Vilnius University Press

Contents lists available at Vilnius University Press

Slavistica Vilnensis
2023, vol. 68(2), pp. 10–23

ISSN 2351-6895 eISSN 2424-6115
DOI: https://doi.org/10.15388/SlavViln.2022.68(2).1

Macarius: A HTR Model for Romanian Slavonic Early Printed Books

Vladimir R. Polomac

University of Kragujevac, Serbia
E-mail: v.polomac@filum.kg.ac.rs
ORCID ID: https://orcid.org/0000-0001-5140-9330

Abstract. The paper describes the process of creating and evaluating the HTR (Handwritten Text Recognition) model for Romanian Slavonic early printed books (first half of the 16th century, Middle Bulgarian Church Slavonic, Cyrillic Script) using the Transkribus software platform, based on the principles of artificial intelligence, machine learning and advanced neural networks. The HTR model was created on the material of Romanian Slavonic early printed books from Târgoviște printing house: the *Liturgikon* from 1508 and the *Teatraevangelion* from 1512 from the oldest printing house managed by hieromonk Macarius, as well as the *Apostle* from 1547 from the printing house managed by Dimitrije Ljubavić. The most important result of the paper is the creation of the first version of the generic HTR model *Macarius* (named in honor of hieromonk Macarius, the first South Slavonic and Romanian printer) with exceptional performance – the percentage of incorrectly recognized characters (including accent marks) is only 2.7%. The research has shown that this HTR model can also be used for the automatic recognition of Romanian Slavonic early printed books published in the second half of the 16th century. HTR model *Macarius* together with Ground Truth data is available to all users of the *Transkribus* platform, which ensures its wider use, as well as the possibility for further improvement of its performance.

Keywords: *Transkribus*, HTR (Handwritten Text Recognition), Romanian Slavonic early printed books, Middle Bulgarian Church Slavonic, hieromonk Macarius, artificial intelligence, machine learning.

Macarius: HTR модел за старе словенске штампане књиге из Румуније

Сажетак. У раду је описан процес креирања и евалуације HTR (Handwritten Text Recognition) модела за старе словенске штампане књиге из Румуније (прва половина XVI века, средњобугарска редакција старословенског језика, ћирилица) помоћу софтверске платформе Transkribus, засноване на принципима вештачке интелигенције, машинског учења и напредних неуронских мрежа. HTR модел је креиран на материјалу старих словенских књига штампаних у Трговишту: на *Служабнику* из 1508. и *Четворојеванђељу* из 1512. године из најстарије штампарије којом је руководио јеромонах Макарије, као и на *Апостолу* из 1547. године из штампарије којом је руководио Димитрије Љубавић. Најважнији резултат рада представља креирање прве верзије генеричког HTR модела *Macarius 1.0.* (названог тако у част јеромонаха Макарија, првог јужнословенског и румунског штампара) са изузетним перформансама – проценат погрешно препознатих карактера (укључујући и акценатске знаке) износи свега 2.7%. Истраживање је показало да се овај HTR модел може користити и за аутоматско препознавање старих словенских штампаних књига

Received: 29.08.2023. **Accepted:** 20.10.2023

из Румуније насталих у другој половини XVI века. HTR модел *Macarius 1.0.* заједно са Ground Trouth подацима доступан је свим корисницима платформе Transkribus, чиме је обезбеђена његова шира употреба, као и могућност даљег усавршавања његових перформанси.
Кључне речи: *Transkribus*, HTR (Handwritten Text Recognition), старе словенске штампане књиге из Румуније, средњобугарска редакција старословенског језика, јеромонах Макарије, вештачка интелигенција, машинско учење.

Macarius: HTR modelis senosioms slavų spausdintoms knygoms iš Rumunijos

Santrauka. Straipsnyje aprašomas HTR (*Handwritten Text Recognition*) modelio kūrimas ir įvertinimas. Modelis skirtas senosioms slavų spausdintoms knygoms iš Rumunijos (XVI a. pirmoji pusė, vidurio bulgarų bažnytinė slavų kalba, kirilica) nagrinėti naudojant *Transcribus* programinės įrangos platformą, paremtą dirbtinio intelekto, informatikos mokslo ir pažangių neuroninių tinklų principais. HTR modelis buvo sukurtas remiantis seniausios Rumunijoje spaustuvės Tirgovišteje seniausių spausdintų slavų knygų medžiaga („Služebnik“, 1508 m., ir „Evangelija“, 1512 m.), taip pat „Apaštalu“, 1547 m., išleistu spaustuvėje, kuriai vadovavo Dimitrijus Ljubavičius.
Svarbiausias darbo rezultatas – sukurta standartinio modelio HTR *Macarius* 1.0 pirmoji versija. Modelis, pavadintas pirmojo pietų slavų ir rumunų spaustuvininko hieromonko Makarijaus vardu, pasižymi išskirtinėmis savybėmis, nes neteisingai atpažintų simbolių procentas (įskaitant kirčio ženklus) siekia tik 2,7 procento. Tyrimai parodė, kad šis HTR modelis taip pat gali būti naudojamas automatiniam XVI amžiaus antroje pusėje spausdintų senų slavų knygų iš Rumunijos atpažinimui. Modelis HTR *Macarius* 1.0. kartu su *Ground Truth* duomenimis prieinamas visiems *Transkribus* platformos naudotojams, o tai užtikrina platesnį jo naudojimą, taip pat galimybę toliau jį tobulinti.
Reikšminiai žodžiai: *Transkribus*, automatinis teksto atpažinimas, senosios slavų spausdintos knygos Rumunijoje, vidurio bulgarų bažnytinė slavų kalba, hieromonkas Makarijus, dirbtinis intelektas.

