DIACRONIA
Impavidi progrediamur!

ARTICLE

doi:10.17684/i14A196en
ISSN: 2393-1140
www.diacronia.ro

Digitising (Romanian) Cyrillic using Transkribus: new perspectives

Constanța Burlacu[1]*, Achim Rabus[2]
[1]*Merton College, University of Oxford, Merton St., OX1 4JD Oxford, United Kingdom*
[2]*Department of Slavic Studies, University of Freiburg, Werthmannstr. 14, 79085 Freiburg, Germany*

Article info

History:
Received September 17, 2021
Accepted September 26, 2021
Published December 12, 2021

Key words:
Transkribus
handwritten text recognition
Romanian Cyrillic
digital humanities

Abstract

In this paper we discuss the application of the software platform Transkribus (*transkribus.eu*), an AI-assisted tool for Handwritten Text Recognition (HTR), to 16th century Romanian manuscript and printed sources using Cyrillic scripts. After an overview of the basic functionality of the HTR technology and Transkribus, we discuss the Romanian and bilingual Slavonic-Romanian sources we used, give an insight on training specific and generic as well as smart (i.e. transliterating from Cyrillic into Latin script) models, evaluate their performance and discuss implications of HTR for philological research in the Digital Age. We conclude with an outlook on future research perspectives.

1. Introduction: what is Transkribus and how does it work?

Transkribus (*transkribus.eu*, Muehlberger *et al.*, 2019) is an AI-assisted tool that can be trained to transcribe manuscripts and early printed books written in a wide variety of languages, styles and scripts. Moreover, it serves as a platform for collaborative work on different sources. Its Handwritten Text Recognition (HTR) technology is considerably more advanced than the traditional Optical Character Recognition (OCR) technology[1], since the recognition process does not only focus on individual characters, but is line-based, taking into account neighbouring graphemes and even words to determine the most likely transcription.

The AI-approach the HTR technology follows is an instance of so-called supervised learning. This means that the HTR engine needs a certain amount of digital high-quality images of the sources one is interested in as well as corresponding, manually corrected diplomatic transcriptions. In multiple epochs[2], the model learns the palæographic and linguistic features of the source in question. For easily legible and regular sources, one can start training a transcription model for a particular source with as few as around 2000 transcribed word tokens. Good results are usually achieved with around 10 000 word tokens, models that have generic capabilities in that they are able to transcribe a variety of different hands or even handwriting styles have been trained on more than 100 000, sometimes millions of tokens. Within Transkribus, HTR models can be shared with colleagues or made publicly available.

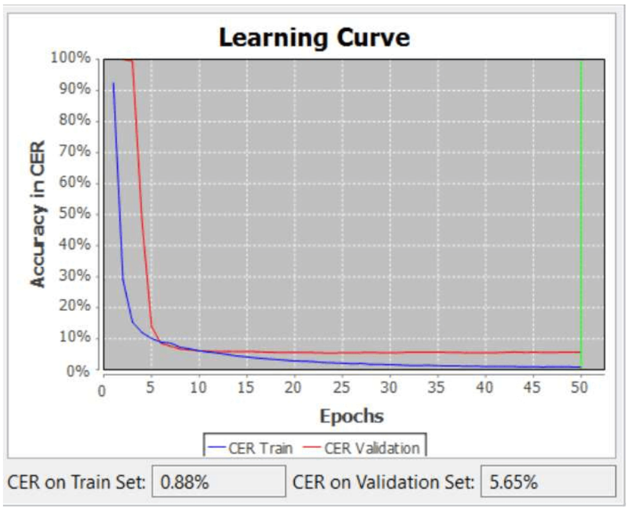
The main quality measure for HTR models is their Character Error Rate (CER). Good models for one specific source reach a CER of below 5%, meaning that less than one in 20 characters (including punctuation marks) is transcribed incorrectly. Usable models that allow for manually correcting the errors produced in less time than transcribing everything manually from scratch have a CER of below 10%.

The typical learning curve of an HTR model looks as shown in Fig. 1. As one can see, during the first five or so epochs, the CER drops drastically, while the improvement during the remaining epochs (in this case with the overall amount of 50) is rather slight.

*Email address: *constanta.burlacu@merton.ox.ac.uk*.

[1] Felix Dietrich compares the step from OCR to HTR to the step from a brute-force algorithm such as the one implemented in Deep Blue, the computer that first defeated the world chess champion as early as 1997, to the hugely more sophisticated approach by AlphaGo defeating the Go champion in 2016 (*readcoop.eu*).

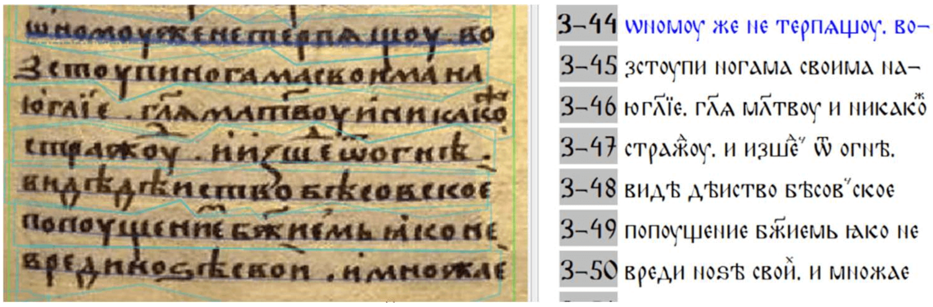
[2] According to *fon.hum.uva.nl*, an epoch is “one complete presentation of the data set to be learned to a learning machine”.



