SLOVO, sv. 72 (2022), 181–192, Zagreb 2021.

Achim RABUS
University of Freiburg
achim.rabus@slavistik.uni-freiburg.de

UDK 003.349.1:004.021
004.352.243
DOI: https://doi.org/10.31745/s.72.5
Izvorni znanstveni članak
Primljen: 23. studenog 2020.
Prihvaćen: 12. listopada 2021.

HANDWRITTEN TEXT RECOGNITION FOR CROATIAN GLAGOLITIC

The paper presents and discusses recent advances in Handwritten Text Recognition (HTR) technologies for handwritten and early printed texts in Croatian Glagolitic script. After elaborating on the general principles of training HTR models with respect to the Transkribus platform used for these experiments, the characteristics of the models trained are discussed. Specifically, the models use the Latin script to transcribe the Glagolitic source. In doing so, they transcribe ligatures and resolve abbreviations correctly in the majority of cases. The computed error rate of the models is below 6%, real-world performance seems to be similar. Using the models for pre-transcription can save a great amount of time when editing manuscripts and, thanks to fuzzy search (keyword spotting), even uncorrected HTR transcriptions can be used for various kinds of analysis. The models are publicly available via the Transkribus platform. Every scholar working on Glagolitic manuscripts and early printings is encouraged to use them.

Key words: Handwritten Text Recognition, Glagolitic script, Digital Humanities, manuscripts, early printings

1. INTRODUCTION

Due to new technological developments, both the Digital Humanities (DH) and research on Artificial Intelligence (AI) have made considerable progress in recent times, making them both relevant to historical and philological disciplines (VAN LIT 2020). One highly promising combination of these two fields is Handwritten Text Recognition (HTR). This AI-supported technology allows the development of models that are capable of transcribing diverse historical and contemporary handwritten scripts and handwriting styles with

an error rate of typically well below 10%, sometimes below 5%. This makes them an important prerequisite for mass digitization and a valuable tool for the pre-transcription of manuscripts intended to be used for traditional philological text editions.

This paper is structured as follows: first, I report on the basic principles of HTR technology and the Transkribus application (TRANSKRIBUS Team at University of Innsbruck 2020). I then elaborate on the specifics of training the HTR models for Glagolitic. The subsequent section is devoted to the application of the models to different sources and a discussion of the results obtained. I conclude the paper with an outlook on the opportunities and limits of Glagolitic HTR for future research.

2. HTR TECHNOLOGY AND TRANSKRIBUS

Computer-assisted Handwritten Text Recognition is a considerably more complex task than the traditional Optical Character Recognition (OCR) used for modern printed texts: as opposed to modern printed texts, handwritten texts contain ample variation among letterforms even within the handwriting of one scribe, not to mention between different scribes. Moreover, the letterforms differ depending on their position within the word. Many handwriting styles are cursive, thus further complicating the computer's task of recognizing individual letters.

In order to tackle these issues, HTR technologies systematically take into account not only individual letters or glyphs, but also the neighboring letters, words, and even entire lines. Compared to traditional OCR technologies, the line-based approach yields a considerably lower Character Error Rate (CER). HTR is based on AI technologies, specifically on neural networks (STRÖBEL; CLEMATIDE; VOLK 2020; INZAUGARAT 2018). These neural networks need to be trained using high-quality images and corresponding diplomatic transcriptions for each line of the handwritten text in the image. This means that training HTR models is an instance of supervised machine learning.

There are several HTR engines and applications on the market, both open- and closed- source, e.g., kraken (KIESSLING 2019), tesseract (KAMLAH; WEIL 2020), or HTR+ and PyLaia, which are integrated into Transkribus. The use of the most HTR engines requires advanced IT knowledge, such as familiarity with command line interfaces, Phyton or the ability to install packages on webservers, rendering them rather unusable for the average humanities scholar

with no IT background. The software package Transkribus is arguably the most user-friendly HTR application currently on the market. It can be installed on all major platforms (Windows, Macintosh, Linux) and features a rather self-explanatory graphical user interface (GUI);[1] most significantly, numerous HTR models for diverse scripts and handwriting (and printing) styles have already been made publicly available. According to READ-COOP (www.readcoop.eu), the European cooperative behind Transkribus, as of February 2021, the Transkribus community has grown to 50,000 registered users from all over the world. While most active users are primarily interested in Western languages and scripts (such as German, Dutch, or English), the number of scholars concerned with historical Slavic documents is increasing.[2]

If no public models for the script and handwriting style in question are available, one needs to train one's own model. Model training in Transkribus is rather straightforward. One needs a certain amount of images with corresponding transcriptions, the so-called 'ground truth'. According to the Transkribus FAQ (https://transkribus.eu/wiki/index.php/Questions_and_Answers), 15,000 transcribed words are sufficient for a first model. In my experience, however, depending on the complexity of the handwriting, decent results can be achieved starting from around 5,000 words of training data.

