Submitted on 2025, April 4th
Accepted on 2025, July 31st

"Studi Slavistici", XXII, 2025, Special Issue: 35-54
DOI: 10.36253/Studi_Slavis-17690
ISSN 1824-7601 (*online*)
Articles

Achim Rabus
Martin Meindl

The Digital Revolution in Slavic Manuscript Studies: HTR Technology and its Impact on Philological Research

1. *Introduction: Philology in the Digital Age*

Like society as a whole, scholarship has entered the digital age quite some time ago. Computers and post-PC devices such as smartphones or tablets are ubiquitous and have become indispensable in our daily work, social interaction, and cultural life. Activities in the field of Paleoslavic studies or historically oriented philology in general, such as the preparation of editions, are also extensively carried out with the help of computers. However, in philology, computers are rarely used to their full potential. The original source – for instance, a historical Slavic manuscript – is indeed digitized using a computer for preparing an edition, but in many cases only for the purpose of subsequent 're-analogization' in the end product of a traditional book. The main advantages of digital data and media, such as searchability, statistical analysis, or automation of certain processes, are often not utilized, or the texts are frequently not prepared in a way that makes them machine-readable and analyzable. This wastes significant potential for a substantial improvement in the quality of philological and linguistic research. In this context, there is sometimes a rather defensive attitude towards digitization concepts in our field, for example – partly justifiably – with regard to the sustainability of digital storage formats. Digitization components in some projects appear more as a necessary evil that has been more or less organically attached to the project plan to meet the requirements of third-party funders, rather than as an integral component that can significantly improve research quality (see **FIGURE 1**).

The intention here is not to proclaim the end of book publications, including book editions. These clearly have their legitimacy in the digital age for a variety of reasons. Rather, the appeal is to also utilize the advantages of digital techniques. This can be achieved, on the one hand, by exploiting the inherent fluidity of digital and online forms of publishing: deliberately releasing beta versions of digital editions – without claiming to have already eliminated all errors – before a possible final book publication. The text is thus accessible to the scholarly public much earlier, the international professional community can immediately work with it and provide feedback on editorial principles and errors; the hermeneutic revision of the edition can thus be partly accomplished through crowdsourcing (an example of this is <http://dhcrowdscribe.com/>, last access: 31.03.2025). On the other hand, an intelligent digital secondary use of editorial data is also advisable (Cleminson 2008, who references a statement made by David Birnbaum in 1995). Data that is primarily used for the creation of book publications can be reused, for example, for the creation

Achim Rabus (University of Freiburg) – achim.rabus@slavistik.uni-freiburg.de
Martin Meindl (University of Freiburg) – martin.meindl@slavistik.uni-freiburg.de
The authors declare that there is no conflict of interest

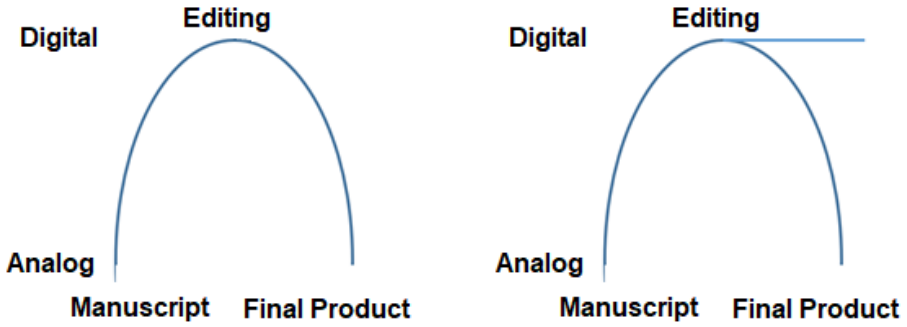


FIGURE 1
Traditional workflow in philology with digital elements only in processing, not in the end product vs. desirable workflow with digital secondary use

of web corpora (Waldenfels, Rabus 2015) or the training of HTR models (Rabus 2019). Successful examples of this approach are <http://www.vmc.uni-freiburg.de> (last access: 31.03.2025) or parts of the *Russian National Corpus* (<http://ruscorpora.ru/new/search-old_rus.html>, last access: 31.03.2025). The great advantage of this digital secondary use is that additional corpus-linguistic investigations are possible with relatively little extra effort compared to the book publication. This allows historically oriented philology to connect with the quantitative turn, which is becoming increasingly prevalent, particularly in synchronic linguistics. The quantitative approach is token-based, not type-based (Levshina 2019), and categorical perspectives are giving way to probabilistic ones. It also allows for the processing of much larger quantities of data than traditional close reading approaches. This allows for a more general overview of a given corpus rather than the zoomed in, detailed analysis of small amounts of text. It follows that within this paradigm, one hundred percent precision and error-free results, the classic goal in traditional editorial philology, while still desirable and worth striving for, are no longer necessarily viewed as the only and most important goal under all circumstances (see also Piper 2020). Possible inaccuracies in the edition or corpus do increase noise in the quantitative paradigm as well, but they do not necessarily impair the overall quantitative-probabilistic statement of the respective investigation (see also Rabus 2024), especially since several statistical methods have been shown to handle noisy data well and do not reflect the errors produced by HTR engines in the results provided (Eder 2013, Franzini *et al.* 2018, Rabus, Thomson 2023, Polomac, Rabus 2025). To conduct such quantitative investigations, the Paleoslavic community needs a broad foundation of digitally available texts and corpora, which ideally should be created efficiently and cost-effectively. We therefore argue that in order to create the necessary data for proper quantitative analysis at a reasonable price and pace, a higher tolerance towards noisy data has to be adopted.