CER on Train Set: 0.88% CER on Validation Set: 5.65%

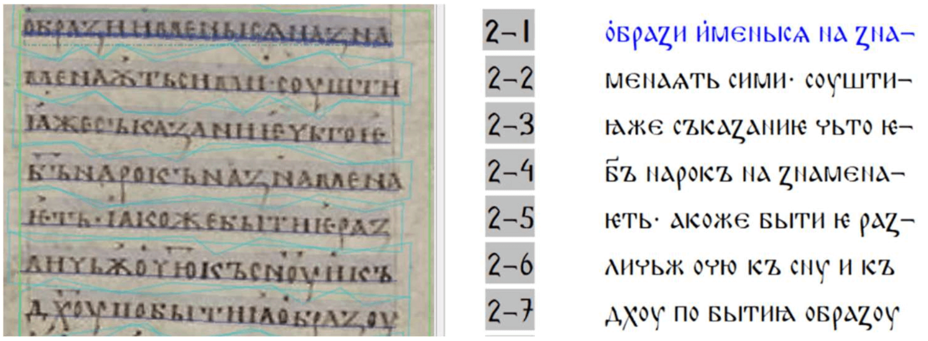
Figure 1: Example of an HTR learning curve in Transkribus

For pre-modern Cyrillic scripts, public Transkribus models have been published, both for Slavonic *ustav* and *poluustav* (Rabus, 2019). They have been trained on hundreds of thousands of word tokens, leading to generic capabilities of the models. This means that, to a certain extent, the models are capable of transcribing sources written in a variety of hands, from different regions and times. Figs. 2–4 give an example of the performance of the publicly available generic models for Church Slavonic.



3-44 ѡномоу же не терпащоу. во-
3-45 зстоупи ногама своима на-
3-46 югл҃їе. гл҃а мл҃твоу и никако҆
3-47 стражоу. и изшеⷣ ѿ огнѣ.
3-48 видѣ дѣиство бѣсовское
3-49 попоущение бж҃иемь ꙗко не
3-50 вреди нозѣ свой. и множае

Figure 2: VMČ, SIN 993, April, 16th century

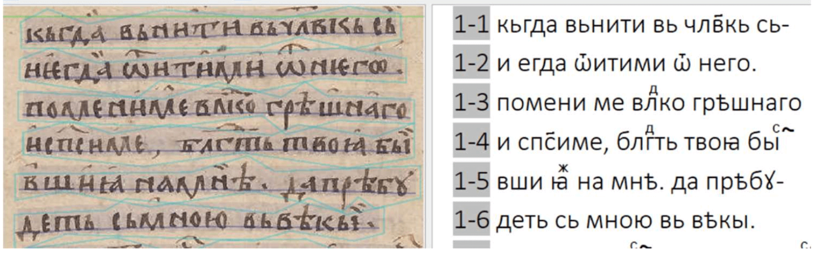


2-1 ȯбраꙁи йменꙑꙗ на ꙁна-
2-2 менаать сими· соушти-
2-3 ꙗже съкаꙁание чьто ѥ-
2-4 б҃ъ нарокъ на ꙁнамена-
2-5 ѥть· акоже быти ѥ раꙁ-
2-6 личьж оч҄ю къ сноу и къ
2-7 дхоу по бытиꙗ обраꙁоу

Figure 3: Izbornik Svjatoslava, 1073

As can be seen, despite some variation in transcription quality, the overall performance of the public models for Slavonic is rather convincing. Within the Transkribus platform, these models can be used by everyone free of charge[3].

[3] Transkribus and all public models can be used free of charge. Upon registration, users get 500 credits, sufficient for transcribing some 400 pages. Users who wish to transcribe more pages can obtain additional credits, see *readcoop.eu*.



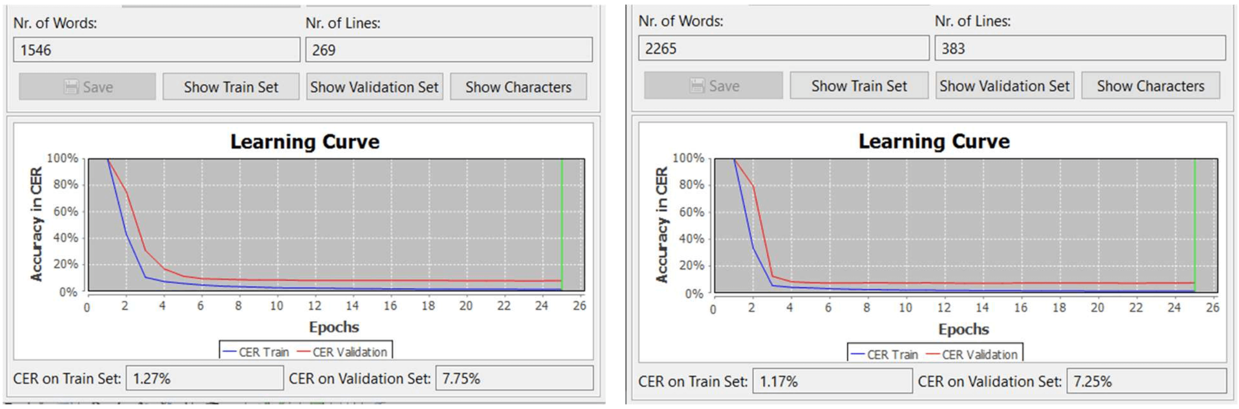
1-1 кьгда вьнити вь чл҃вкь сь-
1-2 и егда ѿитими ѿ него.
1-3 помени ме влⷣко грѣшнаго
1-4 и спс҃име, блгⷣть твоꙗ быⷭ҇
1-5 вши ꙗⷤ на мнѣ. да прѣбꙋ-
1-6 деть сь мною вь вѣкы.

