SCRIPTA
& e- SCRIPTA

e-S C R I P T A

22 / 2022

Serbian Early Printed Books from Venice: Creating Models for Automatic Text Recognition Using *Transkribus*

Vladimir R. Polomac

Abstract: The paper describes the process of creating a model for the automatic recognition of Serbian Church Slavonic printed books from Venice (from Božidar and Vincenzo Vuković's printery) by using the *Transkribus* software platform, based on the principles of artificial intelligence and machine learning. By using the example of *Prayer Book (Euchologion)* (1538–1540) from Božidar Vuković's printery, it has been shown that a successful model for the automatic recognition of individual books (with around 5% of unrecognized characters) can also be trained on the material consisting of approximately 4000 words, and that the increased amount of training material (in our case around 38000 words) leads to the improvement of the model and reduced error rate (between 1–2% of unrecognized characters). The most notable result of the paper is manifested through the creation of a generic model for the automatic text recognition of Serbian Church Slavonic books from Božidar and Vincenzo Vuković's printery. The initial version of the generic model (called *Dionisio 1.0.* by the Božidar Vuković's Italian pseudonym – *Dionisio della Vecchia*) is the first resource for the automatic recognition of the Serbian medieval Cyrillic script, publicly available to all users of the *Transkribus* software platform (see https://readcoop.eu/model/dionisio-1-0/).

Key words: *Transkribus*, Automatic Text Recognition, Serbian Early Printed Books, Artificial Intelligence, Machine Learning, Venice.

1. Introduction

The basis for this paper can be traced in the recent research related to the use of the *Transkribus*[1] software platform for the automatic recognition of Medieval Slavic Cyrillic manuscripts (cf. Rabus 2019a, Polomac and Lutovac Kaznovac 2021).[2] The central part of Rabus 2019a is dedicated to the creation of a model for the automatic recognition of Church Slavonic manuscripts,[3] accompanied by quantitative and qualitative analyses of the results which were obtained by applying these models to manuscripts written in uncial and semi-uncial Cyrillic scripts. The main merit of this paper is manifested in the fact that by using concrete examples the automatic recognition of Serbian Church Slavonic manuscripts is indeed achievable in practice, and that one can automatically obtain the first version of the text in an electronic form with the acceptable error rate – around 4% of incorrectly recognized characters,[4] so that after the manual correction performed by a competent philologist, done within a much shorter time frame and with fewer human and financial resources, the result can

[1] *Transkribus* (https://readcoop.eu/transkribus) is a free-access software platform for the automatic recognition and search of manuscripts developed as a part of the READ project at the University of Innsbruck. Unlike the traditional approach (OCR technology), which focuses on individual letters in the recognition process, *Transkribus* uses the HTR technology, which is based on memorizing and recognizing the image of the entire line of text. Recently developed and implemented into Transkribus, the HTR+ algorithm based on the artificial intelligence and advanced neural networks substantially reduces the time needed to train text-recognition models, with significantly higher accuracy rates. For a more detailed account of the technological background and operational aspects see Mühlberger et al. 2019.

[2] The paper Burlacu and Rabus 2021: 1–9 investigates the potential applications of the *Transkribus* software platform on Romanian Cyrillic manuscripts, while Rabus 2022 focuses on manuscripts and printed books written in Croatian Glagolitic script. *Transkribus* is successfully applied in the projects of transcribing multilingual (Greek-Latin-Church Slavonic) historical dictionaries (see Thompson 2021), as well as in the digitization of Church Slavonic printed books from the Berlin State Library (see Neumann 2021).

[3] The functionality of the *Transkribus* software platform is primarily manifested in the possibility of training one's own model for the automatic text recognition, regardless of the language or script of the manuscript. The training of the automatic recognition model represents an instance of machine learning based on neural networks in which the model in the learning process compares the manuscript's images and the associated letters, words, and lines from the diplomatic edition of the text. A successful model training requires manuscript images to have the highest possible quality and at least 15000 words of a recognized text. The required amount of data is significantly smaller when creating a model for the automatic recognition of old printed books – around 5000 words of the recognized text. For more details about the model training see Mühlberger et al. 2019: 959, Rabus 2019a: 11–14, Rabus 2019b.

