UDC 091(=163.41): 004.352.243
https://doi.org/10.18485/ms_zmsfil.2021.64.2.1
Изворни научни рад

Vladimir Polomac
Tamara Lutovac Kaznovac

AUTOMATIC RECOGNITION OF SERBIAN MEDIEVAL MANUSCRIPTS BY APPLYING THE *TRANSKRIBUS* SOFTWARE PLATFORM: CURRENT SITUATION AND FUTURE PERSPECTIVES

The paper investigates the potentials of applying the model for the automatic recognition of (Russian) Church Slavonic manuscripts to Serbian medieval manuscripts written in various types of Cyrillic scripts by employing the Transkribus software platform. The analysis has shown: (a) that the use of the existing generic model for the recognition of Church Slavonic manuscripts can yield rather good results when applied to Serbian medieval manuscripts written in uncial or semiuncial script, (b) that the manuscripts written in the cursive script require the creation of a separate model, and (c) that the creation of a generic model within the Transkribus platform for the Serbian medieval manuscripts would make the process of digitization substantially faster, which in turn would lead to faster realization of tasks within the existing projects related to Serbian historical corpus linguistics and lexicography .

Key words: Transkribus, Serbian medieval manuscripts, automatic text recognition, information technology, artificial intelligence, machine learning.

У раду се истражују могућности примене модела за аутоматско рашчитавање (руских) црквенословенских рукописа у оквиру софтверске платформе Transkribus на српске средњовековне рукописе писане различитим типовима ћирилице. Анализа је показала (а) да примена постојећег генеричког модела за аутоматско рашчитавање црквенословенских рукописа може дати веома добре резултате на српским средњовековним рукописима писаним уставом или полууставом, (б) да је за рукописе писане брзописом неопходно креирати посебан модел, и (в) да би се креирањем генеричког модела за српске средњовековне рукописе у оквиру платформе Transkribus, процес дигитализације могао значајно убрзати, што би даље могло водити и убрзању рада на текућим пројектима из српске историјске корпусне лингвистике и лексикографије.

Кључне речи: Transkribus, српски средњовековни рукописи, аутоматско рашчитавање текста, информационе технологије, вештачка интелигенција, машинско учење.

1. Introduction. The basis for starting the work on this paper is the article by A. Rabus (2019a), on the potentials of the automatic recognition of Church Slavonic manuscripts by using the Transkribus software platform. In the article, the first of its kind within Slavonic studies, the author starts by providing a brief overview of previous attempts to develop the technology for the automatic recognition of medieval manuscripts,[1] including the technology behind the Transkribus

[1] The paper primarily focuses on rare occasions of applying the OCR technology to the automatic recognition of old Slavonic printed books and printed editions of Church Slavonic manuscripts (the electronic edition of Bdin collection (*Bdinski sbornik*) and RRuDI corpus), but also emphasizes its complete inapplicability to automatic recognition of old Slavonic manuscripts (cf. Rabus 2019a: 10). The possibility of automatic recognition of old Slavonic manuscripts and printed books by using

platform,[2] reserving the central part of his paper to the creation of the model for the automatic recognition of Church Slavonic manuscripts,[3] accompanied by the quantitative and qualitative analysis of the results obtained by applying the said methods to the manuscripts written in various types of Church Slavonic Cyrillic script. The overall significance of the Rabus 2019a paper lies in the fact that by using concrete examples the paper convincingly showed that automatic recognition of Church Slavonic manuscripts by applying the Transkribus is indeed a reality, since the first version of the recognized text in the electronic form that has an acceptable error ratio (about 4% of all characters)[4] is automatically achieved, and after the manual correction done by a competent philologist, investing much less time and human and financial resources, the result is the text of the manuscript in electronic form, suitable for further philological and linguistic investigations. In doing so, this paper provides direction on how to significantly expedite the digitization of medieval Slavonic manuscripts by the use of the Transkribus platform, thus providing a significant impetus for the development of diachronically oriented Slavonic studies.[5] The special importance of the Rabus 2019a paper can be found in the fact that the models for the automatic recognition of Church Slavonic manuscripts are made publicly available through the Transkribus platform, so their potential can be tested on other medieval Slavonic manuscripts as well. The investigation concerning the application of possibilities of such models to medieval Serbian manuscripts written in various types of Cyrillic script represents the overall goal of the current article. Since the models in question are based on

the artificial intelligence based on neural networks was for the first time mentioned in Корниенко – Черепанов – Ясницкий 2008. This paper is more significant in a theoretical rather than practical sense since the level of character recognition accuracy is around 80%, thus requiring a large amount of time for manual corrections to the text, thus rendering the entire recognition process no more economical than the traditional approach (cf. Rabus 2019a: 10).

[2] Transkribus is a free access software platform for the automatic recognition and search of manuscripts which was developed within the READ project at the University of Innsbruck. Unlike the traditional approach that focuses on individual letters (OCR technology), Transkribus uses HTR technology based on memorizing and recognizing the entire image of the line from the text. Recently developed and implemented into Transkribus, HTR+ algorithm is based on the artificial intelligence and advanced neural networks and significantly reduces the time required for the training of text recognition models, with a substantially higher accuracy ratio. For more details, see Rabus 2019a: 10–11.

[3] The functionality of the Transkribus platform is particularly manifested in the potential to train one's own automatic text recognition model, irrespective of the language or script used in the manuscript. The training of the automatic recognition model represents an instance of machine learning based on neural networks in which during the learning process the model compares the manuscript photographs and corresponding letters, words and lines of the text in the diplomatic edition. The successful training of a model requires photographs of the manuscript having the best possible quality and at least 15000 words of previously recognized text. For more details, see Rabus 2019a: 11–14.

[4] Transkribus possesses the possibility to automatically calculate the ratio of incorrectly recognized letters (CER) by comparing the automatically recognized version of the text and manually corrected version. For more details, see Transkribus Glossary at https://readcoop.eu/glossary/character-error-rate-cer/.

[5] Automatic recognition of Serbian (medieval) manuscripts could significantly expedite the work on the current Serbian historical lexicographic projects (*Dictionary of the 12th–18th Century Serbian Language* and *Dictionary of the Slavonic Serbian Language*), as well as the preparation of the electronic historical corpus of the Serbian language.

Old Church Slavonic and Russian Church Slavonic manuscripts written in uncial or semiuncial Cyrillic scripts, our paper starts from the hypothesis that their application to medieval Serbian manuscripts written in the uncial or semiuncial can yield more or less acceptable results,[6] while the manuscripts written in cursive Cyrillic script should require the creation of a separate recognition model. In structuring the paper we proceeded in line with the stated hypotheses. Accordingly, the Section 2 provides a detailed overview of the existing models for the automatic recognition of medieval Slavonic manuscripts, including the analysis of the results of their application to medieval Serbian manuscripts written in various types of Cyrillic script, while the Section 3 provides concluding remarks and the perspectives for further research.

2. Application of existing models for the automatic text recognition to Serbian medieval manuscripts. Two models for the automatic recognition of Church Slavonic Cyrillic manuscripts are available as a part of the Transkribus software platform.[7] The first model, called VMČ_Test_4+, is based on portions of the Russian Church Slavonic manuscript *The Great Reading Menology*, written in semiuncial 16th century Cyrillic script. A total of 173,287 words were used for the training of the model, with CER (Character Error Rates) being 3.82% (more details can be found in Rabus 2019a: 15–19). The other model, dubbed Combined_Full_VKS_2 and based on parts of the Old Church Slavonic *Codex Suprasliensis* (11th century), *The Catecheses of Cyril of Jerusalem* manuscript (11th century) and the Russian Church Slavonic manuscript *Great Reading Menology* (16th century), represents an attempt at creating a generic model suitable for the automatic recognition of different manuscripts written in uncial or semiuncial Cyrillic script. A total of 393,079 words were used for the training of the model, with 3.94% CER (for more details, see Rabus 2019a: 23–27).[8] These models were tested on Serbian medieval manuscripts which are currently the focus of interest of our philological and linguistic investigations: a) on Serbian medieval charters and letters currently being prepared to be used for the development of a specialized electronic corpus (cf. Polomac 2021), as well as b) on the Serbian Church Slavonic manuscript by the name of *Christian Topography of Cosmas Indicopleustes*, which is in the process of preparation for publication in its original graphemic structure together with the associated philological and linguistic studies.[9]

2.1. Automatic recognition of Serbian medieval charters and letters. When choosing the charters and letters to be included in the investigation, we took into account the size of the manuscripts, their philological and cultural significance, state of the manuscripts' preservation and legibility, as well as the availability of

[6] The term *semiuncial* denotes the Resavian type of uncial script (cf. Јерковић 1996).

[7] A. Rabus has also created two publicly accessible models for the automatic recognition of the Glagolitic script: the first model contains approximately 28,000 words from various printed Glagolitic books from Tübingen and Urach (see https://readcoop.eu/model/glagolitic-print/), while the other model comprises approximately 171,000 words from the Breviary of Vid of Omišalj and the Second Beram Breviary (see https://readcoop.eu/model/glagolitic-handwritten-14th-and-15th-century/).