Figure 4: Bdinski Zbornik, 14th century

However, since, as mentioned above, HTR models learn not only palæographic, but also linguistic features (e.g., probabilities of grapheme or even word combinations), the performance of these models on Romanian Cyrillic sources is mediocre to bad. Simply put, they try to find Slavic words or grapheme combinations in Romanian texts, yielding unsatisfying results. Because of that, the need to train specific HTR models for Romanian Cyrillic (and for bilingual sources) arises.

2. Training and evaluating models for Romanian Cyrillic

The models we developed so far for Romanian Cyrillic have been trained on manuscript and printed material coming from the 16th century, *Apostolos* and *Psalter* texts mainly. Initially, the approach has been to develop specific models for certain sources, so that the first manuscript to be analysed was the *Voroneț Codex*, for which two subsequent models have been developed (Romanian_Cyrillic_0.01 and 0.02, see Fig. 5).

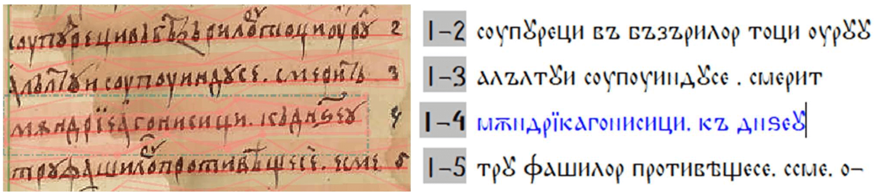


Nr. of Words: 1546
Nr. of Lines: 269
CER on Train Set: 1.27%
CER on Validation Set: 7.75%

Nr. of Words: 2265
Nr. of Lines: 383
CER on Train Set: 1.17%
CER on Validation Set: 7.25%

Figure 5: Learning curves for Romanian_Cyrillic_0.01 (left) and 0.02 (right)

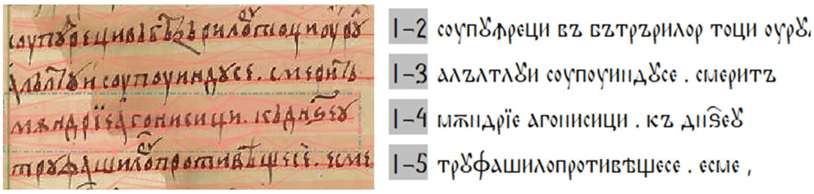
On this occasion about 30 pages have been transcribed, which add up to slightly more than 2000 word tokens to be used as Ground Truth to train a new model. As we can see, the CER is of 7.25%, which, considered the small amount of data provided, makes the model already usable for transcription. The example in Fig. 6 shows the efficacy of the model.



1-2 соүпꙋреци въ бъзърилор тоци оүрꙋꙋ
1-3 алълтꙋи соүпоүиндꙋсе . смерит
1-4 мжндрїкагонисици. къ дисеꙋ
1-5 трꙋ фашилор противѣщесе. ссме. о–

Figure 6: Transcription of *Voroneț Codex*, f. 82r, with Romanian_Cyrillic_0.02

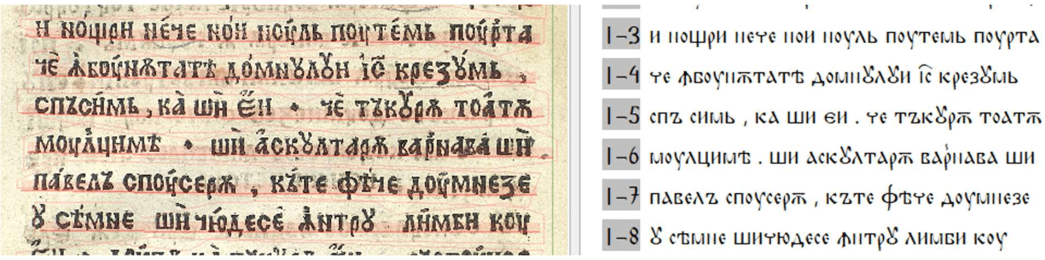
As we can see, this first model has difficulties in identifying superscripts (**соүп8фрецивъ**, **оүр8л**), ligatures (**бътрърилор**), letters such as **к** and **є** in **мжндрїє** and **є смє-**. Additionally, there are various mistakes when it comes to spacing, as for example the first transcribed verb **соупꙋрецивъ** or **мѫндрїе агонисици** in line 4. Nonetheless, the model can be used in order to transcribe further pages of the same source and so improve the initial models, which is how we proceeded further, creating a third model (Romanian_Cyrillic_0.03) by adding about 5000 word tokens, which resulted in a CER of 5.85%. In this new transcription letters are better identified than by the previous model, as well as the ligature **тр** and the superscripts[4], though the model does not recognise the superscript in the last line and considers the two initial words as one **трꙋфашило̑ противѣщесе** (Fig. 7).



