Progress in making available blackletters typefaces and handwritten written heritage using artificial intelligence

Prof. PhDr. Dusan Katuscak, PhD[1]

Summary

The topic of the study is the scientific and methodological context of the European project of basic research READ[2] and the application of the results of this research in Slovakia. The study is part of the ongoing applications of the READ project. It shows the progress in research, applications and experiments that the international community of digital humanities has been doing within the READ-COOP[3] association since 2019. Part of these activities is also the Slovak project of applied research with the acronym SKRIPTOR[4], planned for 2020-2024. Based on information research and selection of the latest information sources, it shows progress in research and applications in the field of optical character recognition OCR[5]. The core of the study is focused on a user and not an IT approach to the use of the Transcribus[6] platform for automatic recognition of texts of historical documents. It describes the experience and knowledge gained in mastering the Transcribus platform, which uses the artificial intelligence of the OCR machine and the HTR +[7] method. The study explains and illustrates the main steps of the experiments, the process of learning the machine up

---

[1] **ORCID:** 0000-0001-7444-1077. Mail: dusan.katuscak@fpf.slu.cz. Silesian University Opava. Faculty of Arts and Sciences in Opava; Department of Czech Studies and Librarianship. State Scientific Library in Banská Bystrica.

[2] **READ** Recognition and Enrichment of Archival Documents, the solution of which took place in the years 2016 - 2019 within the Horizon2020 program. [cit 2.10.2021]. Available from: https://cordis.europa.eu/project/id/674943

[3] **READ-COOP.** [cit 2.10.2021] Available from: About us - READ-COOP (readcoop.eu). In October 2021, the association had 86 members from 24 countries. The only member country from Central and Eastern Europe at that time was Slovakia.

[4] **SCRIPTOR.** Project APVV-19-NEWPROJECT-17816 (2020-2024). Innovative access to the written heritage of Slovakia through a system of automatic transcription of historical manuscripts. [Innovative disclosure of written heritage of Slovakia through the automatic transcription of historical manuscripts]. Research organizations: Matej Bel University in Banská Bystrica (responsible researcher doc. Imrich Nagy PhD); State Scientific Library in Banská Bystrica - partner (guarantor prof. PhDr. Dušan Katuščák, PhD)

[5] **OCR** (Optical Character Recognition)

[6] **Transcribus**. A comprehensive platform for digitization, text recognition supported by artificial intelligence as well as for transcription and retrieval of historical documents - from anywhere, anytime and in any language. With Transkribus Lite it is possible to use Transkribus in the browser of personal computers and smartphones. Many of the functions from the Transkribus Expert client can also be used in Transkribus Lite. The platform integrates tools developed by research groups across Europe, including the Human Rights Pattern and Technology Recognition Group of the Technical University of Valencia and the CITlab University of Rostock Group. In June 2020, it had more than 37,000 registered users. The platform was created in the context of two EU projects tranScriptorium (2013-2015) and READ (2016-2019).

[7] **HTR +** Handwritten Text Recognition. Transcribus HTR + software. HTR + cannot yet immediately run a universal automatic transcription for all types of manuscripts, but must first be trained in a specific typeface and handwriting .... The HTR + machine can be trained in any language and typeface.

to the creation of new models of transcription and the results of automatic transcription of printed fractures and manuscripts by Andrej Kmeť. The study also presents the first new efficient model of transcription of the printed historical printed font of the Slovak blackletters typefaces in the Transkribus platform. First, it explains a unique experiment with the transcription of printed Slovak and Czech fracture texts. The following is a description of the advanced experimental transcription of Andrej Kmeť's manuscript letters. It presents the possibilities of making transcribed collections and documents available on local networks and on the Internet.[8]

Keywords

Digital humanities. OCR. READ-COOP. Transcription platform. Project Script. Andrej Kmet. Blackletters typefaces. Read & Search.

Grant details

SCRIPTOR. Project APVV-19-NEWPROJECT-17816 (2020-2024). Innovative access to the written heritage of Slovakia through a system of automatic transcription of historical manuscripts. [Innovative disclosure of written heritage of Slovakia through the automatic transcription of historical manuscripts]. Research organizations: Matej Bel University in Banská Bystrica (responsible researcher doc. Imrich Nagy PhD); State Scientific Library in Banská Bystrica - partner (guarantor prof. PhDr. Dušan Katuščák, PhD)

Introduction

The most significant progress in research, development and applications in digitization in the social sciences and humanities, ie in digital humanities, has occurred especially in the last ten years. The subject of professional interest is automatic optical font resolution (OCR). OCR of common printed documents has long been sufficiently managed with the help of high-quality OCR tools. Dozens of researchers and experimenters have been working on the more demanding issue of OCR of historical manuscripts and prints using artificial intelligence in recent years alone. Progress was made in the implementation of the READ project, which, as a scientific research project for basic research, was directly subordinate to the European Commission. This project was evaluated annually by independent evaluators. The main output of the project is a usable platform and tool Transcribus. The Transcribus platform is a global innovation focused on the transcription of historical manuscripts and documents. In Central and Eastern Europe, Slovakia is so far the only country that seeks to develop the initiatives of the European Basic Research READ in the project Script applied research (Figure 1).

Digital humanities and the READ project

I consider digital humanities as a common name and a cross-cutting methodology for all applications of information and communication technologies (ICT) in the social sciences and humanities, disciplines and their corresponding practice. This methodology was comprehensively applied in the READ project, which was implemented under the Horizon 2020 program. It was a research and innovation program, contract number в „- 674943. The final evaluation of the project was 12.09.2019 in Luxembourg. The author and coordinator of the project was prof.

[8] The study was translated from Slovak into English using a Google machine translator

Günter Mühlberger from the University of Innsbruck. The READ project was funded by the European Union in the amount of approximately EUR 8.2 million. Funding ended on June 30, 2019. The guarantor of the project is the University of Innsbruck, which since 2016 has been researching the basic technologies of text segmentation, handwriting recognition, keyword search for historical documents and tools for making results available. Teams from the Universities of Valencia, Rostock, the Technical University of Vienna and other research institutions represented in the READ project participated in all areas of research. Cooperation with other partners from Europe and the world has developed. Research and development is ongoing. Tens of thousands of users of the Transcription Bus platform create new models of transcription based on historical manuscript and printed collections of national institutions, especially libraries and archives. Collaborating with the community of researchers centered around the Transcribus platform can be useful for our environment of digital humanities experts.

The common vision of scientists, experts and other users is for publicly available transcription models to gradually become a useful common tool for the automatic transcription of historical documents. It is necessary to reach such a level that it is no longer necessary to create manuscripts for each collection and to print separate models. For users, it should be a kind of "black/hidden box" in which artificial intelligence itself selects from the trained integrated models the most suitable model for the transcription of historical prints, manuscripts, typescripts and other documents that the user wants to study or make available. However, there is a long way to go. It is still necessary to create a number of partial models.

I consider it important that Slovak experts be part of the joint international effort and that the future "black/hidden box" be prepared to provide assistance to all in the transcription of Slovak historical collections and documents. At this stage of development, it is important to focus the international community's attention on preparing transcription models based on larger collections that contain hundreds and thousands of pages.

Current state of research and applications

Existing information sources on OCR, on the one hand, relate to ongoing theoretical research on artificial intelligence itself. The authors of theoretical works are mainly computer scientists and mathematicians. On the other hand, there are works whose authors are from the environment of social sciences and humanities and disciplines, ie digital humanities. They address the topic of OCR and HTR from the user's point of view in terms of the practical applicability of existing OCR tools and platforms. Theoretical or user contributions can be divided into two groups according to whether they deal with OCR of printed or manuscript (HTR) works.

A comprehensive review of the READ project includes a project study (Mühlberger 2016), final research reports (Mühlberger 2019a) and a collective study of READ researchers (Mühlberger et al. 2019b), the first published review of how the HTR + is used by a wide community of experts. which shows the current application of handwriting recognition technology in the cultural heritage sector. A collective study (Mühlberger, et al. 2019b) points to the development of character recognition methods.

Since the middle of the 20th century, the character recognition of printed and handwritten documents has developed together as OCR primarily for printed documents. First, the scanned images of the printed text were converted to machine code and compared with the finished font templates. Printed documents contain characters from predefined, ready-made character files, making comparison easier. However, even OCR machines for printed characters are capable of further "training".

However, manuscript texts present a different problem due to the many differences in manuscripts, hands, and changes in manuscripts over time. Manuscripts have become a new challenge for computer scientists. First, in the 1980s, research and development on handwriting recognition developed using statistical methods. This was followed in the 1990s by research and development of pattern recognition in combination with artificial intelligence and the development of deep neural networks in 2000 and 2010. It was also a period of significant development and capacity building of information and communication technologies that can be used for demanding computing, creation and storage of data files.

Mass digitization projects have been implemented in several developed countries and massive digital repositories and archives of printed and manuscript documents have been created[9]. After mass digitization, it is time to use digital content obtained by digitizing manuscripts. If usable, editable text is to be obtained from scanned images of manuscript documents, it is necessary to use advanced HTR Transcription recognition technology, or the same but commercial Quartex (Adam Matthew Digital 2018).

The project has all the attributes of the digital humanities methodology. These attributes include in particular: (a) the cooperation of researchers; (b) scientificisation in the social sciences and humanities; c) interdisciplinarity; d) teamwork (interinstitutional, interstate, universities, libraries, archives, galleries, museums); (e) strong involvement of IT professionals in research, education and knowledge dissemination; f) artificial intelligence (Hidden Markov Model (HMM)).

