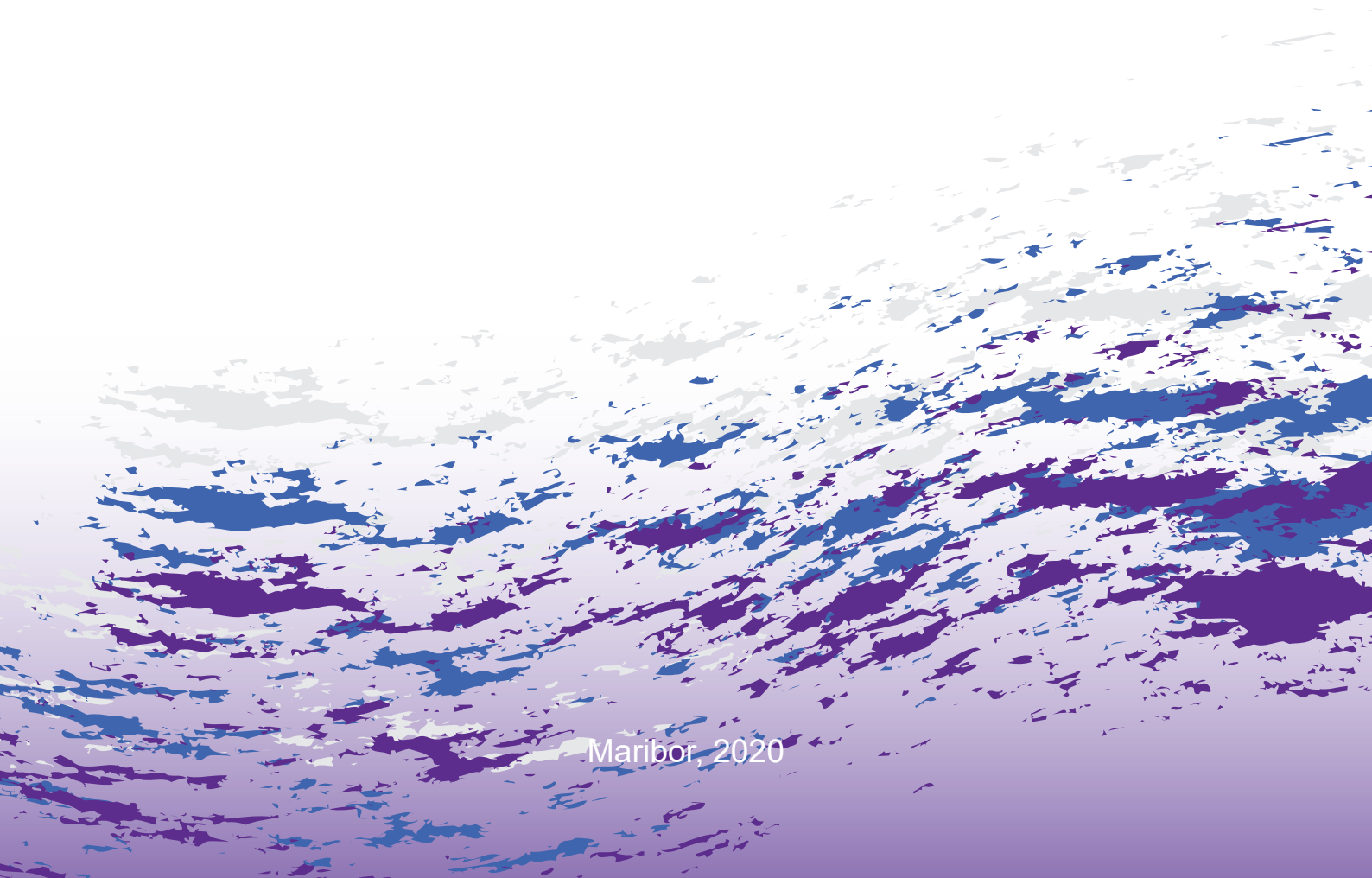
PAM Pokrajinski arhiv Maribor

Moderna arhivistika

Časopis arhivske teorije in prakse
Journal of Archival Theory and Practice

Letnik 3 (2020), št. 1 / Year 3 (2020), No. 1



Maribor, 2020

Pokrajinski arhiv Maribor

Moderna arhivistika

Časopis arhivske teorije in prakse
Journal of Archival Theory and Practice

Letnik 3 (2020), št. 1 / Year 3 (2020), No. 1

Maribor, 2020

Moderna arhivistika

Časopis arhivske teorije in prakse

Journal of Archival Theory and Practice

Letnik 3 (2020), št. 1 / Year 3 (2020), No. 1

ISSN 2591-0884 (online)
ISSN 2591-0876 (CD_ROM)

Izdaja / Published by:
Pokrajinski arhiv Maribor / Regional Archives Maribor

Glavni in odgovorni urednik / Chief and Responsible editor:
Ivan Fras, prof., Pokrajinski arhiv Maribor, Glavni trg 7, SI-2000 Maribor,
telefon/ Phone: +386 2228 5017; e-pošta/e-mail: ivan.fras@pokarh-mb.si