1 Introduction

Recent papers describing the process of creating of the HTR model for Serbian early printed books within the *Transkribus*[1] software platform [Polomac 2022a, 2022b] represent the starting point for the investigation in the current study. The most significant result of these papers is the creation of the publicly available generic HTR model for Serbian early printed books called *Dionisio 2.0*: https://readcoop.eu/model/dionisio-1-0/. This model was trained on the material of Serbian early printed books from the 15th–16th centuries, printed in Cetinje, Venice, Goražde, Gračanica, Mileševa, Belgrade and Mrkša's Church [Polomac 2022b, 159]. The parameters and performance of this HTR model are shown in the following table.

[1] The software platform *Transkribus* (https://readcoop.eu/transkribus/) is a tool for manual and automatic text recognition and search of old manuscripts and printed books, regardless of the time of creation, language or script. This software platform allows its users to train their own HTR model for automatic text recognition. Training the HTR model is an example of machine learning based on advanced neural networks in which the HTR model compares photographs of handwriting and the corresponding letters, words and lines of text in the diplomatic edition. More details about the technological background and the way this platform works can be found in Rabus 2019.

Table 1: Parameters and performance of the model Dionisio 2.0

Engine[1]	Word Count on Train Set[2]	Number of Epochs[3]	CER on Train Set	CER on Validation Set[4]
PyLaia HTR	176,481	72	1.71%	2.40%

Quantitative and qualitative analysis of the performance of the HTR model showed that it could be used to obtain transcripts with a minimal percentage of misrecognized characters (about 2–3%), which usually depends on the quality of the book's photograph, the frequency of use of accented characters and superscript letters, as well as regularity of the use of accent marks in appropriate positions [Polomac 2022b, 159–160]. In the continuation of the research, we were interested in whether the HTR model trained on the material of Serbian early printed books of the 15th–16th centuries can also be used for efficient automatic recognition of the Romanian Slavonic early printed books from the beginning of the 16th century.[6] This question is particularly interesting in view of the fact that the work of the printing house of Đurađ Crnojević in Cetinje from the last decade of the 15th century[7] and the work of the oldest Romanian printing house founded in 1508, probably in the vicinity of Târgoviște in Wallahia,[8] was managed by the same printer – hieromonk Macarius.[9] The

[2] PyLaia HTR is an automatic text recognition software available within the *Transkribus* platform. For more details see *Transkribus Glossary* at https://readcoop.eu/glossary/pylaia/.

[3] The minimum amount of data necessary for a successful training of the HTR model for automatic recognition of old printed books is about 5,000 words, and about 15,000 words for manuscripts. Generic HTR models with high quantitative and qualitative performance (such as the *Dionisio 2.0.* model) can be trained on the material of hundreds of thousands or even millions of words.

[4] The term *epoch* in machine learning stands for „one complete presentation of the data set to be learned to a learning machine" [Burlacu, Rabus 2021, 1].

[5] Character Error Rate (CER) represents the basic quantitative indicator of the success of the HTR model and is obtained by comparing automatically recognized and manually corrected text. The HTR model can be considered successful if the CER is below 5%. For more details cf. *Transkribus Glossary* https://readcoop.eu/glossary/character-error-rate-cer/.

[6] The pioneering paper dedicated to the automatic recognition of the old Romanian Cyrillic script [Burlacu, Rabus 2021] is oriented towards manuscripts to a greater extent, and less so towards old printed books.

[7] The oldest Serbian Church Slavonic and South Slavonic Cyrillic printing house was founded in Cetinje (in today's Montenegro) by ruler Đurađ Crnojević in the last decade of the 15th century. Between 1493 and 1496, five books were printed there, two of which have been preserved in their entirety: the *Octoechos of the Fifth Tone* (completed on January 4, 1494) and the *Psalter with Appendices* (completed on September 22, 1495), two in fragments, probably printed in 1496: the *Octoechos of the Fifth Tone* and *Euchologion (Prayer Book)*, while one book (the *Evangelion*) has not been preserved, but its existence is known to us based on a manuscript copy from the church in Buđanovci from 1548. For more details, see the most recent source [Лазић 2020, 326–330].

