Optimisation of archival processes involving digitisation of typewritten documents

Stančić, Hrvoje; Trbušić, Željko

Source / Izvornik: **Aslib Journal of Information Management, 2020, 72, 545 - 559**

Journal article, Accepted version
Rad u časopisu, Završna verzija rukopisa prihvaćena za objavljivanje (postprint)

https://doi.org/10.1108/AJIM-11-2019-0326

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:131:186434

Optimisation of Archival Processes involving Digitisation of Typewritten Documents

Hrvoje Stančić

Department of Information and Communication Sciences,

Faculty of Humanities and Social Sciences, Zagreb, Croatia, and

Željko Trbušić

Division for the History of Croatian Literature,

Institute for the History of Croatian Literature, Theater and Music,

Croatian Academy of Sciences and Arts, Zagreb, Croatia

Aslib Journal of Information Management, Vol. 72, No. 4, pp. 545-559.
https://doi.org/10.1108/AJIM-11-2019-0326

Abstract:

Purpose – The authors investigate optical character recognition (OCR) technology and discuss its implementation in the context of digitisation of archival materials.

Design/methodology/approach – The typewritten transcripts of the Croatian Writers' Society from the mid-sixties of the 20th Century are used as the test data. The optimal digitisation setup is investigated in order to obtain the best optical character recognition results. This was done by using the sample of 123 pages digitised at different resolution settings and binarisation levels.

Findings – A series of tests showed that different settings produce significantly different results. The best OCR accuracy achieved at the test sample of the typewritten documents was 95.02%. The results show that the resolution is significantly more important than binarisation pre-processing procedure for achieving better OCR results.

Originality/value – Based on the research results, the authors give recommendations for achieving optimal digitisation process setup with the aim of increasing the quality of OCR results. Finally, the authors put the research results in the context of digitisation of cultural heritage in general and discuss further investigation possibilities.

Keywords Digitisation, Optical character recognition, Resolution, Binarisation, Typewritten documents, Archival materials, Cultural heritage

Paper type – Research paper

Introduction

The information retrieval and keyword search capabilities have become the standard of all digital repositories today. Some of them offer full text search of the documents and records they keep and preserve but this is usually just a fraction of vast archival holdings kept in the paper form. Therefore, mass digitisation [1], indexing and full text retrieval of archival materials is the field in which the improvement of process automation can be expected. In a modern-day

personal computer era optical character recognition (OCR) process does not require special-purpose hardware or expensive software to build a system that can be used in the digital repositories (Blostein and Nagy, 2012, p. 3). Special purpose workflows can be easily built depending on the type of archival materials (Blanke *et al.* 2011). Although some can be easily digitised and converted to searchable form using OCR (e.g. printed materials), some of them can be a bit more complicated, like typewritten documents, while for other, like handwritten documents, it can be very difficult and sometimes impossible to efficiently apply either optical or intelligent character recognition (ICR) and achieve meaningful results. It is expected that the handwritten text recognition (HTR) might benefit more from the application of machine learning (ML) and artificial intelligence (AI) approaches (Muehlberger *et al.*, 2019) than the type of archival documents discussed in this work. However, HTR accuracy can also be improved using similar digitisation testing methodology as the one discussed here.

AI is a broader term usually encompassing different technologies such as pattern recognition, text analytics, natural language processing (NLP), named entity recognition (NER), machine learning (ML) etc. AI is based on expert systems which store and transfer information in symbolic containers (Tweedie, 2018). It should be noted that although pattern recognition is mentioned as the technology belonging to the AI realm, and OCR is in its essence a pattern recognition problem, this research is more focused on the digitisation process and its effect on the OCR accuracy rather than whether OCR is or is not influenced by recent advancements in the field of artificial intelligence. Some authors, such as Shank (1991), claim that as the technology, once considered as AI, becomes widely used it cedes to become considered as AI. A good example is the chess programs. Once they were related with AI while today one can find them embedded in toys. Similar is with the OCR. Could it be that AI is something that has not been done yet, or that has not been scaled up to be used as an everyday technology? However, before the AI-related technologies enable more efficient recognition of the early typewritten and older handwritten documents, quality OCR results can be achieved by setting up the optimal digitisation process. But what is the optimal digitisation process, and could it be abstracted so that the same approach is relevant for different archival materials? The goal is to discuss, using the example of quality levels and binarisation procedures, that the appropriate testing methods and their results can greatly improve management effectiveness of a large-scale digitisation project regardless of the technology used. Even though state-of-the-art technology is easy to obtain nowadays, not every institution has the means to update software on a regular basis or pay for per-page OCR service. Thus, a process optimisation methodology can greatly improve cost effectiveness of archival procedures.

Goal of this research is to formulate a testing procedure that is applicable to a variety of similar typewritten documents that can be found in the archives all around the world. The use of optimised procedure unlocks the potential to accelerate the workflow and automate the digitisation process (Blostein and Nagy, 2012), thus saving time and reducing costs of document conversion. Choosing the optimised approach can potentially save days, or even weeks of work in a large-scale digitisation project and produce substantial economic benefits.

Users of the digital repositories often conduct their search and make decisions based on the information that has been delivered to them. One can question if the users have received the full information available in the digital repository, or some parts have not been provided to the users because of the inaccuracy of OCR process (Traub *et al.*, 2015). To minimise the error rate of these systems a methodological, empirically tested approach to OCR is needed. Therefore, the goal of this paper is to test influence of resolution in which documents are digitised and the process of binarisation [2], as the pre-process for OCR, on the results of the OCR in order to develop recommendations for optimal digitisation process setup. Although research used a sample of the twentieth century typewritten texts in Croatian language, the proposed

methodology is relevant for other languages as well. Research investigated if the typewritten documents can be successfully optically recognized using modern OCR software and if the accuracy is good enough to automate the process on a larger scale with minimal user involvement, e.g. without manual pre-processing of every page, or manual correction of the output. Similar tests were conducted on a much larger scale from 1991 to 1996 by the Information Science Research Institute (ISRI) at the University of Nevada, Las Vegas, where they examined then available OCR systems and published their results in yearly reports. Albeit older, their methodology is still valuable today and can be additionally improved upon using the newer technologies available today and by considering different type of archival material. They have developed an evaluation software called The ISRI Analytic Tools for OCR Evaluation (referred to as The ISRI Tools in the text) (Rice and Nartker, 1996) which was used to test the recognition accuracy.