Significance and features of the Transcribus platform

The result of the READ project is mainly the Transkribus platform[10]. The results of the basic research of the READ project are implemented in this platform. The creation of the Transcribus research platform was, in addition to basic research, one of the main goals of the READ project. Approximately € 2.5 million out of € 8.2 million has been invested in the development of this research infrastructure. Follow-up projects are now emerging in which basic and applied research continues. Adopting the Transkribus platform can also have significant economic effects.

According to data from the internal documentation of the READ project, the market prices of manual transcription of historical manuscripts range from 10 EUR to 30 EUR or more for simple English, German, Latin for a specific manuscript. Assuming EUR 15 per page as an average cost, in the READ project, operators generated a monetary value of EUR 4-6 million. These data are an added value and a potential

[9] In 2020, Transkribus received the EU Horizon Impact Award. [cit. 2021-10-14] Available online: https://cordis.europa.eu/article/id/422311-horizon-impact-award-2020-awards-eu-funded-projects-with-the-greatest-societal- impact

[10] If you are interested in transcribing individual shorter documents, you can try using one of the publicly available transcription models with a similar typeface, print, or handwriting.

source of development of the newly established READ-COOP association and a convincing confirmation of the basic concept of research aimed at new knowledge and at the same time to the commercial use of tools that are the results of the application of new knowledge.

Representatives of digital humanities in Slovakia have different attitudes to this initiative. From enthusiastic expressions of agreement and admiration to very reserved and even negative attitudes (such as "it's nothing for us", "we have other worries", "artificial intelligence will not replace us experts"). These are often reactions which, on the one hand, verbally declare an interest in "digitization" and "artificial intelligence", but on the other hand show a lack of knowledge about the issues and possibilities of digitization and the use of artificial intelligence. Attitudes suggest a preference for traditional paradigms of work and research rather than a real effort to seek innovative tools to access and interpret our vast historical written heritage as part of Europe's cultural heritage.

As for the transcription of Slovak, it found itself on the list of languages in the final report on the READ project thanks to my own initiative, without any support and essentially without the interest of national institutions, archives, libraries, museums and academia. It was a job that I devoted about 2,000 hours since 2018 and that I financed until 2020 only from my own resources. The achieved results, know-how and experience led me to the effort to introduce a revolutionary and innovative platform Transkribus in Slovakia and the Czech Republic, especially into the education system, as well as into the practice of memory and fund institutions through research and development projects.

The Transkribus platform is free software (open source) with a guarantee of safe use for registered clients of the platform. Anyone can create their own account and then download an expert client for free or use the simpler Transkibus Lite tool. An API is available to connect clients' computers or mobile devices to the platform. Most software tools are free software that can be obtained from GitHub.

READ-COOP

The READ project ended on 30 June 2019. Subsequently, the international association READ-COOP SCE (Societas Cooperativa Europeae - SCE) was established on 1 July 2019. Its aim is to maintain and further develop the Transcribus platform. Experts and institutions are interested in the continuation and development of the Transkribus service. Today, in 2021, there are more than 80,000 Transkribus users working with this platform on a daily basis.

Figure 1 Extension of the Transcribus platform in Europe (Source: readcoop.eu, as of September 2021)Projekt SKRIPTOR

Slovak experts respond to new trends in OCR and research of historical documents with the SKRIPTOR project (Katuščák and Nagy, et al. 2019). The project has a European and national dimension. The aim of the project is the implementation and dissemination of the latest technological innovations and knowledge about the effective approach of the professional and lay public to the Slovak and foreign written heritage.

The results of the SKRIPTOR project focus on an innovative approach to documents and knowledge. The SKRIPTOR project is a direct follow-up to the recently completed European READ project. The technological and scientific innovations of the READ project are based on the use of artificial intelligence and the methodology of the digital humanities.

The strategic goal of the SKRIPTOR project is to create conditions at the national level for a competent partnership of Slovak researchers with leading European research, to establish and then actively participate in multilateral European scientific cooperation. The SKRIPTOR project is implemented in the field of history and archiving. It also extends into library and information science.

The SKRIPTOR project focuses on modern documents. However, collections that are subject to review and access may also include major recent texts and documents and incunabula, 16th-century printed materials, historical magazines, newspapers, as well as valuable 18th-20th-century materials, etc.

The originality and innovation of the SKRIPTOR project lies in the implementation of new knowledge gained in the excellent European READ research. The main goal of the SKRIPTOR project is to implement the latest knowledge and insights from the research of automatic text recognition of historical documents in Slovakia. The aim of creating new models using the Transkribus platform is to confirm and achieve in our collections a reduction in the price of transcription from 30 euros for manual transcription of a page to less than 1 euro / page for automatic transcription of texts.

In the Script project, we have preliminarily selected the following collections for research and experimental transcription: 1. Slovak and Czech fracture (blackletters typefaces); 2. Andrej Kmeť - personal handwritten correspondence; 3. Martin Lauček - Collectanea; 4. Collection of Isaac Abrahamides of Hrochotsky from 1600 - 1601; 5. Bed of Juraj Schmidelio-Kováčik from 1598 - 1607; 6. Canonical visits of the Banská Bystrica diocese from the 18th - 19th century; 7. Hurban, J. M. manuscript documents; 8. Roman Catholic registries; 9. Land registers of Teresian urban regulation; 10. Parcelling protocols of stable cadastre; 11. Congregational minutes, sedrial protocols; 12. Other collections of documents identified during archival research.

Advances in research

Hodel writes about the progress in print text recognition based on optical print recognition (Hodel et al 2021). Hodel also deals with the most important practical aspect of transcription, namely the question of what is the accuracy or error of transcription. Based on empirical data from the READ research and based on Günter Mühlberger's findings from 2014a, 2019 lists three error classes. It considers it confirmed and verified that: a) if the error rate of CER characters is less than 10%, which is 10 or less errors per hundred characters, then the transcription result is good, legible and, if appropriate, further editing is possible output; b) if the error rate of CER characters is ≤ 5%, then the transcription result is very good; c) if the error rate of CERs is below 3%, then the transcription results can be considered excellent and the error rate of CERs below 2.5% can be considered excellent.

Hodel is talking about transcription without training as a goal. He states that in order to create an optimal universal model of transcription of manuscripts of various hands, styles, fonts, periods, etc., which would not always require the preparation of separate models, it is necessary to have as many excellent models as possible. He believes that these models of transcription should probably be developed for various similar classes of manuscripts, such as the 19th-century *current* script, which is the subject of his attention.

Strobel contributes to progress in the field of optical character recognition (OCR) (Strobel et.al 2020). Based on the analysis of the effectiveness of some OCR systems of printed German historical newspapers (fractures), the authors came to the conclusion that a sufficient training sample (so-called ground truth) is 50 newspaper pages. Their findings are based on comparisons of five OCR systems: *1. ABBYY FineReader XIX10 (FRXIX) from 2005, 2. ABBYY FineReader Server 11 (FRS11) inserted in previous versions into the Transcribus and Transcribus HTR + 4. Kraken, 5. Tesseract.*

Martinek et.al. (2020) presents in his theoretical experimental study a system of segmentation of printed text and OCR. It deals with a set of methods that allow you to perform OCR of historical prints in German based on a small amount of training data. Describes its OCR system that uses recurrent neural networks. It focuses on the partial processes of the OCR system, mainly on the analysis of the page layout, including the segmentation of the text block and lines, and on the OCR itself. The described experiments are aimed at determining the best way to achieve good OCR results for historical German printed documents. They used digitized archival material from the *Porta fontium* project from the Czech-Bavarian border for the experiment. Specifically, it was 10 pages from the newspaper *Ascher Zeitung* from the second

half of the 19th century printed by fracture. They used 7 pages for training, 1 page for validation and 2 pages for evaluation of effectiveness. Another 15 pages were used for page template identification and segmentation training. The authors consider the obtained results to be comparable or even better than the results of several recent systems. In the case of a fracture from a German newspaper, they achieved the following CER values compared to other systems: Porta fontium CER 0.024. Tesseract (deu_frak) CER 0.053. Tesseract (Fracture) CER 0.045. CER transcripts 0.027. It is not known whether Czech experiments, including the Pero OCR application, are aimed at creating a competitive or supportive activity against the Transcription Bus platform and at a specific freely available tool for transcribing historical manuscripts and prints.

Martin Kišš (2018) deals with the topic of recognition of modern printed texts written by fracture in his diploma thesis. He based his research on *TensorFlow*, originally developed by *Google* and available as an open source machine learning platform. Part of his approach is a built-in generator of artificial historical texts. Using this generator, he created an artificial data set on which he trained a neural network for row recognition. He tested this neural network on real historical lines of text and achieved a success rate of 89.0% character accuracy after training.

The research of the Technical University in Brno (Kiš et al 2019) is interesting. As a result of research and experiments, the *Pero OCR*[12] application is now constantly being developed. According to the authors, "they provide the most advanced recognition of the baseline of the text, which is based on convolutional and repetitive neural networks ...". They presented a set of data that will allow future development and evaluation of document analysis for poor quality images. It is primarily intended for line-level text recognition, layout analysis, image recovery, and text binarization. Several of my transcriptions in the *Pero OCR* application showed a good ability of the application to segment and recognize blocks of text and lines of printed antiquity and *Czech fracture*, and usability also for manuscripts. However, a more detailed assessment of the *Pero OCR* application requires more thorough user analysis.

Fracture (blackletters typefaces) transcription

The experiment concerned the application of artificial intelligence to the automatic transcription of Slovak and Czech fractures (so-called Schwabach). Fracture is a type of printed font that has been widely used since the 15th century in Czech and Slovak books, newspapers and magazines in the modern age and later, practically until the 50s of the 20th century.

