HAL
open science

Detecting Articles in a Digitized Finnish Historical Newspaper Collection 1771-1929: Early Results Using the PIVAJ Software

Kimmo Kettunen, Teemu Ruokolainen, Erno Liukkonen, Pierrick Tranouez, Daniel Antelme, Thierry Paquet

To cite this version:

Kimmo Kettunen, Teemu Ruokolainen, Erno Liukkonen, Pierrick Tranouez, Daniel Antelme, et al.. Detecting Articles in a Digitized Finnish Historical Newspaper Collection 1771-1929: Early Results Using the PIVAJ Software. DATech 2019, May 2019, Bruxelles, Belgium. 10.1145/3322905.3322911 . hal-02111142

HAL Id: hal-02111142

https://hal.science/hal-02111142v1

Submitted on 25 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Detecting Articles in a Digitized Finnish Historical Newspaper Collection 1771–1929: Early Results Using the PIVAJ Software

Kimmo Kettunen
The National Library of Finland
DH Research
University of Helsinki
Mikkeli, Finland
firstname.lastname@helsinki.fi

Teemu Ruokolainen
The National Library of Finland
DH Research
University of Helsinki
Mikkeli, Finland
firstname.lastname@helsinki.fi

Erno Liukkonen
The National Library of Finland
DH Research
University of Helsinki
Mikkeli, Finland
firstname.lastname@helsinki.fi

Pierrick Tranouez
LITIS laboratory
University of Rouen Normandy
France
Pierrick.Tranouez@univ-rouen.fr

Daniel Antelme
LITIS laboratory
University of Rouen Normandy
France
Daniel.Antelme@univ-rouen.fr

Thierry Paquet
LITIS laboratory
University of Rouen Normandy
France
Thierry.Paquet@univ-rouen.fr

ABSTRACT

This paper describes first large scale article detection and extraction efforts on the Finnish Digi[1] newspaper material of the National Library of Finland (NLF) using data of one newspaper, *Uusi Suometar* 1869–1898. The historical digital newspaper archive environment of the NLF is based on commercial docWorks[2] software. The software is capable of article detection and extraction, but our material does not seem to behave well in the system in this respect. Therefore, we have been in search of an alternative article segmentation system and have now focused our efforts on the PIVAJ machine learning based platform developed at the LITIS laboratory of University of Rouen Normandy [11–13, 16, 17].

As training and evaluation data for PIVAJ we chose one newspaper, Uusi Suometar. We established a data set that contains 56 issues of the newspaper from years 1869-1898 with 4 pages each, i.e. 224 pages in total. Given the selected set of 56 issues, our first data annotation and experiment phase consisted of annotating a subset of 28 issues (112 pages) and conducting preliminary experiments. After the preliminary annotation and experimentation resulting in a consistent practice, we fixed the annotation of the first 28 issues accordingly. Subsequently, we annotated the remaining 28 issues. We then divided the annotated set into training and evaluation sets of 168 and 56 pages. We trained PIVAJ successfully and evaluated the results using the layout evaluation software developed by PRImA research laboratory of University of Salford [6].

The results of our experiments show that PIVAJ achieves success rates of 67.9, 76.1, and 92.2 for the whole data set of 56 pages with three different evaluation scenarios introduced in [6]. On the whole, the results seem reasonable considering the varying layouts of the different issues of Uusi Suometar along the time scale of the data.

CCS CONCEPTS

• CCS → Applied computing → Document management and text processing → Document capture → Document analysis

• CCS → Information systems → Information retrieval → Document representation → Document structure

• CCS → Information systems → Information systems applications → Digital libraries and archives

KEYWORDS

Document layout analysis, article extraction, historical digitized Finnish newspaper archives, PIVAJ software

ACM Reference format:

Kimmo Kettunen, Teemu Ruokolainen, Erno Liukkonen, Pierrick Tranouez, Daniel Antelme, Thierry Paquet. 2019. Detecting Articles in a Digitized Finnish Historical Newspaper Collection 1771–1929: Early Results Using the PIVAJ Software. In *Proceedings of DATeCH conference (DATeCH2019).*