The decision to use The ISRI Tools in this research, as opposed to some more recent evaluation tools, was made considering two requirements: the software needed to be able to process Croatian texts in a modern operating system environment and it should be openly available tool that provides the most comprehensive data possible. We used an updated port of The ISRI Tools with UTF-8 encoding capabilities that corresponded well to our needs [3]. Other openly available evaluation tools either functioned as a GUI versions of The ISRI Tools or provided less data relevant to our research. Hubert *et al.* (2016) deliver three key reasons why The ISRI Tools are still the most advanced OCR evaluation tool to date: '(1) it presents confusion matrices and accuracy values for single characters and words, (2) it comes with an extensive set of separate tools that each assess and highlight different performance metrics, and (3) it is the only toolkit suite in existence to have been used as a *de facto* standardized assessment tool'.

The usual digitisation project using OCR consists of four stages (Cojocaru *et al.* 2016, p. 109): (1) Image capture and image pre-processing, (2) OCR, (3) Text post-processing, and (4) Quality evaluation. In our investigation we will try to answer questions primarily concerning the first stage of digitisation and pre-processing aiming to achieve better results in the second stage. All stages are further explained by the Optical Character Recognition IMPACT Best Practice Guide (Anderson, 2010a).

Research aimed to investigate two main hypotheses: (H1) the accuracy rates will grow with the improvement of digitisation quality, and (H2) the binarisation will additionally improve the OCR results. Research also investigated a sub-hypothesis: (H2.1) the extremely high and extremely low binarisation will produce illegible results that cannot be efficiently recognized by a machine or read by humans.

Since our goal was to create an automated workflow of optical character recognition based on the type of documents in our repository, we have decided to use a Linux software called ImageMagick [4] for the binarisation. It is a command-line tool that is easily integrated into an existing system and can serve as a basis for an automated image manipulation process. The actual character recognition process was done using two different engines, a proprietary Abbyy FineReader 15 [5] and an open-source Tesseract 4 [6]. The general agreement is that the good OCR accuracy rate is 98-99% of characters successfully recognized (Holley, 2008, p. 5, Strange *et al.* 2014, p. 13) and we were interested to explore if those results could be achievable on the chosen set of documents.

The testing process

Dataset

The test data for this research consisted of 123 A4 sized colour scanned pages that comprises 17 typewritten transcripts of Croatian Writers' Association [7] (hrv. *Društvo hrvatskih književnika, DHK*) board meetings and plenary sessions (119 pages) from 1966 to 1968 with the addition of the typewritten Declaration Concerning the Name and the Position of the Croatian Literary Language (hrv. *Deklaracija o nazivu i položaju hrvatskog književnog jezika*, 4 pages). The documents are held in the archive of the Division for the History of Croatian Literature at the Croatian Academy of Sciences and Arts and represent a small sample from the DHK holding that covers the period from 1900 up until 1970. It is not clear if the documents were all written using the same typewriter or by the same person because information like that was never, or rarely, included. The documents do not have any significant differences in the condition of the paper or the print since they were stored in the same environment and that they are from the same time period. Some of the pages have degradations due to the way they were stored (Figure 1) but generally the documents are in good condition. No tables, figures or images are present in the text, only an occasional handwritten annotation on the margin. This sample is a valid representation of a much greater number of documents that are expected to be found in many archives that hold 20th century records.

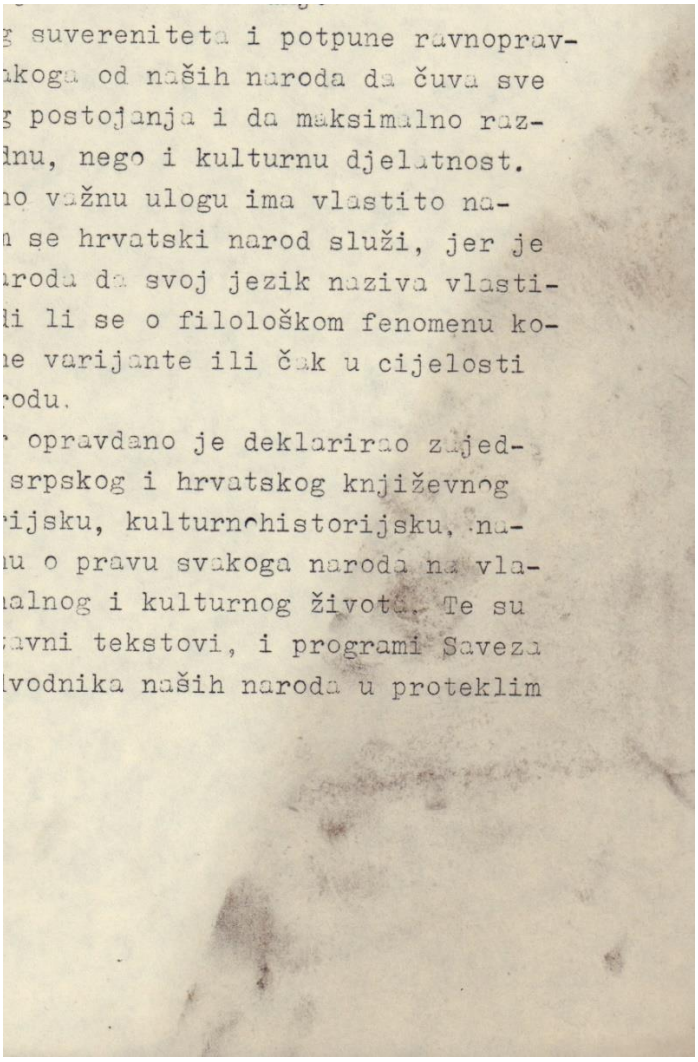

Figure 1. The stains produced by improper storage that can affect OCR success rates

The language of the documents was not a problem for the recognition since both Abbyy and Tesseract can recognize various languages, including Croatian. The specific Croatian characters included in the documents (as compared to documents written in English) are *č*, *ć*, *đ*, *š*, *ž* and bigrams *lj*, *nj* and *dž*. The software had no problems in recognizing them correctly. The character *x*, that is not a part of the Croatian alphabet, was commonly used when something needed to be omitted from the transcript or a mistake was made by the typist and then crossed out (Figure 2).