As soon as the training data are available, the training process is initiated manually. It takes place remotely on the Transkribus servers (physically located in Austria), which means that users do not need a powerful workstation at their disposal to initiate model training. Any business or consumer computer is sufficient to initiate the training process. During training,[3] the algorithm compares the visual information of the handwritten lines with the corresponding transcriptions multiple times. After numerous epochs (Brownlee 2018 discusses terminology with respect to training neural networks), the model learns to identify the specifics of the handwriting and reaches a certain CER. As one can see in the following figure, during the first epochs, the CER drops rapidly, while it takes many additional epochs to reach the lowest possible CER. A typical curve visualizing the training of an HTR model has a hyperbolic shape.

[1] A browser-based lite-version is also available at URL: http://transkribus.eu/lite.

[2] See RABUS 2019 and RABUS.a for an overview of currently available HTR models for different types of (early) Slavic handwriting and RABUS.b for the training of generic models.

[3] Depending on the size of the training data and server load, training of a typical HTR model takes from less than one hour to more than 24 hours.

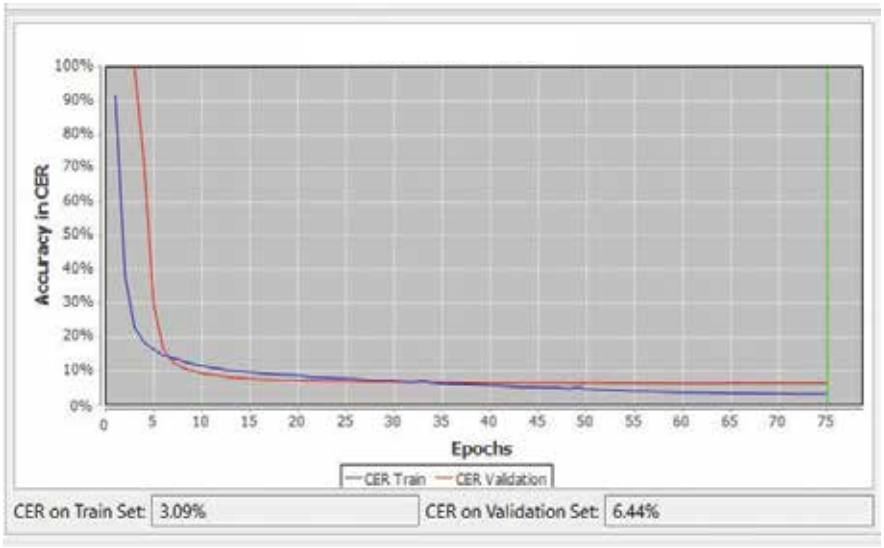

Figure 1: Learning curve of an HTR model in Transkribus
Slika 1. Krivulja strojnoga učenja HTR modela u Transkribusu

3. TRAINING HTR MODELS FOR GLAGOLITIC HANDWRITING AND (EARLY) PRINTING

In order to train models for Croatian (handwritten) Glagolitic, several factors must be taken into account. First, Glagolitic handwriting is rather complex, featuring numerous ligatures and, more importantly, abbreviations. Second, scholars working with Glagolitic sources typically use a transliteration into Latin script. Moreover, they tend to add editorial emendations such as the resolutions of abbreviations in parentheses. A good HTR model for Croatian Glagolitic should be able to transliterate into Latin script and, ideally, make informed guesses about how to resolve abbreviations. This means that the models should have certain 'smart' capabilities and imitate some kind of philological intuition and intelligence.

In order to achieve these goals, a large amount of training data is needed. I chose to follow the 'recycling approach' – instead of transcribing numerous pages from scratch, I used pre-existing transcriptions of Glagolitic manuscripts, uploaded the images to Transkribus, ran layout analysis for line

segmentation and pasted the corresponding transcriptions into the program.[4] For the handwritten model, I used the transcription of the first part of the Second Beram Breviary, provided by Sanja Zubčić and the Breviary of Vid Omišljanin provided by Jagoda and Guido Kappel.[5] The manuscripts were written by different hands, meaning that the model is expected to cope with different handwriting styles sufficiently well. The model for handwritten Glagolitic can be found in Transkribus under the name *Handwritten Glagolitic*.