[1] https://digi.kansalliskirjasto.fi/etusivu?set_language=en
[2] https://extranet.content-conversion.com/dW/_layouts/15/start.aspx#/SitePages/Home.aspx

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
DATeCH2019, May 8–10, 2019, Brussels, Belgium
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7194-0/19/05...$15.00
https://doi.org/10.1145/3322905.3322911

1 Introduction

This paper describes article detection and extraction efforts on the Finnish Digi newspaper material of the National Library of Finland (NLF) using data of one newspaper, Uusi Suometar 1869–1898. It is a common practice that historical newspaper collections are digitized on page level: pages of the physical newspapers are scanned and OCRed and the page images serve as the basic browsing and searching unit of the collection. Searches to the collection are made on page level and results are shown on page level to the user. Page, however, is not any kind of basic informational unit of a newspaper, only a typographical or printing unit. Pages consist of articles or news items (and advertisements or notices of different kind, too), although length and form of them can be quite variable. Thus, separation of the article structure of digitized newspaper pages is an important step to improve usability of digital newspaper collections. As the amount of digitized historical journalistic information grows, also good search, browsing and exploration tools for harvesting the information are needed, as these affect usability of the collection. Contents of the collections are one of the key elements of usefulness of the collections, but also presentation of the contents for the user is important [10, 19]. As Dengel and Shafait [8] formulate it: "availability of logical structure facilitates navigation and advanced search inside the document as well as enables better presentation of the document in a possibly restructured format." Possibility to use article structure will also improve further analysis stages of the content, such as topic modeling or any other kind of content analysis. Information retrieval performance of a newspaper collection should also improve, if its contents are indexed on article level [5]. Several digitized historical newspaper collections have implemented article extraction on their pages. Good examples are for example Italian La Stampa[3], British Newspaper Archive[4], and Australian Trove[5].

The historical digital newspaper archive environment of the NLF is based on commercial docWorks software. The software is capable of article detection and extraction, but our material does not seem to behave well in the system in this respect. We have not been able to produce good article segmentation with docWorks, although such work has been accomplished e.g. in the Europeana Newspaper framework [19]. Thus, we have been in search of suitable software to perform article segmentation. We ended up trying software named PIVAJ developed in the LITIS laboratory of University of Rouen Normandy [11-13, 16, 17].

2 Selection of the article extraction tool

Article extraction on complex layout digitized newspaper pages is not an easy task. Results of the biannual ICDAR competition on historical newspaper layout analysis show that current algorithms segment and label about 80–85 % of the pages correctly at best [1–3, 7]. These results, however, are achieved with a smallish evaluation corpus of less than 100 documents [8] and the results are thus only indicative. Evaluation campaign Maurdor[7] from years 2013 and 2014 is the most comprehensive evaluation task in overall document segmentation and identification. Some results of the Maurdor campaign are shown in [5].

The latest round of ICDAR competition, ICDAR 2017, considered comparative evaluation of page segmentation and region classification methods for documents with complex layouts. Presented results included the results of the evaluation of seven methods: five submitted to the competition, and two state-of-the art systems: commercial ABBYY FineReader® Engine 11 and open-source Tesseract 3.04. In a combined task of segmentation and classification of page contents all but one of the systems remain in the range of 72.5 to 83% correctness. The best performing system, MHS 2017, joint work of HoChiMinh City University of Technology (Ho Chi Minh City, Viet Nam) and Chonnam National University (Gwangju, Republic of Korea), achieves a clearly better performance of 90.6% [7].

Our chosen tool PIVAJ has two parts: an on-line application and an off-line system. The offline system analyzes newspapers' digitization images in order to rebuild their logical structure, from issues to sections to articles. The online system allows for the display of the resulting analysis on the Web, as well as additional functionalities such as transcription corrections. We use only PIVAJ's offline system for newspaper image analysis and article marking.

