.05	91.98	93.31	**92.69**
Tesseract Dalin	87.86	76.74	**93.32**	**91.44**	92.55	91.07
Abbyy-Tesseract Dalin	87.32	72.49	93.13	89.7	93.34	89.92
Abbyy Saldo	87.19	71.33	93.02	89.15	**93.38**	89.72
Tesseract Saldo	87.89	76.73	92.46	89.39	**92.95**	90.21
Abbyy-Tesseract Saldo	87.18	71.96	93.05	89.42	93.27	89.78
Abbyy Saol-hist	87.37	**71.96**	93.13	89.44	**93.47**	89.83
Tesseract Saol-hist	87.84	76.75	92.46	89.38	**92.95**	90.19
Abbyy-Tesseract Saol-hist	87.41	**72.73**	93.19	89.82	93.42	90.03
Abbyy Fem	87.1	70.81	92.81	88.62	93.15	89.2
Tesseract Fem	87.88	76.72	92.45	89.37	**92.96**	90.2
Abbyy-Tesseract Fem	87.19	71.61	92.89	89.03	93.09	89.4
Abbyy All	87.36	**71.88**	93.09	89.52	**93.4**	89.76
Tesseract All	88.03	77.09	**92.57**	89.41	**92.97**	90.16
Abbyy-Tesseract All	87.41	72.35	93.13	89.75	93.34	89.82

Table 2: Evaluation results of 400 pages divided into three time periods. First run without external word lists (baseline), four runs with external word lists, one run for each word list and last run with all four word lists. Each run was performed with three engines. The results marked in bold highlight successful runs that outperform the baseline.

As can be observed in Table 2, Abbyy shows a small improvement on the word level with Dalin, Saol-hist and all word lists for 1818-1906. Interestingly, it also shows a small improvement on the word level with Dalin for 1997-2018. There is some improvement for Tesseract both on the character and word levels, with Dalin for the later period 1907-1996, and minor improvement with all word lists for the same period. Neither of the runs for 1907-1996 has improved over the baseline for Abbyy, Tesseract or Abbyy-Tesseract. This could be explained by the fact that the systems are eager to find a lexical match in the external word lists, and since the external lists get higher priority, wrong words are being replaced. Thus, there is a higher percentage of words that are replaced with incorrect ones for that particular content. Consequentially, CAR is also decreasing. Another explanation of the low performance is the high ratio of out-of-word vocabulary for this period as seen in Figure 4.

Further, Saldo, Saol-hist and all word lists improve the character accuracy of Abbyy and Tesseract for the period 1997-2018. Tesseract shows a minor improvement with Fem on the character level.

6 Summary and Conclusions

We explored the effect of adding curated word lists to improve the two-OCR engine system when digitizing Swedish newspapers from 200 years. We found that the addition of word lists in combination with the two-OCR engine system did not provide the expected significant improvement of the OCR result, even though some combinations proved more successful than others.

[12]The differences in time for the different runs with the two-OCR engine were marginal. It took approximately 157 seconds to process one page. To process the whole material, we run 3 OCR processes simultaneously, which took 6 hours. With additional OCR processes, this time could be reduced accordingly.

One explanation of the OCR results reported in this study might stem from the fact that our method for matching whether a word appears in the word list is rather naive. It simply applies string matching. To make the best out of the word lists, matching should be combined with a Levenshtein-distance method. Another problem could be grounded in the correlation between specific types of OCR errors and images as well as graphical elements in the printed page, resulting in occurrences of segments with no referable word because the system mistook an image for being a word. When we examined the effect of word lists using the segmentation tool we found that external word lists could improve the word accuracy if we consider the confidence score as a factor in deciding whether a word is correct or not. Consequently, words that receive a higher confidence level when found in the word lists, might lead to better word prioritisation and in turn improve the quality of the text as a whole. Our manual inspection showed the system had to "guess" when deciding on the correct word, hence resulting in a more or less random output. Furthermore, a closer look on the source material showed 30% of the material has low print quality. When we studied the source material to get an estimate of the distribution of Antiqua and Blackletter by going through the images manually we learned that for 1818-1906 25% is Antiqua and 75% is Blackletter, for 1907-2018 75% is Antiqua and 25% is Blackletter. Therefore more thorough qualitative analysis combined with quantitative analysis of the relation between OCR errors and the source material could provide the basis for a taxonomy of error types which will further contribute to the development of benchmarks and quality indicators. Segmentation precision also plays a significant role for the amount of character and word errors. In the complete material we identified around 100 segments with considerable difference in sequential tokens between the ground truth and the OCRed data caused by inaccurate segmentation. To address this problem, preprocessing techniques are being tested by Zissor to improve the accuracy of the image processing.

There are however some unexpected variations in the accuracy results that have to be examined in detail in relation to the specific word lists and the given time period of the sample. The quality control of the two engine system verifies that the word lists are taken into consideration during the OCR process but their apparent unpredictable effect on the results have to be further analysed. The external word lists had noticeable effect on the internal confidence values of the OCR programs. The effect of these combined with the possibility to include the rate of correspondence between the results from the two-OCR engine (on the word level), could be used as a variable to indicate the quality of the results, something we will explore later in the project. Another important variable seems to be the scoring scheme which is governed by a set of rules for deciding on which of the systems processed the correct word. The impact of the rule set will be analysed once the consequences of the use of word lists is further examined.

The preliminary results reported here are based on the first quantitative evaluation against the complete ground truth. By studying the results manually we could observe some correlation between specific types of OCR errors and images as well as graphical elements in the printed page. Future work aims to combine these findings with a substantial qualitative analysis to address possible sources of errors resulting from degradation of paper, bad print quality and complexity of layout and typography. The final results will be reported in future publications but these preliminary findings indicate some areas that will receive a more detailed focus as the project progresses.