1-2 соупꙋфреци въ бътрърилор тоци оурꙋ.
1-3 алълтаꙋи соупоуиндꙋсе . смеритъ
1-4 мѫндрїе агонисици . къ диꙋ҇еꙋ
1-5 трꙋфашилопротивѣщесе . есме ,

Figure 7: Transcription of *Voroneț Codex*, f. 82^r, with Romanian_Cyrillic_0.03

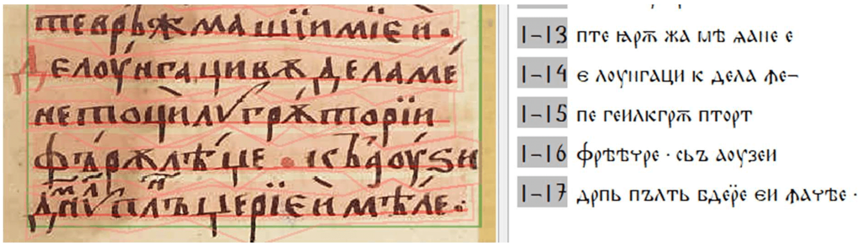
While the development of transcription models for manuscripts requires a high amount of manually checked Ground Truth data, models for printed materials need much less input. In creating a model for the Coresi's 1563 printed *Apostolos*, in fact, we reached quite a low CER (4.46%) with less than 4000 word tokens. The transcription of a previously unprocessed page (p. 70) of the *Apostolos* has the result shown in Fig. 8.



1-3 и нощри нече нои ноуль поутемь поурта
1-4 че ꙟбоунѫтатѣ доминꙋлꙋи і҇с крезꙋмь
1-5 спъ симь , ка ши еи . че тъкꙋрѫ тоатѫ
1-6 моулцимѣ . ши аскꙋлтарѫ вар̑нава ши
1-7 павелъ споусерѫ , къте фѣче доумнезе
1-8 ꙋ сѣмне шичюдесе ꙟитрꙋ лимби коу

Figure 8: Transcription of 1563 *Apostolos* text, p. 70, with BRV12_Romanian Printing 16th c

All letters are recognised correctly and there are only spacing mistakes in lines 4 (**ꙟбоунѫтатѣ**), 5 (**спъсимь**) and 8 (**шичюдесе**). The peculiarity of this printed text is to have very few superscripts and abbreviations, as well as to follow modern editorial rules when it comes to spacing. Therefore, and due to the fact that printed characters have a far more regular outline than handwritten letters, the HTR learning process of this material is much easier than for manuscripts. Indeed, if we transcribe a manuscript source with this model, the result is unusable, as shown in Fig. 9.

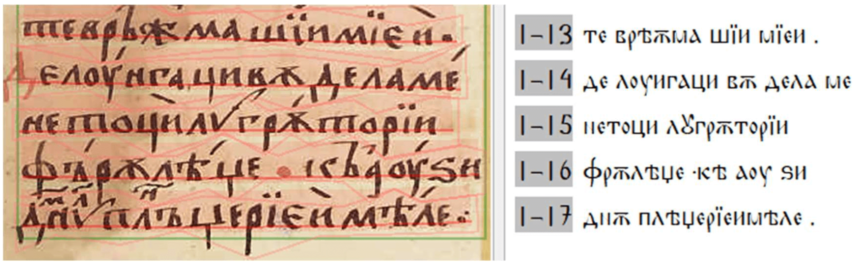


1-13 пте ѩрѫ жа мѣ ѧане є
1-14 є лоунгаци к дела ꙟе−
1-15 пе генлкгрѫ пторт
1-16 фрѣѣчре · сьъ лоузеи
1-17 дрпь пълть бдѐрѐ еи ꙟачѣе ·

Figure 9: Transcription of *Scheia Psalter*, f. 7^r, with BRV12_Romanian Printing 16th c

[4] In these first models we have decided to bring the superscripts down to the line level, though such is not the case for the later models we have developed. Rendering the superscripts faithfully as they are in the original text has been especially easy thanks to Daniel Bunčić, who provided keyboard drivers for typing in Church Slavonic. See *obshtezhitie.net*.

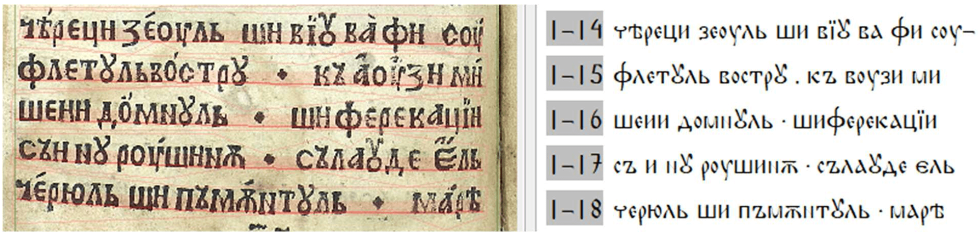
Nonetheless, if we merge the two above mentioned models, Romanian_Cyrillic_0.03 and BRV12_Romanian Printing 16th c, we can obtain a combined model which gives a fairly good transcription result, both for printed and manuscript sources. The same page from the *Scheia Psalter* is transcribed as shown in Fig. 10.