[4] The character error rate (CER) is calculated by comparing the automatically generated text and manually corrected version. For more details see Transkribus Glossary at https://readcoop.eu/glossary/character-error-rate-cer/.

be realized as the electronic form of the manuscript text ready for further philological and linguistic investigations. The specific merit of Rabus 2019a paper can be seen in making models for the automatic recognition of Church Slavonic manuscripts publicly available, so that their performance can be further attested by applying them to other Slavic medieval manuscripts. The results of applying these models to the Serbian medieval manuscripts written in different types of Cyrillic script were presented in Polomac and Lutovac Kaznovac 2021. The authors draw a conclusion that the application of the existing publicly available models for the automatic text recognition (created by A. Rabus) produces relatively good results when applied to Serbian medieval manuscripts written in semi-uncial Cyrillic script, but they also suggest that special models be created for the manuscripts written in cursive Cyrillic script. Knowing that the initial results concerning the application of the *Transkribus* software platform to Slavic medieval Cyrillic manuscripts were quite encouraging, by writing this paper we have also wished to extend the area of research to the Serbian Early Printed Books. The primary goal of the paper is the creation of the model for the automatic recognition of Serbian Church Slavonic books[5] printed during the 16th century in Venice in Božidar Vuković's[6] and his son Vicenzo's[7] printery. Limiting the scope of the paper only to books coming from the Vuković printery is justified by the fact that the printery was responsible for the publication of the majority of the preserved Serbian Early Printed Books from the 16th century,[8] as well as by the fact

[5] No publicly available models for the automatic recognition of Old Serbian/Slavic Cyrillic printed books can be found in the *Transkribus* platform. A. Rabus, a German Slavicist, has developed a model for the automatic recognition of Glagolitic printed books from Tübingen and Urach, along with several models for the recognition of Cyrillic and Glagolitic manuscripts (see Rabus 2019a: 15–19, 23–27, Rabus 2022).

[6] Božidar Vuković was a Serbian merchant from Zeta (Podgorica and the area surrounding Lake Skadar). After his arrival at Venice (in 1516 at the latest) he acculturated his Serbian name to the new environment by creating a Latin and an Italian pseudonym (It. *Dionisio della Vecchia*, Lat. *Dionisius a Vetula*) from his Serbian name and the toponym of Starčeva Gorica (at Lake Skadar), indicating his origin (cf. Lazić 2018a). Božidar Vuković continued his successful trade in Venice, as well as his social and public affairs, which is evidenced by the data from 1534 stating that he received a noble title from Charles V of Habsburg, while in 1536 he was elected the steward (gastaldo) of the Orthodox community (the Brotherhood of Greeks) in Venice (cf. Pešikan 1994: 79–80). Books from his printery were aimed at the Serbian Orthodox Church and its flock under Ottoman rule, yet the motives of his printing business were not only patriotic and religious, but also mercantile and financial (cf. Lazić 2013; 2020b).

[7] For more details about Vicenzo Vuković's life and work see Pešikan 1994: 83–85.

[8] In the first phase of his work Božidar Vuković printed *Hieratikon (Liturgikon)* (1519), *Psalter with Appendices* (1520) and *Prayer Book (Miscellany for Travellers)* (1520), while in the second phase he printed *Prayer Book (Miscellany for Travellers)* (1536), *Octoechos, Mode 5–8.* (1537), *Festal Menaion* (1538) and *Prayer Book (Euchologion)* (1538–1540) (cf. Pešikan 1994: 86–89; Lazić 2018b:

that the equipment from this printery was later also used in other Venetian printeries, namely those of Jerolim Zagurović, Antonio Rampazetto and Bartolomeo Ginammi (see Pešikan 1994: 117–119, 175, 177).[9] On the other hand, the research regarding the applications of the Transkribus software platform for the automatic recognition of the Church Slavonic printed books from the Berlin State Library have shown that the best results are obtained if one creates models which belong to a single printing tradition or a single printery (Neumann 2021: 211).

2. Transcription Process and the Creation of the Special HTR Model

The initial methodological problem was the fact that we had no transcripts of the Serbian Early Printed Books from Venice at our disposal to be used for model training. The first data set for the training of the specific model was obtained by using the generic model for the automatic text recognition of old and modern Church Slavonic printed books currently developed by A. Rabus.[10] At the moment of application, this model contained the material consisting of 92003 words, with a considerably low rate of incorrectly recognized characters (2.39%) after 150 training epochs.[11] The automatic text recognition process was started on *Prayer Book (Euchologion)* (1538–1540) from Božidar Vuković's printery.[12] In the first experiment, when we applied Rabus' generic model for old and modern Church Slavonic printed books to