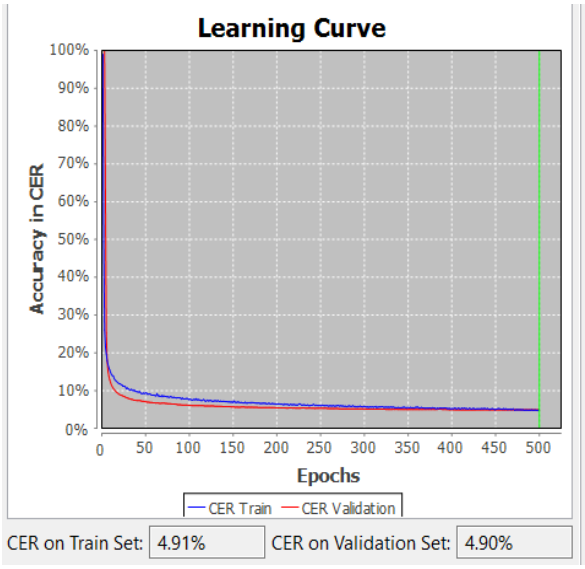


FIGURE 2
Typical learning curve of an HTR model

2. *Mass Digitization and Handwritten Text Recognition*

Promising developments in terms of mass digitization have emerged in recent years in the field of Handwritten Text Recognition (HTR). This technology automatically generates searchable and further processable transcriptions from digitally available images of manuscripts and printings, is based on artificial intelligence methods, and, unlike traditional text recognition methods such as classical Optical Character Recognition (OCR), takes the immediate surroundings within the line to be recognized – i.e., neighboring letters and words – into account during the transcription process. Before manuscripts can be recognized by HTR programs, a so-called model needs to be trained. This model is based on a set of training data which consists of digital images of manuscripts and corresponding diplomatic transcriptions created and checked by human experts. These data are provided to the HTR program in a process of supervised learning for model training. The program trains its recognition performance in numerous iterations (epochs), each time comparing the transcription hypothesis created by the program with the manually created transcription, thus improving the recognition performance. This process corresponds to the typical procedure of training artificial neural networks (**FIGURE 2**. See also Rabus 2019 for a more detailed description of the training process).

Depending on the amount and quality of training data (or Ground Truth, GT), modern HTR programs can achieve a Character Error Rate (CER, i.e., the number of characters still incorrectly recognized after the completion of training) of less than 5% by the end of

the training process for handwritten sources, and significantly lower for old printings[1]. For complex or heterogeneous manuscripts, a CER between 5 and 10% is realistic. However, for a variety of reasons, using the CER of a given model to gauge the model's performance on a certain manuscript is not always straightforward. First, the CER is calculated by taking into account parts of the training data which are set aside during the training (the so-called Validation Set), meaning the model did not see them during training, so they can then be used to evaluate the model's capabilities. As a result, the CER on Validation Set is not a reliable predictor of how well the model performs on other texts that were not part of the training data, especially in cases where their handwriting style or linguistic profile differs from the training data. Second, every error, be it word separation, confusion of similar letters or punctuation, are counted as the same. The question of which errors are more important or more detrimental than others is of course not universal and can vary from project to project. It is therefore important to investigate the models' performance on other texts. Numerous examples of such an analysis in Slavic and non-Slavic contexts can be found in Burlacu, Rabus (2021), Polomac (2022), Polomac *et al.* (2023) Rabus (2019), Rabus (2022) and Rabus (2024), Thompson (2021). Additional examples are shown in **FIGURE 3**.

Here we can see the automatic transcription of a manuscript written in Russian *Poluustav* from the 16th century (*Minei-Čet'i, fevral'* – Moscow, GIM, Sin. 179, f. 610v.). The model for this script type is of high quality, the transcription performance is comparatively good, and errors are mostly limited to word separation (1-4 добрыѧ) and superscript letters (1-3 соуцїй, 1-4 добрыѧ), even in other documents. While the error rate is low it has to be mentioned that, depending on your research question, an error like 1-3 соуцїй instead of соуцїи, which breaks the case alignment with людѣ, might be detrimental for certain research questions.

Besides these specific models we also trained a generic model capable of transcribing other Cyrillic script types such as *Ustav*. The quality does not quite match that of the specialized *Poluustav* model, especially in the area of superscript letters. However, the results are definitely usable as a pre-transcription, as the example from the *Služebnik Varlaama Chutynskogo* (f. 10r) from the 12th/13th century shows (see **FIGURE 4**).

For writing styles that trend towards Skoropis' (**FIGURE 5**), a comparatively low error rate can also be achieved, although models with a smaller scope can hardly be applied to texts with a different writing style than the training data. Still, such models generally produce usable pre-transcriptions, so here too, completely manual transcription would be more time-consuming and financially costly than error correction of the computer-assisted pre-transcription which only contains few errors (1-3 и лѣта, 1-4 мужь нѣкии).

[1] The quality of the training of such models is strongly dependent on the amount of available training data. The hyperbolic progression, i.e., strong decrease in CER in the early epochs and only slight improvements in later epochs, is typical. Overall, even in ideal situations, a CER of 0% is not achieved.

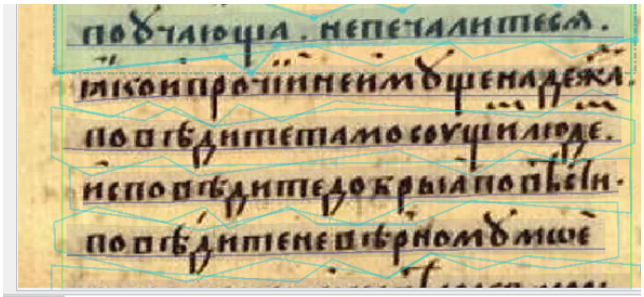


FIGURE 3
Example of the transcription performance of the HTR model VMC_Test-4+ (173,287 tokens, 3.83% CER on Validation Set)

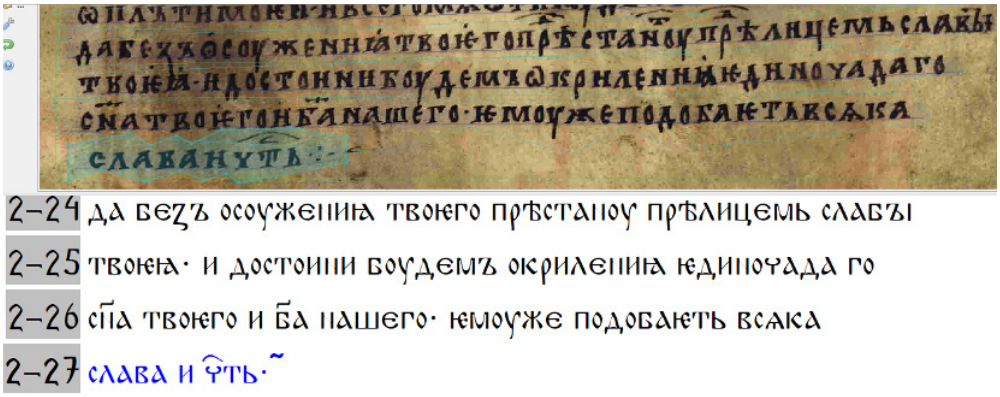


FIGURE 4
Ustav from the 12th-13th century. Example of the transcription performance of the model Combined_Full_VKS_2 (393,079 tokens, 3.94% CER on Validation Set)