1-13 те врѣѫ̄ма шїи мїеи .
1-14 де лоуигаци вѫ̄ дела ме
1-15 нетоци лȣгрѫ̄торїи
1-16 фрѫ̄лѣце ·кѣ доу ѕи
1-17 дииѫ̄ плѣцерїеимѣле .

Figure 10: Transcription of *Scheia Psalter*, f. 7^{r}, with Combined Romanian Cyrillic

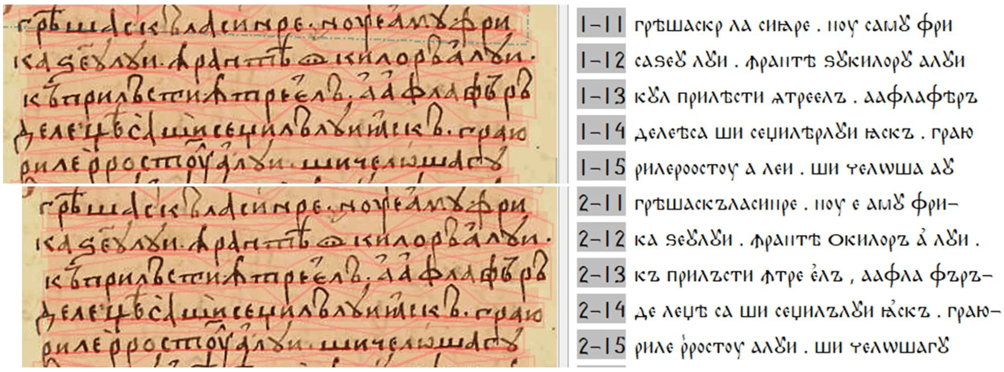
The transcription is way from perfect, for in fact it needs manual correction when it comes both to letter identification and spacing. However, considering the fact that the textual source is completely new to the model and that the size of the latter is fairly small, the provided transcription is a good starting point. Similarly, when applying the combined model to a printed source, the result is a functional transcription, see Fig. 11.



1-14 чѣреци зеоуль ши вїȣ ва фи соу–
1-15 флетȣль вострȣ . къ воузи ми
1-16 шеии доминȣль · шиферекацїи
1-17 съ и нȣ роушинѫ̄ · съладȣде єль
1-18 черюль ши пъмѫ̄итȣль · марѣ

Figure 11: Transcription of 1570 *Psalter*, f. 74^{v}, with Combined Romanian Cyrillic

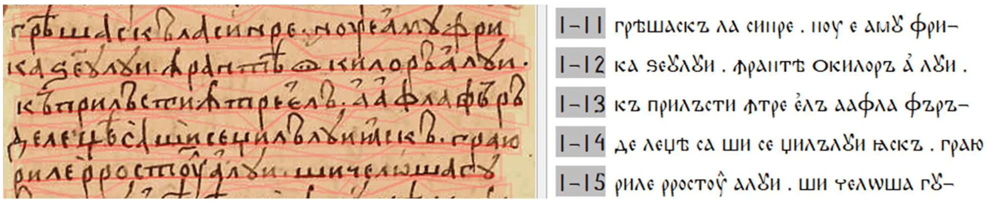
Although aspiring at a combined model which can be used on a wide variety of unseen sources is desirable, so far the Combined Romanian Cyrillic model is too small to be able to perform well on scripts different from the ones it was trained on. To prove the point, we have transcribed a page of the *Hurmuzaki Psalter* (f. 30^{r}) with both the Combined model (CER 5.65%) and a model specifically developed for the psalter text, that is, trained on transcriptions coming from the text in question (CER 10.11%). When the performance of the two models is compared, it is possible to notice how the specialised model, although with a higher CER, gives a better result than the more generic model. In Fig. 12 we see the performance of the Combined Romanian Cyrillic model first, followed by that of the Psaltirea Hurmuzaki 2 model.



1-11 грѣшаскр ла сиѩре . ноу самȣ фри
1-12 casеȣ лȣи . ꙟрантѣ ѕȣкилорȣ аллȣи
1-13 кȣл прилѣсти ꙟтреелъ . аафлафѣръ
1-14 делеѣса ши сецилѣрлȣи ѩскъ . граю
1-15 рилеростоу а леи . ши челѡша аȣ
2-11 грѣшаскъласиире . ноу є амȣ фри–
2-12 ка ѕеȣлȣи . ꙟрантѣ окилоръ ѧ҆ лȣи .
2-13 къ прилъсти ꙟтре ѐлъ , аафла фъръ–
2-14 де лецѣ са ши сецилълȣи ѩ҆скъ . граю–
2-15 риле ѽростоу алȣи . ши челѡшагȣ

Figure 12: Combined Romanian Cyrillic and Psaltirea Hurmuzaki 2 applied to *Hurmuzaki Psalter*, f. 30^{r}