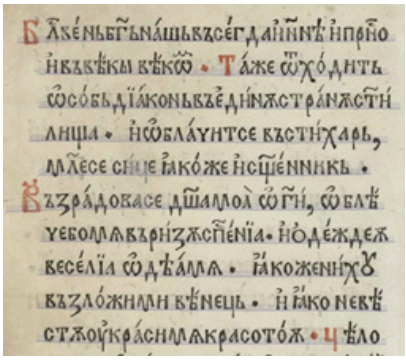
[8] The credits for founding the printing house go to the Wallachian ruler Radul IV (Romanian: Radu cel Mare) (1495–1508) [Erich 2010, 1151]. Between 1508 and 1512, three books were printed: the *Liturgikon* (1508), the *Octoechos* (1510) and the *Tetraevangelion* (1512). The typographical equipment of these books is associated among scholars with Crnojević's printing house in Cetinje, but also

final goal of the paper is to create the first version of a generic model for the automatic recognition of Romanian Slavonic early printed books – *Macarius*, named in honor of the aforementioned hieromonk Macarius, the first South Slavonic and Romanian printer.[9]

2 The process of transcribing and creating a special HTR model

In the first experiment, we tested the performance of the HTR model *Dionisio 2.0* on the oldest Romanian Slavonic printed book – the *Liturgikon* (Romanian: *Liturghier slavonesc*) from 1508.[10] The process of automatic text recognition using this model covered the first 60 pages of the book with 5,239 words. After the manual correction of the automatically recognized transcripts, an automatic statistical analysis was carried out, which showed that the percentage of incorrectly recognized characters on this material is about 7%. This result shows that the HTR model *Dionisio 2.0* trained on the material of Serbian early printed books can be effectively used to obtain the first transcripts of the oldest Romanian Slavonic printed book. Figure 1 presents a qualitative analysis of the performance of the HTR model based on a comparative display of a part of sheet 6v and automatically recognized text using the *Dionisio 2.0* model.

Figure 1. The Liturgikon (1508), part of sheet 6v and automatically recognized text using the Dionisio 2.0 model



1-1 Бл҃вéнь бг҃ъ нáшь въсегдá и нн҃ѣ и прo-
1-2 й въ вѣкы вѣкѡ҃ · тáже ѡ҃хóдить
1-3 ѡсóбь дїáконь въ ѣдúнъ стрáнстú-
1-4 лиця · й ѡ҃блáунтсе въстúхарь,
1-5 мл҃се се сúце íакóже й сі҃йéнникь ·
1-6 Възрáдова се дííа моà ѡ҃ гú, ѡ҃блѣ-
1-7 че бо ми въ рúзь спéнїа · й ѻ҃дéждеж-
1-8 весéлїа ѡ҃дѣáми · íакоже нúхѣ
1-9 възлóжими вѣнець · й íако невѣ-
1-10 стоу҃крáсими красотóж · цѣло-

with printing houses in Venice, as well as with the cyrillic printing house of Schweipolt Fiol in Krakow [Erich 2010, 1151–1153]. The printing house was most likely located in the Dealu monastery near Târgovişte [Erich 2010, 1154]. Middle Bulgarian Church Slavonic is used in all three books.

[9] The high quality of the printing, the similarity in the structure of the text and the repetition of visual elements in the Cetinje and Wallachian books [Simić 2020, 198–200; 2022, 260–261], as well as the graphic and orthographic and linguistic characteristics of the prefaces and afterwords of Cetinje books and the afterwords of Wallachian books serve as a proof that the Cetinje and Wallachian hieromonk are the same person [Savić 2012a, 2012b, 102–104]. We meet the same person in historical sources as the abbot of the Hilandar monastery in Mount Athos between 1525 and 1533 [Savić 2012b, 101–102].

[10] We downloaded the digital photos of the book from the website of the Digital National Library of Romania: http://digitool.bibnat.ro/R.

Based on the previous presentation, we can conclude that the HTR model *Dionisio 2.0* makes most frequent mistakes when recognizing the letters ѫ and ѧ, characteristic of the Middle Bulgarian Church Slavonic in which the *Liturgikon* was written, and unknown in the Serbian Church Slavonic printed books on which the HTR model was trained. So instead of въ ѐди́нѫ стра́нѫ ст҃и́- 3, въ ри́зѫ 7, ми 7, 8, ѻ҆де́ждеѫ 7, невѣ́стѫ оу҆кра́си мѧ 9–10, красото́ѫ 10 the HTR model incorrectly yields въ ѐди́нъ стра́нст҃и- 3, въ ри́зь 7, мѧ 7, 8, ѻ҆де́ждеж- 7, невѣ́стоу҆кра́сими 9–10, красото́ж 10. Other errors refer to the omission of the titlo mark: ннѣ 1 instead of н҃нѣ 1, not recognizing a capital letter: та́же 2 instead of Та́же 2, accent marks: ѿсо́бь 3, ѿблáчит 4, ѿ г҃и 6, ѿблѣ́че 6–7, веселїа 8 уместо ѡ҆со́бь 1, ѡ҆бла́чит 4, ѡ҆ г҃и́ 6, ѡ҆блѣ́че 6–7, весе́лїа 8, spaces between words: ꙗ҆́коже ни́хꙋ 8 instead of ꙗ҆́ко жени́хꙋ 8, and only in one example the letter п҃ро- 1 instead of п҃рно 1.