[8] For the potentials and problems in the creation of a generic model for the automatic text recognition within the Transkribus platform, see Rabus 2019b, Hodel et al. 2021.

[9] This edition and study is being prepared for publication by Tamara Lutovac Kaznovac.

images with proper quality. Having in mind these criteria, the expected choice was made by selecting the two most important 14th century Serbian charters, written in the uncial type of Cyrillic script: *Banjska Chrysobull* (henceforth: BC) (cf. Трифуновић 2011), and also *Dečani Chrysobull* (third version, henceforth DC III) (cf. Ивић – Грковић 1976: 34–37). The study also includes several other charters and letters from 14th–15th centuries written in uncial (semiuncial) and cursive Cyrillic script (more on this in 2.1.3). Our methodological approach involved conducting a respective experiment for each manuscript, as well as quantitative and qualitative analyses of the obtained results. Following the process of the automatic recognition of selected manuscript folios (for BC and DC III) or entire manuscripts (in the case of shorter charters and letters) by using the said models, we conducted a manual correction of the text. By comparing the automatically recognized text with the corrected version of the text, we calculated the ratio of unrecognized characters (CER), and this was followed by the qualitative analysis which especially took into account the performance of the models depending on the photographic image quality.

2.1.1. In the first experiment[10] the performance of the VMČ_Test_4+ and Combined_Full_VKS_2 models was tested when applied to the first ten folios of BH (from 5r to 9v). The statistical overview concerning the ratio of incorrectly recognized characters (CER) is given in the following table:

Table 1: CER in BH (sheets 5r–9v)

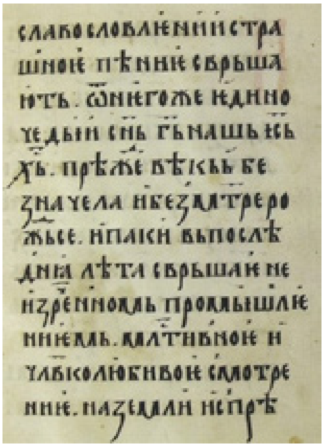
Model	VMČ_Test_4+	Combined_Full_VKS_2
Folio	CER	CER
5r	22,18%	23.79%
5v	20,51%	12,82% %
6r	26,18%	16,73%%
6v	16,95%	13,90 %
7r	19,62%	10,94%
7v	20,08%	8,11%
8r	13,99%	12,94%
8v	20,65%	10,51 %
9r	20,35%	18,95%
9v	21,91%	8.13 %
Mean	20,24%	13,68%

The above overview indicates that the application of the VMČ_Test_4+ model did not produce satisfactory results (mean CER is 20.24%), while the results

[10] In all conducted experiments the automatic recognition was performed by employing the linguistic model. For more details, see https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/.

of applying the generic Combined_Full_VKS_2 model can be estimated as relatively good (mean CER is 13.68%). A comparative representation of the folio 5v photograph and the corresponding automatically recognized text in Table 2 illustratively provides qualitative insight into the performance of both models.

Table 2: VMČ_Test_4+ and Combined_Full_VKS_2 and BH (folio 5v)

VMČ_Test_4+	Трифуновић 2011: 20	Combined_Full_VKS_2
славословлени и стра- шное пѣннѣ сврьша- ють. ѡнегоже єдино- чедыи сн҃ь г҃ь нашь съ хь. прѣжⷣе вѣквь бе- значела и без вагреро- жь се. и паки вь послѣ- дн҃ꙗлѣта врьшаи не- изренномь пролышали. ни киль. валтивное и Гл҃вколюбивое смотре- нии. на зевали испрѣ-		славословлени и стра- шноѥ пѣниѥ сврьша- ють· о нѥгоже ѥдино- єдыи сн҃ь г҃ь нашь сь хь. прѣжⷣе вѣкьь бе- значела и без мт҃ре ро- жь се. и плкивь послѣ- Дн҃ꙗ лѣта сврьшаю не и ꙁренномь промышлѥ- ниѥмь· мⷭ҇тивноѥ и чл҃вколюбивоѥ смотре- не. на земли испрѣ-

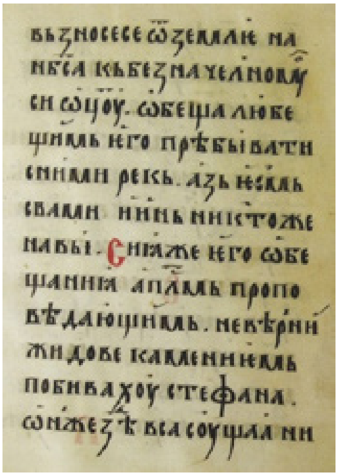
Based on the representation given above, we can conclude that both models make recognition errors most frequently when the pajerak mark is involved, which was expected since the mark was not registered in the model training process: instead of the expected страшꙿноѥ 1/2, безꙿ 6, послѣдꙿнꙗ 7/8, неизр҇енꙿномь 8/9, промышꙿлѥниѥмь 9/10, мⷭ҇тивꙿноѥ 10, сꙿмотрениѥ 11/12, исꙿ прѣ- 12, the VMČ_Test_4+ model incorrectly recognizes страшное 1/2, без 6, послѣдн҃ꙗ 7/8 неизренномь 8/9, пролышали. ни киль. 9/10, валтивное 10, смотрении 11/12, испрѣ- 12, while the generic model recognizes страшноѥ 1/2, без 6, послѣДн҃ꙗ 7/8, не и ꙁренномь 8/9, промышлѥниѥмь 9/10, мⷭ҇тивноѥ 10, смотрене 11/12, испрѣ- 12. A large number of mistakes in both models is attributed to the recognition of blanks between words: instead of славословѥнии 1, без начела 5/6, исꙿ прѣ- 12, both models recognize славословени и 1, безначела 5/6, испрѣ- 12; instead of ѿ нѥгоже 2, мт҃ре рожⷣь 6/7, послѣдꙿнꙗ лѣта 7/8, the VMČ_Test_4+ model recognizes ѡнегоже 2, вагрерожь 6/7, послѣдн҃ꙗлѣта 7/8, while instead of паки вь 8, неизр҇енꙿномь 8/9 the generic model recognizes плкивь 8, не и ꙁренномь 8/9. The recognition of the titlo mark and superscript letters also poses a problem for both models, yet the generic models shows somewhat more successful results: instead of хⷭ҇ь 5, рожⷣ се 6/7 both models recognize хь 5, рожь се 6/7; instead of і҇сь 4, неизр҇енꙿномь 8/9 the VMČ_Test_4+ model renders съ 4, неизренномь 8/9, while the generic model renders сь 4, не и ꙁренномь 8/9; concerning the expected мт҃ре рожⷣь 6/7 and мⷭ҇тивꙿноѥ 10 the VMČ_Test_4+ model renders вагрерожь 6/7 and валтивное 10, while the generic model manages to recognize the titlo mark and a superscript letter in these examples.

The difference in performance of the two models results from a large number of errors when the VMČ_Test_4+ model attempts to recognize the letters ѥ and м: instead of славословѥнии 1, страшꙿноѥ 1/2, пѣниѥ 2, ѿ нѥгоже 2, ѥдиноѵедыи 3/4, свршаѥ 9, промышꙿлѥниѥмь 9/10, м҃лтивꙿноѥ 10, ч҃лвколюбивоѥ 11, сꙿмотрениѥ 11/12, the VMČ_Test_4+ model renders славословени и 1, страшное 1/2, пѣннѣ 2, ѡнегоже 2, єдиноѵедыи 3/4, врьшаи 9, пролышли. ни киль. 9/10, валтивноє 10, Гл҃вколюбивое 11, смотрении 11/12; instead of м҃тре ро҃жь 6/7, промышꙿлѥниѥмь 9/10, м҃лтивꙿноѥ 10, земли 12 there are вагрерожь 6/7, пролышли. ни киль. 9/10, валтивноє 10, зевали 12. The remaining letter recognition errors of this model were registered in a small number of examples: the letter т – instead of м҃тре ро҃жь 6/7 we have вагрерожь 6/7; the letter ь – instead of і҃сь 4, вѣкьь 5 there is съ 4, вѣквь 5; the letter ѵ – instead of ч҃лвколюбивоѥ 11 we have Гл҃вколюбивое 11; the letter ѿ – instead of ѿ нѥгоже 2 there is an incorrect ѡнегоже 2; instead of пѣниѥ 2 we have an incorrect пѣннѣ 2.

The qualitative analysis shows that the performance of the generic model is significantly better that it appears to be judging solely by the CER ratio. As has already been observed, the largest number of errors is associated with the recognition of the pajerak mark and blanks between words. The examples in which the generic model incorrectly recognizes letters are not numerous. This model also has problems recognizing the ligature ѥ in several examples: instead of славословѥнии 1, свршаѥ 9, сꙿмотрениѥ 11/12 there are incorrect славословени и 1, свршаю 9, смотре не 11/12. The remaining letter recognition errors were registered as single instances: the letter а – instead of паки вь 8 we have the incorrect пакивь 8; the letter ѿ – instead of ѿ нѥгоже 2 there is the erroneous о нѥгоже 2; the letter и – instead of сꙿмотрениѥ 11/12 we have смотре не 11/12; the model omitted letters in two examples – ѵ in ѥдиноедыи 3/4 (from the expected ѥдиноѵедыи 3/4) and н in смотре не 11/12 (from expected сꙿмотрениѥ 11/12).