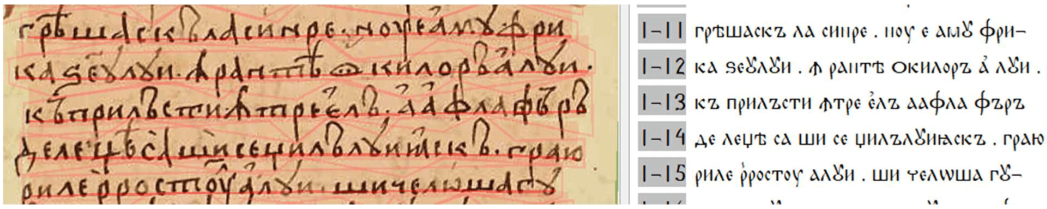
The most prominent discrepancy in performance between the two models is that the Psaltirea Hurmuzaki model recognises better the letter forms, though there are many spacing and word-recognition mistakes, while the result of the Combined model is almost impractical in this instance. The reason for this being that the Combined model has been exposed to and checked against a limited amount of data, coming from two sources only, neither of which is the *Hurmuzaki Psalter*. As soon as we add to the Combined model the GT data used for the Hurmuzaki model, the automatic transcription undoubtedly improves. If we apply the newly Combined model to the same folio of the *Hurmuzaki Psalter*, we obtain the result shown in Fig. 13.



1-11 грѣшаскъ ла сипре . ноу є амꙋ фри–
1-12 ка ѕеꙋлꙋи . ꙟраитѣ окилоръ ӑ лꙋи .
1-13 къ прилъсти ꙟтре ѐлъ аафла фъръ–
1-14 де леџѣ са ши се џилълꙋи ꙗскъ . граю
1-15 риле рростоу̂ алꙋи . ши ҹелѡша гꙋ–

Figure 13: Combined_Romanian_Hurmuzaki applied to *Hurmuzaki Psalter*, f. 30r

All letters are identified correctly, and the only errors are due to word separation in lines 13 **а̂ афла**, where the infinitival marker *a* has not been separated from the verb, 14 **џилълꙋиꙗскъ**, where the word has not been recognised as one, and 15 **а лꙋи . ши ҹелѡшагѡꙋ**-, where the first two words have been taken as one, while the last 6 letters are part of one word which continues in the following line **ҹелѡшагꙋлъ**. The model presents an improvement when compared to the previous two, that is, the recognition and reproduction of the superscript **ʌ** in line 15. Understandably, if the transcription model receives more training data and learns the peculiarities of specific scripts, its performance will outplay that of smaller models. Following this line of thought, we decided to combine all general models developed so far (also the bilingual models discussed below) and see whether a general model with GT amounting to 30 900 word tokens and a CER of 8.31% would perform any better in transcribing the *Hurmuzaki Psalter*. Although the difference in performance on this very page is not significant, especially when it comes to spacing, it is interesting to see that the Combined_mono-bilingual_Romanian model identifies the word **џилълꙋиꙗскъ** in line 14 as such (Fig. 14). This is probably due to the enablement of the *language model* feature present in Transkribus, which puts the accent on the linguistic rather than the palæographic aspect of the training data used in building the model[5].



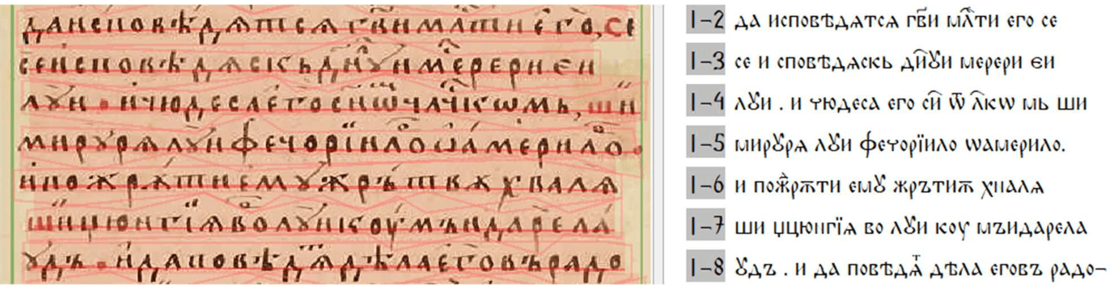
1-11 грѣшаскъ ла сипре . ноу є амꙋ фри–
1-12 ка ѕеꙋлꙋи . ꙟ раитѣ окилоръ ӑ лꙋи .
1-13 къ прилъсти ꙟтре ѐлъ аафла фъръ
1-14 де леџѣ са ши се џилълꙋиꙗскъ . граю
1-15 риле р̆ростоу алꙋи . ши ҹелѡша гꙋ–

Figure 14: Combined_mono-bilingual_Romanian applied to *Hurmuzaki Psalter*, f. 30r

Another important aspect of text production in the 16th century in Romania is its bilingualism – many religious books were written or printed both in Church Slavonic and in Romanian. While the letterform would not vary from one language to another, it is still necessary to develop a new HTR model for these sources. In fact, as mentioned before, when compared to OCR technologies which focus on the recognition of individual letters, HTR presents a form of 'language intelligence', for it analyses the proximity of letters in accordance with their in-line position and tries to guess what is their most likely distribution based on what it learned about the language from GT data. Consequently, because Church Slavonic and Romanian are two distinct languages, the HTR technologies would need to learn some linguistic features

[5] Further information about this function can be found at: *readcoop.eu*.

of both. Similar to the approach taken for the previous models, we decided to recycle the work done up to this point and build on it. The first attempt has been to use the Combined Romanian Cyrillic on the *Bratu Codex*, a bilingual *Apostolos* text from mid-16th century, and so have an initial transcription, which has then been checked manually and brought up the level of Ground Truth data. Unfortunately, the image quality of the source is extremely low, so that a slightly higher amount of word tokens (11 500) has not been enough to obtain a satisfactory CER, which currently adds up to 11.12%. As a matter of fact, the training of a new transcription model for *Ciobanu Psalter*, another bilingual source digitally available in high quality images, had better results, although it was trained on less GT (CER 7.97% with 7800 word tokens). The combination of the two model leads us to a CER of 10.26%, which, when applied on a new source, has the results shown in Fig. 15.