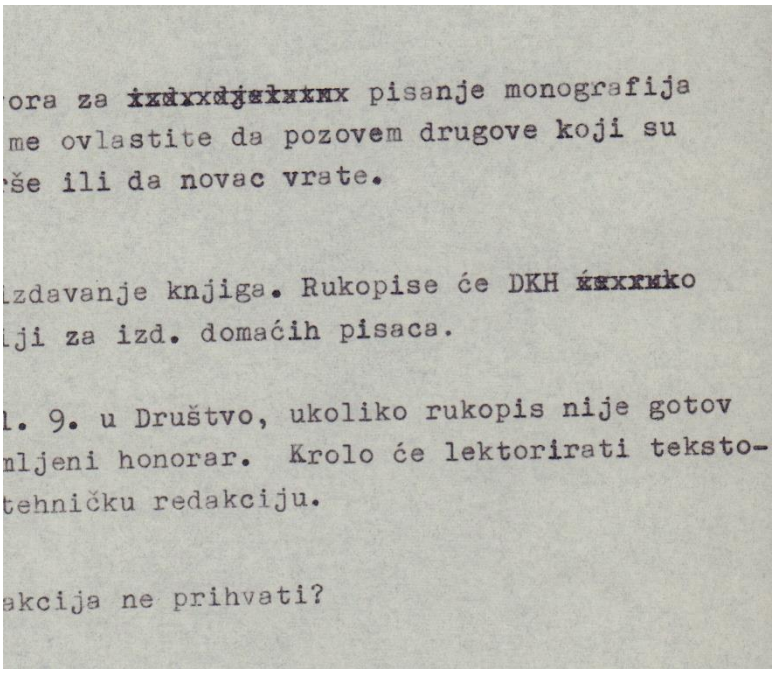

Figure 2. Usage of character *x* for correcting mistakes

Typewritten documents are specific in a way that they can be considered machine-printed, but the quality of the print is not even, and it depends on the key pressure force applied by the typist. The uniformity of character shapes enables the modern OCR software to recognize the text (as opposed to handwriting) but the difference from the industrial printing technologies is that more print anomalies are present (faint, blurred, dark and filled-in characters that can cause problems during the recognition process) causing certain characters not to be recognized.

Methodology

Research started with scanning all 123 pages of the chosen dataset using a flatbed scanner. Each page was scanned as a colour, uncompressed TIFF image in six different quality levels (100, 200, 300, 400, 500 and 600 dots per inch, dpi) thus producing the dataset of 738 pages. The scanner used was a CIS scanner, i.e. not a CCD scanner. This is a potential limitation of research because CIS scanners use hardware-based interpolation for image creation. However, all scanned images were scanned using optical, i.e. not (software-based) interpolated resolution. The predefined scanner settings were used consistently when digitising documents in order to avoid possible uncertainties regarding file compression and file format specific restrictions. For example, the brightness and contrast settings, as well as the file type were used consistently throughout the scanning procedure. The images were not additionally edited in any way. The scanning process was conducted by a professional. Additional care was taken that the pages are not skewed during the positioning on the scanner surface and that the scanned object is not raised from the scanning glass which would produce blurred images. The latter was possible since all the 123 pages of the chosen data set were in good condition, i.e. they were not wrinkled.

The six different quality levels were taken in succession without replacing the scanned object or manipulation of any kind.

For optical character recognition we have used two up-to-date OCR engines: Tesseract version 4.0.0 with Leptonica library version 1.76.0 and Abbyy FineReader 15 (release 4). No additional customized training or language files were used, and the default OCR settings were not changed. In this way the engine itself is observed as a 'black box' system without the need to discuss its internal workings. Although the advantages and disadvantages of this approach can be discussed further (e.g. Abbyy FineReader has much more customisation options than Tesseract), this falls out of the scope of this research. For this reason, the accuracy results are not analysed further than in their relation to the different quality levels and binarisation settings. No special page segmentation options were selected during the recognition and the output was plain .txt format with line separation.

Accuracy measurement was conducted with the aid of The ISRI Tools and the accuracy of the recognized text was calculated on a character and word levels. The values are expressed as accuracy percentages, rather than error rates (CER and WER), according to The ISRI Tools output format. The character accuracy method is the basic method of quantifying OCR accuracy and it consists of counting the number of character insertions, deletions, or substitutions needed to fully correct the text by comparing the OCR output with the ground truth files (Rice and Nartker, 1996, p. 4) prepared by the authors. The word accuracy measurement is a percentage value of correctly recognized words in comparison to the number of words in the ground truth. The ground truth text files preparation followed a user guide of The ISRI Tools (Rice and Nartker, 1996, p. 2): a tilde (~) was used where a character should be rejected for comparison and circumflex (^) as a suspect marker. These markers were calculated in the final accuracy score output of the ISRI tools. The example of a rejected character is shown in Figure 3 where it is not expected that the OCR engine correctly recognized a character since it is an overtype of two different letters. The size of the ground truth was the same as the size of the tested sample: 123 pages.

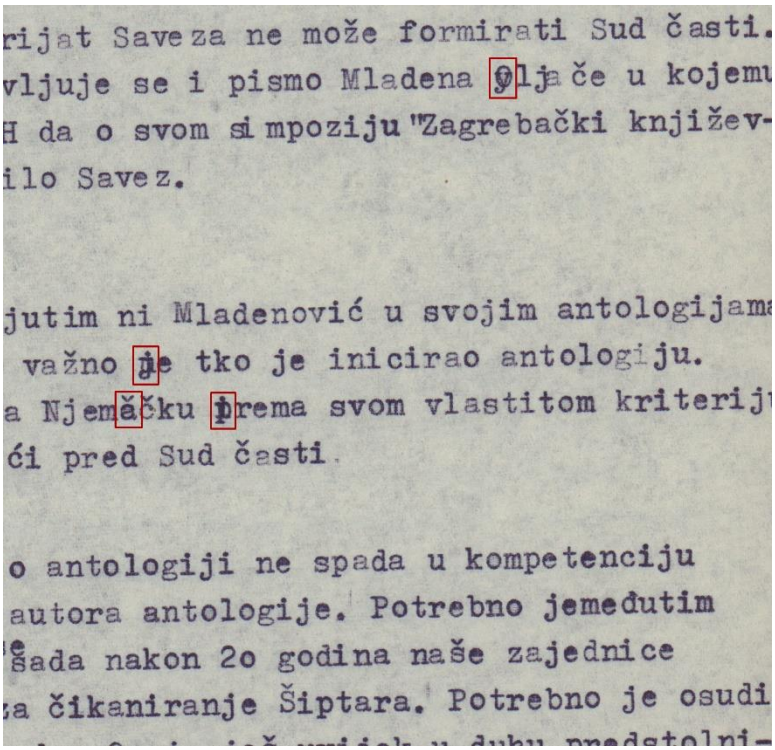

Figure 3. The example of characters not expected to be recognized by the OCR engine