The potential of the generic model can further be illustrated when applied to the folio 7v in which CER is only 8.11% (Table 3).

Table 3: Combined_Full_VKS_2 and BH (folio 7v)

Трифуновић 2011: 24	Combined_Full_VKS_2	Ground Truth
	вьзно се се ѿ землю на	вьзносе се ѿ землѥ на
	нбслкь безнаѵелному	нб҃са кь безнаѵелꙿному҃
	си оц҃оу. обеща любе-	си ѡц҃оу. ѡбеща любе-
	щимь ѥго прѣбывати	щимь ѥго прѣбывати
	с ними рекь. азь ѥсмь	с ними рекь. азь ѥсмь
	свлми. и инь никтоже	с вами. и инь никꙿтоже
	навы. Сиꙗ же ѥго обе-	на вы. Сиꙗ же ѥго ѡбе-
	щаниꙗ. ап҃лмь пропо-	щаниꙗ. ап҃лмь пропо-
	вѣдаюцщимь. невѣрни	вѣдаюцщимь. невѣрꙿни
	жидове каменниѥмь	жидове каменниѥмь
	п҃об҃иваҳоу стефана.	побиваҳоу стефана.
	он же з҃ѣ вса соуцщаа ни-	ѡнꙿ же з҃ѣ вса соуцщаа ни-

The largest number of errors in this folio is again attributed to the recognition of the pajerak mark and blanks between words: instead of безначелꙿномꙋ̑ 2, никꙿтоже 6, невѣрꙿни 9, ѡнꙿ 12 there is the incorrect безначеломꙋ 2, никтоже 6, невѣрни 9, он 12; instead of вьзносе 1, нб҃са кь 2, с вами 6, на вы 7 the generic model renders вьзно се 1, нбслкь 22, свлми 6, навы 7. Expectedly, there are errors connected with the titlo mark and the use of superscript letters: instead of нб҃са кь 2, безначелꙿномꙋ̑ 2, побивахоу 11 we have the incorrect нбслкь 2, безначелномꙋ 2, пⷣоⷪбивахоу 11. A small number of errors in letter recognition can most frequently be associated with the letters ѡ and а: instead of ѡцб҃оу 3, ѡбещанија 7/8, ѡнꙿ 12 the generic model incorrectly renders оцб҃оу 3, обещанија 7/8, он 12; instead of нб҃са кь 2, с вами 6 there are incorrect нбслкь 2, свлми 6. In one example the model makes an error when recognizing the ligature ѥ and the letter є: instead of землѥ 1 and камениѥмь 10 we have the erroneous землю 1 and камєниѥмь 10.

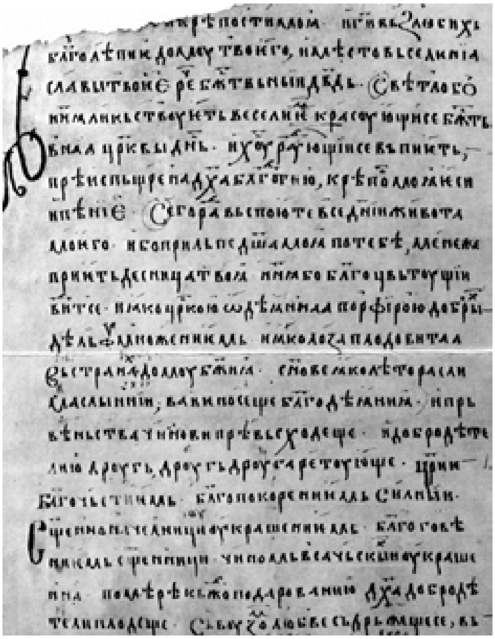
2.1.2. In the next experiment we tested both models on a portion of DC III (in the folios 8r–10v, as well as the folio 76v, which is the initial part of *The Nun Evgeniya's Charter to the Dečani Monastery* (see Младеновић 2007: 391–406). The statistical overview of the incorrectly recognized character ratio (CER) is given in the following table.

Table 4: CER in DC III (8r–10v, 76v)

Model	VMČ_Test_4+	Combined_Full_VKS_2
Folio	CER	CER
8r	22.46%	13.45%
8v	29.64%	15.84%
9r	24.82%	13.23%
9v	23.26%	15.84%
10r	24.97%	12.79%
10v	27.83%	13.79%
76v	19.34%	9.90%

The generic model yielded results almost twice better on average than VMČ_Test_4+ model. The successfulness of the generic model is well evidenced by the comparative representation of the folio 76v and the automatically recognized text (Table 5).

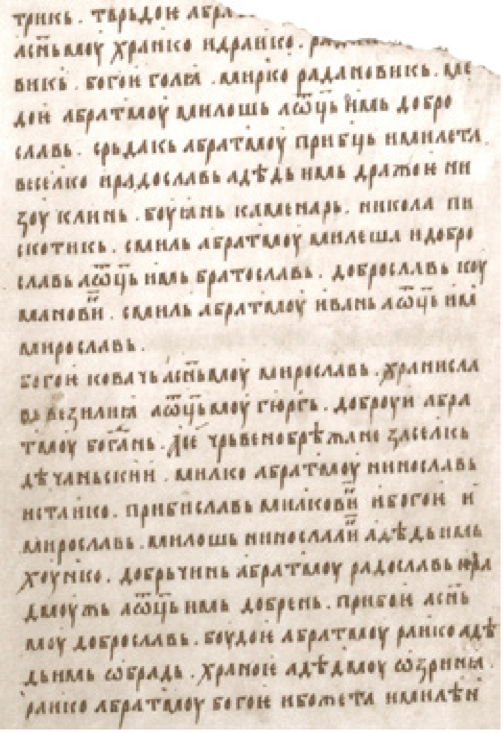
Table 5: Combined_Full_VKS_2 and DC III (folio 76v)

Младеновић 2007: 399	Combined_Full_VKS_2
	...рѣ пости моꙗ. и г҃и възлюбихь бл҃голѣпиѥ҆ домоу твоѥ҆го· и҆ мѣсто вь҆селѥнїа славы твои ѥ̑ⷤ реⷱ҇ бжⷭ҇твьныи дв҃дь. свѣтло бо нныаликьствоуѥть веселиѥ красоуюциⷨ се бжⷭ҇ть. внаа цр҃квы дн҃ь. и хⷪ҇оу раⷣуюцїиⷨ се въпиѥть, прѣиспьцрена дх҃а блⷣ҃тию, крѣпⷪⷨ моꙗ ѥси и пѣнїѥ. Сего раⷣ" вьспоюте все дн҃їи живота моѥго. ибо прильпе дш҃а моꙗ по тебѣ, мене же приѥть десница твоа. нн҃ꙗ бо бл҃гоцвьтоуцїи витсе. и ꙗко цр҃кою ѿ дѣꙗнʼнаа порфїрою добрыⷯ дѣль о множениѥмь и ꙗко лоза плодовитаа Вь страⷯ на домоу бж҃иꙗ. сн҃ове ꙗко лѣтораслихла слыннїи. ваⷭ҇ и ино сеце бл҃годѣꙗниꙗ. и прьвѣньства чинови прѣⷣвьсходеце. и добродѣтелию дроугь дроугь дроуга ретоуюце. цр҃ию Благоуѣ҆стиѥ҆мь҆· благопокорениимь силни· ценно начелници оукрашениимь. бл҃гоговѣѠнниимь сц҃енʼници. чиномь всачьскыи оукрашенна. по мѣрѣ кьжⷣо по дарованию дх҃а добродѣтели плⷪⷮ деце. Съвоузолюбве съдрьжеце сѐ, вь

Based on the given representation, it can be concluded that the generic model most frequently confuses letters и and ѥ in different positions: твои ѥ̑ⷤ 3, ваⷭ҇ и ино сеце 13, благопокорениимь 16, оукрашениимь 17 and бл҃гоговѣѠнниимь 17/18 instead of твоѥѥ 3, ваніѥ носеце 13, бл҃гопокорениѥмь 16, оукрашениѥмь 17 and бл҃гоговѣниѥмь 17/18. In two examples the letters ь and ъ are confused as a part of a preposition and a prefix: възлюбихь 1, Съвоузолюбве 20 instead of вьзлюбихь 1, сьвоузⷪⷨ любве 20. Other errors in the recognition of the letters amount to single instances: твоа 9 instead of твоꙗ 9, ѿ дѣꙗннаа 10 instead of ѡдѣꙗннаа 10, о множениѥмь 11 instead of ѹмножениѥмь 11, хласлыннїи 13 instead of маслынънїи 13, цр҃ию 15 instead of цр҃иѥ 15, ценно 17 instead of Сц҃енно 17. Along with these errors, in a large number of examples we registered the failure to recognize pajerak mark, superscript letters and the titlo (most commonly superscript с under the titlo: дн҃ь 5, хⷪ҇оу 5, крѣпⷪⷨ 6, цр҃кою 10 instead of днⷭ҇ь 5, хⷭ҇оу 5, крѣпⷭ҇о 6, цⷭ҇ркою 10), and the blanks between words. Certain errors occur because the model recognizes accent marks as superscript letters: for instance, твои ѥ̑ⷤ 3, красоуюциⷨ 4, раⷣуюцїиⷨ се 5 and прѣⷣвьсходеце 14 instead of твоѥѥ 3, красоуюци 4, раⷣуюцїи се 5 and прѣвьсходеце 14.