1-2 да исповѣдатса гв҃и мл҃ти его се
1-3 се и сповѣдаскь ди҃ꙋи мерери еи
1-4 лꙋи . и чюдеса его си҃ ѿ л҃кѡ мь ши
1-5 мирꙋра лꙋи фечоріило ѡамерило.
1-6 и пожⷬ҇рѫти емꙋ жрътиѫ҇ хиала
1-7 ши ццюнгїа во лꙋи коу мъидарела
1-8 ꙋдъ . и да повѣдаⷮ дѣла еговъ радо–

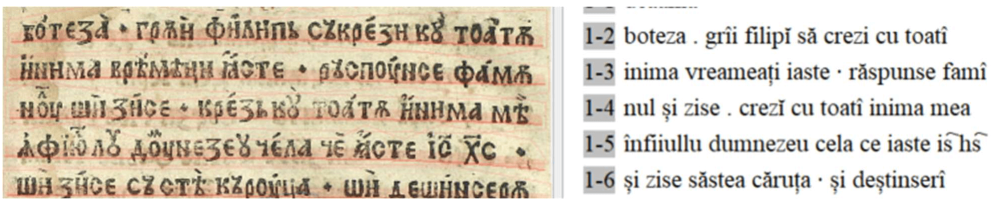
Figure 15: Bilingual_Romanian_Slavic_Cyrillic_0.01 applied to *Voroneț Psalter*, f. 19r

The model struggles with recognising superscripts, both in Romanian and in Slavonic words (see for example lines 5 фечоріилѡ̑ ѡамерилѡ̑ and 4 и чюдеса его сп҃ѿ чл҃ⷱкѡмь), which is a typical problem for Transkribus HTR models. Additionally, there are some errors in identifying certain letters – н is rendered as и in дн҃ꙋи in line 3 and коумънъдаре in line 7, and в is rendered as either и or н in жрътвѫ хвала in line 6. Although the alternation between the two languages is marked by punctuation and spacing, the overall text is written in *scriptura continua*, so that word separation is another source of mistakes for this model. Nonetheless, the combination of high-quality images and more GT data could bring to efficient transcription models for bilingual sources, valuable for any philological work on old Romanian.

The last aspect which we have investigated while using Transkribus on old Romanian texts was the option of transliterating Cyrillic into Latin script. Since the second half of the 20th century, in fact, Romanian scholarship leaned towards transliteration rather than transcription for its critical editions of old texts (Fischer, 1962; Avram, 1964), so that the development of transliterating models might prove beneficial in the field. In order to create Ground Truth for a transliterating model, we chose a standard transliteration table from Romanian Cyrillic to Latin script, used a converter to transliterate the GT data, re-uploaded the data and trained a new, smart model with the ability to transliterate from Cyrillic to Latin script[6]. We are fully aware that the principles of transliteration applied by the model are somewhat controversial. However, our main goal was to provide a proof of concept that transliterating HTR models for Romanian Cyrillic work and that there is no need for one-to-one correspondence between the visual image of the source and the transcribed character. In the future, the scientific community should decide on a generally accepted transliteration system used for transliterating HTR models, for which the rules of the *transcrierea interpretativă fonetică* (interpretative phonetic transcription) result unsuitable. The model's performance is shown in Fig. 16 (Coresi's 1563 *Apostolos*, p. 37).

As with transcribing, when transliterating it is important to render diplomatically the original source, which in turn will give consistency to the GT data used for training a HTR model. Here the model (CER 5.32%) recognises faithfully all letters and spacing (but in lines 5 ꙟ фїю̑ лу/*în fiiul lu* and 6 for съ стѣ/*să stea*) in accordance with the transliterating rules we have used originally, so that ѫ is rendered consistently with *î*, ъ with *ă*, ь with *ĭ*, etc. Additionally, the superscript letters are all correctly identified and brought

[6] The table of correspondence has been taken from *Wikipedia*, while the converting tool is *Protea*.



1-2 boteza . grîi filipĭ să crezi cu toatî
1-3 inima vreameați iaste · răspunse famî
1-4 nul și zise . crezĭ cu toatî inima mea
1-5 înfiiullu dumnezeu cela ce iaste is͡ hs͡
1-6 și zise săstea căruța · și deștinserî

Figure 16: BRV12_Transliterated Romanian applied to 1563 *Apostolos*, p. 37

down to line, so that the transliterated text is fairly easy to read.

3. Implications

We have shown that by using HTR technology and Transkribus a large number of manuscripts or early printed books can be pre-transcribed efficiently. While the error rate of the computer model is higher than that of a trained specialist, the costs are incomparably lower. Similar to traditional (manual) editorial projects where there is at least one complete round of correction after the first transcription, usually by the Principal Investigator of the editorial project, there needs to be at least one arguably more time-consuming round of correction after HTR transcription. Still, as shown in Rabus (in press), the overall cost of an editorial project using Transkribus amounts to roughly one tenth of the overall cost of a traditional project with exclusively manual labour.