Glavna urednica / Editor in chief:
mag. Nina Gostenčnik

Uredniški odbor / editorial board:

- dr. Thomas Aigner, Diözesanarchiv St. Pölten, Avstrija
- dr. Borut Batagelj, Zgodovinski arhiv Celje, Slovenija
- dr. Bojan Cvelfar, Arhiv Republike Slovenije, Slovenija
- mag. Nada Čibej, Pokrajinski arhiv Koper, Slovenija
- Ivan Fras, Pokrajinski arhiv Maribor, Slovenija
- mag. Nina Gostenčnik, Pokrajinski arhiv Maribor, Slovenija
- dr. Joachim Kemper, Institut für Stadtgeschichte Frankfurt am Main, Nemčija
- Leopold Mikec Avberšek, Pokrajinski arhiv Maribor, Slovenija
- dr. Miroslav Novak, Pokrajinski arhiv Maribor, Slovenija
- dr. Rik Opsommer, Stadsarchief Ieper - Universiteit Gent, Belgija
- Darko Rubčić, Državni arhiv u Zagrebu, Hrvaška
- dr. Izet Šabotić, Filozofski fakultet Univerziteta u Tuzli, Bosna in Hercegovina
- mag. Boštjan Zajšek, Pokrajinski arhiv Maribor, Slovenija

Recenziranje / Peer review process:
Prispevki so recenzirani. Za objavo je potrebna pozitivna recenzija. Proces recenziranja je anonimen. / All articles for publication in the conference proceedings are peer-reviewed. A positive review is needed for publication. The review process is anonymous.

Lektoriranje / Proof-reading:
mag. Boštjan Zajšek, mag. Nina Gostenčnik

Prevajanje:
mag. Boštjan Zajšek (slovenščina), mag. Nina Gostenčnik (slovenščina, angleščina)

Oblikovanje in prelom / Design and typesetting:
mag. Nina Gostenčnik

Prejeto / Received: 01. 07. 2020

1.01 Izvirni znanstveni članek *1.01 Scientific Article*

EVALUATING AND IMPROVING OCR EFFICIENCY

Hrvoje Stančić, Ph. D.
Department of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
hstancic@ffzg.hr

Željko Trbušić
Institute for the History of Croatian Literature, Theatre and Music,
Croatian Academy of Sciences and Arts, Zagreb, Croatia
ztrbusic@hazu.hr

Abstract:

The purpose of this research is to establish a method for OCR quality evaluation in different archival situations and stages of document ingest process. The conducted experiments explain the importance of OCR optimization using the example of Croatian typewritten materials. Furthermore, the paper discusses the importance of unique, or distinct, words in document retrieval.

Key words:

digitization, OCR quality, ISRI Tools, archival information systems, ingest

Izvleček:

Ocena in izboljšanje učinkovitosti optičnega prepoznavanja znakov (OCR)

Namen raziskave je vzpostaviti metodo preverjanja kvalitete OCR v različnih situacijah in fazah prevzemanja arhivskega gradiva. Opravljeni preizkusi pojasnjujejo pomembnost optimizacije OCR na primeru hrvaških tipkanih dokumentov. Prispevek obravnava tudi pomembnost enkratnih, ali jasnih, besed pri iskanju dokumentov.

Ključne besede:

digitalizacija, kvaliteta OCR, orodja ISRI, arhivski informacijski sistemi, zajem

1. INTRODUCTION

Archives deal with large amounts of data and one of the primary concerns of archivists is how to control this data and re-use it efficiently. Digitization enables archivists to expedite this process, but it is very important to apply the correct methods – those appropriate to the type of archival materials and the intended outcomes of the digitization project. Recent advancements in the field of archival science are heavily focused on the implementation and use of new technologies and methods of (big) data processing that can be beneficial in speeding up the process of archival ingest, description and dissemination. This article focuses on a part of this process and examines how optical character recognition (OCR) technology helps the automatization of document description with minimal or no human interaction. The goal is to propose the method that can be applied to a specific archival material in the ingest process of a large-scale archival information system resulting in producing more precise search

results. It will also be shown how this method could be adapted to various archival processes using different technological solutions.

Optical character recognition is a well-known technology which enables archives to convert scanned documents into machine-readable format. The resulting optically recognized documents can subsequently be treated in a similar way as if they were digitally born and therefore various additional procedures can be applied (e.g. text mining and big data analytics). In order to get the best possible results and to preserve document authenticity it is important to measure accuracy of the recognized text at various stages of digitization process. The steps in the process of converting documents into machine-readable digital format using OCR are adopted, with small changes, from Cojocaru et al. (2016, p. 109) and are: (1) Image capture and pre-processing, (2) OCR, (3) Text post-processing and (4) Quality evaluation.

Image capture and pre-processing stage deals with acquiring the appropriate digital reproductions and editing of the documents that are going to be recognized. For the image capture part, it is possible to use optical scanners and digital cameras, with various results, and for the pre-processing part there are multiple commercial and open-source tools available today. Common belief is that the higher the resolution the better the recognition rate of documents that are optically recognized (hypothesis 1, H1). This research shows that the proper testing methods need to be applied and that they can provide archives with better end-results in terms of recognition rate, cost of the OCR implementation and long-term sustainability. The pre-processing techniques mostly used when dealing with OCR are image binarization and deskewing with the addition of noise reduction and border removal.