Acknowledgements

The research presented here is funded by the Swedish Research Council (the project *Evaluation and refinement of an enhanced OCR-process for massdigitisation* grant: IN18-0940:1). It is also supported by Språkbanken Text and Swe-Clarin, a Swedish consortium in Common Language Resources and Technology Infrastructure (CLARIN) Swedish CLARIN (grant 821-2013-2003).

References

Chantal Amrhein and Simon Clematide. 2018. Supervised OCR Error Detection and Correction Using Statistical and Neural Machine Translation Methods. *Journal for Language Technology and Computational Linguistics (JLCL)*, 33(1):49–76.

Lars Borin and Markus Forsberg. 2011. A diachronic computational lexical resource for 800 years of Swedish. In *Language technology for cultural heritage*, pages 41–61. Springer, Berlin, Germany.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Rafael C. Carrasco. 2014. An open-source OCR evaluation tool. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH, pages 179–184, NY, USA. Association for Computing Machinery.

Guillaume Chiron, Antoine Doucet, Mickael Coustaty, Muriel Visani, and Jean-Philippe Moreux. 2017. Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information. In *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–4. IEEE.

Simon Clematide and Phillip Ströbel. 2018. Improving OCR quality of historical newspapers with handwritten text recognition models. In *Workshop DARIAH-CH*, Neuchâtel. University of Neuchâtel.

Dana Dannélls, Lars Björk, and Torsten Johansson. 2019. Evaluation and refinement of an enhanced ocr process for mass digitisation. In *Proceedings of Digital Humanities in the Nordic Countries*, pages 112–123, Copenhagen, Denmark. CEUR-WS.org.

Senka Drobac, Pekka Kauppinen, and Krister Linden. 2019. Improving OCR of historical newspapers and journals published in Finland. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, pages 97–102, Brussels, Belgium. ACM.

Steffen Eger, Tim vor der Brück, and A. Mehler. 2016. A comparison of four character-level string-to-string translation models for (OCR) spelling error correction. *The Prague Bulletin of Mathematical Linguistics*, 105:77 – 99.

Ian Gregory, Paul Atkinson, Andrew Hardie, Amelia Joulain-Jay, Daniel Kershaw, Catherine Porter, Paul Rayson, and C.J. Rupp. 2016. From Digital Resources to Historical Scholarship with the British Library 19th Century Newspaper Collection. *Journal of Siberian Federal University, Humanities and Social Sciences*, 9(4):994–1006.

Ido Kissos and Nachum Dershowitz. 2016. OCR error correction using character correction and feature-based word classification. In *12th IAPR Workshop on Document Analysis Systems DAS*, pages 198–203, Santorini, Greece. IEEE.

Mika Koistinen, Kimmo Kettunen, and Tuula Pääkakkönen. 2017. Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur and Antiqua Models and Image Preprocessing. In *Proceedings of the 21st Nordic Conference on Computational Linguistics, NODALIDA*, Gothenburg, Sweden. Association for Computational Linguistics.

William B. Lund, Daniel D. Walker, and Eric K. Ringger. 2011. Progressive Alignment and Discriminative Error Correction for Multiple OCR Engines. In *International Conference on Document Analysis and Recognition*, pages 764–768, Beijing, China. IEEE.

Jie Mei, Aminul Islam, Yajing Wu, Abidalrahman Mohd, and Evangelos E Milios. 2016. Statistical learning for OCR text correction. *arXiv preprint*, abs/1611.06950.

Thi-Tuyet-Hai Nguyen, Mickaël Coustaty, Doucet Antoine, and Nhu-Van Nguyen. 2018. Adaptive Edit-Distance and Regression Approach for Post-OCR Text Correction. In *20th International Conference on Asia-Pacific Digital Libraries, ICADL*, Hamilton, New Zealand. Lecture Notes in Computer Science.

Thi-Tuyet-Hai Nguyen, Adam Jatowt, Mickael Coustaty, Nhu-Van Nguyen, and Antoine Doucet. 2019. Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing. In *Proceedings of the 18th Joint Conference on Digital Libraries*, JCDL, pages 29–38, Champaign, IL, USA. IEEE.

Thi-Tuyet-Hai Nguyen, Adam Jatowt, Nhu-Van Nguyen, Mickaël Coustaty, and Antoine Doucet. 2020. Neural Machine Translation with BERT for Post-OCR Error Detection and Correction. In *JCDL: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, Virtual Event*, pages 333–336, New York, NY. Association for Computing Machinery.

Christian Reul, Uwe Springmann, Christoph Wick, and Frank Puppe. 2018. Improving OCR Accuracy on Early Printed Books by combining Pretraining, Voting, and Active Learning. *J. Lang. Technol. Comput. Linguistics*, 33(1):3–24.

Pratip Samanta and Bidyut B. Chaudhuri. 2013. A simple real-word error detection and correction using local word bigram and trigram. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING)*, pages 211–220, Kaohsiung, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Sarah Schulz and Jonas Kuhn. 2017. Multi-modular domain-tailored ocr post-correction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2716–2726, Copenhagen, Denmark. Association for Computational Linguistics.

Miikka Silfverberg, Pekka Kauppinen, and Krister Lindén. 2016. Data-driven spelling correction using weighted Finite-State methods. In *Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata*, pages 51–59, Berlin, Germany. Association for Computational Linguistics.

Ray Smith. 2007. An Overview of the Tesseract OCR Engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 629–633, Curitiba, Brazil. IEEE.

Uwe Springmann and Anke Lüdeling. 2017. OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus. *Digital Humanities Quarterly*, 11(2).

Zissor. 2017. Zissor Content System. Implementation of dual OCR motors, Phase II. Technical report, Kungliga biblioteket (KB), Stockholm, Sweden.