After acquiring the accuracy results for different quality levels (100-600 dpi) the next step was image binarisation testing. We have used an open-source software ImageMagick and its threshold function in order to binarise the images. The threshold function takes the original image and delivers a binarised version according to the set percentage value. Even though Tesseract applies Otsu binarisation algorithm and Abbyy FineReader incorporates their own set of image pre-processing methods (Intelligent Background Filtering and Adaptive Binarisation) external binarisation was used because of three key reasons: (1) it is suggested as a pre-processing stage in recent previous research projects (Smitha *et al.*, 2016; Koistinen *et al.*, 2017a, b), (2) the preliminary testing results showed different accuracy levels when external binarisation was applied, and (3) binarisation is suggested as an image improvement pre-processing stage in the OCR software developers documentation [8]. Since findings of this research discovered that binarisation is not a necessary procedure due to the internal binarisation of the OCR software, the results represent a step forward in the digitization methodology because current literature suggests using binarisation.

The first step was to determine which level of binarisation can be safely applied to our documents thus testing the sub-hypothesis H2.1. This procedure included extracting the first page of every document in every quality level (108 files in total, a sub-sample) that were binarised using the threshold settings from 0-100% in 10% increments. Upon visual inspection it was concluded that no further testing and recognition is possible on images with the threshold set to 20% and below since it produced almost completely white pages. Similar situation happened with the threshold value set to 90-100% – it produced almost completely black images that have no further use. Therefore, the files pre-processed using those settings were discarded. The process continued with the recognition and accuracy testing of the remaining sample. The 70% and 80% threshold value images took too long to process by Tesseract (some pages took more than 10 minutes to finish) and therefore could not have an additional positive impact on OCR implementation in archival systems and were removed from further testing. FineReader had no such difficulties and it should be further tested for accuracy of higher binarisation levels. Finally, the remaining 30-60% threshold settings were used in the optical character recognition accuracy testing. The sub-sample accuracy results (Table 1) showed that only 50% and 60% binarisation settings results are comparable to the not pre-processed documents and therefore were applied to the whole sample.

Table 1. The effect of binarisation on the character accuracy (preliminary testing)

Binarisation (%)	**Tesseract Character Accuracy (%)**	**FineReader Character Accuracy (%)**
n/a	82.46	93.41
30	8.99	8.28
40	24.88	34.69
50	53.05	64.76
60	77.74	89.13

The second phase involved binarisation of the whole test data using the narrowed down threshold settings (50-60%). The binarised images were then processed by Tesseract and Abbyy FineReader, evaluated by ISRI tools and the results were compared with the results of the not pre-processed images. The whole testing methodology is shown in Figure 4.