The ground truth – the training data for the handwritten Glagolitic model – totals 170,000 word tokens in size, making it a medium-sized model (some of the large public models within Transkribus have more than a million tokens, while others have considerably fewer). If the manuscript the model is to transcribe differs significantly from the handwriting style seen during model training, results may be unsatisfactory. The model has a computed CER of 5.73% meaning that roughly six out of every 100 letters are transcribed incorrectly. Since this includes incorrectly resolved abbreviations and punctuation marks, the real-world performance of the model may even be slightly better.

Although the HTR technology was primarily developed with the goal of deciphering handwritten texts, the AI-boosted line-based approach can be successfully applied to (early) printed sources as well. Since the letters in early printings exhibit less variation than handwritten glyphs (albeit more than the letters used in modern printing), it is possible to achieve a lower CER with a smaller amount of training data. Taking these factors into account, we trained a model for early printed Glagolitic texts, mainly the Urach-Tübingen texts (VORNDRAN 1977). Since no previous transcriptions were available to us, we used the handwritten Glagolitic model for a pre-transcription and manually corrected the errors to create a sufficient amount of ground truth data. The printed Glagolitic model has total of 28,000 tokens and a CER of 3.51%.[6]

---

[4] I would like to thank my student assistants Stefanie Anemüller, Eleonora Hermes-Krukenberg, Clara Lietzmann, and Elena Renje as well as Richard Dean for helping create the ground truth for the models.

[5] The transcription of the first part of the Second Beram Breviary was provided thanks to Milan Mihaljević, leading researcher of Research Centre of Excellence for Croatian Glagolitism, see also URL: https://zci.stin.hr and URL: https://beram.stin.hr. I would like to express my sincere gratitude to all the mentioned colleagues for kindly providing me with their transcriptions. Without their valuable help, it would not have been possible to train HTR models for Glagolitic.

[6] The model is the result of a collaboration between the Department of Slavic Studies at the University of Freiburg and the University Library Tübingen.

In the following section, I assess the real-world performance of the respective models using an array of different sources.

4. APPLICATION OF THE HTR MODELS

The first example to assess the real-world performance of the handwritten Glagolitic model is taken from the First Beram Breviary, f. 9r.

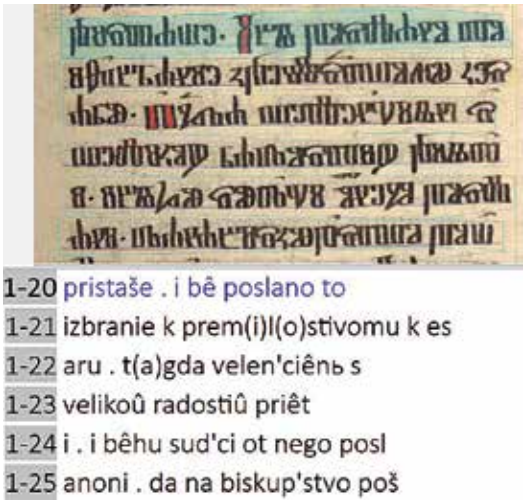

Figure 2. Transcription quality: The First Beram Breviary, general quality
Slika 2. Kakvoća preslovljavanja: Prvi beramski brevijar, opća kakvoća

As one can see, the overall transcription quality is decent. Most letters are recognized correctly, ligatures such as *pr-* (1-21, 1-23), *-gd-* (1-22) or *-tv-* (1-25) are obviously unproblematic for the model. Abbreviations are also resolved, correctly in *t(a)gda* (1-22), but slightly incorrectly in *prem(i)l(o)-stivomu* (1-21), which should have been *prem(i)l(o)st(i)vomu*. The superscript mark representing the front *yer* seems to cause the model problems: sometimes it is rendered correctly, such as in *sud'ci* (1-24), but sometimes it is omitted such as in *priêt(')* (1-23). Apart from this, the main errors in this section are in the area of word separation and hyphenation.[7] Apparently, the model did

[7] This seems to be typical for HTR models and holds also true for the models for Cyrillic Church Slavonic (RABUS 2019).

not see *kes-aru* (1-21f.) often during training. This resulted in the model confusing the first letter of the word with the preposition *k* (which it had correctly separated before in line 1-21) and failing to add a hyphen at the end of the line. Remarkably, there is an interesting hypercorrect error at the beginning of line 1-25. While the correct rendition should be *posl-ani*, the model transcribed this passage with *posl anoni*, adding the letters *o* and *n* for no obvious reason. Apparently, the fact that the model is 'smart' insofar as it has learned to expand abbreviations comes at the price of occasional hypercorrect additions of unnecessary letters.