Eskenazi et al. [9] surveyed and proposed a typology for most document segmentation algorithms between 2008 and 2017. According to this typology, PIVAJ's segmentation part would belong to *group 3*: Layout potentially unconstrained, subgroup hybrid techniques, as it combines methods of *group 2's* segmentation by feature classification (B&W for text and separators, grayscale for pictures) with another higher-level classifier for text level identification. After the segmentation part non-learning algorithms compute the general layout grid and the reading order of the page as was done in [11-12], and finally a recursive regular expression parser combines the parts into sections and articles.

From a user perspective, PIVAJ article annotation system is a machine learning software that is first taught with sample page images that have been marked for layout structure with entities like *title, body, vertical separator, horizontal separator, pictures* etc. Amount of needed training material is not described very clearly in documents related to PIVAJ, as it depends on the size and variability of newspaper's pages. Hebert et al. [11], for example, have used 11 pages for training for a standard, two-column, late 18th and early 19th century newspaper. After training, PIVAJ should be able to detect layout of similar pages, extract and mark articles from the pages, and output the results in ALTO and

[3] http://www.archiviolastampa.it/
[4] https://www.britishnewspaperarchive.co.uk/
[5] https://trove.nla.gov.au/

[7] http://www.maurdor-campaign.org/

METS files. PIVAJ has so far been used e.g. to process 100 000 pages of Journal de Rouen, out of which it detected 550 000 articles [12]. In an evaluation with a very small evaluation data set of 42 documents PIVAJ's accuracy in article extraction was 85.84% [11]. This is in line with results achieved in ICDAR's historical newspaper layout analysis competition, where best systems achieve accuracy of 83–86% [1–2]. Same range of performance is achieved also in ICDAR 2015 and ICDAR 2017 where documents with complex layout were analyzed [3, 7].

3 Selection of experimental data

At the time of writing the whole Digi has 941 newspaper titles with 4 078 876 freely usable pages. The number of the titles shows that if we are going to perform article extraction on the collection, we need to start with a modest sub collection of the most important newspapers to see, whether article extraction is feasible and results good enough for larger scale work. If the results are promising, we need to assess whether our chosen segmentation method scales up so that large-scale article extraction with more newspapers is possible.

To get an overview of usage of newspapers we examined first usage of different titles from Digi's use logs. We studied both title and page load statistics of the 20 most used newspapers during the period of 1.1.2009 and 21.2.2017. During that time there were about 10 million title loads and 20 million page loads for the 100 most used newspapers. The newspapers included both Finnish and Swedish newspapers from different parts of Finland, mostly published in largest cities like Helsinki, Turku and Tampere, but also some newspapers from smaller towns. Figure 1 shows the 20 most used newspapers.

#	Aineistotyyppi	Nimeke (ISSN)	Käyttökerrat	Prosenttiosuus
1	SAN	Åbo Underrättelser (0785-398X)	583715	5,3 %
2	SAN	Finlands Allmänna Tidning (1457-4314)	563285	5,1 %
3	SAN	Hufvudstadsbladet (0356-0724)	458065	4,1 %
4	SAN	Uusi Suometar (1457-4721)	355417	3,2 %
5	SAN	Aamulehti (0355-6913)	331822	3,0 %
6	SAN	Helsingfors Dagblad (1458-0802)	306917	2,8 %
7	SAN	Suomalainen Wirallinen Lehti (1457-4675)	294532	2,7 %
8	SAN	Työmies (fk14802)	236481	2,1 %
9	SAN	Sanomia Turusta (1457-4616)	186407	1,7 %
10	SAN	Päivälehti (1458-2619)	176639	1,6 %
11	SAN	Wiborg (1457-4837)	158094	1,4 %
12	SAN	Suometar (1457-4705)	147320	1,3 %
13	SAN	Oulun Wiikko-Sanomia (1457-4527)	142471	1,3 %
14	SAN	Helsingfors Tidningar (1457-439X)	138902	1,3 %
15	SAN	Nya Pressen (1458-2503)	132940	1,2 %
16	SAN	Hämäläinen (1457-4403)	125164	1,1 %
17	SAN	Åbo Tidningar (1457-4802)	113930	1,0 %
18	SAN	Vasabladet (0356-1844)	111389	1,0 %
19	SAN	Keski-Suomi (1457-4640)	110110	1,0 %
20	SAN	Satakunta (1458-2740)	106215	1,0 %