167–170; Lazić 2020a: 334–344). After Božidar Vuković's death, during the period of 1456 to 1461, his son Vicenzo Vuković reprints *Psalter with Appendices* (1546), *Prayer Book (Miscellany for Travellers)* (1547), *Hieratikon (Liturgikon)* (1554), *Prayer Book (Miscellany for Travellers)* (1560), *Octoechos, Mode 5–8.* (1560) and *Psalter with Appendices* (1561) (see Pešikan 1994: 89–91; Lazić 2018b: 171–173; Lazić 2020a: 336). In Vicenzo Vuković's printery in 1561, Stefan of Scutari (see Pešikan 1994: 197) printed *Lenten Triodion*, and in 1566 Jakov of Kamena Reka (see Pešikan1994: 134) printed *Prayer Book (Miscellany for Travellers)*.

[9] The importance of Vuković printery and later Venetian printeries which used Vuković's equipment is evidenced by the fact that around two-thirds of the preserved stock of the Serbian Early Printed Books from 15th to 17th centuries come from these printeries. For a detailed list of the books printed in Venice see Lazić 2018b: 178–182.

[10] Once again I wish to express my gratitude to A. Rabus for allowing me to use the model, as well as for his support and help in the writing of this paper.

[11] The term *epoch* is used in machine learning to denote "one complete presentation of the data set to be learned to a learning machine" (see Burlacu and Rabus 2021: 1).

[12] The book is stored in the Matica Srpska library (cf. Grbić et al. 1994: 55–56; Pešikan 1994: 146–147), and is also available in the digital form, see http://digital.bms.rs/ebiblioteka/publications/view/5552.

folios 2–16 (a total of 28 pages of the *Prayer Book (Euchologion)* (1538–1540)), we obtained the first transcripts in which the rate of incorrectly recognized characters was 21.88% on average.[13] After the manual correction of obtained transcripts, we had the starting material for the training of the specific model. A model called *Dionisio 0.1.* (by the publisher Božidar Vuković's Italian pseudonym – *Dionisio della Vechia*) was trained only by using the transcripts of the first 28 folios of the book with 4049 words.[14] After 50 training epochs, the unrecognized character rate of the training set was 0.12%, while for the validation set (2 pages) it was 4.94%.[15] In the following stage of the transcription process we applied the *Dionisio 0.1.* model to the folios 16–51 (70 pages of *Prayer Book (Euchologion)* (1538–1540)). Having manually corrected automatically obtained transcripts, we gained more material for the training of a new model, and consequently somewhat better performance. The *Dionisio 0.2* model was trained by using the first 95 pages of the book with 13830 words. After 100 training epochs the unrecognized character rate for the training set was 0.17%, while for the validation set (5 pages) it was 2.67%. In the following stage of the transcription process we applied the *Dionisio 0.2.* model to folios 51–136 (169 pages in total) and after the manual correction of the automatically recognized text we trained the model again, gaining even better performance. The *Dionisio 0.3.* model was trained by using the material consisting of 256 pages of the book with 37902 words. After 100 training epochs the unrecognized character rate for the training set was 0.46%, while for the validation set (13 pages) it was 1.67%, which can be seen in the learning curve shown in the following picture.[16]

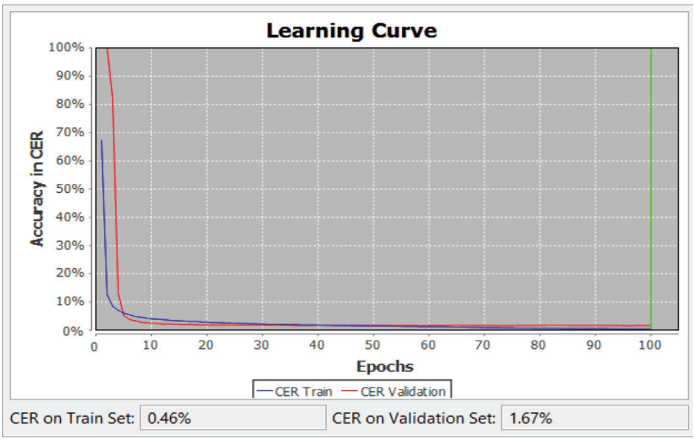
---

[13] This piece of information does not represent a true indicator of the model's success since the largest number of errors is associated with the failure to recognize accents. A more detailed quantitative analysis was not conducted because the model was still in the development stage at the time of writing of this paper.