As part of my education in the subject of digitization at the Silesian University at the Institute of Czech Studies and Librarianship, I used the tools of artificial intelligence Transcribus to prepare probably the first extremely successful transcription of Slovak and Czech printed text - fracture - historical Moravské noviny, Opavský besedník and Slovak publication Jánošík. I prepared transcription models of Slovak and Czech fractures (Table 1). In the exercise set, I achieved an CER error rate of 0.39%. (Character Error Rates). However, a higher value of 0.44% achieved on the validation set (CER on Validation Set) is decisive for the practical use of this model.

[12] Available after registration (e-mail and own password) on the Internet: pero-ocr.fit.vutbr.cz

TRANSKRIPCIA FRAKTÚRY (ŠVABACHU)								
DÁTUM	Metóda OCR	Cvičný súbor		Validačný súbor		Presnosť CER		ID modelu
		str	riadky	str	riadky	**Cvičný**	**Validačný**	
20210824	OCR base 29418	7	8092	1	888	0,20%	0,91%	36160
20210905	OCR base 29418	9	11231	4	1179	0,18%	1,07%	36358
20210912	OCR base 29418	17	20805	5	2252	0,39%	0,44%	36550
20210913	OCR base 36550	7	2462	3	276	0,03%	1,78	36607

Table 1 Results of fracture transcription (swab) of Slovak and Czech historical presses

From now on, we are able to transcribe a fracture in Slovak and Czech historical presses with an accuracy of about 99%. In our case, the accuracy is 99.56%. The error rate is 0.44%.

After logging in, the results of the Czech text fracture transcription are available in the Transkribus Lite (Transkribus) platform in the FRAKTURA_CZ collection (114429, Owner) and on the Internet in the beta version of the Read & Search browser Moravské noviny - Schwabacher (transkribus.eu). Newspaper Opava Besedník: Opava Besedník (1) - Schwabacher (transkribus.eu). Transcription of the Slovak fracture text: Jánošík - Schwabacher (transkribus.eu).

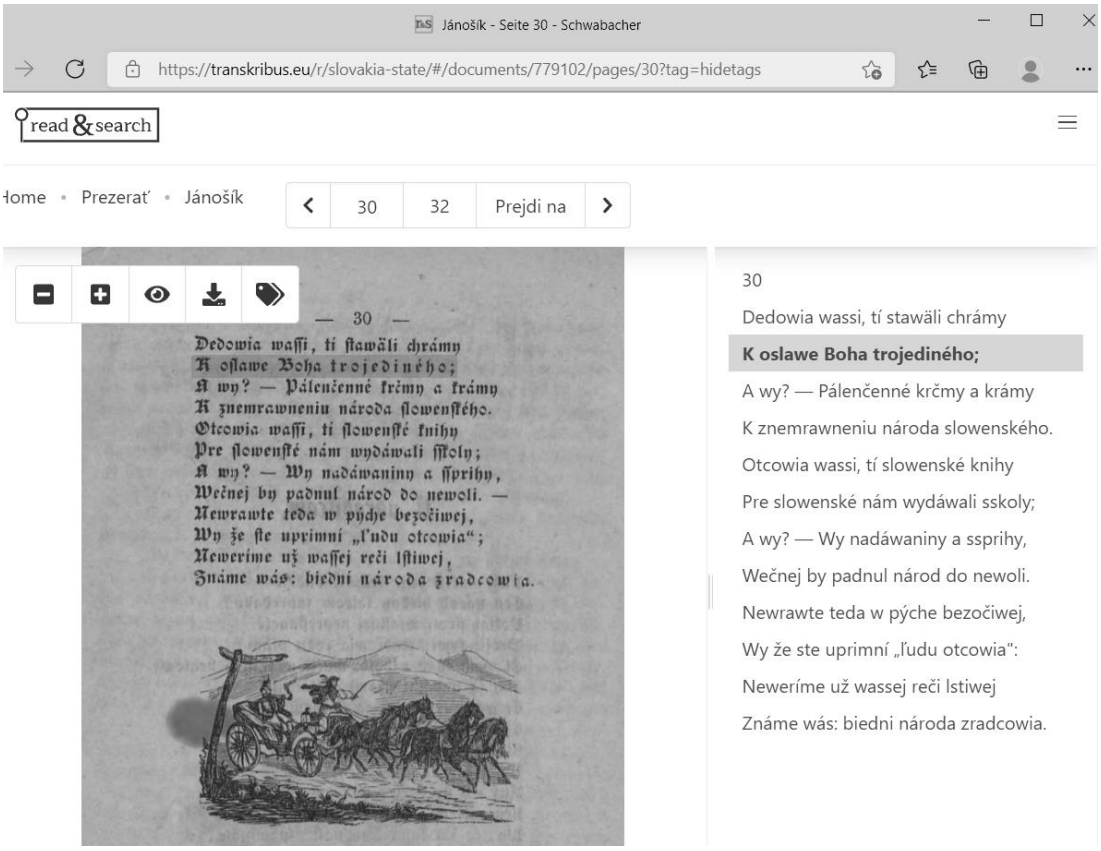


Figure 2 Representation of the transcription of a fracture next to an image (from the Slovak publication Jánošík J. N. Bobulu, 1862) Source: author

Figure 3 Representation of fracture overlap directly in the image (from the Slovak publication Jánošík J. N. Bobulu, 1862). Source: author

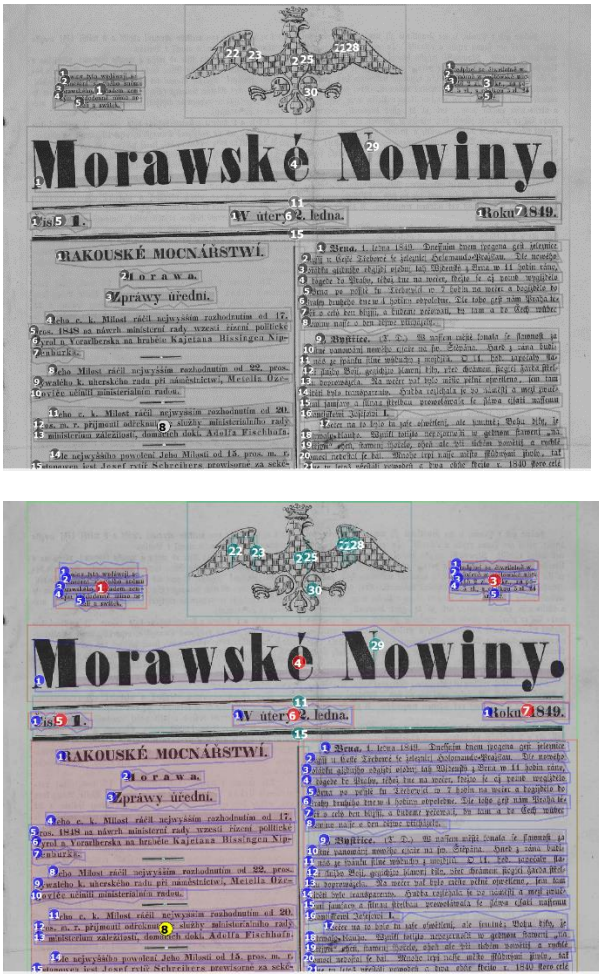
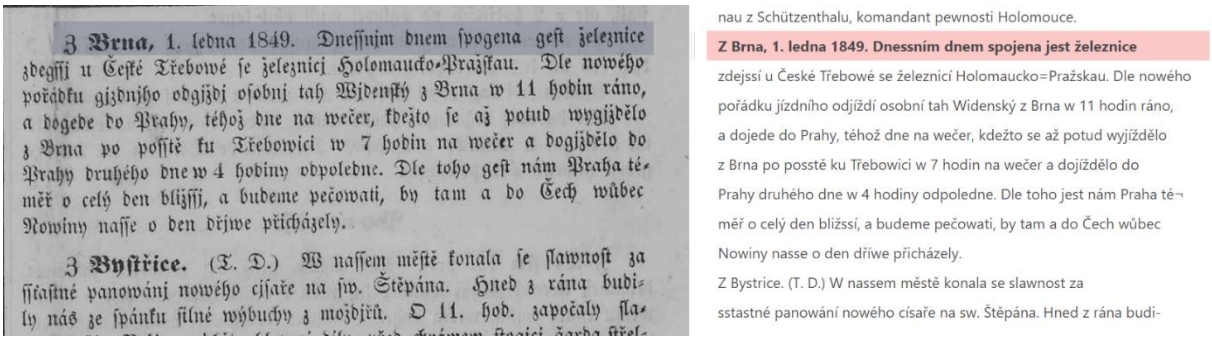


Figure 4 Segmentation of the Moravské noviny page (1849). Headline, graphics, text blocks, paragraphs, lines (Source: author)



Z Brna, 1. ledna 1849. Dnessnim dnem spogena geſt zeleznice zdegſſi u Ceſké Třebowé ſe zeleznici Holomaucko-Prazſkau. Dle nowého pořádku gizdnjho odgiżdj oſobnj tah Widenſký z Brna w 11 hodin ráno, a dogede do Prahy, téhoż dne na wečer, kdežto ſe aż potud wygiżdělo z Brna po poſſtě ku Třebowici w 7 hodin na wečer a dogiżdělo do Prahy druhého dne w 4 hodiny odpoledne. Dle toho geſt nám Praha tě- měř o celý den bliżſſj, a budeme pečowati, by tam a do Čech wůbec Nowiny naſſe o den dřjwe přicházely.