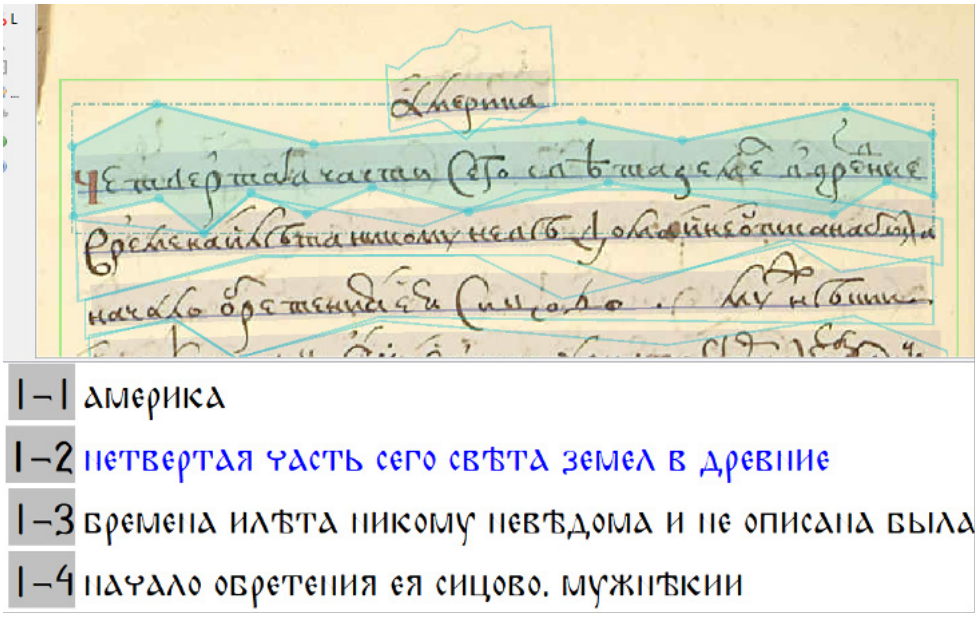
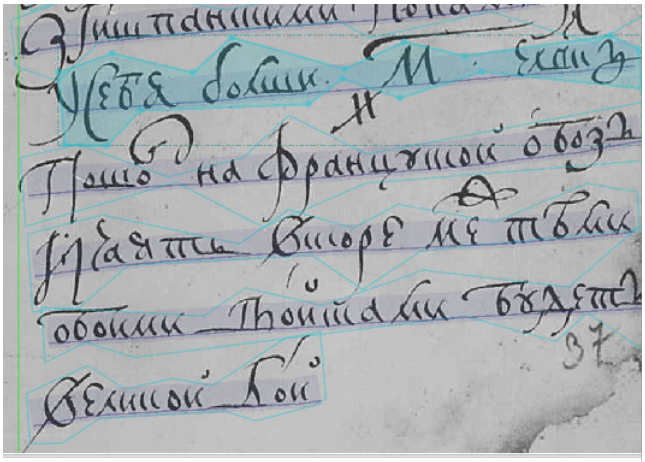


FIGURE 5
Kosmografija, RGB 173.1, 101, 17th century, p. 44, Model Kosmografija_0.2
(3,614 tokens, 9.99% CER on Validation Set)

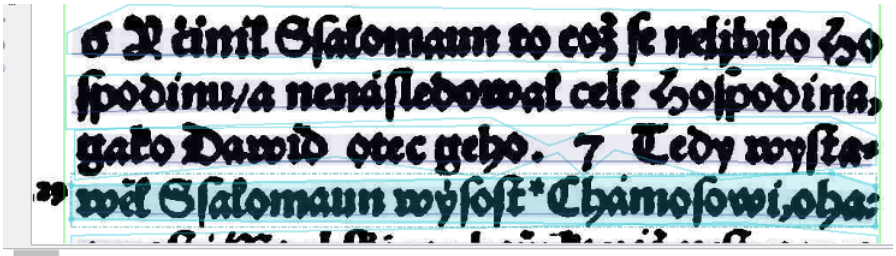
For more cursive script variants, the Russian *Vesti-Kuranty* from the second half of the 17th century can be cited as an example (see **FIGURE 6**). Here, depending on the legibility of the writing, a certain decline in recognition quality can be observed. Nevertheless, the result is far from unusable: while the correct recognition of ѹ҃вкъ in line 1 is impressive, for example, французжское instead of французской is suboptimal. Here, the inclusion of additional appropriate training data would likely yield even better results. Examples of successful mass digitization efforts beyond the Slavic field can be found, for instance, at <https://www.transkribus.org/sites> (last access: 01.04.2025) or <http://prhlt-kws.prhlt.upv.es/bentham/> (last access: 31.03.2025). They demonstrate how institutions such as libraries and archives can efficiently and largely automatically provide a large number of sources to the public through mass digitization, which can be accessed much more effectively through search functions than through digital images alone.

When using HTR data for linguistic analysis, meaningful quantitative investigations can be made even with just the uncleaned transcription of a manuscript, as demonstrated in Petrov, Rabus 2023. Deliberately, a workflow is tested here that dispenses with manual post-correction. This allows quantitative, token-based investigations with a tolerable amount of noise to be conducted without great time expenditure for transcription, in principle with any digitally available manuscript for which a corresponding HTR model exists.