In light of this, one must wonder whether the advantages of the 'smart' capabilities of the model actually outweigh its disadvantages. In order to assess this issue, the following section from the First Beram Breviary (98v) serves as an example:

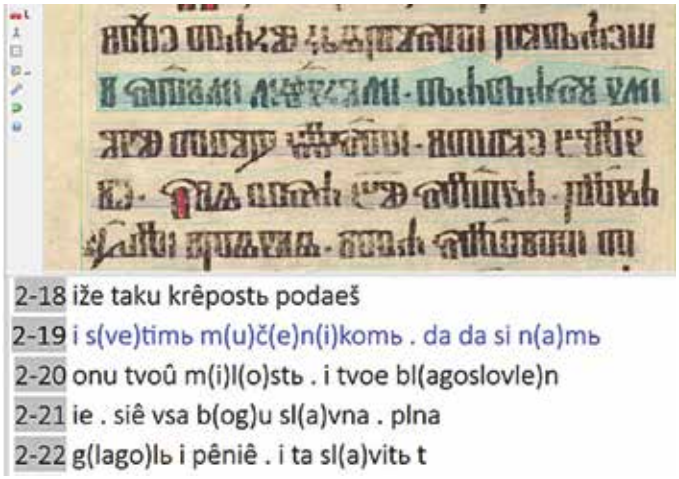

Figure 3. Transcription quality: The First Beram Breviary, abbreviation
Slika 3. Kakvoća preslovljavanja: Prvi beramski brevijar, razvezivanje kratica

Numerous abbreviations appear in this section. Sometimes, the model had to add just one letter and the corresponding parentheses, such as in *sl(a)vna* (2-21); in other cases, it had to add several individual letters in the middle of a word separated by other letters (such as in *m(u)č(e)n(i)komь*, 2-19); finally, it had to add numerous coherent letters in the correct sequence (such as in

bl(agoslovle)n). Overall, it appears that the model copes rather well with the task of adding the correct expansions of the abbreviations.

In situations where the handwriting is slightly different and the contrast of the manuscript is worse, the model's capability to expand abbreviations correctly deteriorates. This becomes obvious when looking at a section of another manuscript, the First Vrbnik Breviary (1v):

Figure 4: Transcription quality: The First Vrbnik Breviary, errors with abbreviations
Slika 4. Kakvoća preslovljavanja: Prvi vrbnički brevijar, pogreške pri razvezivanju kratica

While *b(la)ž(e)nu* (2-7) is correct, *sl(o)v(e)n-* (2-7) is not. Moreover, the model did not attempt to expand *dvu* (2-8) or *istice* (2-9, which should have been *i s(ve)tice*). Nevertheless, even here, most of the expansions are correct. It is consequently reasonable to assume that the 'smart' capabilities are advantageous and outweigh the detrimental effects of hypercorrect additions of letters. However, a broader, quantitative study would be necessary to analyze this question in greater detail.

As mentioned before, the printed Glagolitic model (available in Transkribus under the name *Glagolitic printings*) has been trained using a smaller (and somewhat less consistent) amount of ground truth. For this reason, the overall real-life performance of the model is not significantly better than the model for handwritten Glagolitic, which one might assume due to the fact that the letters of early printed sources are easier to read and more regular than those in handwritten sources. The following is an example of the Catechism of 1561 (from the *Symbolum Nicaenum*):

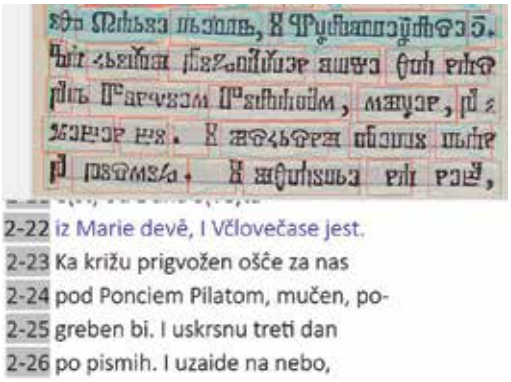

Figure 5. Transcription quality: Catechism (1561)
Slika 5. Kakvoća preslovljavanja: Katekizam (1561.)