Z Bystřice. (T. D.) W naſſem městě konala ſe ſlawnoſt za ſſtaſtné panowánj nowého cjſaře na ſw. Štěpána. Hned z rána budi- ly nás ze ſpánku ſilné wýbuchy z mozdjřů. O 11. hod. započaly ſla-

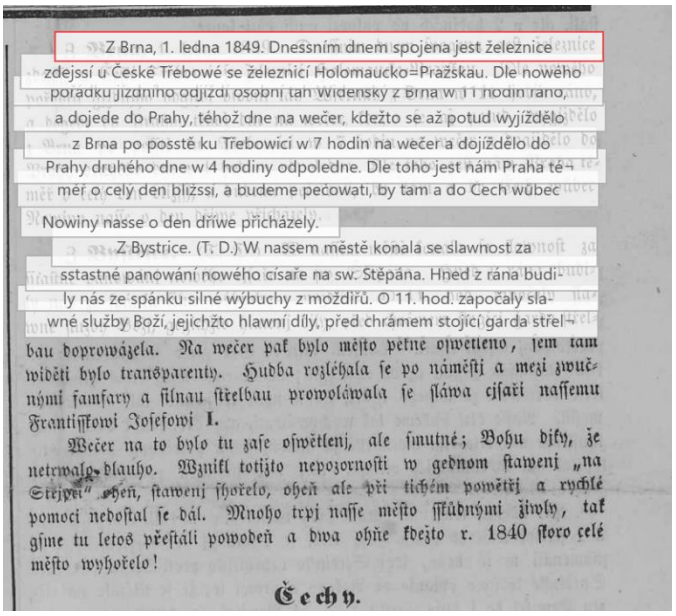
nau z Schützenthalu, komandant pewnosti Holomouce.

Z Brna, 1. ledna 1849. Dnessním dnem spojena jest železnice

zdejssí u České Třebowé se železnicí Holomaucko=Pražskau. Dle nového pořádku jízdního odjíždí osobní tah Widenský z Brna w 11 hodin ráno, a dojede do Prahy, téhož dne na večer, kdežto se až potud wyjíždělo z Brna po posstě ku Třebowici w 7 hodin na večer a dojíždělo do Prahy druhého dne w 4 hodiny odpoledne. Dle toho jest nám Praha tě- měř o celý den bližssí, a budeme pečowati, by tam a do Čech wůbec Nowiny nasse o den dříwe přicházely.

Z Bystrice. (T. D.) W nassem městě konala se slawnost za sstastné panowání nového císaře na sw. Štěpána. Hned z rána budi-

Figure 5 Moravské noviny Transkribus Lite - transcription next to the picture



Z Brna, 1. ledna 1849. Dnessním dnem spojena jest železnice zdejssí u České Třebowé se železnicí Holomaucko=Pražskau. Dle nového pořádku jízdního odjíždí osobní tah Widenský z Brna w 11 hodin ráno, a dojede do Prahy, téhož dne na večer, kdežto se až potud wyjíždělo z Brna po posstě ku Třebowici w 7 hodin na večer a dojíždělo do Prahy druhého dne w 4 hodiny odpoledne. Dle toho jest nám Praha tě- měř o celý den bližssí, a budeme pečowati, by tam a do Čech wůbec Nowiny nasse o den dříwe přicházely.

Z Bystrice. (T. D.) W nassem městě konala se slawnost za sstastné panowání nového císaře na sw. Štěpána. Hned z rána budi- ly nás ze spánku silné wýbuchy z moždířů. O 11. hod. započaly sla- wné služby Boží, jejichžto hlawní díly, před chrámem stojící garda střel- bau doprowázela. Na wečer pak bylo město pěkně osswětleno, jem tam widěti bylo transparenty. Hudba rozléhala se po náměſtj a mezi zwuč- nými fanfary a ſilnau ſtřelbau prowoláwala se ſláwa cjſaři naſſemu Frantiſſkowi Joſefowi I.

Wečer na to bylo tu zaſe oſwětlenj, ale ſmutné; Bohu djky, že netrwalo dlauho. Wznikl totižto nepozornoſti w gednom ſtawenj „na Střjpri“ oheň, ſtawenj ſhořelo, oheň ale při tichém powětřj a rychlé pomoci nedoſtal ſe dál. Mnoho trpj naſſe město ſſkůdnými žiwly, tak gſme tu letos přeſtáli powodeň a dwa ohně kdežto r. 1840 ſkoro celé město wyhořelo!

Čechy.

Figure 6 Moravské noviny Transkribus Lite - transcription overlaps the text of the image

An experiment with a collection of letters by Andrej Kmeť

Historians, linguists, archivists, librarians, documentarians and everyone else who comes into contact with manuscript texts have been dreaming about the automatic transcription of manuscript texts for decades. Gradually, the automatic transcription of manuscripts becomes a reality. Behind it is massive international basic research in the field of artificial intelligence and thousands of hours of work.

I published signal information about working with the Transkribus platform in 2018 in a blog and on my Facebook status. I was surprised by the declared interest in this job. This is understandable, because many historians, linguists, librarians, educators are increasingly educated in the use of new technologies in their work, and I understand that the innovations that make their work easier are important.

The transcribus, of course, does not replace the professional and scientific erudition of historians and archivists. Therefore, their reserved attitude is understandable. Artificial intelligence does not compete with experts. It helps them. Automatic transcription can be just one step in the scientific work of experts. This is followed by historical research of the text and context of transcribed texts and information, editing of texts obtained by transcription, identification of entities, creation of keywords,

metadata that are discovered in the text (dates, names of persons, names of geographical units, corporations, etc.).

The purpose of more extensive transcription using the top Transcribus platform is to facilitate the reading and access to unique collections, documents, archival units, which are usually found in archives in only one copy. This is the difference between the presence of units in libraries and archives. The archives contain unique, authentic original documents, collections, archive units, while the libraries contain document titles, which often have hundreds to thousands of copies. Unique archives need to be made available. The path to access leads through their transcription.

After the transcription of historical texts and manuscripts, digital content can be edited, interpreted, used and made available for wider use in public information systems and services. In addition, the transcribed original text, for example in Latin, Hungarian, German, or another language, can be at least approximately further automatically translated into another language. This quite significantly changes the nature of the work of archivists and historians.

The result of my work are transcription models of different quality. An overview of the models is given in Table 2.

MODELY TRANSKRIPCIE RUKOPISOV ANDREJA KMEŤA											
DÁTUM	Metóda	**Model**	**Cvičný súbor**		**Validačný súbor**		**Presnosť CER súbor**		**Počet cyklov**	**CER/WER**	
		ID	strany	riadky	strany	riadky	cvičný	validačný	(epochy)	znaky	slová
20190125	CITlabHT+	**10135**	125	22549	26	3497	1,15%	3,37%	200	5,97%	21,60%
20190201	CITlabHT+	**10410**	152	29905	46	4499	1,27%	2,97%	200	6,19%	22,13%
20190205	CITlabHT+	**10548**	166	29411	46	4573	1,37%	1,84%	200	5,91%	21,87%
20201012	CITlabHT+	**26809**	111	18071	98	2921	0,44%	7,25%	500	6,08%	21,87%
20210410	CITlabHT+	**31888**	119	19291	13	3126	1,15%	5,16%	200	3,77%	12,27%
20210821	**CITlabHT+GT**	**36009**	**185**	**28672**	**26**	**4703**	**1,87%**	**5,79%**	**200**	**2,48%**	**7,73%**

Table 2 Overview of experiments with transcription models of Andrej Kmeť's handwritten correspondence [13]

[13]Explanations to the table

Datum-Date*:* Model Creation Date (YYYYMMDD)

Metóda-Method: Manuscript transcription method chosen (HTR +)

ID: The model identification number in my collections and among all Transribus models on the remote server

Cvičný súbor - Train file: Number of pages and number of lines that were manually rewritten and used to learn (train) the machine in the Transcribus platform. We rewrote 211 pages for the exercise. Of these, 185 are used for training and 26 for validation (verification)

Validačný súbor - Validation file: The number of pages and lines that were selected from the total number of rewritten pages to verify learning accuracy.

Presnosť CER - CER Accuracy: CER (Character Error Rates) Accuracy. The error rate of the characters in the input file and in the validation file. For manuscripts, it is practically impossible for manual transcription to be 0.0%.

Počet cyklov - Number of cycles: Number of cycles, so-called epochs that the machine used for learning (training).

CER / WER: Values express real practical, user accuracy resp. Character Error Rates (CER) and Word Error Rates (WER) in 6 models from 2019-2021, which I own

.

For comparison, I tested all the models listed in the table on one, the most accurately prepared double page in *FINAL* quality in the collection ID 115514. This is a letter from Andrej Kmeť Ľ.V. Rizner (Document ID 621673). The error rate of words is *de facto* irrelevant, because a wrong character (eg punctuation) also causes the error rate of a word in most cases.

The average recalculated error rate of characters in six models is CER 5.0%, while I created 5 of them on training sets and pages of different quality, which were mainly in the status *In Progress*. However, for the practical transcription of hundreds of other pages, it will be best to use the 36009 model, which I created from 185 pages of the training set and 26 pages of the validation set. It appears that the lowest CER accuracy values in the validation file do not mean that the models that are in the first five rows of the sixth column in Table 2 and are not created on the *ground truth* pages are the most suitable for further transcription.

For the final preparation of this model, I used well-prepared pages in *ground truth* quality. In terms of the accuracy of transcription of other letters by Andrej Kmeť, I consider the results of Model 36009 with CER values of 2.48% and WER 7.73% to be the best. In the future, based on further experience, I will consider providing this model of my free use for similar manuscript collections.

We continuously organize and publish the results of document transcription and also on the Internet through a tool developed by the READ-COOP team called READ & SEARCH. Public access to the documents is via the Read & Search page - https://transkribus.eu/r/slovakia-state/#/, the interface of which I translated into Slovak.