The qualitative analysis of errors indicates the necessity of training a special HTR model for the automatic recognition of the Romanian Slavonic early printed books. After the manual correction of the automatically recognized transcripts of the first 60 pages of the *Liturgikon* (1508), we obtained a minimal set of Ground Truth data[11] for training the first version of the *Macarius* HTR model. The parameters and performance of the HTR model *Macarius 0.1* are shown in the following table.

Table 2: Parameters and performance of the HTR model Macarius 0.1

Engine	Base Model	Word Count on Train Set	Number of Epochs	CER on Train Set	CER on Validation Set
PyLaia HTR	Dionisio 2.0	5,026	250	0.20%	3.51%

The previous table shows how by training the special HTR model *Macarius 0.1* based on only 5,026 words (i.e. 60 pages) of the *Liturgikon* (1508) and the basic HTR model[12] *Dionisio 2.0.*, the percentage of incorrectly recognized characters is an exceptional at 3.51%. In the further course of transcription using the HTR model *Macarius 0.1* transcripts of the next 70 pages (from pages 61 to 130) of the *Liturgikon* (1508) were automatically recognized. After the manual correction of the automatically recognized transcripts, a more extensive set of Ground Truth data was obtained for training the second version of

[11] The term Ground Truth Data in machine learning refers to completely accurate data used to train the model. In our case, these would be exact transcripts of digital photographs of the manuscript or early printed book. For more details on this term, see *Transkribus Glossary* at https://readcoop.eu/glossary/ground-truth/.

[12] Learn more about the training process using the basic HTR model in *Transkribus Glossary* at readcoop.eu/glossary/base-models/.

the *Macarius* HTR model. The parameters and performance of the HTR model *Macarius 0.2* are shown in the following table.

Table 3: Parameters and performance of the HTR model Macarius 0.2

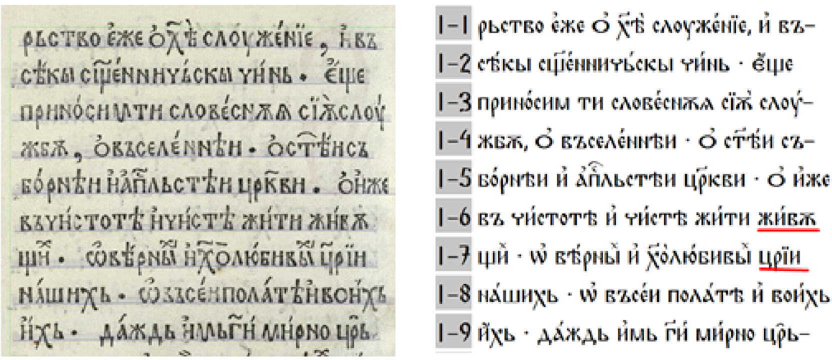
Engine	Base Model	Word Count on Train Set	Number of Epochs	CER on Train Set	CER on Validation Set
PyLaia HTR	Dionisio 2.0	10,535	250	0.10%	2.60%

The previous table shows how increasing the set of Ground Truth data when training the second version of the HTR model leads to a decrease in the percentage of incorrectly recognized characters. In the continuation of the transcription, we applied the HTR model *Macarius 0.2* to the rest of the *Liturgikon* (pp. 131–258). After the manual correction of the automatically recognized transcripts, we obtained an even more extensive set of Ground Truth data for training the third version of the *Macarius* HTR model. The parameters and performance of the HTR model *Macarius 0.3* are shown in the following table.

Table 4: Parameters and performance of the HTR model Macarius 0.3

Engine	Base Model	Word Count on Train Set	Number of Epochs	CER on Train Set	CER on Validation Set
PyLaia HTR	Dionisio 2.0	19,972	250	0.10%	2.00%

Figure 2. The Liturgikon (1508), part of sheet 44v and the automatically recognized text using the Macarius 0.3 model



The quantitative analysis shows that by increasing the training set, the percentage of incorrectly recognized characters decreased to only 2%. The real

success of the HTR model and the nature of the errors is best evidenced by the comparative display of a part of sheet 44v and the automatically recognized text in the figure 2.

As the figure shows, the HTR model *Macarius 0.3* successfully recognizes not only letters, but also accent marks, punctuation marks, titlo marks and superscripts, as well as spaces between words. On the given sample, the HTR model made only two mistakes: in the example жи́вѫ 6, the space between words is not marked (it should be жи́вѫ¬ 6), and in the example цр҃ïи 7, the titlo mark is missing (it should be ц҆р҃ïи 7).

3 The application of the HTR model *Macarius 0.3* to other books printed in Târgovişte

In the following experiment, we tested the performance of the HTR model *Macarius 0.3* in the automatic text recognition of the *Octoechoes* (Romanian: *Octoih slavonesc*) from 1510 and the *Tetraevangelion* (Romanian: *Evangheliar slavonesc*) from 1512, the other two books published in the oldest Romanian printing house in Târgovişte under the direction of the printer hieromonk Macarius.[13] The results of the experiment are shown in the following table.