The second stage in the procedure deals with choosing the appropriate character recognition engine. There are various solutions available today and they fall into two categories: commercial and open source. If the decision is made to choose one option over the other, it should be supported by higher accuracy levels and improved long-term sustainability in terms of cost and time-management. Commercial software should offer higher accuracy (hypothesis 2, H2) and better customer support, it should be easier and faster to implement and maintain, but its cost on the long run could be substantial (e.g. when dealing with millions of documents in a situation where the price is determined by the amount of data processed). Open source software is free to use and modify, but it requires advanced technical knowledge for successful implementation and offers no costumer support whatsoever. It is, therefore, crucial to test different solutions for their performance on archival materials and make a long-term sustainability projection. It is of no use to implement a costly software that one can apply to a small part of the archives' holdings (due to per-page cost of document processing) or to implement an open source engine that will give much lower accuracy. This research examines text recognition as a standard archival procedure that can and should be incorporated at all levels and not just in a small digitization project.

The third stage in the OCR procedure, text post-processing, uses various methods of text clean-up, for example automatic dictionary supported text validation or a more disruptive method like crowdsourced campaigns. This stage is not in the focus of this research, but it can be equally important for the large-scale OCR implementation in archives.

The fourth, quality evaluation stage is concerned with choosing the appropriate method for measuring and evaluating the effectiveness of the OCR workflow in place. Those procedures were in focus of this research. Therefore, The ISRI Analytic Tools for OCR Evaluation[1] (The ISRI Tools), i.e. the ported version (Santos, 2014) maintained at its corresponding GitHub page[2], were used as a tools to quantify accuracy of the recognized text. The suite consists of 17 programs that deliver statistical data and can serve as a standardized measurement tool. Despite their age, these tools were the perfect match for experiments and evaluation as they (1) support modern operating system environment and language of the texts used, (2) are easy to use and openly available, and (3) deliver statistical data that can't be matched by any other tool in existence (at least to our knowledge and as well as according to Hubert et al., 2016). In the research, a character accuracy measurement program *accuracy* and a word accuracy measurement program *wordacc* were used. Both methods use a ground truth (GT) file, which is a 100% correctly transcribed text file, and compares it to the recognized text delivering multiple statistical data. The character accuracy report is computed by counting the necessary deletions, substitutions and insertions (Levenshtein distance[3]) of characters needed to correct the recognized text (Rice and Nartker, 1996, p. 4) while the word accuracy report is generated by counting the misrecognized words. If a stopword file is used, the *wordacc* program will also be able to count the stopwords, non-stopwords and distinct non-stopwords accuracy that can be very beneficial in the assessment of keyword search capabilities of the recognized text (Figure 1). Stopwords are described as words that contain less data relevant for searching (Lujo, 2010, p. 40) while on the other hand distinct non-stopwords are characterized as relevant words whose occurrence in the text is scarce (Rice, Jenkins and Nartker, 1995, p. 12). Both concepts are important to understand when dealing with the projection of search capabilities of recognized text. For example, recognition accuracy of the Croatian word *ovo* (*this*) does not affect document retrieval greatly and can be marked as a stopword, while the accuracy of the non-stopword *sastanak* (*meeting*) must be 100% if one wants to retrieve all documents referring to the held meetings (of a board of directors, perhaps). If the latter word would appear only once in a document or a set of documents, it can also be marked as a distinct non-stopword. The example shown in Figure 1 lists a total of 301 words in a one-page sample (line 3) from which there are 137 distinct non-stopwords occurring only once in a page with 94.16% accuracy rate (line 38). The selected page will not be retrieved if one of the 8 misrecognized words is searched for, resulting in an incomplete list of documents. Therefore, the distinct word accuracy performance measure is a counter argument to the statement that full-text searchable documents are resilient to OCR errors because of the redundancy of text (Rice, Jenkins and Nartker, 1995, p. 12). A positive example is found in the words that occur three times on a given page: six different words are found, and the accuracy level is 100% (line 40) which means that every word is correctly recognized at least once. Also, 10 words are found twice on a given page, but one is missed (line 39). This means that even if one of those words is entered in the search engine, the example document would be retrieved even though one occurrence of a word was not correctly recognized.