усебя болши .≠М. чл҃вкъ
пошол на французское обозъ
и і чаять вскоре меж тѣми
обоими войсками будетъ,
великой бой

FIGURE 6
Vesti-Kuranty 1676, f. 37r., Model Kuranty_1672_0.6
(42,177 tokens, 9.24% CER on Validation Set)



1-1 ó Y činił Sſałomaun to což ſe neljbiło Ho
1-2 ſpodinu / a nenáſledował cele Hoſpodina
1-3 gako Dawid otec geho. ¶ Tedy wyſta=
1-4 wěł Sſałomaun wýſoauſt Chamoſowi,oha=

FIGURE 7
Transcription of the *Kralice Bible*, Model Blackletter_German_Czech_0.1
(293,849 tokens, 0.54% CER on Validation Set)

In the aforementioned study, for example, the most frequent bigrams at the word level are presented, which occur in the March volume of the Menologion from the 16th century of the Moscow Theological Academy (Collection 173.I, 92-2) available only in manuscript form (the statistically most significant are к немоу, с҃го дх҃а, к тебѣ), as well as the distribution of the spelling of <ы> vs. <и> after velars in East Slavic monuments following the Second South Slavic Influence. Even detailed orthographic analysis is possible using uncorrected HTR data as has been shown in Polomac, Rabus 2025. On the basis of high quality HTR transcriptions (CER ranging from 0.69-1.44%) of early printed Serbian books orthographic details such as the distribution of <ѣ> and <ь> have been analyzed.

HTR technology is also suitable for the digitization of old printings, as the example of the Czech Kralice Bible (1613, p. 317) shows (see **FIGURE 7**).

This model consists of a large amount of training data in German and a small amount in Czech, which explains why the recognition quality, despite the very large model with a very low CER on Validation Set, is still in need of improvement: the specific Czech characters, letter combinations and words were seen too rarely during model training compared to specifically German characters, letter combinations and words. This shows why it is important to test models on one's own data which has not been seen during training. We call this the model's 'real-world-performance' since this evaluation provides a more realistic estimate of how the model will perform on new data since computed Character Error Rates can sometimes be misleading.

Beyond the transcription of manuscripts and old printings, there is also the possibility of (re-)digitizing printed editions using HTR technology, as the example in **FIGURE 8** shows.

Even if the use of HTR in this regard seems less spectacular compared to complex manuscripts, this is another relevant task where HTR tools can be applied. The digitization of traditional editions for which no digital source data are available for digital secondary use, carried out in this way, is also a worthwhile endeavor that can serve mass digitization, further processing of texts through the creation of dictionaries or corpora, and making numerous texts available.

Finally, models for the angular Croatian Glagolitic script have also been trained (Rabus 2022). Interestingly, due to the nature of the training data, the model can directly perform a transliteration from Glagolitic to Latin script, which can further increase the accessibility of Croatian cultural heritage for non-specialists (see **FIGURE 9**).

Also notable here is the automatic resolution of abbreviations, which the model has learned from the corresponding training data, and which shows that it is capable of reproducing philologically correct decisions, although this comes at the cost of hypercorrect resolutions. The correct transcription of ligatures also presents no major problem for these models (see, for example, *primi* in the penultimate line and *prinosimь* in the last line).

For the Paleoslavic or generally philological-historical target group, which partly has no explicit affinity for digital methods, a recommended HTR program is the software Transkribus (transkribus.eu, TRANSKRIBUS Team at University of Innsbruck 2020, Muehlberger *et al.* 2019), with which the preceding transcription examples were also created,

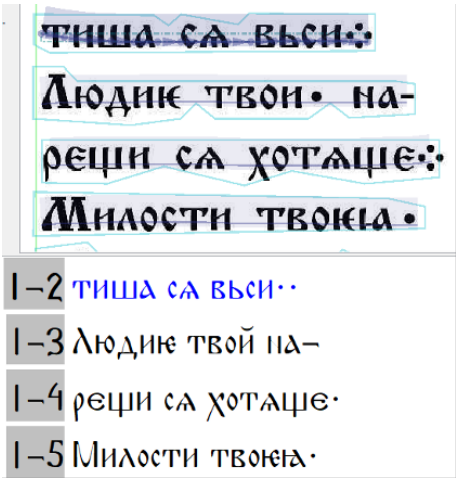


FIGURE 8
Redigitalization of the edition of the *Učitelno evangelie* (Tichova 2012), Model Generic_ChSl_printings_0.1 (75,332 tokens, 1.54% CER on Validation Set)

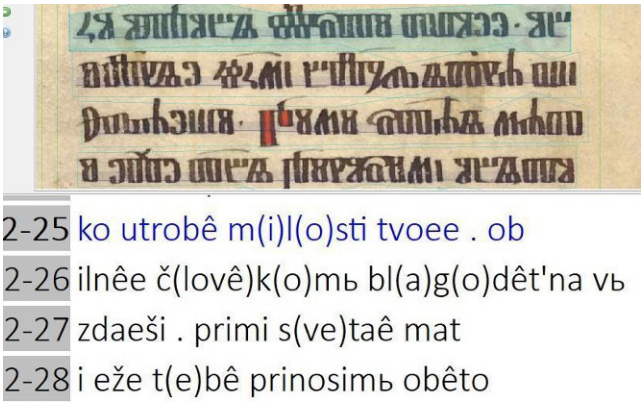


FIGURE 9
1. beramski brevijar, 15th century, f. 92v, Model Glagolitic_0.03 (140,455 tokens, 6.11% CER)

and the mentioned models were trained. Transkribus has the advantage that many HTR models are already available and that it features a user-friendly, browser-based graphical interface, usable by laypeople on all devices with browsers, and thus does not come with a high entry barrier such as the installation of alternative operating systems, setting up servers, or learning how to use the command line. These entry barriers are often so high for the target group discussed here that the attempt to use computer-assisted handwriting

recognition is abandoned, and digitization is carried out in the traditional manual way. Still, it must be mentioned that Transkribus is a paid (freemium) service in comparison to open-source solutions such as, e.g., eScriptorium/Kraken which can, in principle, produce similar results (Rabus, Thompson 2023).