[14] All models described in this paper have been created by using the HTR+ engine. The performance comparison of models trained on the same material by the HTR+ engine and PyLaia engine will be left for future investigations.

[15] In the machine learning process, the evaluation of a model's performance is made by using a validation set, consisting of a smaller part of the entire material prepared for the model training (between 2% and 10%), but which has not been seen during the training.

[16] The learning curve shows that during the first 10 training epochs the ratio of unrecognized characters decreases sharply both in the training set and the validation set. As the training process progresses, we can see that the ratio of unrecognized characters decreases more slowly, becoming more stable and maintaining the same level after 50 epochs until the end, which is expected since this is the number of training epochs recommended by *Transkribus*. It is not necessary to have a special computer equipment for the model training because the entire process is performed by Transkribus servers in Austria. The usual time spent for the training of models presented in this paper was between 2–3 hours.



Picture 1: *The Dionisio 0.3. model learning curve*

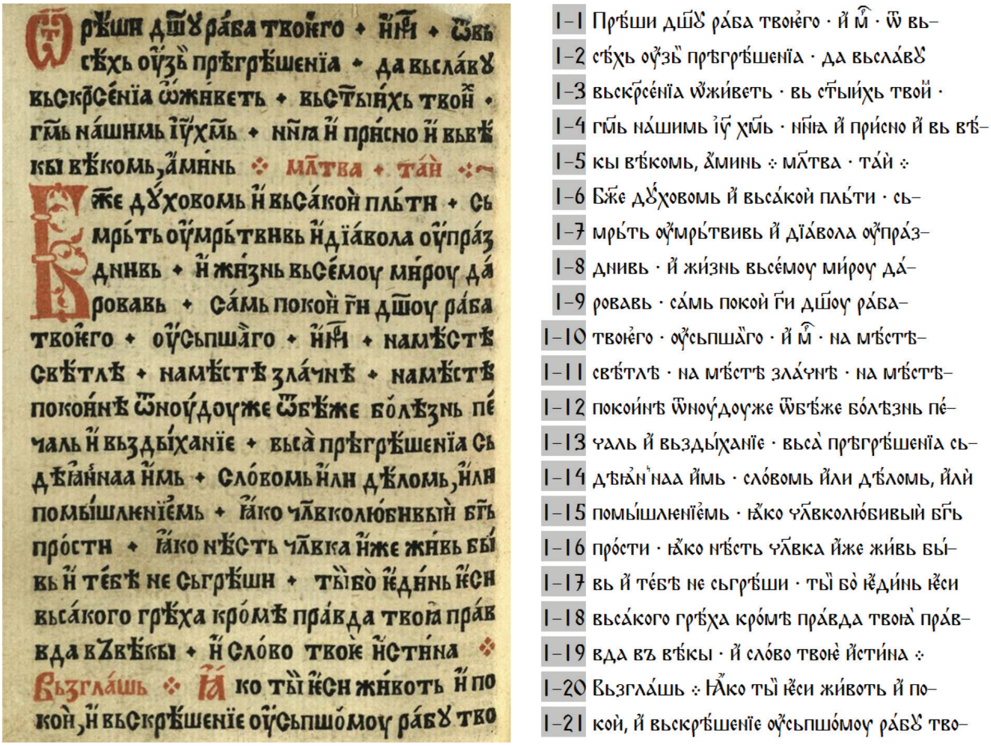
The results of the training process for the three models which were used clearly indicate that the success rate of the model depends on the amount of the training material (see Table 1).

Table 1: Performance of the *Dionisio* model and the amount of training data

Model	Word count	Number of pages	Training Set CER	Validation Set CER
Dionisio 0.1.	4049	28	0.12%	4.94%
Dionisio 0.2.	13830	95	0.17%,	2.67%
Dionisio 0.3.	37902	295	0.46%	1.67%

To illustrate the success of the *Dionisio 0.3.* model, we can use the image of the folio 85v and the automatically recognized text shown in the Picture 2.