---

[1] *The ISRI Tools for OCR Evaluation were created at the Information Science Research Institute, University of Nevada, Las Vegas in the early 90's. They conducted testing of then available OCR systems from 1992 to 1996 and open-sourced the tools in 2005.*

[2] *Available at https://github.com/eddieantonio/ocreval.*

[3] *After the Russian mathematician Vladimir Levenshtein.*

```
UNLV-ISRI OCR Word Accuracy Report Version 6.1
----------------------------------------------

     301   Words
      17   Misrecognized
  94.35%   Accuracy

Stopwords
   Count   Missed   %Right   Length
      31        1    96.77        1
      41        2    95.12        2
      10        0   100.00        3
      22        2    90.91        4
      11        0   100.00        5
       2        0   100.00        7
       1        0   100.00        8
     118        5    95.76    Total

Non-stopwords
   Count   Missed   %Right   Length
       9        4    55.56        1
       3        2    33.33        2
       7        1    85.71        3
       8        0   100.00        4
      23        0   100.00        5
      17        0   100.00        6
      14        0   100.00        7
      22        2    90.91        8
      37        0   100.00        9
      20        2    90.00       10
      12        0   100.00       11
       8        1    87.50       12
       1        0   100.00       13
       2        0   100.00       14
     183       12    93.44    Total

Distinct Non-stopwords
   Count   Missed   %Right   Occurs
     137        8    94.16        1
      10        1    90.00        2
       6        0   100.00        3
       1        0   100.00        8
     154        9    94.16    Total
```

Figure 1. The ISRI Tools word accuracy report (one-page sample) showing the stopwords, non-stopwords and distinct non-stopwords categories.

2. METHODOLOGY

The research is focused on performance of two different OCR engines used with a set of typewritten documents from the mid-20th century scanned at six different quality levels. The testing method consists of 4 steps: (1) Dataset extraction, (2) Ground truth (GT) and stopwords preparation, (3) Accuracy measurement method selection, and (4) Results comparison.

2.1 Dataset

Testing was conducted on a dataset of 123 typewritten pages from the Croatian Writers' Society (hrv. *Društvo hrvatskih književnika, DHK*) archival holdings recovered from the Department for the History of Croatian Literature at the Croatian Academy of Sciences and Arts. The DHK archive spans the time period from 1900 (the foundation of the society) up until 1971 and consists mainly of handwritten and typewritten materials. The typewritten documents occur more prominently after 1945 and in time completely replace handwriting as a main device for the creation of formal documents and correspondence. This reflects the situation that occurs in almost every part of life (personal and formal) in that time period which opens up this research to the vast collection of typewritten documents found in the archives worldwide. The tested documents were extracted from the collection of transcripts of board meetings and plenary sessions held during the period from 1966 to 1968. The pages were scanned at six different quality levels (100-600 dots per inch (dpi) with the increment of 100 dpi) in order to test the effect of resolution on the accuracy. During this phase great care was taken that the documents were not removed or tampered with in any way while scanning at different quality levels was performed, assuring the credibility of the results.

The second variable introduced in data collection phase was using two different OCR engines – Tesseract 4 and Abbyy FineReader 15 (release 4) – for the recognition process. Both are modern OCR systems that are used in multiple environments (culture, science, banking etc.) and can successfully recognize text in many different languages (including Croatian, the main language of our documents). Tesseract is an open source engine freely available for download and run form the Linux terminal[4] (different GUI applications are also available). The default OCR settings were used in combination with the official Croatian language trained data files (needed to recognize language-specific symbols). Abbyy FineReader[5] is a proprietary software that has been on the market for more than 20 years and its use is widespread in the industry. The latest version offers multiple functionalities of a complete document management system, but this research explored only the OCR part of the software, i.e. its accuracy. The default OCR settings were used, apart from the language which was set manually for all the pages that were processed. The output for both engines was set to plain .txt files with line separation.

2.2 GT and stopwords preparation

The ground truth (GT) was prepared according to The ISRI Tools guidelines. The tools enable marking characters in two different ways: using the tilde (~) and a circumflex (^). Tilde serves as a wildcard to mark characters the OCR engine is not meant to recognize and that should not be included in the final accuracy score. Circumflex is used as a suspect marker for characters that are hard or near impossible to recognize. During

[4] *Tesseract is maintained and can be downloaded at: https://github.com/tesseract-ocr/.*

[5] *A copy of FineReader 15 was donated by DigitalMedia Ltd., Croatia for the purpose of this research.*

the preparation of GT for this research, a tilde was used more often, and it served to mark the characters that were an overtype (overlapping of two different letters) producing an illegible symbol. Overall, the created ground truth should represent the original text as much as possible, even the mistakes made by the typist or unusual symbols used and the line/paragraph separation should be true to the original document. The tools do not include blank lines or added white space in the analysis, but the newline information is computed and shown as <\n> in the accuracy report.

The possibility of using a dictionary as a ground truth (so that time consuming manual transcription is avoided) is considered, but the limitation is too severe. Words can be marked positively by the system if the generated text value is contained in the dictionary, but the dictionary word may not necessarily correspond to the one in the text.