A relatively good result is also recorded by the generic model in the folios 8r–10v DC III (CER is 14.15% on average). This special mention is due to the fact that the generic model yields approximately the same results in the folio 76v written in the Serbian Church Slavonic and the folios 8r–10v which mostly consist of the proper names of people from the monastery grounds.

Table 6: Combined_Full_VKS_2 and DC III (folio 8r)

Ивић – Грковић 1976: 143	Combined_Full_VKS_2
	трикь. тврьдоѥ абра- леньмоу хранкоⷨ и драико. ра- викь. богоѥ голꙗ. мирко рада новикь. ме- доѥ абра тмоу милошьлѡць имь добро- славь. срьдакь абраⷥ тмоу прибць и милета. веселко и радосла вьлдѣдь имь дражоѥ ни зоуклинь. боуꙗнь каменарь. николапи- скотикь. смиль абратмоу милеша и добро- славьлѡць имь братославь. доброславькоу- манови. смиль абра тмоу и вань а ѡц҃ь им ....рославь. богоѥ ко влⷣчьа сн҃ьмоу мирославь. храни сла- вь везили ꙗ а ѡц҃ь моугюргь. доброу и абра- тмоу богань· Дее чрьвено брѣжа не за селкь дѣчаньскни. милко а братмоу ни но славь- истлико. прибиславь милкови и богоѥ и мирославь. милошьни послалилдѣдь имь хоунко. добрьчинь абратвлоу радославь кра- дмоужь а ѡц҃ь имь добрень. прибоѥ а сн҃ь моу доброславь. боудоѥ абратва оура икоа дѣ- дь имь ѡбрадь. храноѥ адѣ дмоу ѿ зринꙗ. ра и коа братмоу богоѥ и боже та ими лѣн

Based on the comparative representation of the folio 8r and corresponding automatically recognized text (Table 6) it can be concluded that the generic model most frequently does not recognize the letter а (confuses it with л): леньмоу 2, лѡць 3, 9, ко влⷣчьа 12, лдѣдь 5, 17, истлико 16 instead of а сн҃ь моу 2, а ѿц҃ь 3, 9, ковачь 12, а дѣдь 5, 17, и сталко 16, as well as the letter ѿ (confuses it with the letters ѡ and о): ѡц҃ь 9, ѡц҃ь 10, 13, 19 instead of ѿц҃ь 9, 10, 13, 19, letter м: ....рославь 11, абратвлоу 18, абратва оура икоа 20 instead of мирославь 11, а братꙿ моу 18, а брат моу раико а 20, and the letter с: леньмоу 2, Дее 14 instead of а сн҃ь моу 2, А се 14. Other errors in the recognition of letters are restricted to individual instances: Дее 14 instead of А се 14 (capital letter), милошьни послалилдѣдь 17 instead of милошь нинослалиⷦ҇ а дѣдь 17, ѿ зринꙗ 21 instead of ѡзринꙗ 21. One example can be singled out: крадмоужь 20/21 instead of радмоужь 20/21, in which the initial letter is crossed out. Like the folio 76v, the folios 8r–10v contain the largest number of errors that are associated with the recognition of pajerak mark (it occurs quite frequently), superscript letters and titlo mark. What is characteristic of the folios 8r–10v is that the model does not recognize blanks between words, probably because in the course of training there was no opportunity for it to gain insight into the specific onomastic vocabulary of the charter.

2.1.3. Encouraged by the relatively good results of the generic model applied to DC III, we conducted additional experiments with several charters and letters from the 14th–15th centuries written in different types of Cyrillic script.

A statistical overview of the results when applying the generic model to the letters from the 14th–15th centuries written in cursive script is given in the following table.

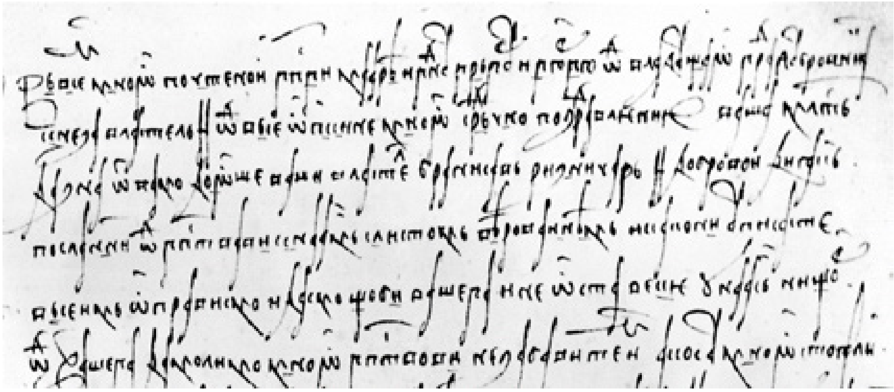
Table 7: Combined_Full_VKS_2 and cursive scripts of 14th–15th century[11]

Letter	CER
Emperor Stefan Uroš V's Letter to Dubrovnik (around 1358)	51.15%
Letter of Jerusalem Metropolitan Mihailo to Dubrovnik (1386)	43.46%
King Tvrtko I Kotromanić's Letter to Dubrovnik (may 1389)	38.44%
Letter of Dubrovnik to Lady Mara Branković and her Sons (1402)	55.20 %
Letter from Dubrovnik about the Settlement of Nikša Sorkočević's Debt (1419)	50.48%
Letter of Turkish Sultan Murad II to Dubrovnik (1431)	56.33%

The data above confirm the initial hypothesis that the Combined_Full_VKS_2 generic model is not suitable for the automatic recognition of the charters and letters written in cursive Cyrillic script, which is expected since the model was trained exclusively by using the manuscript material written in uncial or semiuncial. An extraordinarily high percentage of errors indicates that it is necessary to train a separate model for the automatic recognition of manuscripts written in cursive script.

The lack of usefulness of the generic model in the process of recognizing cursive Cyrillic script is illustratively evidenced by the comparative representation of the photographs of the letter which was sent to Lady Mara Branković and her sons in 1402 from the Dubrovnik office (cf. Стојановић 1929: 146–147) and the corresponding automatically recognized text in the following table.

Table 8: Cursive Cyrillic script and Combined_Full_VKS_2

Letter from Dubrovnik to Lady Mara Branković and Her Sons (1402)


[11] The experiment involved photographic images from the repository of *The History of the Serbian Language* scientific project conducted by the Faculty of Philosophy in Novi Sad.

Combined_Full_VKS_2
Веленно ѡно чтено и инини порѣ и н҃нѣ иреноⷨ и чюⷣно ѿ плачено ѡ прора. прочни се недоу поутельнѡⷨ все ѿ псеи немно ѿ срочно посрѣвлени, помохⷭ҇лть ндѡⷨ хоⷨ. но аще и ции властеⷨ сⷣрници рна ни чтрь. и о пробои. чинъ. посланни нанесенїю листовь вѣрованнѣхь и скони оннѡне- вьсеихь ѡ прⷪ҇ци и схо искощови оушенїи не ѡсновенїе но сеⷨ. нищоѵ. а҃і. ѿвашего. холихохноѡрнилова не ревнⷯи теи. ꙗкоⷤ соно ѿ ст҃оель

The statistical overview of the results of applying the generic model to the 14th–15th century charters written in uncial script has been given in the following table.