Figure 1. 20 most used newspaper in the Digi collection 2009-2019. Column labels from left to right are: type of data (newspaper), title (ISSN), number of loads, per centage of usage.

As our article extraction trial material we chose newspaper Uusi Suometar (US), which is the fourth most used newspaper in the statistics and the first newspaper in the list in Finnish. The first three titles were published in Swedish, which was the major language of publication in Finland until 1880. Uusi Suometar started to appear in year 1869 and was published with the same name until the end of year 1918. After that its name was changed to Uusi Suomi. Uusi Suometar was first published twice a week, later on its publication frequency increased to six times a week and from 1913 it became a daily.

Layout of Uusi Suometar follows the normal Manhattan style [14] and the publication starts with three columns. Later on columns were added in periods of 3–5 years until in 1894 the layout consisted of nine columns. From 1894 onwards, the number of columns varied from six to seven. This pattern of increasing columns can be found in most of the main newspapers of the same period, e.g. in *Aamulehti, Työmies, Päivälehti, Hämäläinen* and *Satakunta*. As for advertisements, Uusi Suometar did not contain ads or similar announcements in the early years. However, the number of advertisements increased fast and by 1890 they formed a prominent portion of the content. Advertisements bring complexity to the layout of a newspaper page as they can include more graphical elements, odd fonts and can be set on the page more unevenly than news articles.

Out of the available Uusi Suometar material, we established a data set containing 56 issues sampled from years 1869-1898 with 4 pages each, i.e. 224 pages in total. Given the selected set of 56 issues, our first annotation and experiment phase consisted of annotating a subset of 28 issues (112 pages) and conducting preliminary experiments. The motivation for this phase was to gain experience with the PIVAJ annotation toolkit, to detect and solve potential problems with the annotation practices, and to familiarize ourselves with the PIVAJ experiment pipeline (learning from ground truth and applying the learned model). In order to run the experiment pipeline, we divided the 28 issues into a training and a development set of 21 and 7 issues (84 and 28 pages), respectively. After the preliminary annotation and experimentation resulting in a consistent practice, we fixed the annotation of the first 28 issues (112 pages) accordingly. Subsequently, we annotated the remaining 28 issues (112 pages). We then divided the annotated set to training and test sets of 168 and 56 pages, respectively.

As mentioned above, the number of columns per page in US issues varies from 3 to 9. Because we could not be sure if PIVAJ would be able to learn simultaneously from issues with such a varying number of columns, we wanted to try out learning and evaluating an individual model for issues with a shared amount of columns. Therefore, we selected an equal amount of issues with the number of columns 3–9 from the years 1869–1898. We note that this distribution does somewhat differ from a uniformly sampled set of issues since some column numbers appear more frequently in the collection than others. In

retrospect, however, as it turned out that PIVAJ indeed has no problem handling varying number of columns, the straightforward uniform selection criterion would have been perfectly justifiable and is thus recommended for future work.

In order to study if number of columns had an impact on PIVAJ, we first trained an individual model on the 3 issues containing 3 columns in the training set and evaluated the model on the issue in the development set with 3 columns. Subsequently, we repeated the same for the rest of the column numbers from 4 to 9. As expected, PIVAJ was able to learn and provide meaningful predictions (measured by visual inspection) on the development issues. We then trained a single model using all the 21 issues with varying column numbers and evaluated the resulting model on all the 7 issues in the development set. Our hypothesis was that in case the PIVAJ would have difficulties handling the variety of number of columns, we would observe anomalies during training or, more importantly, in the predictions on the development set. However, we saw no evidence of this, that is, PIVAJ again provided meaningful predictions and no undesirable behavior on the development set (measured by visual inspection). Therefore, we adopted the single model approach for the primary experiments conducted on all 56 issues discussed in the next section.