A stopword file that was prepared included 1,198 unique Croatian words extracted from the list created by Lujo (2010) that initially contained many duplicate words (due to the collecting method and the specific Croatian grammar) which were identified and removed. Even though different Croatian stopword lists are available, Lujo's work was used since its idea corresponded well with the purpose of this research which is the automatization of text recognition process.[6]

2.3 Accuracy measurement method

The ISRI Tools' two main accuracy measurement programs, character accuracy and word accuracy, were applied to all samples and the accuracy percentages were extracted[7]. This research covers those results, but the focus is moved to the distinct non-stopword accuracy levels. This measurement was identified as an important part of the investigation of retrieval capabilities of recognized text. To improve the methodology and to add additional evaluation possibilities, a precision and recall system (P/R) is considered. Distinct non-stopwords are presented as the 'relevant data' while everything else (stopwords and non-stopwords) is considered as the 'total data'. Even though this concept can be argued further (important information is also found in the non-distinct words), this research tries to present the distinct non-stopwords as the data that cannot be replaced and should not be lost in the document recognition process, while the stopwords and non-stopwords are redundant and there is a high possibility that they will be retrieved when the rest of the document or similar documents are included in the search process. Precision is calculated by dividing the correctly recognized distinct non-stopwords by the total number of correctly recognized words in the sample data. Recall is calculated by dividing the correctly recognized distinct non-stopwords with the total number of distinct non-stopwords in the documents. In this way the distinct non-stopwords are examined through the P/R evaluation principle (Figure 2).

$$Precision = \frac{\text{correct distinct non-stopwords}}{\text{correct total words}}\ x\ 100\%$$

$$Recall = \frac{\text{correct distinct non-stopwords}}{\text{total distinct non-stopwords}}\ x\ 100\%$$

Figure 2. Precision and recall system adapted for distinct non-stopword measurement.

[6] *The engine and the full list can be found at https://bitbucket.org/trebor74hr/text-hr/src/default/.*

[7] *The ISRI Tools express the accuracy in percentages of correctly recognized text, in recent literature values CER (character error rate) and WER (word error rate) can also be found.*

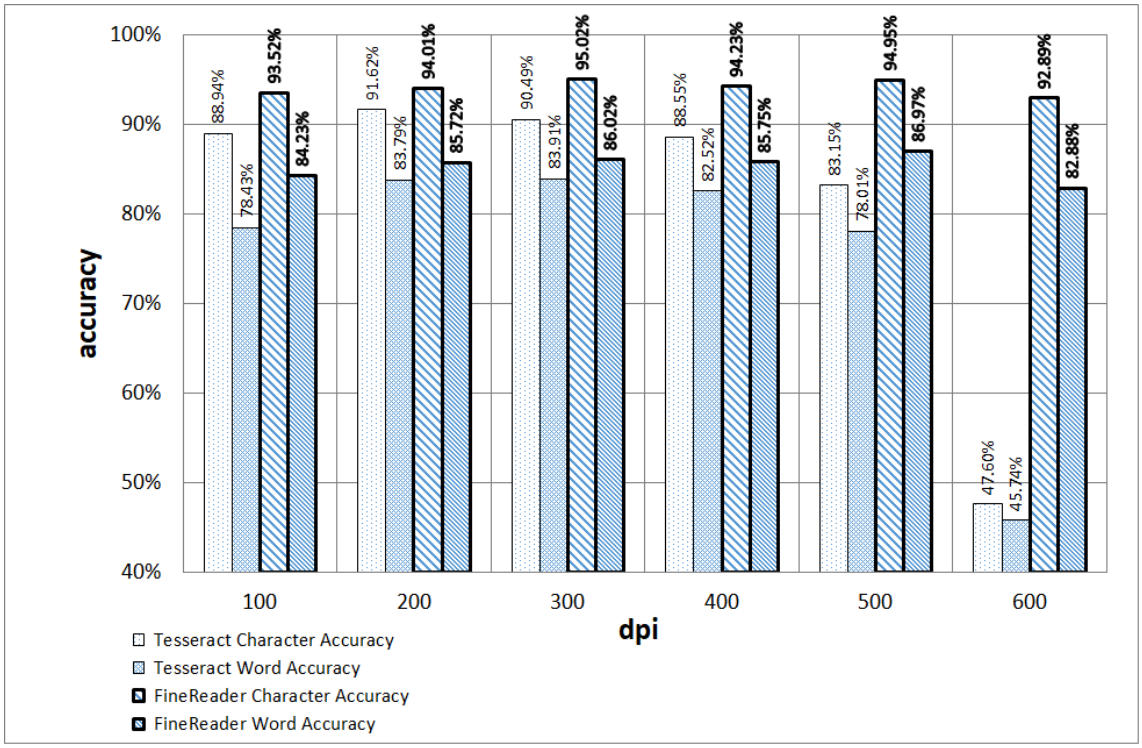
2.4 Results comparison

The last part of the assessment is the results collection and comparison. This step involves dividing the results into categories (in this case into different quality levels and engines) that can subsequently be analyzed, compared and that conclusions can be drawn. The categories should reflect the variables used in the text recognition process and should lead to new knowledge regarding the OCR recognition on a specific type of documents, making the process more efficient.