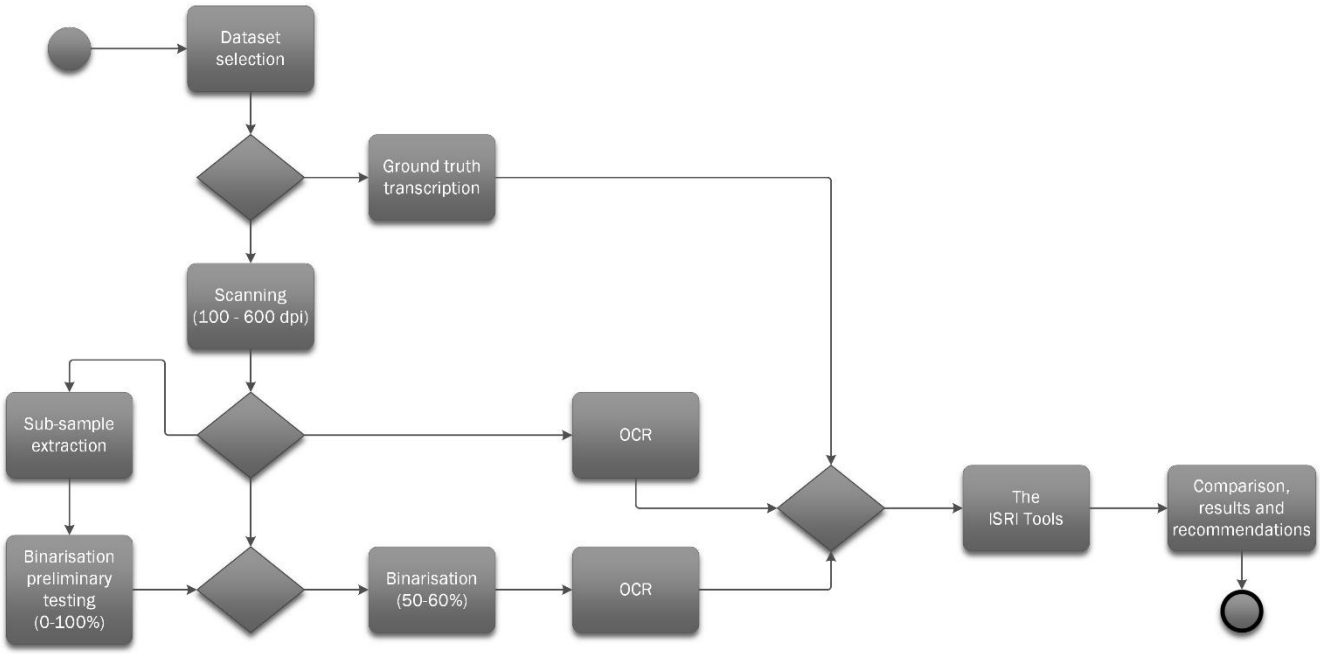

Figure 4. Flowchart of the testing process

Results

During the research phase several indicators were monitored. The average time needed to scan 123 A4 sized pages at different resolutions (100-600 dpi) in colour shows that almost identical time is needed to scan pages at 100 and 200 dpi. The group of resolutions 300-500 dpi also showed almost identical times. At the highest resolution (600 dpi) scanning time was more than doubled from the previous group (Figure 5). The total time of scanning all 123 pages at all tested resolutions was 150 minutes and 14 seconds. If those timings are scaled up to the process of digitisation of e.g. 100,000 pages the time needed for scanning at the resolution of 100-200 dpi would be around 160 hours, at 300-500 dpi around 310 hours while at 600 dpi it would take around 780 hours. Therefore, this information is very important for planning of digitisation projects as it greatly effects the labour cost of a scanner operator. The approximate cost of operating a book scanner 120 hours per week (three shifts) is approximately 1,440 USD and has remained more or less flat in the last few decades (Blostein and Nagy, 2012). Using the optimised scanning quality levels (300-500 dpi), we can save 470 hours per 100,000 pages – almost 4 weeks of work, or 6,000 USD. The change of resolution in the scanning software and the change of pages at the flatbed scanner was not measured but it is clear that calculating that time in as well would only add to the overall time and costs of document scanning. It is also worth mentioning that 100,000 pages is a small set. For example, the United States Department of Energy has over 300 million classified typewritten documents in its archive (Cannon *et al.*, 1999).

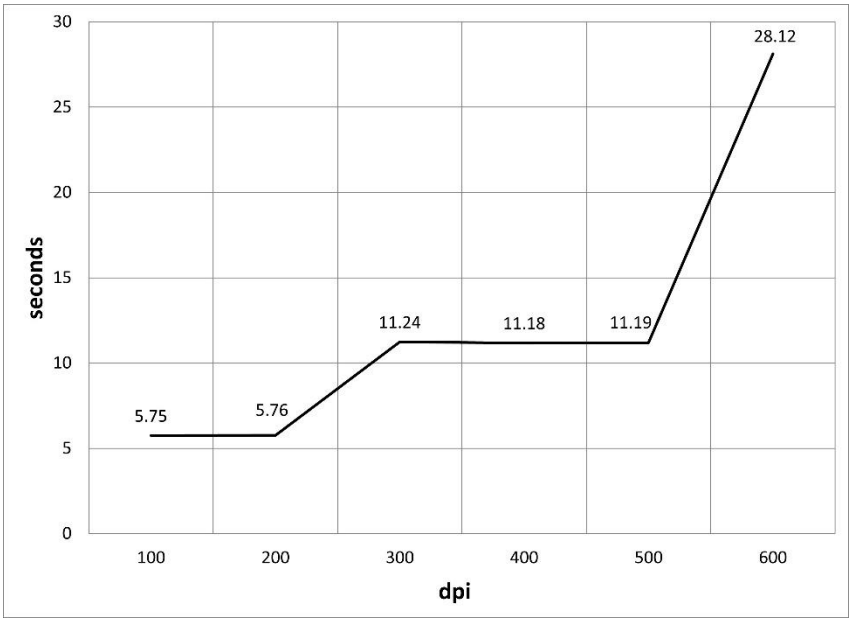

Figure 5. Average scan time

The average image size of the A4 scanned pages in uncompressed TIFF file format ranged from 0.91 MB (at 100 dpi) to 24.81 MB (at 600 dpi) (Figure 6). This information is important for several reasons. Firstly, the size of the resulting collection is significantly increasing as the scanning resolution increases – its size at the 100 dpi is 112.49 MB compared with 3.05 GB at the 600-dpi resolution. If one would be conducting a mass digitisation of e.g. 100,000 pages the difference in size would be 9.15 GB (100 dpi) vs. 248.11 GB (600 dpi). Secondly, the higher resolution the longer time is needed for conducting OCR. Thirdly, as it would be shown later, higher resolution does not automatically mean the higher OCR accuracy rate. Therefore, it would be crucial for any digitisation project to determine the optimal digitisation setup. Current storage costs are fairly low, and it could be argued that size of the resulting database does not add significantly to the cost of the digitisation project, but this information becomes much more valuable regarding long-term preservation as e.g. some of the afore mentioned U. S. Department of Energy documents could be subpoenaed in the year 3000, according to Nagy *et al.* (1999).

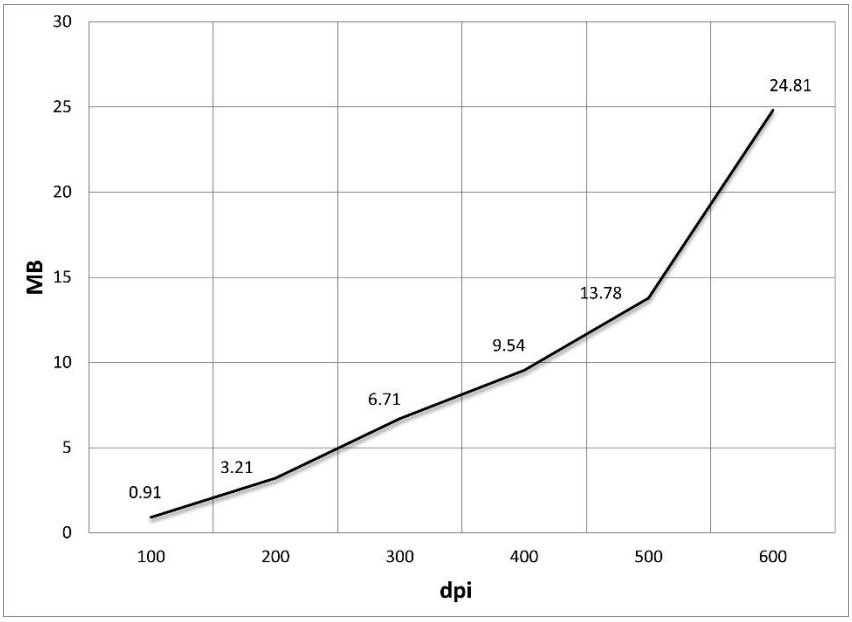

Figure 6. Average image size

The average Tesseract character accuracy results without binarisation pre-processing show that similar values were achieved using 100-400 dpi (between 88 and 91%) while the results were 5% lower at the 500 dpi and significantly lower (around 36%) for the highest (600 dpi) resolution. Word accuracy measurements follow that trend only its values are consistently lower. The reason for the significant drop of OCR quality with the increase of the resolution can be associated with the increase of details present at the higher resolution images. Since the chosen dataset was consisting of documents from 1960s, the paper contained tiny dots or artefacts scanned and present at the 600 dpi images and most likely their presence impeded the OCR process.