4 Results of the experiments

4.1 Evaluation

We trained PIVAJ successfully, ran the experiments and evaluated the physical segmentation results using the layout evaluation software developed by PRImA research laboratory of University of Salford [6][8], which is applicable subsequent to converting the ALTO XML structure used by PIVAJ to PAGE XML. The PRImA software has been employed for evaluation of the biannual ICDAR competitions (2011/13/15/17). However, the specifics of the official competition evaluation are not publicly available and, thus, the competition evaluation is not replicable. Therefore, we instead followed the evaluation presented in [6] with three evaluation scenarios:

The *General recognition* is used to measure the pure segmentation performance. Therefore, misclassification errors are ignored completely. Miss and partial miss errors are considered worst and have the highest weights. The weights for merge and split errors are set to 50%, whereas false detection, as the least important error type, has a weight of only 10%. This is named as general recognition in the output of the evaluation software.

The *Text structure* scenario evaluates region classification, in the context of a typical OCR system, focusing primarily on text but not ignoring the non-text regions. Accordingly, this profile is similar to the first but misclassification of text is weighted highest and all other misclassification weights are set to 10%. This is named as text structure in the output of the evaluation software.

The third scenario, *Indexing*, is based on the OCR profile but focuses solely on text, ignoring non-text regions. This is named as indexing in the output of the evaluation software.

Note that the fourth scenario, Images & Graphics, included by [6] was not included in the PRImA output and is therefore excluded.

4.2 Results

We show our results along the three evaluation scenarios explained in chapter 4.1. and used in [6]. PRImA's evaluation tool does not give an average counting for the results of one newspaper issue, only results page by page. The average results reported here have been calculated from the page by page results. Figure 2 shows area weighted success rates for the three different scenarios.

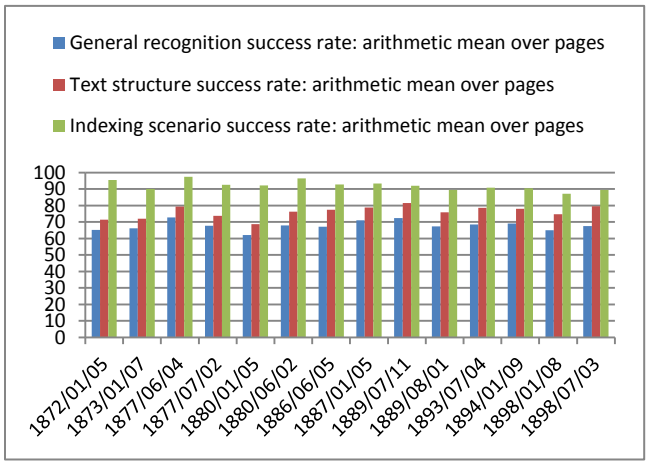

Figure 2. Area weighted success rates for the three evaluation scenarios. Mean average figures for the issues.

On average the three evaluation scenarios get success rates of 67.9, 76.1, and 92.2 for the whole data set of 56 pages.

These scenarios only evaluate the physical segmentation results of PIVAJ on our material; PIVAJ also computes text level (Title 1 to 3, body text), reading order and finally, article extraction.

A more detailed inspection of the results shows the following:

If the pages contain longer sections that consist of short news items of few lines, the news are not well extracted, only the larger section, if there are no clear, e.g. bolded, starts for the short news items.

Figure 3, however, is an example of a quite successful extraction of a complex page: it gets recognition rates of 81.86, 89.95, and 96.54.