3. *Consistency of HTR Quality*

Another possible concern we want to address is the issue of HTR consistency, that is, how strongly the quality of an HTR transcription varies across a single document. One might be concerned that after initially evaluating, say, 10 pages of an HTR transcription and coming to the conclusion that the transcription quality is satisfactory, one transcribes the entire document using HTR only to realize that the transcription quality in other parts of the document is mediocre. In order to tackle this issue, we used data from the Ostroh Bible (1581) which is the first full printed Bible edition in Church Slavonic and was printed at the Ostroh Academy. We used these data to evaluate how much the CER varies across multiple books of the Bible and how this consistency is influenced by the overall HTR quality.

The texts used consist of several books of the Ostroh Bible: the preface(s), Leviticus, Numbers, Deuteronomy, Joshua, Judges and Kings 1-4. For our analysis we used and compared four different transcriptions: one GT version (<http://dic.feb-web.ru/slavonic/corpus/0/bible1581/index.htm>, last access: 31.03.2025) and three different HTR versions of varying quality. We chose to use multiple HTR versions to better gauge how different error rates influence the noise in the HTR data. The three HTR versions (in the following called 'HTR 1', 'HTR 2' and 'HTR 3') were transcribed using three different models on the Transkribus platform:

1. HTR 1: Mar_KoP_OB_0.03, 334,551 Tokens, CER 3,8%, trained on the data from HTR 3, other texts from the Ostroh Printery as well as training data from the Ostroh Bible itself. It is therefore not only specialized in Ukrainian Church Slavonic and typographic specificities of the Ostroh Printery, but also the Ostroh Bible.
2. HTR 2: Mar_KoP_OB_0.02, 320,489 Tokens, CER 2,8%, trained on the data from HTR 3 and other texts from the Ostroh Printery. It is therefore specialized in the Ukrainian Church Slavonic and typographic specificities of the Ostroh Printery, but not the Ostroh Bible itself.
3. HTR 3: Generic_ChSl_old_and_modern_printings_0.3, 309,370 Tokens, CER 2,6%, generic East Church Slavonic print model.

After the transcription minor preprocessing was performed in order to align the transcription principles of the GT version and the HTR transcriptions. This is because some characters were represented differently in the GT version than the models had learned based on the training data used for training. For example, the allographs <ꙁ> and <з> were unified as <з> since this is the only version used in the GT version. Additionally, various

non-Unicode characters were replaced by their Unicode counterparts, accents and breathing marks were removed, Latin characters erroneously used in the HTR transcription[2] were replaced with their Cyrillic counterparts and the placeholders which were used in the GT transcription instead of proper superscript letters were replaced by their superscript counterparts. All these changes were made so that the calculated error rates would be accurate, and differences in the transcriptions which are due to differences in transcription policies would not be counted as actual errors. Besides these changes, the transcription was left untouched, that is, none of the actual transcription errors were corrected.

We then calculated the CER per page based on the Levenshtein (Levenštejn 1965) distance for all three HTR transcriptions. The average CER values are: HTR 1: 4.47%, HTR 2: 11.6% and HTR 3: 12.1%. In order to evaluate the variance of the page-based CER, we first created a violinplot (**FIGURE 10**).

The plot shows the distribution of different page-based CER values in each dataset. The y-axis represents the CER, while the violin plot (the blue outline surrounding the boxplot) represents the shape of the distribution: A wider section of the outline indicates a higher frequency of that particular value, whereas narrower sections correspond to less frequent values. This makes it possible to quickly compare different distributions across datasets (Hintze, Nelson 1998).

The boxplots within the blue outlines display the median as well as the lower and upper quartiles. Quartiles divide a distribution into four equal parts (Everitt, Skrondal 2010: 348). In other words, the lower quartile is the value that is greater than 25% of the data, while the upper quartile is the value that is greater than 75% of the data. The central 50% of the values lie between these two quartiles. The median, represented by a line inside the box, is the value that lies in the middle of the dataset.

First of all, it is apparent that there is a big discrepancy between HTR 1 on the one hand and HTR 2 and 3 on the other, when comparing the overall transcription quality. This is due to the fact that HTR 1 has seen parts of the Ostroh Bible during training and is therefore familiar with its linguistic and graphical specificities, which does not hold for HTR 2 and 3. It is interesting though that HTR 2 and 3 do not seem to differ much, which suggests that the inclusion of other texts from the Ostroh Printery (as in HTR 2) does not increase the transcription quality by much. Additionally, the variance in HTR 1 appears to be much smaller than in HTR 2 and 3, which is depicted in the plot by the wider peaks in the midranges of HTR 1, while HTR 2 and 3 appear much flatter and longer, representing the wider range of CER values. Additionally, the outliers (the highest and lowest values which are outside the whiskers of the plot) are much closer to the mean in HTR 1 than in HTR 2 and 3.

[2] These Latin characters are used by the model due to them being present in the GT used for training. The GT data used some non-Unicode characters and Latin characters to represent certain characters that, at the time of their creation, were not represented in Unicode yet. These are known systematic flaws and can therefore be easily removed before using the HTR transcription before further analysis.

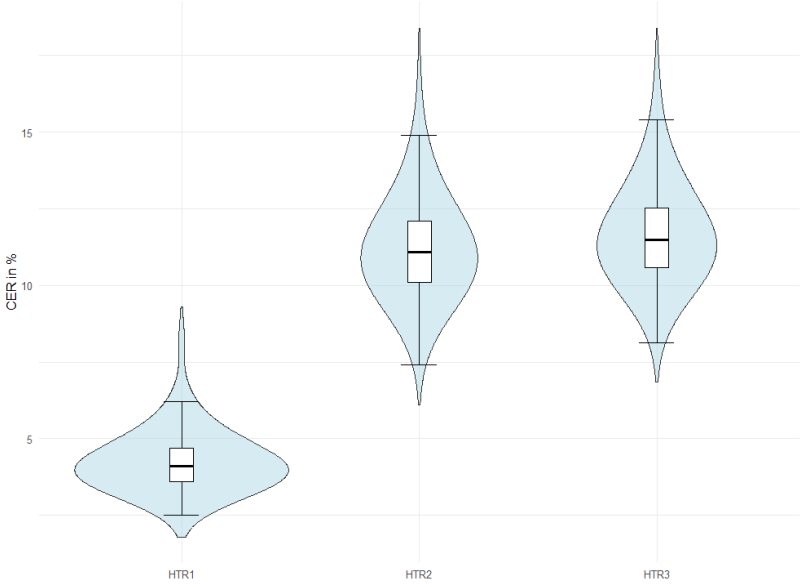


FIGURE 10
Violinplot showing the distribution of the page-based CER for the three transcriptions HTR 1, HTR 2 and HTR 3

After this initial overview of the quality distribution of these HTR transcriptions, we wanted to test how many pages are needed to get a reliable sample to reproduce the average CER. That is to say, how many pages would a philologist have to evaluate in order to get a reliable overview of the transcription quality and is this amount influenced by the overall transcription quality?