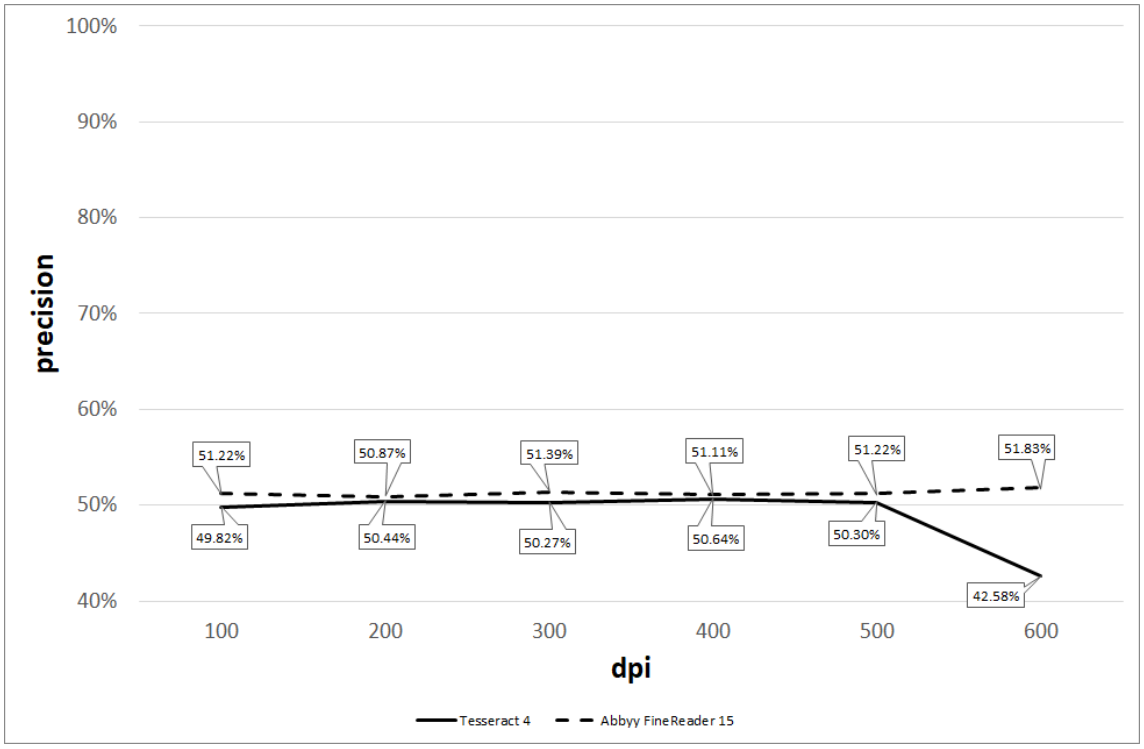
3. RESEARCH RESULTS

On average, Abbyy FineReader performed much better than Tesseract and scored higher on every quality level (Graph 1), confirming H2. The accuracy results of FineReader are more consistent and it has achieved higher scores in both word and character accuracy levels. It can be observed that even though Tesseract achieves a solid performance, its accuracy plummets at the 600-dpi quality level. It is possible that at higher quality levels the scans contain more noise and it confuses the engine. FineReader had no such problems, but the average accuracy is still lower on 600-dpi than it is on all other quality levels. This disproves the statement made in H1. Testing at different quality levels show that both Tesseract and FineReader achieve best results at resolutions between 200 and 500 dpi. In the case of FineReader, there is not a significant difference in the accuracy levels between those quality levels (<2%), but the results obtained using Tesseract have a higher deviation rate.

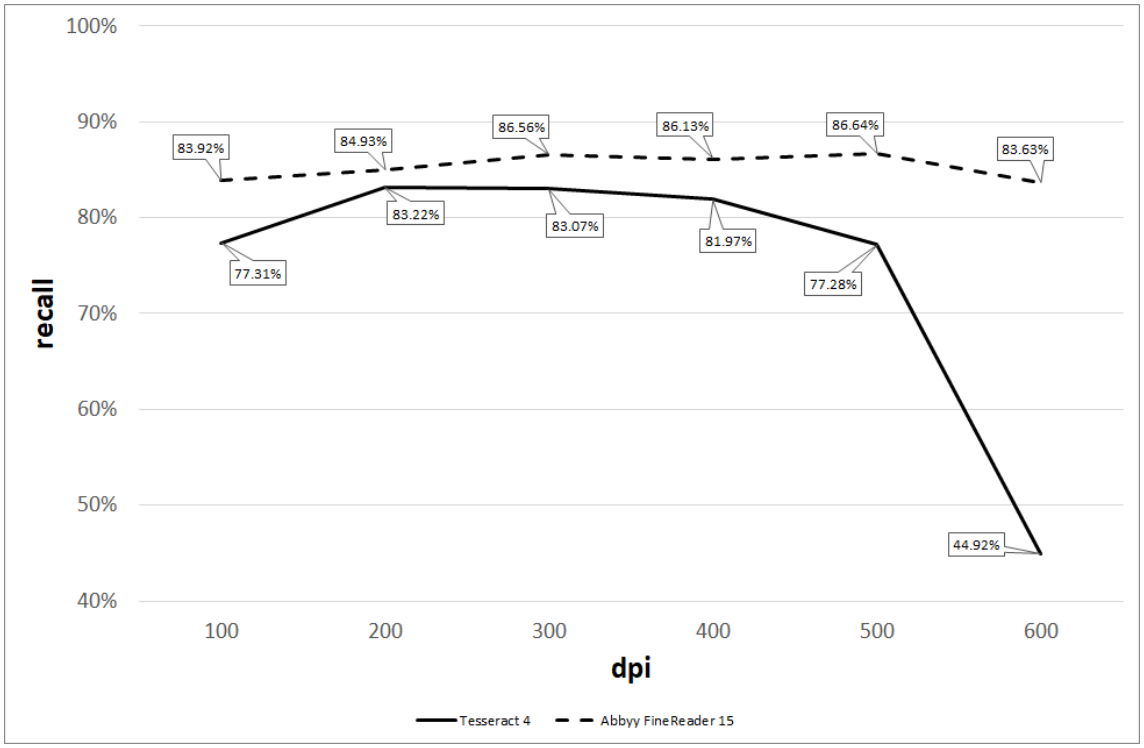
Graph 1. Accuracy of Tesseract 4 and Abbyy FineReader 15 on a 123-page sample of typewritten documents measured at six different quality levels.