[8] https://www.primaresearch.org/tools

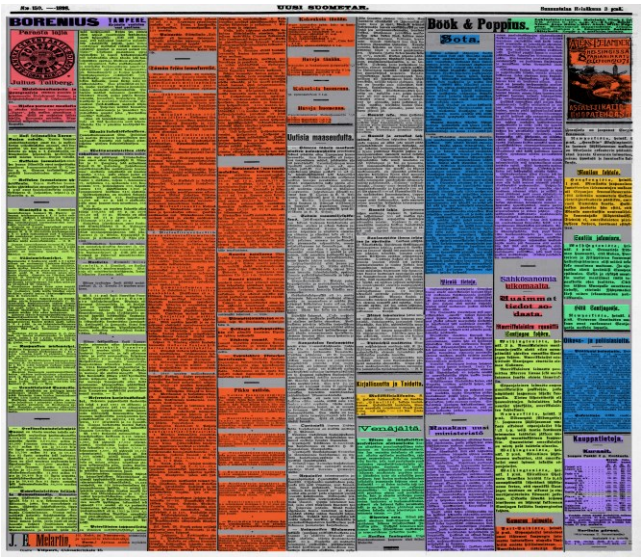

Figure 3. Example output of PIVAJ for a complex page

Advertisements or different announcements get a clearly worse recognition, and they are usually on the fourth page. Layout of the advertisement pages may also vary quite a lot: some are more regular than others in the respect that they do not contain odd graphical elements. One of the worst examples from our data is shown in figure 4. It gets recognition rates of 57.4, 73.02, and 78.73, clearly below the averages.

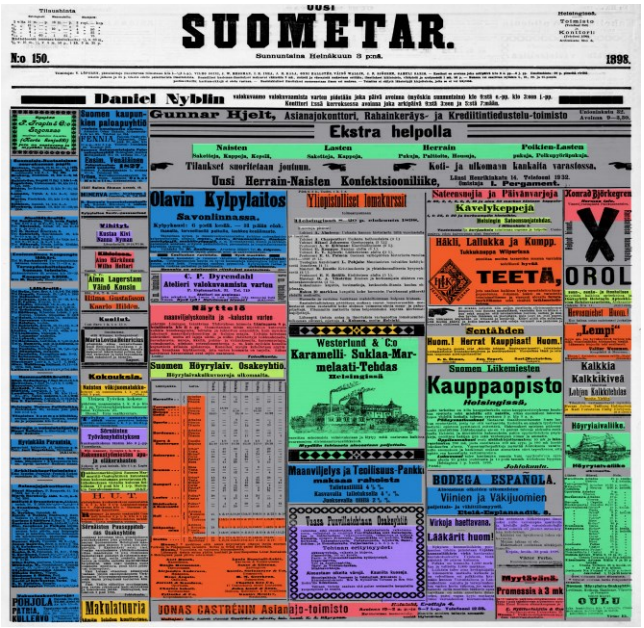

Figure 4. Output of PIVAJ from a page containing lots of advertisements and odd graphical elements

In general, the increasing number of columns along the time line of the data does not cause clear deterioration in performance of PIVAJ. Thus our "one model" tactics for the training of PIVAJ seems justifiable.

We also compared the number of articles PIVAJ marks to the number of articles marked by docWorks in the same page data. docWorks is able to mark 150 articles in the 56 pages, as PIVAJ marks 1013. The page data of docWorks are taken out of our current Digi presentation system, and thus the results cannot be easily compared in a more detailed manner. The distribution of found articles in different issues is shown in Figure 5.

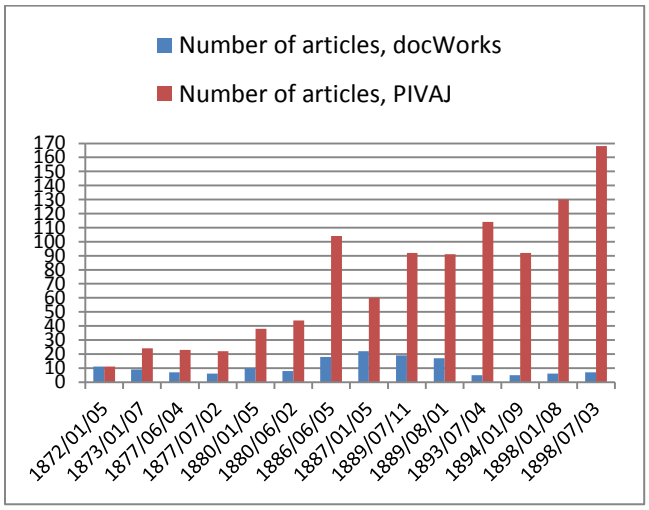

Figure 5. Number of articles found in the 14 issues.