The model copes well with ligatures such as in *treti* (2-25) or *uzaide* (2-26). However, it omits the *l* in *Včloveč(a)l se* (2-22), confusing it with *a*. Nevertheless, the text is easily readable.

The next example is from Trubar's translation of the New Testament (1562/1563), 1st Corinthians 13:1–3:

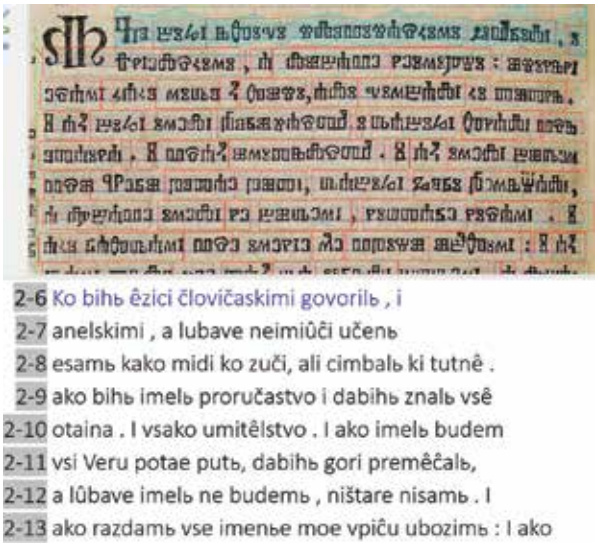

Figure 6. Transcription quality: printed New Testament (1562/1563)
Slika 6. Kakvoća preslovljavanja: tiskani Novi testament (1562./1563.)

As is apparent, initials are usually not recognized; this is typical for automatic HTR, because the algorithms are unable to recognize correctly that the initial letter belongs to the line in question. Apart from this, the recognition quality of the text is decent, with some errors possibly due to inconsistent transcription in the training data (2-7 *anelskimi*), while the reasons for other errors (2-7 *učenь*) remain unclear. Apart from the overall good performance with respect to ligatures, the 'smart' capabilities of this model cannot be evaluated using this passage, since there are no abbreviations. Generally, there are considerably fewer abbreviations in the printed sources than in the handwritten ones.

5. CONCLUSION AND OUTLOOK

The analysis of the real-world performance of HTR models for both handwritten and printed Glagolitic has shown that, although the models are far from producing error-free results, they are actually usable and can save considerable time and money if used for pre-transcription in editorial projects.[8] This holds even though Transkribus has recently switched to a freemium business model, meaning that every user can transcribe around 400 pages for free; afterwards, they will be charged per page transcribed by an HTR model (see https://readcoop.eu/transkribus/credits/). Nevertheless, these costs are incomparably lower than having to transcribe hundreds of pages manually. It may well be possible that the incomparably lower costs of HTR as opposed to manual labor will make the difference as to whether or not a project can be realized at all (RABUS.a). HTR technology may even open up new, previously unknown research opportunities. Since the correct transcriptions of each individual word are often saved internally even in those cases where the final transcription provided by the model is incorrect, the correct transcriptions can be found using the keyword spotting feature implemented in Transkribus. This leads us to recognize the fact that, in the digital age, absolute precision in transcriptions, while still desirable, is not always a necessity. It is possible to work with automatically transcribed texts without post-correction in a quantitative paradigm (RABUS; PETROV). The

[8] To provide an example, Transkribus has been successfully used to produce a pre-transcription of the Glagolitic editio princeps 1483 prepared by staff members of the Old Church Slavonic Institute.

remaining errors produce noise in the data, but do not make linguistic (and other kinds of) analysis impossible.

The handwritten and printed Glagolitic models are publicly available via the Transkribus platform. They can be used free of charge (with the restrictions mentioned above). I would like to encourage any scholar who studies Glagolitic cultural heritage to make ample use of these models and to contact me without hesitation if technical difficulties arise. It is in our common interest to use the recent advances in the Digital Humanities to "revolutionize access to handwritten documents" (www.readcoop.eu). I hope to have shown that, for Glagolitic cultural heritage, there are now tools available that deserve to be tested.

LITERATURE

BROWNLEE, J. 2018. What is the Difference Between a Batch and an Epoch in a Neural Network? URL.: https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/ (17. 4. 2019.)

INZAUGARAT, E. 2018. Understanding Neural Networks: What, How and Why? – Towards Data Science. URL.:https://towardsdatascience.com/understanding-neural-networks-what-how-and-why-18ec703ebd31 (15. 4. 2019.)