Transcription workflow

Based on my own experience, I understand transcription as a complex process, which presupposes in particular commitment, availability of financial resources and infrastructure. The main processes are:

Preparation. In particular: Information archival research (heuristics), identification of possible collections and documents, solving the conditions of availability of collections and documents, quantification and selection of documents for transcription (number of pages and homogeneity of manuscripts), agreement with the owner or administrator of the collection on the place and method of scanning

Scanning. In particular: scanning, photographing documents, naming and organizing directories and files on a computer, archiving source files (TIFF, RAW) and backing up derived files (JPG, PDF, PNG, etc.)

Installation of a professional client and work with the Transkribus platform. In particular: getting acquainted with the Transribus documentation, choosing the image format for Transribus, quality control and preparing images for uploading to Transribus, choosing the method of uploading files, creating your own collection, uploading selected files to the Transribus platform for collection

Manual transcription. In particular: selection of samples of pages for manual transcription according to the specifics of the manuscript, decision on sharing the collection with collaborators and their role, manual transcription of the sample for the training set

Page segmentation and metadata. In particular: segmentation of pages or entire files, quality control and correction of manual transcription and segmentation, document metadata, page metadata, structural metadata, comments, KWS

Creation of a transcription model. In particular: learning the machine for the transcription model, checking the quality and efficiency of the model and correcting the training set, restarting the model creation and checking the quality of the model, selecting *ground truth* quality pages, using the model to transcribe all segmented pages in the collection

Making available and using transcription results. In particular: exporting results in different ways and in different formats, editing and correcting transcription results in Transcriptibus Lite, using a transcription model, making transcription results available on a local network or publishing transcription results online for use on the Internet via Read & Search.

Collection selection

For the experiment, I chose a collection of manuscript, mostly Slovak correspondence by Andrej Kmeť, stored in the library of the Slovak National Museum in Martin, with the prior kind consent of the museum director. Several letters are in Latin, Hungarian and parts of the letters are also in German and Czech. These are letters from Andrej Kmeť[14] (SNM, Martin) from 1841–1908. In the field of scientific approach to correspondence of scholars in modern times in the spirit of digital humanities methodology, the most comprehensive source of knowledge is undoubtedly the international research initiated and led by Howard Hotson in 2014 - 2018 (Hotson 2019). In this study, we are only interested in correspondence as an extensive manuscript material that is suitable for experiments with automatic transcription.

Picture 7 Archive box with leaves for Andrej Sokolík (Photo: author)

The personality of Andrej Kmeť, including the processing of parts of his correspondence, is systematically dealt with by Karol Hollý, and he also mentions other sources concerning Kmeť's manuscript inheritance (Hollý 2013, 2019).

[14] **Andrej Kmeť** (November 19, 1841, Szénásfalu, Austrian Empire (today Bzenica, Slovakia) - February 16, 1908, Turócszentmárton (today Martin, Slovakia)) was a Slovak botanist, ethnographer, archaeologist, and geologist.[1] He identified several new species of plants and created a herbarium with 72,000 specimens. He was one of the first researchers who carried on modern archaeological excavations in Central Europe. In 1892, he founded the Slovak Learned Society (Slovak: *Slovenská učená spoločnosť*), which later became nucleus of the Slovak Academy of Sciences. He was also known for his bitter criticism of alcoholism. Andrej Kmeť was interred in the National Cemetery in Martin. (Source: Wikipedia)

Scanning

Scanning, ie scanning, more accurate photography, took place on May 23 - 30, 2018 in the Library of the Slovak National Museum in Martin. I used the ScanTent device (scanning tent) and the freely available DocScan application for scanning. I used ScanTent intentionally to verify the entire proposed Transribus workflow. It is known that many archives already have parts of the collections more or less scanned. The devices I have chosen are important in cases where the collections have not yet been scanned. It is known that ordinary scientists and users are not allowed to extract archival material from study archives. Amateur photography of pages with smartphones or cameras is problematic for larger files (thousands of pages). Therefore, ScanTent and DocScan is a possible and affordable option that is acceptable with certain practical reservations (format, focus, quality). However, it should be noted that in this case it is a question of photography and not scanning in the true technological sense of the word. In the future, I would definitely use a professional scanner for scanning and scanning in the highest achievable quality.

Figure 8 ScanTent scanning tent

I scanned the complete contents of the five boxes. Some leaves were on several pages, there are incomplete pages, vacancies, etc. One image could contain multiple pages of manuscript. In the shooting phase, images are created, not pages, unless the pages are scanned. Sometimes it is more appropriate to scan the sheets according to the pages, individually, because if the sheet is scanned as a double-sided, sometimes the order of the pages in the subsequent image processing must be arranged. postprocessing. However, in the next step of text segmentation, it is possible to arrange the individual pages as blocks of text in the correct order. The individual pages in Andrej Kmeť's letters did not follow each other, so on the scanned image there were, for example, pages 3 and 1, on the next 2 and 4.

The recording time was about 15 - 20 hours. Scanning was in single-sheet mode, not "series" (with automatic page-by-page scanning), as the handwriting material is on separate sheets of various formats. Part of the material consists of original sheets, part photocopies. In particular, the original sheets are often on brittle paper that would require preservation. Business cards and similar smaller paper sizes -

DocScan asked to "move closer", I solved by underlining a clean A4 page under the missing parts of the sheet. Some sheets were damaged (corner was missing, damaged sides of the sheet. In this case, the system reported “no page found.” I solved this by underlaying the white side as a pad under the sheet and under the missing parts, then DocScan focused.

I had to shoot some components again, as I did not pay the necessary attention to focusing at first. DocScan focuses on the sheet area in several places. Focus is indicated by red and green marks. When the focus is satisfactory, “OK” is displayed, then the shutter button can be pressed. For shooting, I used a Samsung Galaxy 6 mobile phone with the Android operating program with which DocScan worked at the time. The process of transferring data from Samsung (Android) to MacBook Air (iOS operating system) was unclear to me at first. DocScan software is also currently available for the iOS operating system. Finally, I used a Windows computer and downloaded pictures from Pictures from Samsung to another computer. I consider the use of the DocScan system and the Samsung mobile phone to be an absolutely emergency solution, because in further work, especially during segmentation, I discovered a relatively large number of blurred parts of pages. As parts of the page were blurred, the segmentation was inaccurate and subsequently not even transcribed. In the future, I would recommend using high-quality professional scanners for large valuable collections and scanning itself in the highest achievable quality.

When scanning, the DocScan system can be connected directly to the server and the Transkribus platform (in Innsbruck or Rostock) and scan and transfer images directly from the scan to the Transkribus platform. I did not use this opportunity. I considered it necessary to check the accuracy and quality of the scan. Some Transribus operations required the use of Preview, Adobe Acrobat, File Zilla, and so on. I used the tools to adjust the text orientation, eliminate duplicates, arrange pages in the file, etc.

Scanned digital content (images) was: a) ready for further processing in DocScan software (content identification, metadata), b) uploaded without modification on CD ROM for use in SNM at the discretion of SNM and Archive management, c) images were ready for upload to the Transkribus platform and for further processing in the Transkribus software. This was followed by recording, segmentation and transcription of the manuscript text.

I divided the digital content as it is in the archive boxes. So I burned 5 compact discs (CDs), which I handed over in an untranscribed protocol to the then director of the Ethnographic Museum in Martin, dr. Maria Halmova. Collection administrators and archivists can now use digital content and publish it in its entirety. In addition, they can insert one compact disc into each box. They can decide who to allow access to the collection on disk, or again allow them to work with relatively fragile paper original archive sheets. I make the transcribed content available gradually through Read & Search software, which acts as a "software as a service (SaaS)".

Upload digital image files

Scanned images can be processed either locally or edited after import to a remote Transcribus server. Before importing to the server and before using the Transkribus

platform, it is necessary to register, download the Transkribus platform for an expert or Transkribus Lite, in which, however, it is not possible to create your own transcription models. Then you need to create your own private collection, which is available only to the person who created it, unless he decides to share it with other users. It is possible for the "transcriber" to allow access to certain operations, such as students, operators, subcontractors. It can provide access to your own collection for training sample preparation, post-transcription editing, etc. Automatic transcription is performed exclusively on the remote server using the Transcription bus infrastructure. It is possible to work locally with your own documents and collections as needed.

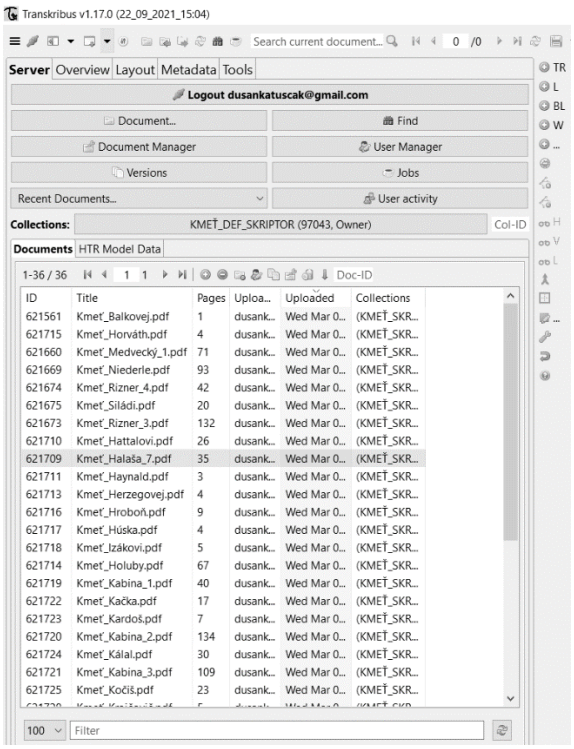


Figure 9 Transcription bus expert. Uploaded collection files KMEŤ_DEF_SKRIPTOR (97043, Owner). Source: author

Before importing files, you need to create your own collection with your files for transcription. One-time uploading and importing of images is possible up to 500 MB in size. If the volume of imported images is larger, the images can be divided into several files and imported one at a time. Larger image files can also be uploaded, imported using an FTP client, such as WinSCP, via the URL or DFG Viewer METS. Images can be uploaded as PDF and JPG, TIFF and more. The collection of imported images created by scanning Andrej Kmeť's letters has 11.7 GB at a resolution of 300 dpi. I did not evaluate the effectiveness of the resolution in scanning in relation to the accuracy of automatic conversion in Transcribus, although, hypothetically, this relationship may be significant.