Table 9: Combined_Full_VKS_2 and charters from 14th–15th century written in uncial script

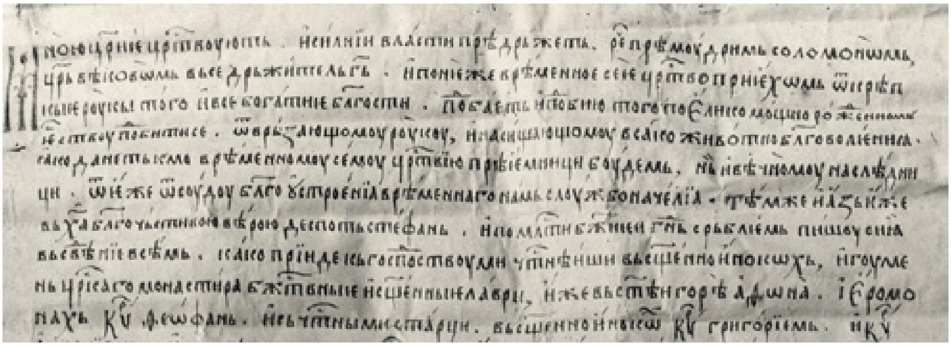
Charter	CER
Dečani Chrysobull (first version, 1330–1331) (lines 1–19)	17.55%
Dečani Chrysobull (first version, 1330–1331) (lines 20–49)	24.46%
Emperor Dušan’s Charter to the St. Sava Cell in Karyes (1348) (lines 1–35)	12.80%
Emperor Dušan’s Charter to the St. Sava Cell in Karyes (1348) (lines 36–64)	13.91%
Despot Stefan Lazarević’s Charter to Despotess Yevpraksia (1404–1405)	9.99%
Lady Mara Branković and her Sons’ Charter to Dubrovnik (1405) (lines 1–27)	18.80%
Lady Mara Branković and her Sons’ Charter to Dubrovnik (1405) (lines 28–43)	27.36%
Despot Stefan Lazarević’s Charter to the Mileševa Monastery (1413)	11.20%
Despot Stefan Lazarević’s Charter to the Great Lavra Monastery (1414–1415)	6.78%

The ratio of unrecognized characters to other charters of 14th–15th centuries written in uncial Cyrillic script when the generic model is applied is mostly positioned within the ranges recorded in the BC and DC III folios. The exceptions are the first version of the Dečani Chrysobull and Lady Mara Branković and her Sons’ Charter to Dubrovnik (1405), where the CER is higher than in the other charters. Along with the expected errors related to the recognition of the pajerak mark, superscript letters and titlo mark in both charters, the high CER in the first version of the Dečani Chrysobull can also be explained by the lower quality of the photograph, which in turn led to problems in recognizing entire portions of the text, while in the case of Lady Mara Branković and her Sons’ Charter to Dubrovnik problems can be attributed to popular vocabulary which the model had no opportunity to familiarize itself with during the training, and, just like in the case of DC III, this led to a greater number of errors concerning the blanks between words.

The potential of the generic model in relation to the recognition of charters written in the Serbian Church Slavonic and uncial Cyrillic script is best illustrated through the example represented by Despot Stefan Lazarević’s Charter to Great Lavra Monastery (1414–1415). A comparative representation of the image of the first ten lines of the charter and the automatically recognized text is given in the following table.

Table 10: Despot Stefan's Charter to Great Lavra Monastery (1414–1415) (lines 1–10)

Младеновић 2007: 283



Combined_Full_VKS_2

оуноюцрⷭ҇нїецрⷭ҇твоують. и силнїи власти прѣдръжеть, реⷱ҇ прѣмоудримь соломонѡмь
цр҃ь вѣковѡмь вьседрьжитель г҃ь. и понеже врѣменное сеи црⷭ҇тво приѥхѡмь ѿ крѣп-
кыє роукы того и всєбогатии бл҃гости. поⷣбⷪаеть и поⷣбⷪию того по єлико мощно рожⷣенꙿномꙋ
ествоупоⷣбⷪити се. ѿврьзающомоу роукоу, и часищающомоу всако животно бл҃говолениꙗ.
ꙗко да не тъкмо врѣменномоу семоу црⷭ҇твїю прѣемници боудемь, нъ и вѣчномоу наслѣдни
ци. ѿ еже ѿ соудоу бл҃гоꙋстроенїа врѣменнаго намь слоужбо начелїа. тѣмꙿже и азь иже
вь х҃а бл҃гочьстиною вѣрою деспоть стефань. и по мⷭ҇ти бж҃нїе и г҃ньсрьблемь пишоу сиꙗ
вь свⷣѣнїе всѣмь. како прїиде і҃сь госпоⷣствоу миⷱ҇тнѣиши вь сщ҃енꙿно и нокѡхь, и гоуме-
нь црⷭ҇каго мона стира бжⷭ҇твныи и сщ҃енныи лаврⷮи, иже вьстѣи горѣ адѡна. їеромо-
нахь у деѡфань. и сь ⷱ҇тными старци. вь сщ҃енꙿно и нокѡⷨ. въ григорїемь. И

The largest number of errors is attributed to the failure to recognize the ligature ѥ: црⷭ҇нїе 1, понеже 2, сеи 2, крѣпкые 2/3, все богатии 3, ествоу 4, бл҃говолениꙗ 4, прѣемници 5, еже 6, бж҃нїе 7, срьблемь 7, бжⷭ҇твныи 8, сщ҃енныи 8 instead of цр҃иѥ 1, понѥже 2, сеѥ 2, крѣпкыѥ 2/3, всебогатиѥ3, ѥⷭ҇ствоу 4, бл҃воволѥниꙗ 4, прѣѥмници 5, ѥже 6, бж҃иѥ7, срьблѥмь 7, бжⷭ҇твныѥ8, сщ҃енныѥ 8. In a smaller number of instances the generic model confuses the letters ь and ъ: прѣдръжеть 1, тъкмо 5, нъ 5 instead of прѣдрьжеть 1, тькмо 5, нь 5, the letters н and и: црⷭ҇нїе 1, бж҃нїе 7 instead of цр҃иѥ 1, бж҃иѥ 7, the letters ч and н: часищающомоу 4 instead of насищающомоу 4, as well as н and в: бл҃гочьстиною 7 instead of бл҃гочьстивою 7. The model was quite successful in recognizing superscript letters, the titlo mark and blanks between words: cf. црⷭ҇твоують 1, црⷭ҇тво 2, црⷭ҇твїю 5, бл҃гоꙋстроенїа 6, etc. Certain errors represent an instance of hypercorrection[12]: поⷣбⷪаеть и поⷣбⷪию 3, поⷣбⷪити се 4 instead of поⷣбаеть и поⷣбию 3, поⷣбити се 4, рожⷣенꙿномꙋ 3 and тѣмꙿже 6 instead of рожⷣенномꙋ 3 and тѣмже 6.

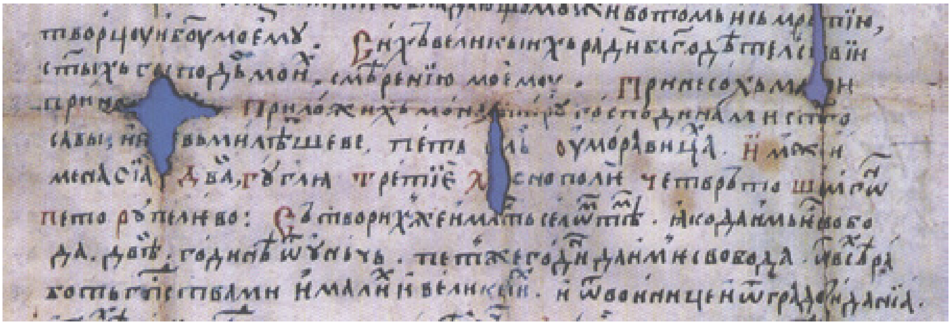
An excellent result was also recorded with the generic model used in the other two charters of Despot Stefan Lazarević: to Despotess Yevpraksia (1404–1405) (CER 9.99%) and to Mileševa Monastery (1413) (CER 11.20%). When com-

[12] For more information on hypercorrection as an atypical error in the process of applying the generic model, see Rabus 2019b: 12.

pared to to Despot Stefan’s Charter to the Great Lavra Monastery (1414–1415), the somewhat higher CER can be explained by a larger number of errors in less legible places where the charter was folded or errors which are due to damaged parts of the text. As an illustration for this claim we can use the example from Despot Stefan’s Charter to the Mileševa Monastery from the following table 11.

Table 11: Despot Stefan’s Charter to the Mileševa Monastery (line 20–27)

Младеновић 2007: 432



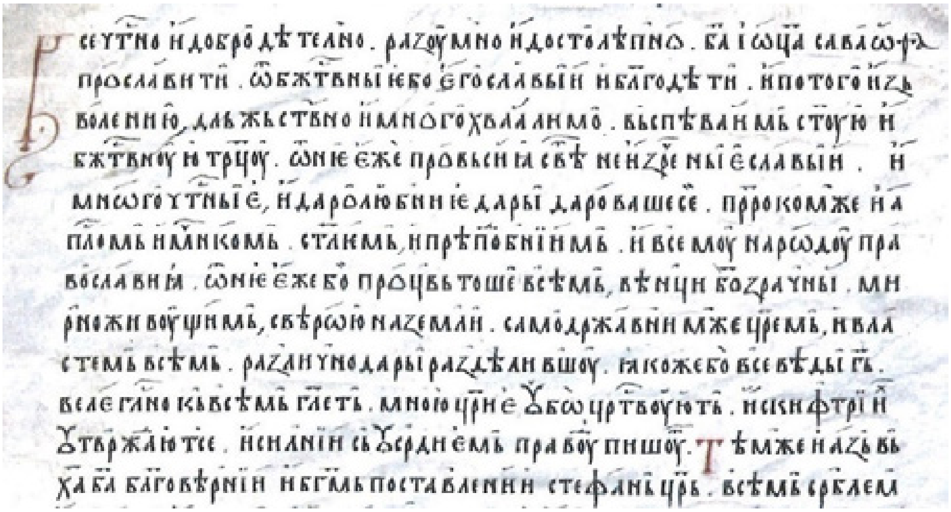
Combined_Full_VKS_2

творцоу и б҃оу моемꙋ. Сихъ великыихь ради бл҃годѣтелствїи
ст҃ыхь бо подь мой, смѣренїю моемоу. Принесохь мате
при на приложихь· мона и и рꙋгосподи нами ст҃го
савы, нивьми лѣшеве. петь вль оуморавица. Имжїи
мена сїардвагꙋ гл҃ꙗ. третїє сно поле четь рⷣо тош мгѡ
петорꙋпелѥво. въ творихⷱ же и мать селѡⷨ тѣⷨ. ꙗко да имь свобо-
да. двѣ̈. годинъ ѿчньчь. петⷱ же годи да ими свобода. и всѣ̆ раⷣ"
боть гн҃ства ми и мали и великый. и ѡ воинице и ѡ градоу и данїа.