The results of Abbyy FineReader 15 text recognition are much higher than the ones produced by Tesseract. On average, summing up all the quality levels of non-binarised sample, both character and word accuracy levels are more than 10% higher, and more importantly the accuracy rates are consistent even at high dpi levels that Tesseract struggles with. The specific dpi accuracy rates for both OCR engines are shown in Figure 7.

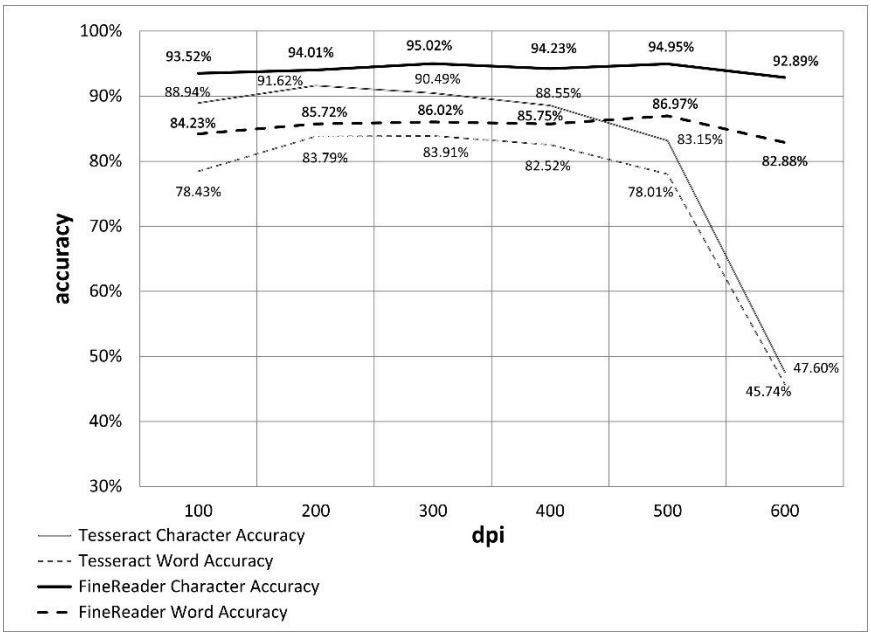

Figure 7. Average OCR accuracy rates of non-binarised sample

Binarisation yielded little success in improving optical character recognition accuracy results both for Tesseract and Abbyy. Accuracy applying 50% binarisation level is much lower across almost all image quality levels, except at the highest dpi level for character accuracy using Tesseract (Figure 8). The 60% binarisation improved the 400-600 dpi character accuracy and 600 dpi word accuracy results using Tesseract, and 600 dpi character and word accuracy using Abbyy FineReader (Figure 9). If the rest of the results of all quality levels are considered, the average accuracy levels are still lover for binarised samples and the highest scores of both word and character accuracy are achieved using the non-binarised documents.

If we compare the performance of Abbyy FineReader and Tesseract on binarised samples it is evident that both character and word accuracy results of Abbyy FineReader are higher and more consistent – similar to the pattern of non-binarised samples – and therefore establish Abbyy FineReader as a more reliable out-of-the-box OCR engine for recognizing typewritten documents.