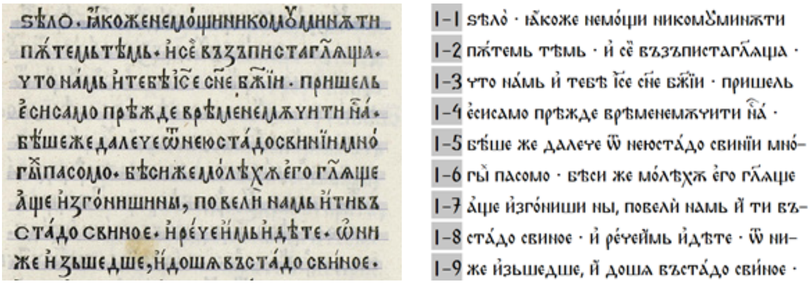
Table 5: The application of the Macarius 0.3 model to other printed books from Târgovişte

Book	Character Error Rate (CER)
Octoechoes (1510)	9.05%
Tetraevangelion (1512)	4.81%

The previous table shows how the percentage of misrecognized characters in the *Octoechoes* from 1510 and the *Tetraevangelion* from 1512 differs significantly. The higher percentage of misrecognized characters in the *Octoechoes* from 1510 can be explained by the poor quality of the photographs (black and white photos with low contrast) used in the experiment. The percentage of misrecognized characters in the *Tetraevangelion* from 1512 is less than 5%, which can be considered exceptional since the HTR model *Macarius 0.3* did not have the chance to go through the material from this book during training. The qualitative performance analysis of the HTR model *Macarius 0.3* on the *Tetraevangelion* from 1512 is illustrated by a comparison of a part of sheet 24r and the automatically recognized text in the following figure.

[13] The photos of the mentioned books were taken from the website of the Digital Library of Bucharest http://digitool.bibmet.ro:8881/R?RN=351072892, section *Opera tipografică românească* (1508–1830).

Figure 3. The Tetraevangelion (1512), part of sheet 24r and the automatically recognized text using the HTR model Macarius 0.3



1-1 ѕѣлѡ̀ · ꙗ҆́коже немо́щи никомꙋ҆минѫти
1-2 пѫтемь тѣмь · и҆ сё възъпистагл҃аща ·
1-3 что на́мь и҆ тебѣ і҆с҃е сн҃е бж҃їи · пришель
1-4 ѐсисамо прѣжде врѣменемѫчити ны҃а́ ·
1-5 бѣ́ше же далече ѿ неюста́до свинїи мно́–
1-6 гы҃ пасомо · бѣси же мо́лѣхѫ ѐго гл҃аще
1-7 а҆́ще и҆згѡ́ниши ны, повелѝ намь и҆҆ ти въ–
1-8 ста́до свиное · и҆ ре́чеи҆мь и҆дѣте · ѿ ни–
1-9 же и҆зьшедше, и҆ доша въста́до сви́ное ·

Based on the comparative overview, it can be concluded that the HTR model *Macarius 0.3* makes most frequent mistakes when recognizing spaces between words. Thus, instead of немо́щи никомꙋ҆минѫти 1, възъпистагл҃аща 2, ѐсисамо 4, врѣменемѫчити 4, неюста́до 5, въ– 7, ре́чеи҆мь 8, и҆ доша 9, въста́до 9 it should be не мо́щи никомꙋ҆ минѫти 1, възъписта гл҃аща 2, ѐси само 4, врѣмене мѫчити 4, нею ста́до 5, въ 7, ре́че и҆мь 8, и҆доша 9, въ ста́до 9. It seems interesting that the HTR model rarely makes mistakes when recognizing accent marks: only ѕѣлѡ̀ 1 instead of the expected ѕѣлѡ 1. In one example, we noticed some errors in the recognition of accents, spaces between words and the recognition of ѡ: ѿ ни– 8 instead of the expected ѡ҆ни 8.

The exceptional performance of the HTR model *Macarius 0.3* on the *Tetraevangelion* from 1512 enabled us to use the Ground Truth data from this book to train a new version of the model. After the automatic recognition and manual correction of 140 pages of the *Tetraevangelion* from 1512, we doubled the amount of Ground Truth data for training the *Macarius 0.4* model. The parameters and performance of the new version of the model are shown in the following table.

Table 6: Parameters and performance of the HTR model Macarius 0.4

Engine	Base Model	Word Count on Train Set	Number of Epochs	CER on Train Set	CER on Validation Set
PyLaia HTR	Dionisio 2.0	39,688	250	0.20%	2.20%

4 Creating and evaluating the *Macarius 1.0* generic model

In the further course of research, we were interested in the performance of the HTR model *Macarius 0.4.* on Romanian Slavonic early printed books published in Dimitrije Ljubavić's printing house in Târgovişte at the end of the

first half of the 15th century.[14] Since the typographic characters of the old Serbian printing house from Goražde were used during the printing of books in Dimitrije Ljubavić's printing house in Târgovişte, it was also interesting to compare the performance of the HTR model *Dionisio 2.0* and *Macarius 0.4* on the *Prayer Book* from 1545 and the *Apostle* from 1547. The results of the experiment are shown in the following table.