Graph 2. Precision of distinct non-stopwords shown across different quality levels and OCR engines.



Graph 3. Recall of distinct non-stopwords shown across different quality levels and OCR engines.



The average presence of distinct non-stopwords in tested documents is 51.40% (on a page level). That means that every other word on a page (including stopwords) is a distinct word and falls into the category of relevant unique data. Precision of distinct non-stopwords is shown in Graph 2. Both engines perform similarly in this respect yielding results close to 50%, except for the 600-dpi level of the recognition made by Tesseract is significantly lower. Recall is shown in Graph 3 and the results are much higher and more consistent for Abbyy FineReader than for Tesseract, reflecting the total word accuracy results shown in Graph 1.

4. DISCUSSION

The fact that proprietary OCR engine is shown to be more accurate can justify the initial cost of the software and the potential costs of the long-term sustainability. The open source engine falls behind in the accuracy measurements, but its usefulness should not be neglected. It enables archives to implement small-scale digitization text recognition projects effortlessly and without any initial costs. It can serve as a potential solution in the situation where OCR was not used as a method of obtaining data during document ingest and as a testing tool for a larger implementation.

The accuracy measurements serve as a guidance in the creation of a sustainable text recognition workflow since they are arguably the most important part of the process, especially if data preservation is considered. It is important to annotate the resulting documents with accuracy measurement values and those annotations should be available to users with clear explanations of their meanings.

The recall of distinct non-stopwords follow similar pattern of average total word accuracy and, as a measurement device, it can be more important than the total word accuracy. The errors produced by the OCR in a high accuracy situation can easily be neglected if a recall of distinct non-stopwords of 100% is achieved. A 99%-word accuracy means there is 1 error in every 100 words and if the average total word count per page in this research is 224, that would, in a hypothetical situation, produce 2 incorrect words per page. If both of those words are categorized as distinct non-stopwords, the search queries would not be able to identify the document based on the search query that includes those two words. In contrast, if the 2 incorrect words are not distinct, the search can still be successful because there are other exact references on the same page. In conclusion, both total word accuracy and recall of distinct non-stopwords should be considered when calculating the effectiveness of the OCR system.

The precision measure of the distinct non-stopwords acquired in this research show that there is an inherit flaw in the application of this performance measure in the context of text accuracy. The problem is that the precision does not apply to misrecognized words. Both engines have a similar precision which does not correspond well to the overall better results of FineReader shown in the other testing protocols (character accuracy, word accuracy, and recall).

5. CONCLUSION

The focus of this research was oriented more on the possible solutions for acquiring the best OCR transcriptions in an automated way, i.e. building a process which involves less human interaction and produces more quality results. The primary concern of building a sustainable OCR system in archives is to ensure that the information produced by the system is authentic and complete. This is hard to achieve since, without manual correction, the resulting recognized text will almost certainly contain errors and

information will be missing or incorrect. The possible solution is to annotate the collection of documents with predicted accuracy results as presented in this paper so that the users are aware of the shortcomings but can still benefit from the full-text search capabilities and automatically extracted metadata.

It is not expected that archives will have the manpower or the funds to manually transcribe millions of pages or even only correct the mistakes generated by the OCR system. On the other hand, digitized textual documents that contain encoded text are much more useful and versatile, e.g. for big data analytics. Using the proposed method of calculating accuracy, precision and recall taking into account (distinct) (non-)stopwords, i.e. incorporating usage of the tools such as The ISRI Tools in the digitization process, the archives can increase quality of the OCR achieving the best possible transcription results in the most cost-effective way. Substantial savings can be achieved in a large-scale implementation using lower dpi levels because less time is used for the image acquiring process (where human interaction is almost mandatory) with no data loss. The average scan time at 300-500 dpi levels for an A4 document used in this research was around 11 seconds and at 600 dpi it was increased to 28 seconds[8]. It is not advised to scan the materials for OCR with the highest setting because it would more than double the processing time and achieve lower accuracy results. On a scale of 1 million pages it would increase the duration of the work for almost 200 days (if the scanning was performed at three 8-hour shifts). Even though the increase in the cost of computing resources is almost insignificant today (according to Blostein and Nagy, 2012), it is important to mention that the file size almost doubles at 600-dpi compared to the 500-dpi quality level.