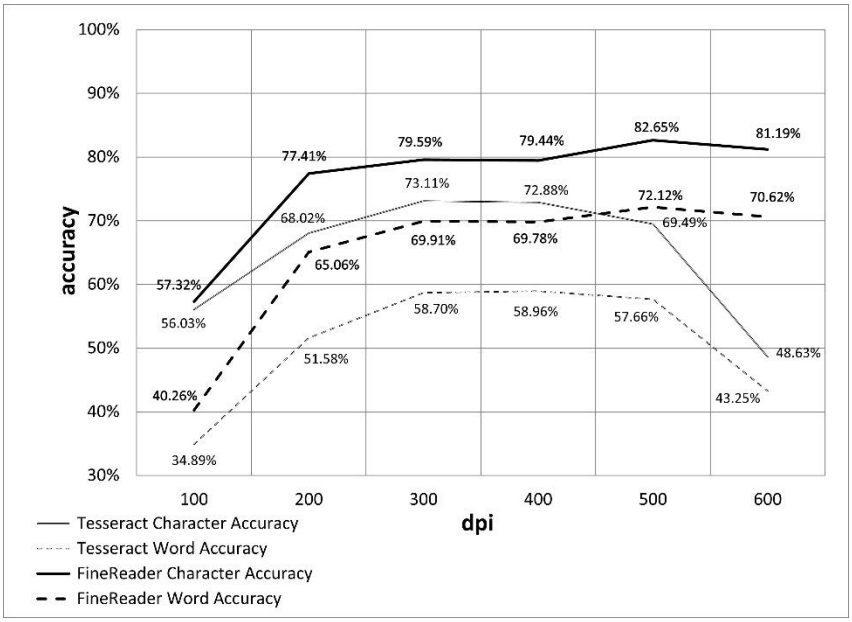

Figure 8. Average OCR accuracy rates applying 50% binarisation level

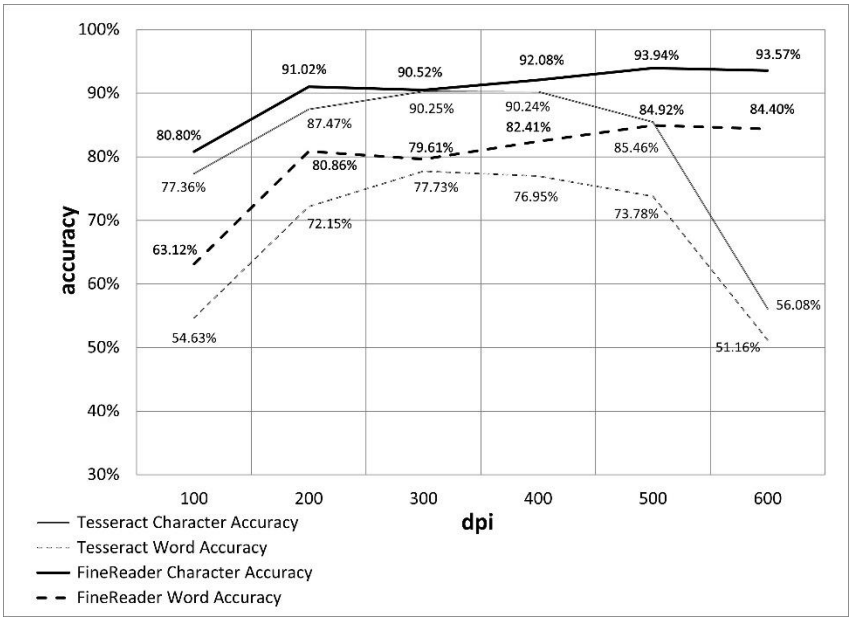

Figure 9. Average OCR accuracy rates applying 60% binarisation level

Discussion and future work

The achieved results led us to reject the hypothesis H1 since the accuracy rates did not grow with the improvement of the digitisation quality. It could be argued that the cut-off point was at the 200 dpi for Tesseract and 300 dpi for Abbyy FineReader – even a bit lower than we would expect. From that point on the accuracy level was decreasing, especially for Tesseract at the resolution higher than 500 dpi. From this we can recommend not to use higher levels of resolution because not only the OCR results will be lower, but more time would be unnecessarily used for scanning and OCR process, and more storage space would be needed thus significantly increasing the overall costs of digitisation projects.

The achieved results led us to also reject the hypothesis H2 since the pre-processing of the digitised images by the binarisation method did not improve the OCR results. Not only it did not improve it, but more time was used for the process of binarisation achieving on the average lower OCR accuracy results. The best result of the study was achieved using a non-binarised image (95.02%), and the overall recognition rate is higher on the original scans vs. the 60% binarisation. Therefore, we are advising not to apply the binarisation method as a pre-processing method in the case of the digitisation of materials like those used in this research. The only advantage of binarisation is found at the highest dpi level, where it improved the results slightly both for Abbyy FineReader and Tesseract. In the scenario of obtaining a pre-made 600 dpi typewritten scanned images collection we could recommend using 60% binarisation as a pre-processing method.

The achieved results of the preliminary testing of the binarisation led us to accept the sub-hypothesis H2.1 since they confirmed that the extremely high and extremely low binarisation produced illegible results that cannot be efficiently recognized by the software or even read by humans. The results of the binarisation threshold set at the levels of 20% or lower and 70% or higher produced white or almost white, or black or almost black images not suitable for the OCR processing.

Overall, the results of this study show that the typewritten documents from the 1960s can be successfully recognized by a modern OCR system and that the methodological approach to setting up a digitisation project could result in the optimised process considering quality, time, human effort and costs. The Croatian language did not pose any difficulties in the recognition process and it is most likely that the other languages would produce similar results using similar dataset. Even though the best achieved result in this research can only be classified as 'average' (Holley, 2008) we argue that it is worth implementing this process on a larger scale. The setup using Tesseract in this research is an example of a cost-effective OCR system, at least in terms of initial costs, that is rather straightforward to implement in various digital repositories. It could easily become a part of large-scale digitisation projects because, we believe, the benefits of OCR and the achieved results shown here are much greater than the drawback of not having any search capabilities at all. Of course, it would be advisable to make a disclaimer concerning the limitations of the OCR at the digital repository search page and the end users should be aware of this fact when conducting queries. Abbyy FineReader achieved significantly higher accuracy rates and it was more consistent at higher quality levels, but its implementation costs should be considered when planning a long-term document conversion project. Overall, by applying the OCR implementation methodology discussed here, higher accuracy results with decreased costs can be achieved in less time needing less human effort and less storage capacities. Thus, a digitisation system set up as recommended produces social (e.g. increased number of indexed documents available to researchers online) and economic (e.g. cheaper per-page document recognition rates for archives) benefits.

We decided not to include an extensive list of all the aspects that determine good OCR quality and methods of improvement (Holley, 2008, Koistinen *et al.*, 2017a, b) since they would slightly move the focus of the paper away from research of the dependence of the scanning quality and binarisation on the OCR process. The digitisation process could have also been discussed in greater detail regarding the quality control (Nagy, 2007) and future studies might consider other file formats and compression algorithms that we did not use in this research. Next, we will investigate emerging cloud-based text recognition and document management solutions (e.g. Google Cloud Vision and Amazon Textract) and analyse them comparatively with the results presented here.

Notes

1. The spelling *digitization* is also common, but we are consistently using *digitisation* according to the IMPACT project Glossary for the Mass Digitisation of Text & OCR (Anderson, 2010b).
2. The spelling *binarisation* is used following the Glossary for the Mass Digitisation of Text & OCR (Anderson, 2010b) as opposed to *binarization*.
3. The ISRI Analytic Tools for OCR Evaluation port is available for download and maintained at https://github.com/eddieantonio/ocreval
4. This free software can be downloaded at: https://www.imagemagick.org/. It is available for multiple platforms including Linux, Windows, Mac OS, Android OS and iOS.
5. FineReader 15 is the latest instalment of the OCR software produced by Abbyy. This copy was graciously donated for research purposes by DigitalMedia Ltd., Croatia.
6. Tesseract begun as a PhD project in HP Labs, Bristol, and was developed there between 1984 and 1994. Today it is an open-source software maintained by Google (Smith, 2007). Current version of Tesseract can be found at: https://github.com/tesseract-ocr/
7. Croatian Writers' Association was founded in 1900 and it is still active today.
8. Improving the quality of the output, Tesseract documentation, https://tesseract-ocr.github.io/tessdoc/ImproveQuality (accessed 30 October 2019).