Table 7: Performance comparison of the Dionisio 2.0 and Macarius 0.4 models on books from Dimitrije Ljubavić's printing house

Book/Model	*Dionisio 2.0*	*Macarius 0.4*
Prayer Book (1545)	CER 5.65%	CER 9.61%
Apostle (1547)	CER 10.22%	CER 7.51%

Based on the previous table, it can be concluded that the HTR model *Dionisio 2.0* shows better results when automatically recognizing the *Prayer Book* from 1545, and the HTR model *Macarius 0.4* shows better results when automatically recognizing the *Apostle* from 1547. Such results, somewhat unexpected, can be explained by the fact that the *Prayer Book* from 1545 was printed in Serbian Church Slavonic, while the *Apostle* from 1547 was printed in Middle Bulgarian Church Slavonic. This experiment also demonstrates how the parallel use of HTR models trained on different Church Slavonic recensions can represent a reliable indicator for language identification in the process of automatic text recognition.

Table 8: Structure and amount of data for training the generic HTR model Macarius 1.0

Book	Word Count
Liturgikon (1508)	19,972
Tetraevangelion (1512)	19,696
Apostle (1547)	17,147

The quantitative performance of the HTR model *Macarius 0.4* on the books from Dimitrije Ljubavić's printing house in Târgovişte points to the necessity of training a new version of the HTR model that will also include the material from the *Apostle* from 1547. By manually correcting the automatically recog-

[14] After the first Romanian printing house in Târgovişte stopped working, the printing business was renewed at the end of the first half of the 16th century by Dimitrije Ljubavić, the grandson of Božidar Ljubavić [Cazacu 1995, Лазић 2020, 340], the founder of the Serbian printing house in Goražde (1519–1523). Two books were printed in Dimitrije Ljubavić's printing house in Târgovişte: the *Prayer Book (Euchologion)* from 1545 and the *Apostle* from 1547.

nized transcripts of 108 pages of the *Apostle* from the year 1547, we obtained a new set of Ground Truth data (about 17,000 words) for training the first version of the generic HTR model *Macarius 1.0*. The structure of this model and the amount of training data are shown in the following table.

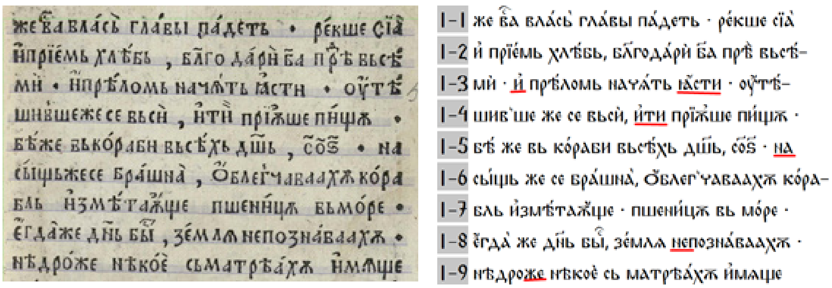
The parameters and performance of the generic HTR model *Macarius 1.0* are shown in the following table.

Table 9: Parameters and performance of the generic HTR model Macarius 1.0

Engine	Base Model	Word Count on Train Set	Number of Epochs	CER on Train Set	CER on Validation Set
PyLaia HTR	Dionisio 2.0	56,815	250	0.30%	2.70%

The qualitative performance of the HTR model *Macarius 1.0* can be illustrated by the comparative overview of a part of sheet 52r of the *Apostle* from 1547 and the automatically recognized text in the next figure.

Figure 4. The Apostle (1547), part of sheet 52r and the automatically recognized text using the Macarius 1.0 model



As can be seen in the figure, the HTR model does not always recognize accent marks and the space between words: so instead of и̏ 3, ꙗ̑сти 3, и̂ти̑ 4 there is the incorrect и̏ 3, ꙗ̑сти 3, и̂ти 4; instead of на- 5, не позна́ваахѫ 8, нѣдро́ же 9 there is the incorrect на 5, непозна́ваахѫ 8, нѣдроже 9. These errors do not significantly affect the quality of the transcript, which can be used for further philological and linguistic research even without any manual correction.[15]

The performance of the HTR model *Macarius 1.0.* was tested on the samples of several books printed in the printing houses of Philip Pictor Moldoveanu, diacon Coresi, diacon Lorinţ and diacon Şerban[16] starting from the end

[15] On the usability of uncorrected HTR transcripts of the Church Slavonic language for linguistic analysis, for more details see [Rabus, Petrov 2023].

[16] The digital photos of the books used in this experiment were taken from the website of the Digi-

of the first half of the 16th century. The quantitative performance indicator of this HTR model compared to the basic HTR model *Dionisio 2.0.* is shown in the following table.