My experience shows that before importing, it is advisable to check digital images, their quality, sharpness, translucency from the opposite side of the sheet, completeness, page orientation, etc. After some experience, I imported PDF files via faster, simple WinSCP software.

Segmentation

After importing files to the server, automatic segmentation must be performed on the server. When segmenting text and images, the client must be connected to the application on the server. Segmentation means that the image of the handwritten text of the document, which is still on the server as an image, is automatically divided into blocks, areas, lines of text. If necessary, manual corrections can be made. These include, for example, arranging, joining and dividing blocks, expanding a polygon, adjusting the baseline below the line, segment boundaries, and the like. Segmentation is key to transcription itself. High-quality scanned pages with sharp handwriting are usually segmented flawlessly. However, sometimes it is necessary to carefully check or adjust the manual order of text regions (TR-Text regions), the order of lines (Lines reading orders), lines and polygons created by the machine (artificial intelligence) after segmentation.

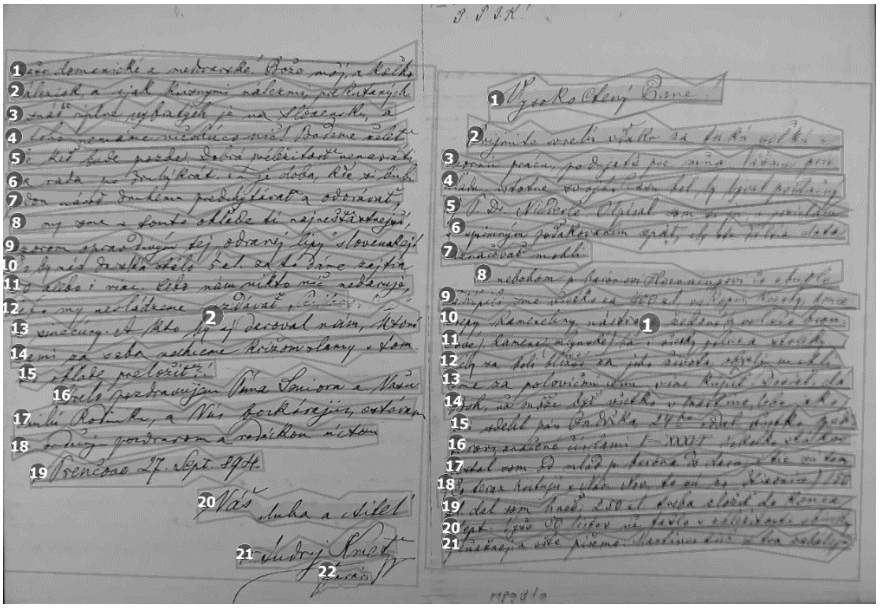


Figure 10 Segmented double page of Andrej Kmet"s letter (two blocks of text, line numbers) Source: author

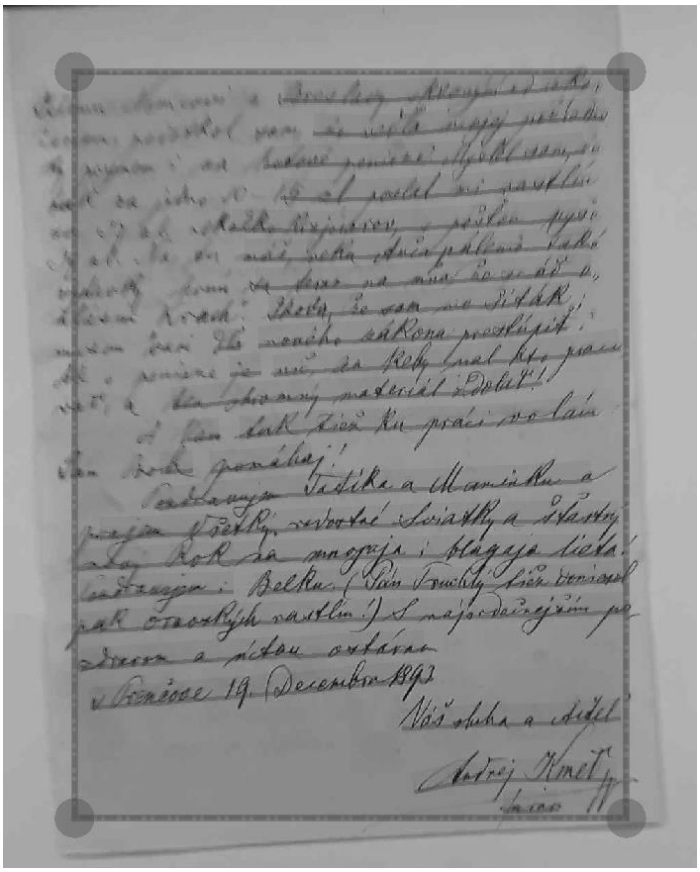
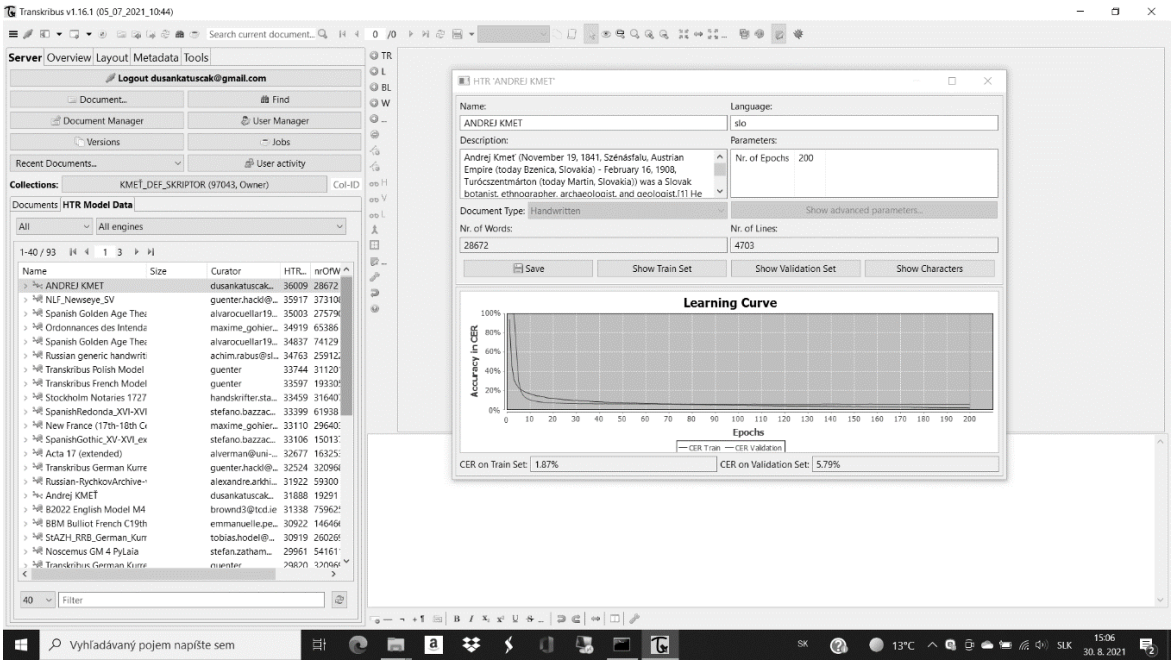


Figure 11 Inaccurate segmentation of sheet ID 621730 probably due to incorrect scanning. Source: author



Obrázok 1 Výsledku tréningu modelu ANDREJ KMEŤ ID 36009

HTR machine training

The Transcribus machine is practiced, actually learning first on the sides that are selected into the practice file. The machine repeatedly, e.g. in 50 cycles, reads each page of the candle file, and gradually identifies characters that cannot be

unambiguously identified or that were caused by incorrect transcription of pages in the *Ground truth* file.

The transribus first creates a model on the sides of the training set. Characters that the machine considers to be incorrect are included among the incorrect characters in the training file. This is the CER on Train Set statistic value. The HTR machine must first be trained for the hand. As a rule, the learning machine should see 100 examples of each character found in the document, which is usually about 50 pages of the training set (Mühlberger et al. [2016]).

After training the model on the pages that have been selected for the training file, Transcribus will automatically use the learned model created on the training file pages to validate it on the pages selected for the verification file. Verification file, so-called validation set is used for practical testing of the model. The machine accesses the text in the verification file repeatedly each time, as if doing so for the first time, applying the model that it "learned" in the training file. At the end of this process, we have a model for automatic manuscript transcription. The most important value for evaluating the transcription accuracy of the created model is the value that expresses the error rate of character transcription in the verification file. This is the CER value on Validation Set.