Except for the example бл҃годѣтелствїи 20, in which the model succeeded in reconstructing the letter т from a damaged part of the charter, in other places where the charter was damaged or folded errors occurred as expected: бо подь 21 instead of господь 21, мате 21 instead of малоє 21, мона и и рꙋгосподи 22 instead of монастирꙋ господина 22, нивьми лѣшеве 23 instead of иже вь милѣшеве 23, вль 23 instead of сель 23, Имжїи 23 instead of Им҄жи 23, тош мгѡ 24 instead of шемгѡⷩ҇ 24, и всѣ̆ 26 instead of ѿ всѣ̆ 26, градоу и данїа 27 instead of градозиданїа 27.

The qualitative analysis concerning the performances of the generic model applied to Emperor Dušan’s Charter to the St. Sava Cell in Karyes (book number: Hil 31) (Skopje, 1348) also reveals excellent recognition results, despite CER in the first part of the charter (lines 1–35) being 12.80%, and 13.91% in the second part. To illustrate this claim, we can compare the photographic image of the first twelve lines of the charter and the automatically recognized text in the table that follows.

Table 12: Emperor Dušan’s Charter to the St. Sava Cell in Karyes (lines 1–12)

Живојиновић 2008: 59–70

Combined_Full_VKS_2
се ч҇ⷭтно и добродѣтелно. разоумно и достолѣпно. б҃а ї ѡ҃ца сав҇ⷭа ѿ фа- прославити. ѿ б҇ⷭжтвныи бо єго славы и и б҃лгодѣти. и по тогⷪⷤ изь- волению, дльжьствно и многⷪⷨ хваа анꙿмⷪⷨ. вь спѣваимь с҃тоую и б҇ⷭжтвноую т҇ⷪрцоу. ѿ не єже провьси ꙗⷤ свѣ неизре҇ⷭчны є҇ⷭ славы йⷯ. и мнѡгоꙋтны є· й дародюбниѥ̓ дарь дароваше се· прроком же й а̑- п҇ⷭломь и мнкомь. стаѥмь и прѣпⷣⷪбнїимь· и всемоу нарѡдоу пра- вослави ꙗ. ѿне є҇ⷭ же бо процвьтошⷣе всѣмь, вѣнци бо зрачны. мн рно живоущимь свѣрѡю на земли. самодржа вꙿ нимꙿ же ц҃ремь̓· й вла- стемь всѣмь· различно дарь раздѣливꙿшоу· ꙗ̑коже бо все вѣды г҃ь велег҃лно къ всѣмь г҃леть. мною ц҃ри є҇ⷭ оvбѡ ц҇ⷭртвоують. и скифтрїйⷯ отвръжⷣаю҃тсе· й сианїй сь̓ ѡ̑срдиємь̓ правоу пишоу· Тѣмꙿже й а̑зь̓ вь̓ х҃а б҃а б҃лговѣрнїи и б҃гмь поставлении стефань ц҃рь. всѣмь сⷣрблем

The above representation indicates that the largest number of errors are result from the failure to recognize the pajerak mark: добродѣтелно 1, разоумно 1, достолѣпно 1, дльжьствно 3, прроком же 5, прѣпⷣⷪбнїимь 6, всѣмь 7, 10, вѣнци 7, бо зрачны 7, мнрно 7/8, земли 8, самодржа вꙿ нимꙿ же 8, различно 9, раздѣливꙿшоу 9, все 9, Тѣмꙿже 11, б҃лговѣрнїи 12, поставлении 12, etc. – instead of добродѣтелⸯно 1, разоумⸯно 1, достолѣпⸯно 1, дльжьствⸯно 3, прр҃окомⸯ же 5, прѣпⷣⷪбⸯнїимь 6, вⸯсѣмь 7, 10, вѣнⸯци 7, бⷪзрачⸯны 7, мирⸯно 7/8, земⸯли 8, самодржавⸯнимⸯ же 8, различⸯно 9, разⸯдѣливⸯшоу 9, вⸯсе 9, Тѣмⸯже 11, б҃лговѣрⸯнїи 12, поставⸯлении12, etc. The CER level is also affected by a large number of examples in which it is necessary to remove superscript accent marks since the model often renders them hypercorrectly: й 5x2, 8, 11x2, даролюбниѥ̓ 5, а̑п҇ⷭломь 5/6, ц҃ремь̓ 8, ꙗ̑коже 9, сианїй 11, сь̓ ѡ̑срдиємь̓ 11, а̑зь̓ 11, вь̓ 11, etc. The hypercorrectness is also evident in the process of recognizing the superscript letters: саваⷭ҇ ѿ фа 1, тогоⷤ 2, многоⷨ хваа анⷱ҆моⷨ 3, трⷪ҇цоу 4, провьси ꙗⷤ 4, неизреⷱ҇ны еⷭ҇ 4, иⷯ 4, прѣⷣпⷪ҇бнїимь 6, ѿне еⷭ҇ 7, процвьтошⷣе 7, цⷬ҇и еⷭ҇ 10, сⷬ҇блем 12 instead of саваѡда 1, того 2, многохваалимо 3, тр҇цоу 4, провьсиꙗ4, неизре҇ныѥ 4, и 4, прѣⷣп҇об҆нїимь 6, ѿнѥеже 7, процвьтоше 7, цр҇иѥ 10, ср҆блем 12. Errors in the recognition of letters are mostly associated with the ligature ѥ: instead of бⷭ҇жтвныѥ 2, ѿнѥеже 4, 7, there is бⷭ҇жтвныи 2, ѿне еже 4, ѿне еⷭ҇ же 7, aslo with letter ꙋ: instead of ꙋбѡ 10, ꙋтвр҆жⷣают҆ се 11, ꙋсрдиемь 12 there is оѵбѡ 10, отвръжⷣаютсе 11, ѻсрдиемь 12, letter л: instead of хваалимо 3, ст҇лѥмь 6, сил҆нии 11 there is хваа анⷱ҆моⷨ3, стаѥмь 6, сианій 11, letter и: instead of хваалимо 3, мир҆но 7/8 there is хваа анⷱ҆моⷨ 3, мнрно 7/8 and letter ы: instead of дары 4, 8 there is дарь 4, 8. Other errors: instead of саваѡда 1 there is саваⷭ҇ ѿ фа 1, instead of велегⷧ҇но 10, кь 10 there is велег҃лно 10, къ 10.

2.2. Automatic recognition of the *Christian Topography* of Cosmas Indicopleustes (1649). The Serbian Church Slavonic translation of Cosmas Indicopleustes' *Christian Topography* (1649) is preserved as a part of a more extensive manuscript which is located in the vault of the Holy Trinity Monastery near Pljevlja (Montenegro), under the book register number Pljevlja 79. The first part of the manuscript (folios 1–101) contains the *Hexameron* by John the Exarch, while the *Christian Topography* (henceforth CT) can be found in the second part of the manuscript, to folio 240.[13] The manuscript was copied in uncial Cyrillic script in 1649 by Gavrilo Trojičanin, the most renowned calligrapher of the time, using a Russian model, in the Holy Trinity Monastery near Pljevlja, and it was ornamented by miniatures by Andrija Raičević, an icon painter and miniaturist (cf. Јагић 1922: 1; Ракић 2016: 407). For the purposes of this research, we used the photographic images stored on microfilm from the Department for Archeography of the National Library of Serbia.[14] Although the quality of the photographs was not ideal, the application of the special VMČ_Test_4+ model and the Combined_Full_VKS_2 generic model rendered substantially better results than those obtained with Serbian medieval charters and letters. The experiment involved 9 folios, in which the average CER value was 6.43% for the VMČ_Test_4+ model and 5.34% for the Combined_Full_VKS_2 generic model. If CER calculations are applied only to unrecognized letters, then the results become even more impressive: CER is reduced to 2.67% with the VMČ_Test_4+ model, and to 1.42% with the Combined_Full_VKS_2 generic model. It is important to mention that the CER values do not vary substantially from one folio to another, meaning that both models yield consistent results regardless of the analysed folio. The variations in CER values depend mostly on the frequency of superscript letters in the individual folios. In other words, the higher the use of the superscript letters, the higher the CER value. What particularly needs to be mentioned are various types of errors in the process of recognizing superscript letters: a) complete omission of an superscript letter (in a large number of instances): обра instead of обраⷥ, сло instead of слоⷡ҇, възпросившем instead of възпросившеⷨ, хⷭ҇ртїанскы instead of хⷭ҇ртїанскыⷯ, новы instead of новыⷯ, etc.;