For this purpose, we chose an approach using randomly created samples. For every one of our three HTR transcriptions, we created five groups of samples, each group containing 1000 samples. The first group contains 1000 samples each with the CER values of 10 pages, the second group contains 1000 samples with the CER values of 20 pages, the third group of 30, the fourth of 40 and the fifth 50 pages[3]. We then calculated the average CER of every sample. Thus, each group now contains 1000 average CERs based on 10, 20, 30, 40 and 50 pages. The distribution of these average CERs is given in **TABLE 1**. This allows us to test how reliably a random sample of the selected size (10, 20, 30, 40, 50 pages) can reproduce the average CER of the whole data set (4.47%, 11.6% and 12.1%) tested on 1000 samples each.

[3] Due to the right-skewed distribution of the original CER (see **FIGURE 10**) we implemented stratified sampling which reproduces the original distribution.

HTR 1

Samples Size	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Reference
10 Pages	3.87	4.12	4.22	4.24	4.34	4.85	4.47
20 pages	3.78	4.11	4.20	4.23	4.32	5.02	
30 Pages	3.86	4.11	4.20	4.23	4.32	4.98	
40 Pages	3.78	4.11	4.20	4.23	4.32	5.05	
50 Pages	3.81	4.11	4.20	4.23	4.32	4.99	

HTR 1

Samples Size	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Reference
10 Pages	10.39	10.98	11.14	11.17	11.34	12.29	11.60
20 pages	10.47	10.99	11.16	11.18	11.34	12.16	
30 pages	10.51	10.98	11.17	11.18	11.34	12.12	
40 Pages	10.27	10.99	11.16	11.18	11.35	12.10	
50 Pages	10.57	11.00	11.16	11.17	11.33	12.06	

HTR 1

Samples Size	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Reference
10 Pages	10.72	11.46	11.62	11.64	11.81	12.66	12.10
20 pages	10.87	11.45	11.62	11.63	11.80	12.49	
30 Pages	10.85	11.45	11.62	11.64	11.82	12.49	
40 Pages	1084	11.46	11.62	11.64	11.79	12.57	
50 Pages	10.92	11.44	11.59	11.63	11.79	12.76	

TABLE 1
Distributions of sampled CERs based on 1000 samples containing, 10, 20, 30, 40 and 50 pages each, for HTR 1, 2 and 3

As can be seen, the calculated CER values based on samples are quite consistent for all three HTR versions, even though some deviance from the original CER to the mean of each test set can be observed: HTR 1: 0.23/0.24%, HTR 2: 0.42/0.43%, HTR 3: 0.46/0.47%. Even the sampled CER values for HTR 3 are very consistent and show that the model does not produce a lot of random outliers. While the CER distribution is very consistent, it needs to be pointed out that most of the sampled CER values are lower than the actual CER based on the entire data set. So, even though the CER reproduction is consistent in the sense that there is not a lot of variance between samples, the values are off by a small amount (roughly 0.2–0.6% below the reference CER). Still, the difference between reference and sampled

data remains small, even for HTR 3 which has the highest CER of all data sets and the reference CER still lies within the range of all sampled CERs.

This experiment goes to show that even a small test set of roughly 20-30 pages can be enough to give a reliable overview of the transcription quality. Therefore, it is enough for a researcher to inspect a smaller part of their HTR transcription to get a good estimation of the overall transcription quality. It is not necessary to perform a CER calculation by creating GT data. Instead, it is advisable to simply qualitatively evaluate parts of the transcription since this will also reveal which errors the model is prone to make.

4. *Strategic Considerations*

From a project management perspective, and thus cumulatively from a strategic perspective for the further development of the field, the advantage of automatic computer-assisted pre-transcriptions becomes clear beyond philological considerations. This is shown by the following – admittedly very rough – calculation:

A full project staff position for three years costs approximately 245,000 euros according to the rates of the German Research Foundation in 2025. If during the usual project duration of three years, the holders of this project position create a transcription of a larger manuscript of about 500 pages with an error rate of about 0.5%, the project leader still needs numerous hours for final control and elimination of the remaining errors; for the calculation conducted here, we set this at 50 hours. If, instead of the staff intended for this purpose, the HTR software performs the pre-transcription, the calculation changes accordingly: Transkribus requires several hours for layout analysis and transcription. However, this is done server-side, meaning that the respective process is initiated on the local computer but then carried out on a server in Austria, so that the local computer can meanwhile perform other tasks or even be turned off. The manual correction of the layout analysis is estimated to take about ten hours. The subscription costs for a team to use the extended Transkribus functions amount to – as of 2025 – 2,500 euros. The error rate of the transcription by the model is self-evidently significantly higher than that of specifically trained staff. Realistically, if a suitable model is available, about 4% can be expected. If this leads to the project leader being occupied with corrections for a total of about 200 hours, additional costs will be incurred until the goal of an error-free edition is achieved. Overall, the costs described here can be listed as in **TABLE 2** (with high estimated labor costs of 100 € per hour for the project leader).