KAMLAH, J.; S. WEIL. 2020. Automatische Texterkennung von Druckwerken mit Tesseract. http://zenodo.org/record/3734046#.YboDr71Kilo (3. 4. 2021)

KIESSLING, B. 2019. Kraken – an Universal Text Recognizer for the Humanities. *Digital Humanities Conference*, Utrecht, 9–12 July, 2019. URL.: https://dev.clariah.nl/files/dh2019/boa/0673.html.

MIHALJEVIĆ, M. (ed.). 2018. *Drugi beramski brevijar: Hrvatskoglagoljski rukopis 15. stoljeća.* V. Badurina Stipčević, I. Botica, M. Dimitrova, M.-A. Dürrigl, I. Hristova Šomova, K. Kuhar, M. Mihaljević, S. Požar, A. Radošević, A. Šimić, M. Šimić, J. Vela, J. Vince, J. Vučković, S. Zubčić, M. Žagar (trans.). Zagreb: Staroslavenski institut.

RABUS, A.a. Automatische computergestützte Transkription paläoslavistischer Quellen und ihre Folgen für Korpuslinguistik und Editionsphilologie. A. M. Bruni; V. S. Tomelleri; G. Ziffer (eds.). *Humboldt-Kolleg 2020 Venedig: Proceedings*. (accepted for publishing)

RABUS, A.b. Training Generic Models for Handwritten Text Recognition Using Transkribus: Opportunities and Pitfalls. *Proceedings of the Dark Archives Conference*. Oxford. (accepted for publishing)

RABUS, A. 2019. Recognizing Handwritten Text in Slavic Manuscripts: A Neural-Network Approach Using Transkribus. *Scripta & e-Scripta 19*: 9–32.

RABUS, A.; I. PETROV. Linguistic Analysis of Church Slavonic Documents – a Mixed-Methods Approach. *Scando-Slavica* (accepted for publishing)

STRÖBEL, P. B.; S. CLEMATIDE; M. VOLK. 2020. How Much Data Do You Need? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 3551–59. URL.: https://www.aclweb.org/anthology/2020.lrec-1.436.pdf (2. 11. 2020.)

TRANSKRIBUS Team at University of Innsbruck. 2020. Transkribus. URL.: https://transkribus.eu/Transkribus/. (15. 10. 2020.)

VAN LIT, L. W. C. 2020. *Among Digitized Manuscripts: Philology, Codicology, Paleography in a Digital World.* Handbook of oriental studies Handbuch der Orientalistik. Section one, The Near and Middle East volume 137. Leiden – Boston: Brill.

VORNDRAN, R. 1977. *Südslawische Reformationsdrucke in der Universitätsbibliothek Tübingen: Eine Beschreibung der vorhandenen glagolitischen, kyrillischen und anderen Drucke der Uracher Bibelanstalt.* Contubernium 24. Tübingen: Mohr.

Sažetak

Achim RABUS

STROJNO PREPOZNAVANJE RUKOPISNOG TEKSTA
ZA HRVATSKU GLAGOLJICU

U radu se predstavljaju nedavni pomaci u tehnologiji prepoznavanja rukopisnoga teksta (HTR) namijenjenoj hrvatskoglagoljskim rukopisnim i ranim tiskanim knjigama. Nakon opisivanja općih načela strojne obuke HTR modela, iznose se značajke načela strojnoga učenja u platformi *Transkribus*, pogotovo modeli korištenja latinice u preslovljavanju glagoljskih tekstova. Pri tome se u većini slučajeva ispravno preslovljavaju ligature i razrješuju kratice. Dobivena čestota pogrešaka je manja od 6%, poput uobičajene čestote pogrešaka kada preslovljavanje provode stručne osobe. Primjena HTR modela u prvom stadiju preslovljavanja može uštedjeti puno vremena pri pripemi i uređivanju rukopisa za objavu, zahvaljujući pretraživanju (pretrazi po ključnim riječima), pa čak i neispravno HTR preslovljavanje može biti korišteno za različite raščlambe. Modeli su javno dostupni posredstvom platforme *Transkribus*. Potičemo sve znanstvenike koji obrađuju glagoljske rukopise i rane tiskane knjige da se njima koriste.

Ključne riječi: strojno prepoznavanje rukopisnoga teksta, glagoljica, digitalna humanistika, rukopisi, rane tiskane knjige

Achim RABUS
University of Freiburg
achim.rabus@slavistik.uni-freiburg.de