According to a certain algorithm, a sample of pages (data set, so-called dataset) is selected from the imported collection, which is used to learn the machine and create a model for a certain type of manuscript. To do this, you need to show the machine the correct examples of text. The machine learns font and word patterns according to the training set. If the collection of texts is from several hands, it is necessary to select an appropriate size of the practice and test sample. The selection of pages can also be done automatically according to a certain algorithm, so that the sample is prepared according to certain pages and contains about 20,000 words. The training file is created directly in the expert editor of the Transkribus client, both locally and on the server. Basically, it is necessary to carefully and very accurately rewrite the manuscript in the editor according to the lines, not to correct anything. The text should be rewritten according to contemporary language and grammar, also with errors and according to other instructions and manuals available for this operation. Sequence of parts of the text, tagging, selection and editing of keywords, descriptive metadata, etc. determined by the author of the transcription and the creator of the transcription model. The transcription result is then visible and evaluated on the test set. If the result is satisfactory, additional files or the entire collection can be automatically transcribed. Simply, after completing the machine learning process and creating the model, the model is available to the owner, who can use it or share it with other users and apply it to any document. Correct and incorrect reading data become the basis of the model.

Automatic transcription

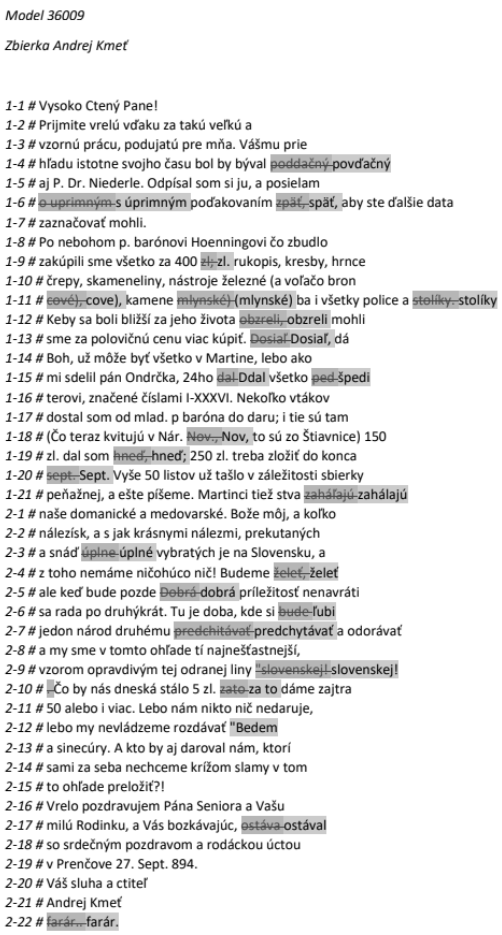
Automatic transcription serves as a basis for scientific editing, in which it is possible to correct text, explicitly add additional data, contextual information, decrypt data, specify tags, give notes, metadata, annotations, accent corrections, abbreviations, lowercase and uppercase letters, paleographic processing, ligatures, etc.

I did the automatic transcription after starting the training and testing. I used my own transcription model and started transcription using HTR +.

The result of learning in the automatic transcription of Andrej Kmeť's manuscript text was initially an excellent result of 1.37% in the training dataset and 1.76% in the test dataset (CER - Character Error Rates). The training set contained 29,411 words and 4,573 lines. I used the model for other sheets and corrected them so that they were in *ground truth* (GT) quality. An overview of experimental models is in Table 2.

In the process of getting acquainted with the Transkribus platform and despite my trial and error, I switched from an error rate of 22.81% in 2018 to an error rate of 1.76% in 2019 with an HTR machine. Transcription efficiency improved significantly after the HTR + machine became available. At first, I only worked with practice sets that were not in *ground truth* (GT) quality. The basic training transcribed set had 50 pages. I relatively easily enlarged this basic file to 185 pages by transcribing additional pages with the older model. I corrected them and added them to the training set. I tried to correct the new parties as accurately as possible into the quality of ground truth.

Finally, I created the mentioned model no. From the GT quality pages. 36009, which can achieve good to excellent transcription results depending on quality, scans, font sharpness, handwriting quality and segmentation quality.



Model 36009

Zbierka Andrej Kmeť

1-1 # Vysoko Ctený Pane!
1-2 # Prijmite vrelú vďaku za takú veľkú a
1-3 # vzornú prácu, podujatú pre mňa. Vášmu prie
1-4 # hľadu istotne svojho času bol by býval ~~poddačný~~ povďačný
1-5 # aj P. Dr. Niederle. Odpísal som si ju, a posielam
1-6 # ~~o uprimným~~ s úprimným poďakovaním ~~späť,~~ späť, aby ste ďalšie data
1-7 # zaznačovať mohli.
1-8 # Po nebohom p. barónovi Hoenningovi čo zbudlo
1-9 # zakúpili sme všetko za 400 ~~zl,~~ zl. rukopis, kresby, hrnce
1-10 # črepy, skameneliny, nástroje železné (a voľačo bron
1-11 # ~~cové),~~ cove), kamene ~~mlynské)~~ (mlynské) ba i všetky police a ~~stolíky.~~ stolíky
1-12 # Keby sa boli bližší za jeho života ~~obzreli,~~ obzreli mohli
1-13 # sme za polovičnú cenu viac kúpiť. ~~Dosiaľ~~ Dosiaľ, dá
1-14 # Boh, už môže byť všetko v Martine, lebo ako
1-15 # mi sdelil pán Ondrčka, 24ho ~~dal~~ Ddal všetko ~~ped~~ špedi
1-16 # terovi, značené číslami I-XXXVI. Nekoľko vtákov
1-17 # dostal som od mlad. p baróna do daru; i tie sú tam
1-18 # (Čo teraz kvitujú v Nár. ~~Nov.,~~ Nov, to sú zo Štiavnice) 150
1-19 # zl. dal som ~~hneď,~~ hneď; 250 zl. treba zložiť do konca
1-20 # ~~sept.~~ Sept. Vyše 50 listov už tašlo v záležitosti sbierky
1-21 # peňažnej, a ešte píšeme. Martinci tiež stva ~~zaháľajú~~ zahálajú
2-1 # naše domanické a medovarské. Bože môj, a koľko
2-2 # nálezísk, a s jak krásnymi nálezmi, prekutaných
2-3 # a snáď ~~úplne~~ úplné vybratých je na Slovensku, a
2-4 # z toho nemáme ničohúco nič! Budeme ~~želeť,~~ želeť
2-5 # ale keď bude pozde ~~Dobrá~~ dobrá príležitosť nenavráti
2-6 # sa rada po druhýkrát. Tu je doba, kde si ~~bude~~ ľubi
2-7 # jedon národ druhému ~~predchitávať~~ predchytávať a odorávať
2-8 # a my sme v tomto ohľade tí najnešťastnejší,
2-9 # vzorom opravdivým tej odranej liny ~~slovenskej!~~ slovenskej!
2-10 # „Čo by nás dneská stálo 5 zl. ~~zato~~ za to dáme zajtra
2-11 # 50 alebo i viac. Lebo nám nikto nič nedaruje,
2-12 # lebo my nevládzeme rozdávať "Bedem
2-13 # a sinecúry. A kto by aj daroval nám, ktorí
2-14 # sami za seba nechceme krížom slamy v tom
2-15 # to ohľade preložiť?!
2-16 # Vrelo pozdravujem Pána Seniora a Vašu
2-17 # milú Rodinku, a Vás bozkávajúc, ~~ostáva~~ ostával
2-18 # so srdečným pozdravom a rodáckou úctou
2-19 # v Prenčove 27. Sept. 894.
2-20 # Váš sluha a ctiteľ
2-21 # Andrej Kmeť
2-22 # ~~farár.~~ farár.

Figure 13 Comparison of automatic transcription with the correct text

Preliminarily, I can state that some of the transcription errors relate to punctuation. A detailed analysis of the causes of inaccuracies will be the subject of further research, as well as research into the correlation between scan quality and segmentation with respect to transcription quality.

Further research

In further research, it will be appropriate to focus on the following areas: a) selection and standard description of larger Slovak manuscripts of European and national significance, b) digitization of selected historical documents according to the experimental plan to confirm or improve known procedures and values with regard to the following segmentation process text and automatic transcription (correlation between different conditions and quality of scanning and transcription, c) thorough analysis and description of the results of text segmentation, d) sharing digital documents with archives and other institutions that will be able to use them at their discretion as a replacement for paper documents, e) creation of models, training and analysis of automatic transcription models according to modern and modern collections and languages (especially Slovak, Czech, Hungarian, Latin, German, Polish), f) verification and evaluation of usability of finished, available transcription models from research in READ project, g) to get acquainted with the best the most common practice of automatic recognition of texts of historical documents in Europe, especially in Germany, Austria, Spain, Hungary, Great Britain, Finland, the Netherlands, Serbia, the use of information and experience in Slovakia, h) automatic transcription of a substantial part of the manuscript collection and virtualization, virtual one digital presentation of volumes located in geographically diverse places (Slovak National Library in Martin, Slovak National Archive in Bratislava, University Library in Bratislava, Országos Széchenyi Konyvtár in Budapest), i) research into the possibilities of increasing the efficiency of recognition of manuscript texts and historical texts documents through the Transkribus system and related tools, j) making transcribed and interpreted collections available to the general public via a digital repository, k) creating documentation that will be used for archives, libraries, academic institutions as well as individuals for automatic transcription of texts.

Conclusion

The case study describes the experience with the Transcribus platform. The described experiments confirm that Slovak manuscripts and historical prints can be effectively and automatically transcribed, while the accuracy of transcription can be very good to excellent. The most important thing is to scan documents in the highest achievable quality on professional scanners. Another condition for the excellent quality of transcription is careful segmentation and preparation of a model with *ground truth* aspects. The transcription results are readable and can be exported in various formats - DOC, TXT, PDF, TEI, METS, further edited, edited, corrected, translated and otherwise used. In the experiment, I achieved an accuracy of 94.21% for Andrej Kmeť's handwriting and 5.79% for a character error rate (CER) of 5.79%. In the transcription of the printed fracture, I achieved an accuracy of 99.56% with a character error rate of 0.44%.