We thus find that the costs for transcription with Transkribus compared to a traditional manual transcription amount to about 10%[4]. In terms of time, a similar ratio can be

[4] A similar ratio is likely to prevail in scholarly cultures where labor is paid less on average. While the percentage costs for Transkribus software would then increase, since the hourly calculated costs of project leaders and the annual costs for transcription staff are largely proportional, this carries little weight. Overall, it can be seen that the Transkribus transcription costs according to this

Cost type	Traditional, without HTR	With HTR (Transkribus)
Transcription staff	245,000 €	–
Final review by project leader	5,000 €	20,000 €
Manual correction of layout analysis	–	1,000 €
Transkribus subscription costs	–	2,500 €
Total	250,000 €	23,500 €

TABLE 2
Exemplary total calculation of a traditional project compared to the use of the HTR program Transkribus

expected, since the automatic transcription by the computer hardly requires any time and can effectively be done over the weekend with the computer turned off. The higher resource expenditure for manual correction of a computer pre-transcription compared to a pre-transcription by humans in no way leads to costs even approaching those of a completely manual creation of a transcription. What is even more evident – the costs for using the software are extremely manageable by comparison, so that these costs represent more of a psychological than an actual hurdle. If one wishes to use open-source tools such as eScriptorium/kraken instead of Transkribus, one can save on the subscription costs. However, one must reckon with higher hardware or administration costs. Overall, the overall cost ratio as compared to a purely manual approach to transcription should be similar in that case as well.

5. *Conclusion*

From what has been said, it follows that HTR technology based on AI represents an important advancement for Paleoslavic studies and other philological disciplines. HTR programs such as Transkribus or eScriptorium enable the efficient pre-transcription of manuscripts or re-digitization of analog editions, thereby allowing, on the one hand, significantly more material to be digitized with the same available resources, or, on the other hand, enabling projects that could not be carried out previously due to lack of funds. Even if programs like Transkribus – unlike various programs like eScriptorium/kraken (<https://github.com/mittagessen/kraken>, last access: 31.03.2025), which are open source – incur medium-term costs for transcription per page, these bear no relation to the costs incurred with traditional manual methods and also to the costs that would be associated with learning to use command-line programs. This significant cost reduction creates entirely new possibilities for editorial philology and historical corpus linguistics from a quantitative perspective.

calculation are in the single-digit percentage range. Thompson (2024: 407) does not calculate with absolute costs but with the time saved through HTR and arrives at a time saving of 60-70%.

The AI-based HTR models thus take over the preliminary work, freeing up resources of human experts in the field of Paleoslavic studies, which can be used for interesting, creative, and modern Paleoslavic research. AI does not, as skeptics might argue, make human labor and expertise superfluous, but rather supports human researchers in the application of their skills and therefore facilitates research.

We conclude this contribution with the wish that all researchers (not only) from the field of Paleoslavic studies who are interested in AI-supported pre-transcription join forces, recycle data, open up new manuscripts through HTR, ideally train new HTR models, thereby significantly enlarging the accessible data and research base and facilitating new approaches in Paleoslavic studies. By working together, we will be able to create more refined models which will in turn speed up research and open up new ways of engaging with Slavic scripture altogether.

Abbreviations

HTR:	Handwritten Text Recognition
OCR:	Optical Character Recognition
CER:	Character Error Rate
GT:	Ground Truth
GIM:	Gosudarstvennyj istoričeskij muzej
RGB:	Rossijskaja gosudarstvennaja biblioteka
NUK:	Narodna in univerzitetna knijžnica

Manuscripts/Printings

Minei-Čet'i, fevral':	Moscow, GIM Sin. 179, 16th century.
Služebnik Varlaama Chutynskogo:	Moscow, GIM Sin. 604 late 12th-early 13th century.
Vesti-Kuranty:	Moscow, RGADA 155, 8, 1676.
Kosmografija:	Moscow, RGB 173.I, 101, 17th century.
Menologion:	March – Moscow, RGB 173.I, 92-2, 1st quarter of 16th century.
Czech Kralice Bible:	*Biblj ſwatá. To geſt, Kniha w njž se wſſecka Pjſma S. Starého y Nowého Zákona obſahugj*, 1613.
Prvi beramski brevijar:	Ljubljana, NUK Ms 161, 1390–1420.
Ostroh Bible:	Moscow, RGB MK Kir. 2°, 1581.

Electronic Sources

Bentham Papers Free Text Search: <http://prhlt-kws.prhlt.upv.es/bentham/> (last access: 01.04.2025).

Biblija Ostrožskaja (1581): <http://dic.feb-web.ru/slavonic/corpus/o/bible1581/index.htm> (last access: 01.04.2025).

DH Crowdscribe: <http://dhcrowdscribe.com/> (last access: 31.03.2025).

Kraken: <https://github.com/mittagessen/kraken> (last access: 02.04.2025).

Nacional'nyj korpus russkogo jazyka: <http://ruscorpora.ru/new/search-old_rus.html> (last access: 31.03.2025).

TRANSKRIBUS: Team at University of Innsbruck 2020, Transkribus, <transkribus.eu> (last access: 02.04.2025).

Transkribus Sites: <https://www.transkribus.org/sites> (last access: 01.04.2025).

Velikie Minei Čet'i Corpus: <http://www.vmc.uni-freiburg.de> (last access: 31.03.2025).

Literature

Burlacu, Rabus 2021: C. Burlacu, A. Rabus, *Digitising (Romanian) Cyrillic using Transkribus: New Perspectives*, "Diacronica", XIV, 2021, pp 1-9.

Cleminson 2008: R. Cleminson, *XSLT and the Analysis of Critical Variants*, in: H. Miklas, A. Miltenova (eds.), *Slovo: Towards a Digital Library of South Slavic Manuscripts. Proceedings of the International Conference, 21-26 February 2008*, Sofia, 2008, pp. 227-233.

Eder 2013: M. Eder, *Mind Your Corpus: Systematic Errors in Authorship Attribution*, "Literary and Linguistic Computing", XXVIII, 2013, 4, pp. 603-614.

Everitt, Skrondal 2010: B.S. Everitt, A. Skrondal, *The Cambridge Dictionary of Statistics*, New York 2010.

Franzini *et al.* 2018: G. Franzini, M. Kestemont, G. Rotari, M. Jander, J.K. Ochab, E. Franzini, J. Byszuk, J. Rybicki, *Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm*, "Frontiers in Digital Humanities", V, 2018, 4, pp. 1-15.

Hintze, Nelson 1998: J.L. Hintze, R.D. Nelson, *Violin Plots: A Box Plot-Density Trace Synergism*, "The American Statistician", LII, 1998, 2, pp. 181-184.

Levenštejn 1965: V.I. Levenštejn, *Dvoičnye kody s ispravleniem vypadenij i vstavok simvola*, "Problemy peredači informacii", I, 1965, 1, pp. 12-25.