Taking this factor into account, it is not too presumptuous to state that HTR technology can be a game changer for the Humanities in the Digital Age. For the first time in history, it allows for the mass digitization of a huge amount of previously unedited sources in a small amount of time and using considerably fewer financial resources than in traditional projects.

Moreover, there is the exciting perspective of using HTR for digitising whole archives and being able to conduct full-text searches in different manuscript (see, e.g., *transkribus.eu*). Additionally, following the quantitative turn, new and interesting ways of research can be developed that do not rely on manual post-correction of HTR results, opening up completely new perspectives on historical textual data (e.g., Camps *et al.*, 2020).

4. Conclusion and outlook

In this paper, we have presented our first experiments with Handwritten Text Recognition for Romanian Cyrillic using the Transkribus platform. Even though the current models are not perfect yet and sometimes commit rather simple errors, we hope to have shown the potential of this technology for Romanian philology in the Digital Age. In the future, it will be a challenging task to train models for different handwriting styles such as for cursive Cyrillic and Latin scripts.

Since the quality and versatility of Transkribus HTR models crucially depends on the amount of Ground Truth data used for training, it is of utmost importance to create models that are orders of magnitude larger than the ones discussed in this paper. This can best be done as a collaborative task. Therefore, we appeal to all scholars concerned with Romanian Cyrillic to join forces, to recycle transcriptions/transliterations originally made for other purposes (e.g., for creating printed or online editions), and thus to create collectively GT data in a fast and efficient way. We encourage everyone interested in AI-assisted Handwritten Text Recognition to get in touch and jointly explore new possibilities to create and apply new HTR models for Romanian (Cyrillic). We strongly believe that HTR is a cornerstone for modern digital philology and express our sincere hope that the Romanian scientific community will engage in furthering this cause for the benefit of us all.

Bibliography

A. Primary sources

A.1. Manuscripts

Bratu Codex, Al. Gafton's critical edition available online.

Ciobanu Psalter, The Slavonic-Romanian Book of Psalms, Moldavia, 1573–1585, Rom. ms. no. 3465 BAR, digital copy available online.

Hurmuzaki Psalter, 16th century, Rom. ms. no. 3077 BAR, digital copy available online.

Scheia Psalter, 16th century, Rom. ms. no. 449 BAR, digital copy available online.

Voroneț Codex, 16th century, Rom. ms. no. 448, digital copy available online.

Voroneț Psalter, 16th century, Rom. ms. no. 693, digital copy available online.

A.2. Printed editions

Psalter: [Brașov, deacon Coresi, 1570], accessible online, CRV 16.

Apostolos: [Brașov, deacon Coresi, 1563], accessible online, CRV 12.

B. Literature

Avram, A. (1964). *Contribuții la interpretarea grafiei chirilice a primelor texte romînești*, in "Studii și cercetări lingvistice", **XV** (1–5).

Camps, J-B., Thibault Clérice, T., & Ariane Pinche, A. (2020). *Stylometry for Noisy Medieval Data: Evaluating Paul Meyer's Hagiographic Hypothesis*, [online].

Fischer, I. (1962). *Principii de transcriere a textelor românești*, in "Limba română", **IX** (5), p. 577–581.

Muehlberger, G., Seaward, L., Terras, M., Ares Oliveira, S., Bosch, V., Bryan, M., Colutto S. *et al.* (2019). *Transforming Scholarship in the Archives Through Handwritten Text Recognition*, in "Journal of Documentation", **75** (5), p. 954–976, Crossref.

Rabus, A. (2019). *Recognizing Handwritten Text in Slavic Manuscripts: A Neural-Network Approach Using Transkribus*, in "Scripta & e-Scripta", **XIX**, p. 9–32.

Rabus, A. (in press). *Automatische computergestützte Transkription paläoslavistischer Quellen und ihre Folgen für Korpuslinguistik und Editionsphilologie*, in "Proceedings Humboldt Kolleg Venice 2020".

C. Transkribus models for Romanian Cyrillic

Model Name	**Sources**	**Word Tokens**	**CER**
BRV12_Romanian Printing 16th c	1563 Apostolos, printed	3951	4.46%
Romanian_Cyrillic_0.01/0.02/0.03	Voroneț Codex	5013	5.84%
Combined Romanian Cyrillic_manuscript	1563 Apostolos + Voroneț Codex	8964	5.65%
Psaltirea Hurmuzaki 1/2	Hurmuzaki Psalter	11 439	10.11%
Combined_Romanian_Hurmuzaki	Hurmuzaki Psalter + 1563 Apostolos + Voroneț Codex	20 310	6.78%
CB_Bilingual Romanian-Slavonic	Bratu Codex	11 553	11.12%
PCb_Bilingual Romanian-Slavonic	Ciobanu Psalter	7865	7.97%
Bilingual_Romanian_Slavic_Cyrillic_0.01	Bratu Codex + Ciobanu Psalter	19 418	10.26%
Combined_mono-and_bilingual_Romanian_0.01	Hurmuzaki Psalter + 1563 Apostolos + Voroneț Codex + Bratu Codex + Ciobanu Psalter	39 728	8.31%