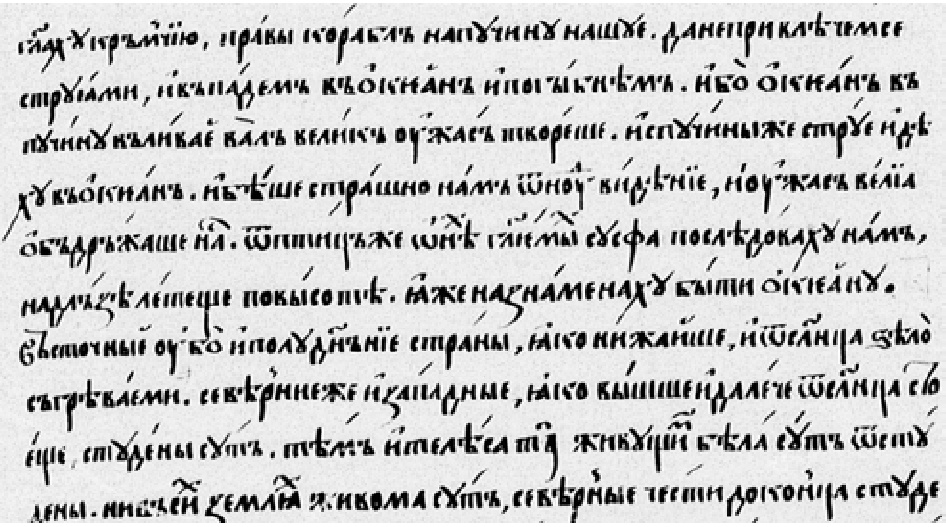
[13] For more details on the manuscript consult Мошин 1958: 254; Станковић 2003: 26. For more information about the paleographic and linguistic particularities of the manuscript, see Јагић 1922. On the miniatures see Ракић 2016: 407–417.

[14] The image has been obtained with the blessing of His Grace Atanasije (Rakita), whom we hereby express our sincerest gratitude.

b) recognizing an superscript letter as the one which is lowered into the text line: e.g. рєч҇ instead of рєⷱ҇, пач҇ instead of паⷱ҇, etc., c) incorrect recognition of a superscript letter: e.g. слоⷨ instead of сло҇ⷡ, бывающиⷯ instead of бывающиⷨ҇, свойⷯ instead of свойⷨ҇, живꙋщийⷯ instead of живꙋщийⷨ҇, нарицаєⷮ instead of нарицаєⷨ҇, etc., and d) hypercorrection, that is the occurrence of a superscript letter in the positions where it would be expected, but where it was not found in the text itself: e.g. ꙗвлꙗєⷮ instead of ꙗвлꙗє, силныⷯ instead of силны, оутврьжⷣаєⷮ instead of оутврьжⷣає, сътворⷮи instead of сътвори, etc. Along with the errors related to the recognition of superscript letters, both models make errors when it comes to the recognition of blanks between words, while the errors related to the recognition of the titlo, punctuation, and regular letters occur infrequently.

The extraordinary performance of the generic model when performing the recognition in CT is evidenced by the comparative representation of the first ten lines from the folio 116r (in which the CER is only 3.78%) and the automatically recognized text in the table 13.

Table 13: CT (folio 116r, line 1–10) and Combined_Full_VKS_2


Combined_Full_VKS_2
гл҇ахꙋ кръмчїю. правы кораблъ на пꙋ чинꙋ нашꙋє. да нє привлѣчєм сє стрꙋꙗми, и въ падемъ въ окєанъ и погыбнѣмъ. ибо окєанъ въ пꙋ чинꙋ въ ливаєⷮ валъ вєликъ оужасътворєшє. испꙋчины жє трꙋєⷣ и дѣ хꙋ въ окианъ. и бѣшє страшно намъ ѿноуⷣ видѣнїє, и оужасъ вєлїа обьдръжашє наⷭ҇ . ѿ птицъ жє ѡнѣⷯ гл҇ємы сꙋсфа послѣдовахꙋ намъ. на длъзѣлєтєщєⷨ по высотѣ. ꙗжє назнамєнахꙋ быти окїєанꙋ. въсточныє оубо и полꙋднънїє страны, ꙗко нижаишє, и ѿ сл҇нца ѕѣло съгрѣваєми. сє вѣрни єжє и западныє, ꙗко вышⷱшє и далєчє ѿ сл҇нца сто- єщє. стꙋдєны сꙋтъ. тѣмъ и тєлѣ сата живꙋщийⷯ бѣла сꙋтъ ѿстꙋ дєны. ни въ сиⷯ зємлꙗ. живома сꙋтъ, сє вѣрныє чєсти до конца стꙋдє

Judging from the above representation, it can be concluded that the generic model most frequently manifested problems with the recognition of the ligature ѥ: окканъ 2 instead of окѥанъ 2, окианъ 4 instead of окѥанъ 4, гл҃емы 4 instead of гл҃ѥмы 4, окїеанꙋ 6 instead of окѥанꙋ 6. When errors concerning letters are considered, only one individual instance was recorded: the failure to recognize the letter с in трꙋеⷣ 3 instead of стрꙋе 3. Among other errors the most numerous ones are those connected with the recognition of blanks between words: пꙋ чинꙋ 1, въ падемъ 2, въ ливае 3, ѹжасътвореше 3, и дѣхꙋ 3/4, се вѣрни еже 8, телѣ сата 9, ѿстꙋдены 9/10, въ си 10 instead of пꙋчинꙋ 1, въпадемъ 2, въливае 3, ѹжасъ твореше 3, идѣхꙋ 3/4, севѣрние же 8, телѣса та 9, ѿ стꙋдены 9/10, въси 10. The extraordinarily low CER level in this folio is especially affected by the infrequency of superscript letters. The following errors were recorded: a) a superscript letter was not recognized: гл҃ѥмы 4 instead of гл҃ѥмыⷯ 4, землꙗ 10 instead of землꙗⷯ 10, b) incorrect superscript letter: живꙋщйⷯ 9 instead of живꙋщйⷨ҇ 9, c) hypercorrection: въливаеⷮ 2 instead of въливае 3, трꙋеⷣ 3 instead of стрꙋе 3, длъзѣлетещеⷨ 6 instead of длъзѣлетеще 6. Along with these examples, we need to mention the instances in which the model performed the recognition of the superscript letters correctly: ѿноѹⷣ 4, наⷭ҇ 5, ѡнѣⷯ 5 and сйⷯ 10. The model made errors in all three examples that contained the pajerak mark in the photograph: кръмчїю 1, севѣрние 8, конца 10 instead of кръмꙿчїю 1, севѣрꙿние 8, конꙿца 10, and in one example there was an occurrence of hypercorrected вышꙿше 8 instead of вышше 8. The titlo used for the purpose of abbreviation was mostly well recognized: гл҃ахꙋ1, гл҃емы 5, сл҃нца 7, 8, in contrast to one example where it was omitted: полꙋднънїе 7 instead of полꙋдн҃ънїе 7. Punctuation (comma and full stop) was also recognized excellently, with only a single instance of a full stop instead of a comma at the end of the fifth line.

3. Concluding remarks. The conducted research has confirmed the initial hypothesis that the application of the existing models for the automatic recognition of Church Slavonic Cyrillic manuscripts can also be quite successful in an overall sense when applied to Serbian medieval manuscripts written in the uncial or semiuncial script, while the application to Serbian medieval manuscripts written in cursive script renders transcripts which are not usable, thus showing that there is a need to create a special model for the recognition of the Serbian cursive Cyrillic script. Among the investigated manuscripts written in uncial or semiuncial scripts, the best results were obtained by applying the existing models to the Serbian Church Slavonic manuscript Cosmas Indicopleustes' *Christian Topography* (1649), which can probably be attributed to the fact that both models for the automatic recognition contain materials from chronologically close manuscripts written in the same script. The success of the application of the existing models to Serbian medieval charters written in the uncial or semiuncial script often varies depending on the quality of the images and the preservation state of the manuscript. While the application of the VMČ_Test+ model mostly produced unsatisfactory results, the application of the generic model resulted in quite usable transcripts. A special mention is due to the manuscript DC III, in which the generic model manifested approximately the same recognition results both in the parts of the manuscript written in the Old Serbian and the parts written in the Serbian Church Slavonic language. Although the ratio of unrecognized characters (CER) was above 10% in many charters, the qualitative analysis has shown that the benefit of the transcripts obtained by applying the generic model can be deemed to be quite satisfactory, especially if we take into account the fact that the model was unable to gain insight into some of the special characters (such as the pajerak mark) during the training process and that the recognition errors are most frequently related to the blanks between words, superscript letters and titlos, and much more rarely to individual letters. The benefit of the generic model is especially evident when used for the automatic recognition of the more voluminous manuscripts written in uncial or semiuncial script. The transcripts of the parts of the manuscripts obtained by applying the generic model can be manually corrected and then used for subsequent training of the generic model for the purposes of enhancing its performance or for the purposes of training a special model for the recognition of the remainder of a voluminous manuscript. (e.g., BC, DC III or CT)[15]. By employing this procedure (cf. Rabus 2019b: 13), in a relatively short period of time we can obtain considerable amounts of data (photographs and corresponding transcripts of Serbian medieval manuscripts), which can be used to create special models for individual voluminous manuscripts, and ultimately a generic model for Serbian medieval manuscripts, thus significantly expediting the work on the current projects involving the Serbian historical corpus linguistics and lexicography.