Table 10: The application of the HTR model Macarius 1.0 and Dionisio 2.0 on The Romanian Slavonic early printed books from the end of the first half of the 16th century

Book (Year)	Printer, Town	*Macarius 1.0*	*Dionisio 2.0*
Tetraevangelion (1546)	Filip Pictor Moldoveanu, Sibiu	5.57%	12.18%
Octoechos (1557)	Diacon Coresi, Braşov	3.79%	8.06%
Triod (1578)	Diacon Coresi, Braşov	9.86%	13.66%
Tetraevangelion (1579)	Diacon Lorinţ, Alba Iulia	6.60%	17.23%
Liturgikon (1588)	Diacon Şerban, Braşov	5.27%	10.78%

As shown in the table above, the HTR model *Macarius 1.0.* can also be successfully used for the automatic recognition of Romanian Slavonic early printed books from the end of the first half of the 16th century. The percentage of misrecognized characters, as we can see, generally hovers around an excellent 5–6%, while in the books printed by diacon Coresi, the percentage of misrecognized characters varies, from an exceptional 3.79% on the *Octoechos of the Fifth Tone* from 1557, to somewhat worse, but still a very good 9.86% on the *Triod* from 1578. The difference in the performance of the HTR model mostly depends on the degree of preservation of the book and the quality of the photographs, as well as on the arrangement and frequency of accent marks in the book. The quantitative performance listed in the table shows how the HTR model *Macarius 1.0.* can automatically obtain transcripts of Romanian Slavonic early printed books from the second half of the 16th century, usable for further philological and linguistic research even without manual correction (based on the example of the *Octoechos of the Fifth Tone* from 1557 with a percentage of misrecognized characters between 3–4%). On the other hand, manual correction of automatically recognized transcripts can be obtained in a much shorter time and with fewer human and financial resources. The comparison of the performance of *Macarius 1.0.* with the performance of the basic model *Dionisio 2.0* clearly indicates the justification of creating a special HTR model for the Romanian Slavonic early printed books.

tal Library of Bucharest http://digitool.bibmet.ro:8881/R?RN=351072892, section *Opera tipografică românească* (1508–1830). For basic information about the printing activity of Philip Pictor Moldoveanu, diacon Coresi, diacon Lorinţ and diacon Şerban see [Petrov 2015, 126–134].

5 Concluding remarks

The conducted research showed that using the *Transkribus* software platform, based on the principles of artificial intelligence, machine learning and advanced neural networks, extremely effective HTR models can be created for the Romanian Slavonic early printed books from the 16th century. The most important result of the paper is the creation of the publicly available generic HTR model for Romanian Slavonic early printed books, called *Macarius 1.0* in honor of hieromonk Macarius, the first South Slavonic and Romanian printer. This HTR model was created on the material of Romanian Slavonic early printed books from the Târgovişte printing house: the *Liturgikon* from 1508 and the *Tetraevangelion* from 1512 from the oldest printing house managed by hieromonk Macarius, as well as the *Apostle* from 1547 from the printing house managed by Dimitrije Ljubavić. Quantitative performance evaluation of the *Macarius 1.0.* HTR model showed that the percentage of misrecognized characters on Romanian Slavonic early printed books from Târgovişte is 2.7%, which can be rated as an exceptional result. Qualitative evaluation of the model's performance indicates that recognition errors are mostly related to accent marks and spacing between words. The research also showed that this HTR model can be used for the automatic recognition of Romanian Slavonic early printed books from the end of the first half of the 16th century in the printing houses of Philip Pictor Moldoveanu, diacon Coresi, diacon Lorinţ and diacon Şerban. The percentage of misidentified characters on books from these printers varies from exceptional (about 3–4%), through excellent (about 5–6%) to very good (about 9–10%). By manually correcting the automatically recognized transcripts of books from the second half of the 16th century using this HTR model, it is possible to relatively quickly obtain new Ground Truth data that will enable the creation of a new version of the HTR model capable of automatic recognizing of Romanian Slavonic early printed books as a whole. For the continuation of research in this direction, the fact that the HTR model *Macarius 1.0* together with Ground Truth data is available to all users of the *Transkribus* software platform remains of utmost importance, enabling its wider use, as well as the possibility for further improvement of its performance.

References

BURLACU, C., RABUS A., 2021. Digitising (Romanian) Cyrillic using Transkribus: new perspectives, *Diacronia*, 14. 1–9. https://doi.org/10.17684/i14A196en

CAZACU, M., 1995. Dimitrije Ljubavić (c. 1519–1564) et l' imprimerie slave dans l' Europe Sud-est au XVIe siècle. Nouvelles contributions, *Anuarul Institutului de Istorie A. D. Xenopol*, 32. 187–207.

ERICH, A., 2010. The Appearance of Printing Activity in the Romanian Space: an Integral Part of a European Phenomenon, *WSEAS Transaction on Information Science and Applications*, 9(7). 1150–1159.

PETROV, I., 2015. *Od inkunabułów do pierrwszych gramatyk. Konteksty rozwoju bułgarskiego języka literackiego (koniec XV – początek XVII wieku)*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

POLOMAC, V., 2022a. Serbian Early Printed Books from Venice: Creating Models for Automatic Text Recognition using Transkribus, *Scripta & e-Scripta*, 22. 11–29.

POLOMAC, V., 2022b. Serbian Early Printed Books: Towards Generic Model for Automatic Text Recognition using Transkribus. In FIŠER, D., ERJAVEC, T. (eds.). *Proceedings of the Conference on Language Technologies and Digital Humanities*. Ljubljana: Institute for Contemporary History, 154–161.