The rate of incorrectly recognized characters from this page of the book is merely 0.78%. The *Dionisio 0.3.* model makes errors recognizing the initial Ѿ in Ѿрѣши 1 (in the incorrect form of Прѣши 1), as well as when recognizing the superscript ѫ in твоѭ 3 (as the incorrect твоѫ 3). The remaining errors are connected with the spaces between words: instead of вь слáвѫ 2 there is the incorrect вьслáвѫ 2, instead of рáба 9, мѣстъ 10, 11 there is the incorrect рáба- 9, мѣстъ- 10, 11. The error related to the recognition of the punctuation mark in the fifth line (instead of ⁘ there is the incorrect ⁖) has occurred because the model had no chance to see the mark during the training process. The comparative representation of the image and automatically recognized text shows that *Dionisio 0.3.* model successfully recognizes not only letters, but also



Picture 2: *Prayer Book (1538–1540) and Dionisio 0.3*

accent marks, punctuation marks (comma and full stop), titlos, superscript letters, initials and spaces between words.

We can conclude from the conducted experiment that the efficient models for the automatic recognition of Serbian Church Slavonic Printed Books (with less than 5% of incorrectly recognized characters) can also be trained by using transcripts which contain fewer than the recommended number of 5000 words.[17] In our *Prayer Book* (1538–1540) case, this means that by using only the transcripts of around 5% of the book (28 pages), a model can be created which will automatically recognize the remaining part of the book (540 pages in total) with approximately 5% of incorrectly recognized characters. What is especially important for the successful automatic recognition is the fact that the performance of the model can be enhanced – the more transcripts used for training, the fewer the number of incorrectly recognized

[17] The same results were obtained by Neumann 2021: 212 on the material consisting of Church Slavonic printed books from the Berlin State Library.

characters. In this way the researcher automatically obtains highly reliable first transcripts which require noting but the final manual correction, after which they can be used for further philological and linguistic investigations, and this, when compared to the traditional approach (manual recognition), requires incomparably less time to form a substantially larger corpus.

3. Application of the *Dionisio 0.3.* model to other volumes from the Vuković printery

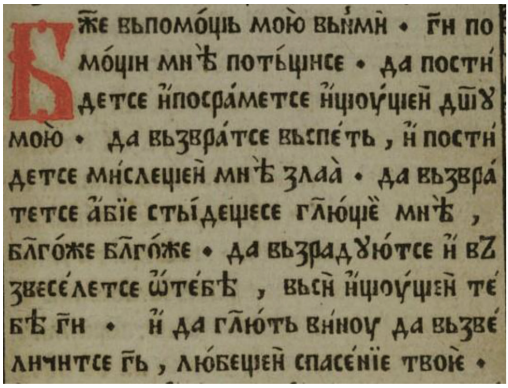
In the next part of the research we tested the performance of the *Dionisio 0.3.* model applied to automatic recognition of other Serbian Church Slavonic books from the printeries of Božidar and Vicenzo Vuković. The experiment encompassed ten pages from each of the following Venetian volumes:[18] *Psalter with Appendices* (1519–1520), *Prayer Book (Miscellany for Travellers)* (1536) and *Festal Menaion* (1538) from Božidar Vuković's printery; *Prayer Book (Miscellany for Travellers)* (1547) and *Hieratikon (Liturgikon)* (1554) from Vicenzo Vuković's printery. The initial hypothesis that the model trained on the material from a single book can also successfully be used for the recognition of other books from the same printery has been confirmed by the results of the quantitative analysis presented in the following table.

Table 2: *Dionisio 0.3.* model and other books from Vuković printery

Book	Character Error Rate (CER)
Psalter with Appendices (1519–1520)	8.99%
Prayer Book (Miscellany for Travellers) (1536)	6.86%
Festal Menaion (1538)	5.45%
Prayer Book (Miscellany for Travellers) (1547)	4.66%
Hieratikon (Liturgikon) (1554)	9.37%

[18] All books are available in digital form through the internet presentation of the Matica Srpska library; see at http://digital.bms.rs/ebiblioteka/publications/index/collection:4. For more details about the books see Grbić et al. 1994: 4–5, 8, 18–36, 62, 69–86.

As the table above shows, the best result in text recognition is obtained on the material from *Prayer Book (Miscellany for Travellers)* (1547) from Vicenzo Vuković's printery. As an illustration for the success of the model we can use the comparative presentation of the folio 32r (lines 6–15) and the automatically recognized text in the following image.