If we compare our results to those of ICDAR competition we can note that the results show the same pattern as in [6] with regards to general recognition, full text recognition and text indexing: the first is the lowest, the second clearly better and the third the best. The main difference is that our figures are lower by a magnitude of 10% units with the first two scenarios and by a magnitude of 5% units with the third scenario.

5 Conclusion

This paper described first large scale article detection and extraction efforts on the Finnish Digi newspaper material of the National Library of Finland (NLF) using data of one newspaper. We have used PIVAJ machine learning based platform developed at the LITIS laboratory of University of Rouen Normandy [11-13, 16, 17].

As training and evaluation data for PIVAJ we chose one central newspaper from our collection, Uusi Suometar. We established a data set that contains 56 issues of the newspaper from years 1869-1898 with 4 pages each, i.e. 224 pages in total, and annotated the data with PIVAJ's annotation tool. We divided the annotated set into training and evaluation sets of 168 and 56

pages. We trained PIVAJ successfully and evaluated the results using the layout evaluation software developed by PRImA research laboratory of University of Salford [6].

The results of our experiments show that PIVAJ achieves success rates of 67.9, 76.1, and 92.2 for the whole data set of 56 pages with three different evaluation scenarios introduced in [6]. The mean quality of the results is lowered by PIVAJ's behavior on advertisement heavy pages. LITIS is currently working on new solutions to overcome this weakness. On the whole, the results seem reasonable considering the varying layouts of the different issues of Uusi Suometar along the time scale of the data.

Our future plans include further testing of PIVAJ with more data of Uusi Suometar. Especially we need more experience in the scalability of the software. The publication history of Uusi Suometar from 1869 to 1918 includes 86 060 pages, so producing article extraction even for one main newspaper is a considerable task. We shall also evaluate working of the developed PIVAJ model on other newspapers: although the layouts of many newspapers seem to be very similar, it is not clear, whether one trained PIVAJ model is enough to cover a substantial set of other newspapers, too. The producer of the PIVAJ software, LITIS Laboratory, ran some cross-newspaper experiments between Le *Journal de Rouen*, *Le journal des Débats* and *Le Gaulois* using models trained on one title on the other titles. What seems important is that newspapers are similar enough at the pixel level: separators should have roughly the same width or height expressed in pixels (thus depending both on the physical dimensions of the paper and the density of the digitizing), proportions between the different sizes of titles should be approximately the same etc. Change in the number of columns is not in general an obstacle for PIVAJ. On the other hand, the statistical nature of this tool makes it sometimes fail on pages that seem very similar to the human eye. LITIS is working at improving this.

ACKNOWLEDGMENTS

Work done at The National Library of Finland was funded by the European Regional Development Fund and the program Leverage from the EU 2014-2020.

REFERENCES

[1] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher (2011). Historical Document Layout Analysis Competition. Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR2011), Beijing, China, September 2011, 1516-1520.

[2] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher (2013). ICDAR2013 Competition on Historical Newspaper Layout Analysis – HNLA2013. DOI: 10.1109/ICDAR.2013.293

[3] A. Antonacopoulos, C. Clausner, C. Papadopoulos, S. Pletschacher (2015). ICDAR2015 Competition on Recognition of Documents with Complex Layouts – RDCL2015. DOI: 10.1109/ICDAR.2015.7333941.