REFERENCES

Hodel, Tobias, David Schoch, Christa Schneider, Jake Pursell. General Models for Handwritten Text Recognition: Feasibility and State-of-the-Art. German Kurrent as an Example. *Journal of Open Humanities Data* 7: 13 (2021), 1–10. http://doi.org/10.5334/johd.46

Polomac, Vladimir. Towards Fundamental Principles for Creating the Electronic Corpus of Serbian Medieval Charters and Letters. *Scripta&e-Scripta* 21 (2021): 41–53.

Rabus, Achim. Recognizing Handwritten Text in Slavic Manuscripts: a Neural-Network Approach Using Transkribus. *Scripta&e-Scripta* 19 (2019a): 9–32.

Rabus, Achim, Training Generic Models for Handwritten Text Recognition Using Transkribus: Opportunities and Pitfalls. *Proceeding of the Dark Archives Conference*, Oxford, 2019b, in print.

Transkribus Glossary: https://readcoop.eu/glossary/<15.07.2021>

*

Живојиновић, Драгић. Скопска хрисовуља цара Душана за келију Св. Саве Јерусалимског у Кареји (Хил. 31). Стари српски архив 7 (2008): 59–70.

Ивић, Павле, Милица Грковић. Дечанске хрисовуље. Нови Сад – Београд: Институт за лингвистику – БИГЗ, 1976.

Јагић, Ватрослав. Козма Индикоплов по српскому рукопису г. 1649-е. Палеографско-језичка студија. Споменик СКА XLIV (1922): 1–39.

Јерковић, Вера. Полуустав у српским повељама од краја XIV века и током XV века. Зборник Матице српске за филологију и лингвистику XLII (1996): 89–113.

Корниенко, Сергей Иванович, Федор Михайлович Черепанов, Леонид Нахимович Ясницкий. Распознавание текстов рукописных и старопечатных книг на основе нейросетевых технологий. Валерий Дмитриевич Соловьев, Виктор Аркадьевич Баранов (ред.). Современные информационные технологии и письменное наследие: от древних текстов к электронным библиотекам. *El' Manuscript 08. Материалы Международной научной конференции*. Казань: Издательство Казанского государственного университета, 2008, 155–156.

[15] Special models trained for the recognition of each of the more voluminous manuscripts can achieve extremely high accuracy, with CER under 1% (cf. Rabus 2019b: 2–3).

Младеновић, Александар. Повеље и писма деспота Стефана. Београд: Чигоја штампа, 2007.

Мошин, Владимир. Ћирилски рукописи манастира Св. Тројице код Пљеваља. Историски записи XIV/1–2 (1958): 235–260.

Ракић, Зоран. Шестоднев Јована Егзарха и Хришћанска топографија Козме Индикоплова. Душан Оташевић, Зоран, Ракић, Ирена Шпадијер (ур.). Свет српске рукописне књиге (*XII–XVII* век). Београд: Галерија Српске академије наука и уметности, 2016, 407–417.

Станковић, Радоман. Рукописне књиге манастира Свете Тројице код Пљеваља: водени знаци и датирање. Београд: Народна библиотека Србије, 2003.

Стојановић, Љубомир. Старе српске повеље и писма, књ. 1. Београд: Српска краљевска академија, 1929.

Трифуновић, Ђорђе. Повеља краља Милутина манастиру Бањска. Књига прва: Фототипија изворног рукописа. Београд: ЈП Службени гласник, 2011.

Владимир Поломац
Тамара Лутовац Казновац

АУТОМАТСКО РАШЧИТАВАЊЕ СРПКИХ СРЕДЊОВЕКОВНИХ РУКОПИСА ПОМОЋУ СОФТВЕРСКЕ ПЛАТФОРМЕ *TRANSKRIBUS*: СТАЊЕ И ПЕРСПЕКТИВЕ

Резиме

Софтверска платформа *Transkribus* (https://readcoop.eu/transkribus/), недавно развијена на Универзитету у Инсбруку (Аустрија), представља алат за ручно и аутоматско рашчитавање и претраживање старих рукописа и штампаних књига, независно од времена настанка, језика или писма. Кључна предност Транскрибуса у односу на друге сродне апликације огледа се у могућности да корисник сам креира сопствени модел за аутоматско рашчитавање текста. Тренирање модела за аутоматско рашчитавање текста представља пример машинског учења заснованог на напредним неуронским мрежама у коме модел упоређује фотографије рукописа и одговарајућа слова, речи и линије текста у дипломатичком издању. За успешно тренирање модела неопходно је обезбедити што квалитетније фотографије рукописа и најмање 15000 речи рашчитаног текста. За аутоматско рашчитавање старословенских и црквенословенских ћириличких рукописа у оквиру Транскрибуса доступна су два модела која је развио немачки слависта А. Рабус: први модел, назван VMČ_Test_4+, заснован је на деловима рускословенског рукописа Великие Минеи-Четьи, писаног полууставном ћирилицом XVI века; други модел, назван Combined_Full_VKS_2, заснован на деловима старословенског Супрасаљског кодекса (XI век), рукописа Катихизиса Кирила Јерусалимског (XI век) и рускословенског рукописа Великие Минеи-Четьи (XVI век), представља покушај креирања генеричког модела за аутоматско рашчитавање различитих црквенословенских рукописа писаних уставном или полууставном ћирилицом.

Основни циљ нашег рада представља истраживање могућности примене Рабусових модела за аутоматско рашчитавање српских средњовековних рукописа писаних различитим типовима ћирилице. Наведени модели тестирани су на српским средњовековним рукописима који су тренутно у фокусу наших филолошких и лингвистичких истраживања: на српским средњовековним повељама и писмима који се приређују за потребе изградње специјализованог електронског корпуса, као и на српскословенском рукопису Хришћанске топографије Козме Индикоплова (1649) који се припрема за објављивање у оригиналној графији уз пратеће филолошке и лингвистичке студије. Проведено истраживање потврдило је почетну хипотезу да примена постојећих модела за аутоматско рашчитавање црквенословенских ћириличких рукописа у начелу може бити веома успешна и на српским средњовековним рукописима писаним уставом или полууставом, док се применом на српске средњовековне рукописе писане брзописом добијају неупотребљиви транскрипти, што упућује на потребу креирања специјалног модела за рашчитавање српске брзописне ћирилице. Међу испитиваним рукописима писаним уставом или полууставом најбоље резултате добили смо применом постојећих модела на српскословенски рукопис Хришћанске топографије Козме Индикоплова, што је вероватно у вези са чињеницом да оба модела за аутоматско рашчитавање садрже материјал из хронолошки блиских

рукописа писаних истим писмом. Успех примене постојећих модела на српске средњовековне повеље писане уставом или полууставом најчешће варира у зависности од квалитета фотографије и очуваности рукописа. И док је примена модела VMČ_Test+ углавном дала незадовољавајуће резултате, применом генеричког модела на свим испитиваним повељама добили смо веома употребљиве транскрипте. Иако је проценат непрепознатих карактера (CER) у многим повељама био изнад 10%, квалитативном анализом показано је да се вредност транскрипата добијених применом генеричког модела може сматрати веома задовољавајућим, посебно ако се има у виду чињеница да модел током тренинга није имао прилике да види неке специфичне карактере (нпр. пајерак) и да се грешке у рашчитавању најчешће односе на размак међу речима, надредна слова и титле, а знатно ређе и на појединачна слова. Вредност генеричког модела посебно долази до изражаја приликом аутоматског рашчитавања обимнијих рукописа писаних уставом или полууставом. Транскрипти дела рукописа добијени помоћу генеричког модела могу се ручно кориговати, а затим искористити за поновно тренирање генеричког модела у циљу побољшања његових перформанси или за тренирање специјалног модела за рашчитавање остатка обимног рукописа. Овим поступком у релативно кратком временском року можемо доћи до великих количина података (фотографија и одговарајућих транскрипата српских средњовековних рукописа) помоћу којих можемо креирати специјалне моделе за појединачне обимније рукописе, а у крајњем исходу и генерички модел за српске средњовековне рукописе, што би могло значајно убрзати рад на текућим пројектима из српске историјске корпусне лингвистике и лексикографије.

University of Kragujevac
Faculty of Philology and Arts
Department of Philology
Jovana Cvijića bb, 34000 Kragujevac, Serbia
v.polomac@filum.kg.ac.rs
tamara.kaznovac@filum.kg.ac.rs

(Примљено: 20. августа 2021;
прихваћено: 28. октобра 2021)