In terms of perception, understanding and use of the transcribed text in general, according to the authors of the Transcribus platform, the following applies: a) if the error rate of "words" is strictly calculated and if the error rate of words is up to 30%, then the text is still understandable and usable if the error rate of "characters" is strictly calculated, and if the error rate of characters is up to 15%, then the text is still understandable and usable for humans.

The Transcribus platform is a great tool for conscientious and patient researchers, who will be significantly facilitated by fine-tuning transcription through editing and

proofreading of results. The platform is not, and rarely will be, intended only for "clickers", ie users who are used to "clicking" more than innovating.

List of bibliographic references

[KATUŠČÁK, D. - NAGY, I. - BÔBOVÁ, M. - KUNC, P. - KURHAJCOVÁ, A. - MALINIAK, P. - MIKUŠKOVÁ, M., NIŽNÍKOVÁ, L. - POLÁKOVÁ, I. - SNOPKOVÁ, B. - TOMEČEK, O.] 2019. SKRIPTOR Projekt APVV-19-NEWPROJECT-17816 (2020-2024). *Inovatívne sprístupnenie písomného dedičstva Slovenska prostredníctvom systému automatickej transkripcie historických rukopisov.* [Innovative disclosure of written heritage of Slovakia through the automatic transcription of historical manuscripts]. Organizácie: Univerzita Mateja Bela v Banskej Bystrici (zodpovedný riešiteľ doc. Imrich Nagy PhD) a Štátna vedecká knižnica v Banskej Bystrici – partner (garant prof. PhDr. Dušan Katuščák, PhD

ADAM MATTHEW DIGITAL (2018), "Handwritten text recognition: artificial intelligence transforms discoverability of handwritten manuscripts", [cit. 2.10.2021]. Dostupné z: www.amdigital.co.uk/products/handwritten-text-recognition

HODEL, T. - SCHOCH, D. - SCHNEIDER, C. A. - PURCELL, J., 2021. Všeobecné modely rozpoznávania rukou písaného textu: uskutočniteľnosť a najmodernejšie. Nemecký Kurrent ako príklad. In: Journal Open Humanities Data, 7, s. 13. Doi: http://doi.org/10.5334/johd.46.

HOLLÝ, KAROL., 2013. *Veda a slovenské národné hnutie: snahy o organizovanie a inštitucionalizovanie vedy v slovenskom národnom hnutí v dokumentoch 1863 –1898.* Bratislava: Historický ústav SAV v Typoset Print s. r. o., 2013.

HOLLÝ, KAROL., 2015. *Andrej Kmeť a slovenské národné hnutie: Sondy do života a kreovanie historickej pamäti do roku 1914.* Bratislava: Veda – Historický ústav SAV, 2015. 279 s. ISBN 978-80-224-1480-7

HOTSON, HOWARD – WALLNIG THOMAS (eds.), 2019. *Reassembling the Republic of Letters in the Digital Age*. Göttingen : Göttingen University Press, 2019. 470 s. [COST Action IS1310; 2014 – 2018. ISBN: 978-3-86395-403-1. DOI: https://doi.org/10.17875/gup2019-1146. [cit. 3. 10. 2021]: Dostupné z: https://www.univerlag.uni-goettingen.de/handle/3/isbn-978-3-86395-403-1

KATUŠČÁK, D., 2008. Súčasný stav formovania stratégie digitalizácie na Slovensku. In: *Kolokvium knihovních a informačních pracovníků zemí V4+.* 6. – 8. července 2008, Brno, ČR. Elektronický zborník, s. 30-46.

KATUŠČÁK, D., 2021. Pochybná hodnota za veľa peňazí? In: *Kultúrny kyslík*. 2021, č. 2, s. 14-17. ISSN 1339-6919. [cit. 3. 10. 2021]. Dostupné z: kulturny_kyslik_2021_2.pdf (ikp.sk)

KATUŠČÁK, D. - KATUŠČÁK, M., 2011c. Základná koncepcia národného projektu digitálna knižnica. In: *Knižnica*, roč. 12, 2011, č. 2, s. 6-10. [cit. 2.10.2021] Dostupné z: Knižnica 2_2011.indd (snk.sk)

KATUŠČÁK, DUŠAN ET. AL., 2011a. *Digitálna knižnica a digitálny archív.* Národný projekt. Operačný program informatizácie spoločnosti OPIS2. Implementácia 2010-2015. Martin: Slovenská národná knižnica, 2011. [Kompletný projekt k žiadosti o nenávratný finančný príspevok zo štrukturálnych fondov Európskej únie ca 4000 s.]

KATUŠČÁK, DUŠAN, 2011b. Národný projekt digitálna knižnica a digitálny archív. In. *Bulletin Slovenskej asociácie knižníc*. Bratislava : SAK, 2011. 38 s. [Opis projektu] Dostupné na: http://dusan.katuscak.net/2011/12/02/digitalna-kniznica-a-digitalny-archiv-opis2/

KATUŠČÁK, DUŠAN, 2011d. Situační zpráva o národním projektu SNK Digitální knihovna a digitální archiv. In: *12. konference Archivy, knihovny, muzea v digitálním světě 2011.* Praha : SKIP, 30. listopadu a 1. prosince 2011 v konferenčním sále Národního archivu v Praze, Archivní 4, Praha 4 - Chodovec. [cit. 2.10.2021] Dostupné z: http://old.skipcr.cz/dokumenty/akm-2011/Katuscak.pdf

KIŠŠ, MARTIN, 2018. *Rozpoznávání historických textů pomocí hlubokých neuronových sítí.* Brno, 2018. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Michal Hradiš, Ph.D.

KIŠŠ, M. - HRADIŠ, M. - KODYM, O. 2019. Brno Mobile OCR Dataset. V: Medzinárodná konferencia o analýze a uznávaní dokumentov (ICDAR) 2019. September 2019, s. 1352 – 1357. ISSN: 1520-5363. [1907.01307] Brno Mobilný OCR Dataset (arxiv.org)

MARTÍNEK, J. - LENC, L. - KRÁL, P., 2020. *Building an efficient OCR system for historical documents with little training data.* Neural Comput & Applic 32, 17209–17227 (2020). [cit. 2.10.2021] https://doi.org/10.1007/s00521-020-04910-x

MKSR, 2019. Revízia výdavkov na kultúru. Priebežná správa. Október 2019 Kap. 4.4 Projekt digitalizácie, s. 75-78. [cit. 2.10.2021] Dostupné 28.9.2021: Revizia_vydavkov_na_kulturu_priebezna_sprava_compressed.pdf (gov.sk)

MKSR, 2020. Revízia výdavkov na kultúru. Záverečná správa. Júl 2020. Kap. 4.9 Digitalizácia kultúrneho dedičstva, 132-139. [cit. 2.10.2021] Dostupné 28.9.2021: Revizia_vydavkov_na_kulturu_-_zaverecna_sprava_compressed.pdf (gov.sk)

MÜHLBERGER, G., 2016. READ (Recognition and Enrichment of Archival Documents) - 2016-2019. [Projektová štúdia]. [cit 6.10.2021.] Dostupné z: https://www.academia.edu/22653102/H2020_Project_READ_Recognition_and_Enrichment_of_Archival_Documents_-_2016-2019

MÜHLBERGER, 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. - ZAGORIS, K., 2019. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5), 954–976. DOI: https://doi.org/10.1108/JD-07-2018-0114

MÜHLBERGER, G. - ZELGER, J. - SAGMEISTER, D., 2014. User-driven correction of OCR errors: combining crowdsourcing and information retrieval technology. In: Antonacopoulos, A. & Schulz, K. U. (Eds.), DATeCH'14: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, Madrid, Spain, 19–20 May 2014 (pp. 53–56). New York, NY: Association for Computing Machinery. DOI: https://doi.org/10.1145/2595188.2595212

MÜHLBERGER, G. - COLUTTO, S. – KAHLE, P., [2016 Preprint] *Handwritten Text Recognition (HTR) of Historical Documents as a Shared Task for Archivists, Computer Scientists and Humanities Scholars. The Model of a Transcription & Recognition Platform (TRP).* [cit. 5.10.2021]. Dostupné z: Günter Mühlberger | Univerzita v Innsbrucku - Academia.edu

POOLE, ALEX H., 2017. The Conceptual Ecology of Digital Humanities. In: *Journal of Documentation*, roč. 73, 2017, č. 1, s. 91 – 122. [cit. 3. 10. 2021]: Dostupné z: https://www.academia.edu/27862789/The_Conceptual_Ecology_of_Digital_Humanities

STROBEL, P.B. - CLEMATIDE, S. - VOLK, M., 2020. How Much Data Do You Need? About the Creation of a *Ground truth* for Black Letter and the Effectiveness of Neural OCR. In: *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020),* pages 3551–3559 Marseille, 11–16 May 2020 c European Language Resources Association (ELRA)

Acknowledgments:

PhDr. Maria Halmová, Mgr. Viera Varínská and PhDr. Anne Peťová for helping to scan the manuscripts of Andrej Kmeť at the Ethnographic Museum in Martin

Olga Kuchtová from Banská Štiavnica for help in finding out information about the life and conditions of Andrej Kmeť's work in Prenčov

Mgr. Maria Bobova, PhD. from the State Scientific Library in Banská Bystrica for help and cooperation in manual transcription and segmentation of pages for the training model and transcription of letters by Andrej Kmeť

Lucia Valjentová, a student of librarianship from the 4th year of the Institute of Czech Studies and Librarianship of the University of Silesia in Opava for help in transcribing a Czech fracture

Aleš Drahotušský for providing a newspaper from the Digital Library of the State Scientific Library in Ostrava