[4] P. Barlas, S. Adam, C. Chatelain, T. Paquet (2014). A typed and Handwritten text block segmentation system for heterogenous and complex documents. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.800.4721&rep=rep1&type=pdf

[5] F. Buhr, B. Neumann (2014). Evaluation of Retrieval Performance in Historical Newspaper Archives comparing Page-level and Article-level Granularity. https://kogs-www.informatik.uni-hamburg.de/publikationen/pub-buhr/Newspaper-Retrieval.pdf

[6] C. Clausner, S. Pletshacher, A. Antonacopoulos (2011). Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods. 2011 International Conference on Document Analysis and Recognition (ICDAR). DOI: 10.1109/ICDAR.2011.282

[7] C. Clausner, A. Antonacopoulos, S. Pletschacher (2017). ICDAR2017 Competition on Recognition of Documents with Complex Layouts – RDCL2017. https://ieeexplore.ieee.org/document/8270160

[8] A. Dengel, F. Shafait (2014). Analysis of the Logical Layout of Documents. In Doerman, D., Tombre, K. (eds.) Handbook of Document Image Processing and Recognition, 177–222. Springer. DOI 10.1007/978-0-85729-859-1

[9] S. Eskenazi, P. Gomez-Krämer, J.-M. Ogier (2017). A Comprehensive survey of mostly textual document segmentation algorithms since 2008. Pattern Recognition 64, 1–14.

[10] N. Fuhr, G. Tsakonass, T. Aalberg, M. Agosti, P. Hansen, S. Kapidakis, C.-P. Klas, L. Kovács, M. Landoni, A. Micsik, C. Papatheodorou, C. Peters, I. Sølvberg (2007). Evaluation of digital libraries. International Journal on Digital Libraries 8(1), 21–38.

[11] D. Hebert, T. Palfray, T. Nicolas, P. Tranouez, T. Paquet (2014). Automatic article extraction in old Newspapers Digitized Collections. In Proceeding DATeCH '14 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, 3–8. http://dl.acm.org/citation.cfm?id=2595195

[12] D. Hebert, T. Palfray, T. Nicolas, P. Tranouez, T. Paquet (2014). PIVAJ: displaying and augmenting digitized newspapers on the Web Experimental feedback from the "Journal de Rouen" Collection. In Proceeding DATeCH '14 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, 173–178. http://dl.acm.org/citation.cfm?id=2595217

[13] D. Hebert, T. Paquet, T. Nicolas (2011). Continuous CRF with multi-scale quantization feature functions Application to structure extraction in old newspapers. 2011 International Conference on Document Analysis and Recognition, Beijing, 2011, 493–497. DOI: 10.1109/ICDAR.2011.10

[14] K. Kise (2014). Page Segmentation Techniques in Document Analysis. In Doerman, D., Tombre, K. (eds.) Handbook of Document Image Processing and Recognition, 135–175. Springer. DOI 10.1007/978-0-85729-859-1

[15] V. Märgner, H. El Abed (2014). Tools and Metrics for Document Analysis Systems Evaluation. In Doerman, D., Tombre, K. (eds.) Handbook of Document Image Processing and Recognition, 1011–1036. Springer. DOI 10.1007/978-0-85729-859-1

[16] T. Palfray, D. Hebert, T. Nicolas, P. Tranouez, T. Paquet (2012). Logical segmentation for article extraction in digitized old newspapers. https://arxiv.org/ftp/arxiv/papers/1210/1210.0999.pdf

[17] P. Tranouez, S. Nicolas, J. Lerouge, T. Paquet (2015). PIVAJ: an article-centered platform for digitized newspapers. http://www.imaging.org/site/PDFS/Reporter/Articles/REP30_3_ARCH2015_TRANOUEZ.pdf

[18] M. Willems, R. Atanassova, Rossitza (2015). Europeana Newspapers: searching digitized historical newspapers from 23 European countries. Insights 28(1).

[19] H.I. Xie (2008). Users' evaluation of digital libraries (DLs): Their uses, their criteria, and their assessment. Information Processing and Management, 44(3), 1346–1373.