6. FUTURE RESEARCH

Future research will broaden the comparative analysis by addition of AI-based text recognition engines (Google Cloud Vision, Amazon Textract) and by involving other measuring devices, such as F-measure.

SOURCES AND LITERATURE

Blostein, D. and Nagy, G. (2012). Asymptotic cost in document conversion. In Viard-Gaudin, C. and Zanibbi, R. (Eds.) *Document Recognition and Retrieval XIX: Proceedings of SPIE*. Washington: SPIE, p. 82970N.

Cojocaru, S. et al. (2016). Optical Character Recognition Applied to Romanian Printed Texts of the 18th-20th Century. *Computer Science Journal of Moldova*, 24(1), pp. 106-117.

Hubert, I. *et al.* (2016). Training & Quality Assessment of an Optical Character Recognition Model for Northern Haida. In Calzolari, N. et al. (Eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Paris: European Language Resources Association (ELRA), pp. 3227-3234.

Lujo, R. (2010). *Lociranje sličnih logičkih cjelina u tekstualnim dokumentima na hrvatskome jeziku*. MA work. University of Zagreb, Faculty of Electrical Engineering and Computing.

Rice, S. V., Jenkins, F. R. and Nartker, T. A. (1995). *The Fourth Annual Test of OCR Accuracy*. Techical Report 95-04. Information Science Research Institute, University of Nevada, Las Vegas.

[8] *A standard desktop flatbed scanner was used but comparable relative timings should be achieved using more professional scanners, too.*

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.

Santos, E. A. (2019). OCR evaluation tools for the 21 st century. In Arppee, A. et al. (Ed.s) *Proceedings of the 3rd Workshop on Computational Methods for Endangered Languages Papers*, Vol 1. ACL, pp. 23-27.

POVZETEK

OCENA IN IZBOLJŠANJE UČINKOVITOSTI OPTIČNEGA PREPOZNAVANJA ZNAKOV (OCR)

Dr. Hrvoje Stančić
Filozofska fakulteta Univerze v Zagrebu, Hrvaška
hstancic@ffzg.hr

Željko Trbušić
Hrvaška akademija znanosti in umetnosti, Zagreb, Hrvaška
ztrbusic@hazu.hr

Tehnologija za optično prepoznavo znakov (OCR) je na trgu že nekaj desetletij in se prvenstveno uporablja za konverzijo skeniranih dokumentov v formate, ki jih je mogoče urejati in ki jih računalnik prepozna. OCR je običajno implementiran v arhivske informacijske sisteme kot del procesa zajema, kjer je mogoče ustvariti avtomatičen potek dela za izvzemanje informacij iz prejetih dokumentov z minimalnim človeškim vložkom ali celo brez njega. To pohitri proces konverzije analognih dokumentov v digitalno obliko, polega tega pa omogoča iskanje po celotnem tekstu in ustvarjanje metapodatkov.

Proces digitizacije, ki vsebuje postopek OCR, sestavljajo štiri faze: 1. digitizacija in predhodna obdelava; 2. OCR; 3. popravki; 4. evalvacija. Vsaka faza vsebuje elemente, katerih učinkovitost je mogoče testirati za pridobitev najboljših rezultatov OCR. Testiranje je razdeljeno na dva dela: testiranje pravilnosti znakov in testiranje pravilnosti besed. V ta namen so uporabljena analitična orodja ISRI za evalvacijo OCR. Pravilnost znakov je izračunana s štetjem potrebnih vstavljenih, dodanih in izbrisanih znakov v OCR-besedilu ob primerjavi z izbirnim tekstom. Pravilnost besed je izpeljana iz števila besed, ki so bile pravilno prepoznane v primerjavi s skupnim številom besed v dokumentu. Obe vrednosti sta predstavljeni v odstotkih.

Rezultati testiranja so bili pripravljeni na vzorcu arhivskega gradiva – 123 strani tipkanega dokumenta iz arhivskega fonda Hrvaške akademije znanosti in umetnosti, ki je bil skeniran na šestih različnih nivojih kakovosti (100–600 dpi) in na dveh različnih OCR-pogonih – odprtokodnem Tesseract, verzija 4.0, in komercialnem Abbyy FineReader 15, verzija 4. Rezultati testov so pokazali, da je bilo komercialno orodje učinkovitejše, nivo pravilnosti kaže, da je bilo v povpreču za 12,38 % boljše pri pravilnosti znakov in 9,86 % pri pravilnosti besed. Kar se tiče nivoja dpi, smo pričakovali, da bodo pri višjih kvalitetah boljši rezultati, vendar se to pri navedenih programih ni zgodilo, še posebej ne pri Tesseractu, ki je pokazal padec pravilnosti pri najvišji nastavitvi (600 dpi). Najboljši rezultati pri Tesseractu so bili pri kvaliteti 200–300 dpi, pri Abbyy pa pri 300–500 dpi. Metodologijo je mogoče uporabiti za različne tipe arhivskega gradiva, rezultate OCR pa je mogoče optimizirati na različnih fazah procesa digitizacije.