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

RABUS, A., Petrov, I., 2023. Linguistic Analysis of Church Slavonic Documents: A Mixed-Methods Approach, *Scando Slavica*, 69/1. 25–38. https://doi.org/10.1080/00806765.2023.2189617

SIMIĆ, V., 2020. Early cyrillic printed books and the migration of decorative forms between the Adriatic and the Danube around 1500, *Zograf*, 44. 189–204. https://doi.org/10.2298/ZOG2044189S

SIMIĆ, V., 2022. Between Venice and the Danube. Hieromonk Makarije and His Cyrillic Incunabula at the Turn of the Sixteenth Century. In PAYNE, A. (ed.). *The Land between Two Seas: Art on the Move in the Mediterranean and the Black Sea 1300–1700*. Leiden: Brill, 252–270.

ЛАЗИЋ, М., 2020. Инкунабуле и палеотипи: српскословенске штампане књиге од краја 15. до средине 17. века. In ПУЗОВИЋ, В., ТАТАЛОВИЋ, В. (ур.). *Осам векова аутокефалије Српске православне цркве*, *књ. 2*. Београд: Свети Архијерејски Синод Српске православне цркве, Православни богословски факултет, 325–344.

САВИЋ, В., 2012а. Записи штампара свештеномонаха Макарија. Језик, писмо, правопис. In СТОЈАНОВИЋ, Ј. (ур.). *Српско језичко наслеђе на простору данашње Црне Горе и српски језик данас*. Подгорица: Матица српска – друштво чланова у Црној Гори, 159–179.

САВИЋ, В., 2012б. Два значајна записа о штампару Макарију, *Прилози за књижевност, језик, историју и фолклор*, 78. 101–112.

References (Transliteration)

BURLACU, C., RABUS A., 2021. Digitising (Romanian) Cyrillic using Transkribus: new perspectives, *Diacronia*, 14. 1–9. https://doi.org/10.17684/i14A196en

CAZACU, M., 1995. Dimitrije Ljubavić (c. 1519–1564) et l' imprimerie slave dans l' Europe Sud-est au XVIe siècle. Nouvelles contributions, *Anuarul Institutului de Istorie A. D. Xenopol*, 32. 187–207.

ERICH, A., 2010. The Appearance of Printing Activity in the Romanian Space: an Integral Part of a European Phenomenon, *WSEAS Transaction on Information Science and Applications*, 9(7). 1150–1159.

LAZIĆ, M., 2020. Inkunabule i paleotipi: srpskoslovenske štampane knjige od kraja 15. do sredine 17. veka. In PUZOVIĆ, V., TATALOVIĆ, V. (ur.). *Osam vekova autokefalije Srpske pravoslavne crkve*, *knj. 2*. Beograd: Sveti Arhijerejski Sinod Srpske pravoslavne crkve, Pravoslavni bogoslovski fakultet, 325–344.

PETROV, I., 2015. *Od inkunabułów do pierrwszych gramatyk. Konteksty rozwoju bułgarskiego języka literackiego (koniec XV – poszątek XVII wieku)*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

POLOMAC, V., 2022a. Serbian Early Printed Books from Venice: Creating Models for Automatic Text Recognition using Transkribus, *Scripta & e-Scripta*, 22. 11–29.

POLOMAC, V., 2022b. Serbian Early Printed Books: Towards Generic Model for Automatic Text Recognition using Transkribus. In FIŠER, D., ERJAVEC, T. (eds.). *Proceedings of the Conference on Language Technologies and Digital Humanities*. Ljubljana: Institute for Contemporary History, 154–161.

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

RABUS, A., Petrov, I., 2023. Linguistic Analysis of Church Slavonic Documents: A Mixed-Methods Approach, *Scando Slavica*, 69/1. 25–38. https://doi.org/10.1080/00806765.2023.2189617

SAVIĆ, V., 2012a. Zapisi štampara sveštenomonaha Makarija. Jezik, pismo, pravopis. In STOJANOVIĆ, J. (ur.). *Srpsko jezičko nasljeđe na prostoru današnje Crne Gore i srpski jezik danas*. Podgorica: Matica srpska – društvo članova u Crnoj Gori, 159–179.

SAVIĆ, V., 2012b. Dva značajna zapisa o štamparu Makariju, *Prilozi za književnost, jezik, istoriju i folklor*, 78. 101–112.

SIMIĆ, V., 2020. Early cyrillic printed books and the migration of decorative forms between the Adriatic and the Danube around 1500, *Zograf*, 44. 189–204. https://doi.org/10.2298/ZOG2044189S

SIMIĆ, V., 2022. Between Venice and the Danube. Hieromonk Makarije and His Cyrillic Incunabula at the Turn of the Sixteenth Century. In PAYNE, A. (ed.). *The Land between Two Seas: Art on the Move in the Mediterranean and the Black Sea 1300–1700*. Leiden: Brill, 252–270.

Vladimir Polomac, PhD in philology, full professor at the Department of the Serbian Language, Faculty of Philology and Arts, University of Kragujevac (Serbia).

Владимир Поломац, PhD, редовни професор на Катедри за српски језик, Филолошко-уметнички факултет, Универзитет у Крагујевцу (Србија).

Vladimiras Polomacas, filologijos mokslų daktaras, Kragujevaco universiteto Filologijos ir menų fakulteto Serbų kalbos katedros profesorius (Serbija).