References

Anderson, N. (2010a), "IMPACT best practice guide: optical character recognition - part 1", available at: http://www.impact-project.eu/uploads/media/IMPACT-ocr-bpg-pilot-s1_01.pdf (accessed 30 October 2019).

Anderson, N. (2010b), "IMPACT workflow resource: glossary for the mass digitisation of text and OCR" available at: https://www.digitisation.eu/download/website-files/WorkflowResources/GlossaryfortheMassDigitisationofText_OCR-ImpactWorkflowResource_01.pdf (accessed 30 October 2019).

Blanke, T., Bryant, M. and Hedges, M. (2012), "Ocropodium: open source OCR for small-scale historical archives", *Journal of Information Science*, Vol. 38 No. 1, pp. 76-86.

Blostein, B. and Nagy, G. (2012), "Asymptotic cost in document conversion", in Viard-Gaudin, C. and Zanibbi, R. (Ed.) *Document Recognition and Retrieval XIX: Proceedings of SPIE, January 2012*, 82970N.

Cannon, M., Hochberg, J. and Kelly, P. (1999), "QUARC: a remarkably effective method for increasing the OCR accuracy of degraded typewritten documents", in: Doermann, A. (Ed.) *1999 Symposium on Document Image Understanding Technology, April 1999, Annapolis, United States*, University of Maryland, pp. 154-158

Cojocaru, S., Colesnicov, A., Malahov, L. and Bumbu, T. (2016), "Optical character recognition applied to Romanian printed texts of the 18th-20th century", *Computer Science Journal of Moldova*, Vol. 24 No. 1(70), pp. 106-117.

Holley, R. (2008), "How good can it get? Analysing and improving OCR accuracy in large scale historic newspaper digitisation programs", *D-Lib Magazine*, Vol. 15 No. 3/4, pp. 1-13.

Hubert, I., Arppe A., Lachler, J. and Santos, E.A. (2016), "Training and quality assessment of an optical character recognition model for Northern Haida", *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris*, European Language Resources Association (ELRA), pp. 3227–3234., available at:

http://www.lrec-conf.org/proceedings/lrec2016/pdf/39_Paper.pdf (accessed 11 February 2020).

Koistinen, M., Kettunen, K. and Kervinen, J. (2017a), "How to improve optical character recognition of historical Finnish newspapers using open source Tesseract OCR engine" in: *8th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, November 2017*, Springer-Verlag, Berlin-Heidelberg, pp. 279-283.

Koistinen, M., Kettunen, K. and Pääkkönen, T. (2017b), "Improving optical character recognition of Finnish historical newspapers with a combination of Fraktur and Antiqua models and image preprocessing" in: Tiedeman, J. and Tahmasebi, N. (Ed.) *Proceedings of the 21st Nordic Conference of Computational Linguistics, May 2017, Gothenburg, Sweden*, Linköping University Electronic Press, pp. 23-24.

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto, S., Déjean, H., Diem, M., Fiel, S., Gatos, B., Greinoecker, A., Grüning, T., Hackl, G., Haukkovaara, V., Heyer, G., Hirvonen, L., Hodel, T., Jokinen, M., Kahle, P., Kallio, M., Kaplan, F., Kleber, F., Labahn, R., Lang, E.M., Laube, S., Leifert, G., Louloudis, G., McNicholl, R., Meunier, J.-L., Michael, J., Mühlbauer, E., Philipp, N., Pratikakis, I., Puigcerver Pérez, J., Putz, H., Retsinas, G., Romero, V., Sablatnig, R., Sánchez, J.A., Schofield, P., Sfikas, G., Sieber, C., Stamatopoulos, N., Strauß, T., Terbul, T., Toselli, A.H., Ulreich, B., Villegas, M., Vidal, E., Walcher, J., Weidemann, M., Wurster, H. and Zagoris, K. (2019), "Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study", *Journal of Documentation*, Vol. 75 No. 5, pp. 954-976.

Nagy, G., Nartker, T.A. and Rice, S.V. (1999), "Optical character recognition: an illustrated guide to the frontier", in: Lopresti, D. P. and Zhou J. (Ed.), Document Recognition and Retrieval VII, December 1999, San Jose, United States, SPIE - The International Society for Optical Engineering, Vol. 3967, pp. 58-69.

Nagy, G. (2007), "Digitizing, coding, annotating, disseminating, and preserving documents", in: Majumder, P., Mitra, M. and Parui, S.K. (Ed.), *Proceedings of the 2006 international workshop on Research issues in digital libraries, Kolkota, India, 2006*, ACM, New York.

Rice, S.V. and Nartker, T.A. (1996), "The ISRI Analytic Tools for OCR Evaluation Version 5.1", Technical Report 96-02, Information Science Research Institute, University of Nevada, Las Vegas.

Shank, R.C. (1991), Where's the AI?, *AI Magazine*, Vol. 12 No. 4, pp. 38-49.

Smith, R. (2007), "An overview of the Tesseract OCR engine", in: *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02, ICDAR '07, 23-26 September 2007, Curitiba, Brazil*, IEEE, Washington, pp. 629-633.

Smitha, M.L., Antony, P.J. and Sachin D.N. (2016), "Document image analysis using ImageMagick and Tesseract-ocr", *International Advanced Research Journal in Science, Engineering and Technology*, Vol. 3 No. 5, pp. 108-112.

Strange, C., McNamara, D., Wodak, J. and Wood, I. (2014), "Mining for the meanings of a murder: the impact of OCR quality on the use of digitized historical newspapers", *DHQ: Digital Humanities Quarterly*, Vol. 8 No. 1.

Traub, M.C., Van Ossenbruggen, J. and Hardman, L. (2015), "Impact analysis of OCR quality on research tasks in digital archives", in: Kapidakis, S., Mazurek, C. and Werla, M. (Ed.), *Research and Advanced Technology for Digital Libraries, 19th International Conference on*

Theory and Practice of Digital Libraries, TPDL, 14-18 September 2015, Poznań, Poland, Springer-Verlag, Berlin-Heidelberg, pp. 252-263.

Tweedie, M. (2018), "6 technologies behind AI", available at: https://codebots.com/ai-powered-bots/6-technologies-behind-ai (accessed 30 October 2019).

Corresponding author

Željko Trbušić can be contacted at: ztrbusic@gmail.com