Levshina 2019: N. Levshina, *Token-based Typology and Word Order Entropy: A Study Based on Universal Dependencies*, "Linguistic Typology", XXIII, 2019, 3, pp. 533-572.

Muehlberger *et al.* 2019: G. Muehlberger, L. Seaward, M. Terras, S.A. Oliveira, V. Bosch, M. Bryan, S. Colutto, H. Déjean, M. Diem, S. Fiel, B. Gatos, A. Greinoecker, T. Grüning, G. Hackl, V. Haukkovaara, G. Heyer, L. Hirvonen, T. Hodel, M. Jokinen, P. Kahle, M. Kallio, F. Kaplan, F. Kleber, R. Labahn, E. M. Lang, S. Laube, G. Leifert, G. Louloudis, R. McNicholl, J. Meunier, J. Michael, E. Mühlbauer, N. Philipp, I. Pratikakis, J. Puigcerver Pérez, H. Putz, G. Retsinas, V. Romero, R. Sablatnig, J.A. Sánchez, P. Schofield, G. Sfikas, C. Sieber, N. Stamatopoulos, T. Strauß, T. Terbul, A.H. Toselli, B. Ulreich, M. Villegas, E. Vidal, J. Walcher, M. Weidemann, H. Wurster, K. Zagoris, *Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study*, "Journal of documentation", LXXV, 2019, 5, pp. 954-976.

Petrov, Rabus 2023: I.N. Petrov, A. Rabus, *Linguistic Analysis of Church Slavonic Documents: A Mixed-Methods Approach*, "Scando-Slavica", LXIX, 2023, 1, pp. 25-38, DOI: 10.1080/00806765.2023.2189617.

Piper 2020: A. Piper. *Can we be wrong? The Problem of Textual Evidence in a Time of Data*. Cambridge 2020.

Polomac 2022: V. Polomac, *Serbian Early Printed Books from Venice: Creating Models for Automatic Text Recognition Using Transkribus,* "Scripta & e-Scripta", XXII, 2022, pp. 11-29.

Polomac *et al.* 2023: V. Polomac, M. Kurešević, I. Bjelaković, A. Colić Jovanović *Digitizing Cyrillic Manuscripts for the Historical Dictionary of the Serbian Language Using Handwritten Text Recognition*, "Slověne", XII, 2023, 1, pp. 295-316.

Polomac, Rabus 2025: V. Polomac, A. Rabus, *Serbian Early Printed Books from Venice: A Quantitative Approach to Orthographic Variations,* "Studi Slavistici", XXI, 2025, 2, pp. 37-60.

Rabus 2019: A. Rabus, *Recognizing Handwritten Text in Slavic Manuscripts: a Neural-Network Approach using Transkribus*, "Scripta & e-Scripta", XIX, 2019, pp. 9-32.

Rabus 2022: A. Rabus, *Handwritten text recognition for Croatian glagolitic*, "Slovo: časopis Staroslavenskoga instituta u Zagrebu", LXXII, 2022, 1, pp. 181-192.

Rabus 2023: A. Rabus, *Training generic models for Handwritten Text Recognition using Transkribus: Opportunities and Pitfalls*, in: S.A. Pink, A.J. Lappin (eds.), *Dark Archives*, 1: *Voyages into the Medieval Unread and Unreadable, 2019-2021,* Oxford 2023, pp. 183-208.

Rabus 2024: A. Rabus, *Tolerating Imperfection: Uncorrected Transkribus Transcriptions in Church Slavonic Studies*, in: L. Taseva, A. Rabus, I.P. Petrov (ur.), *Učitelnoto evangelie na Konstantin Preslavski i južnoslavjanskite prevodi na chomiletični tekstove (IX-XIII v.). Dokladi ot Meždunarodnata naučna konferencija v Sofija 25-27 april 2023 g. // Constantine of Preslav's Uchitel'noe Evangelie and the South Slavonic Homiletic Texts (9th-13th Century). Proceedings of the International Scientific Conference in Sofia, April 25-27, 2023*, Sofija 2024, pp. 452-467.

Rabus, Thompson 2023: A. Rabus, W. Thompson, *Performance of Generic HTR Models on Historical Cyrillic and Glagolitic: Comparison of Engines*, "Scripta & e-Scripta", XXIII, 2023, pp. 11-34.

Thompson 2021: W. Thompson, *Using Handwritten Text Recognition (HTR) Tools to Transcribe Historical Multilingual Lexica*, "Scripta & e-Scripta", XXI, 2021, pp. 217-231.

Thompson 2024: W. Thompson, *Epifanii Slavinetskii's Greek-Slavonic-Latin Lexicon between East and West,* Heidelberg 2024 (= Empirie und Theorie der Sprachwissenschaft, 8).

Tichova 2012: M. Tichova, *Starobălgarskoto učitelno evangelie na Konstantin Preslavski*, Freiburg i. Br. 2012.

Waldenfels, Rabus 2015: R. von Waldenfels, A. Rabus, *Recycling the Metropolitan: Building an Electronic Corpus on the Basis of the Edition of the Velikie Minei Čet'i*, "Scripta & e-Scripta", XIV-XV, 2015, pp. 27-38.

Abstract

Achim Rabus, Martin Meindl
The Digital Revolution in Slavic Manuscript Studies: HTR Technology and its Impact on Philological Research

The paper highlights the recent advances of computer-assisted manuscript transcription using Handwritten Text Recognition (HTR) programs such as Transkribus. Numerous examples showing the capabilities of current HTR models with respect to different Slavic scripts and handwriting styles are presented and ways to use automatically transcribed sources for multiple purposes are discussed. We demonstrate that the transcription quality is stable throughout an entire document and that researchers can gauge the quality of their HTR transcription based on a limited number of pages. Afterwards, a calculation is conducted showing that the use of HTR as an instrument of pre-transcribing manuscripts and printings makes the overall process of transcribing significantly cheaper, thus making projects possible that could not be conducted without HTR technology due to financial reasons. The paper is concluded with an appeal to share training data and make ample use of these new advancements in HTR.

Keywords

Handwritten Text Recognition; Mass Digitization; Artificial Intelligence in Philology; Corpus Linguistics; Multiple Use of Data.