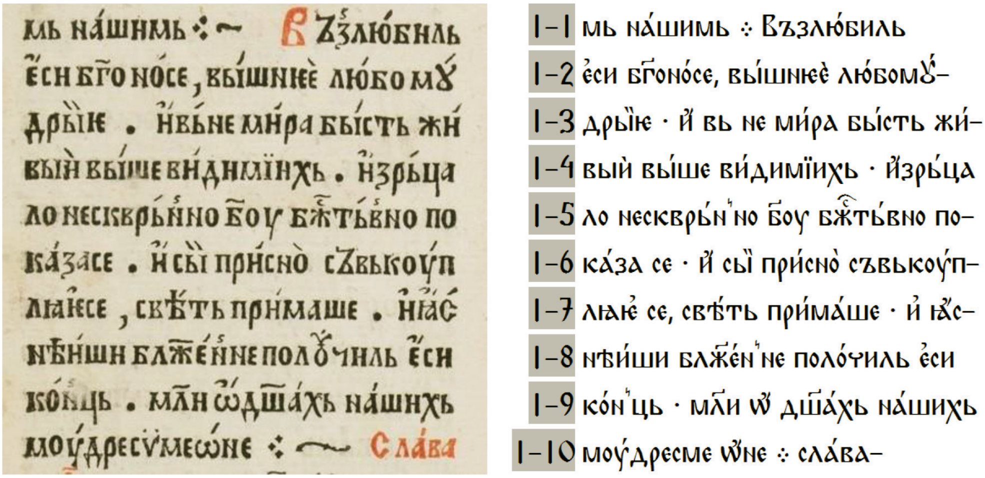


1-6 Бже вь помощь мою вьнмли · ги по–
1-7 мощи мнѣ потьщи се · да пости–
1-8 дет се и посрамлет се ищоущеи дшу
1-9 лою · да вьзврат се вьспѣть, и пости–
1-10 дет се мислеще мнѣ злаа · да вьзвра–
1-11 тет се аѣїѣ стыдеще се гл҃ющѣ мнѣ,
1-12 бг҃оже бг҃оже · да вьрадуют се и въ
1-13 звесѣлет се ѡ тѣбѣ вьсн ищоущеи тѣ
1-14 бѣ ги · и да гл҃ютъ вино҃у да вьзвѣ–
1-15 личит се гь, любещеи спасенїе твое ·

Picture 3: *Prayer Book* (*Miscellany for Travellers*) (1547) and the *Dionisio 0.3.* model

As the table shows, a conclusion can be drawn that the highest number of errors is to be attributed to the recognition of accents. Hence, instead of ищоущеи 8, мою 9, постидет 9/10, мислеще 10, гл҃ющѣ 11, вьсн 13, любещеи 15, твое 15, the *Dionisio 0.3.* model incorrectly renders ищоущеи 8, лою 8, постидет 9/10, мислеще 10, гл҃ющѣ 11, вьсн 13, любещеи 15, твое 15. Several errors are related to the recognition of spaces between words: instead of възвесѣлет 12/13, бг҃о же 12x2, тѣбѣ 13/14 the model renders въ звесѣлет 12/13, бг҃оже 12x2, тѣбѣ 13/14. It is interesting that the model also has problems with the recognition of the letter м: instead of вьнмн 6, посрамет 8, мою 8, there are incorrect forms вьнмли 6, посрамлет 8, лою 8.

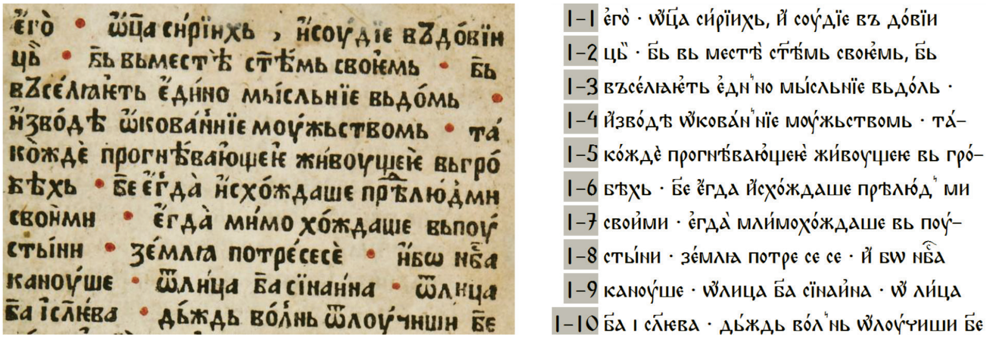
Similar performance of the *Dionisio 0.3.* model has also been observed in *Festal Menaion* (1538) from Božidar Vuković's printery. To illustrate the successfulness of the model on the following picture we can use the comparative representation of a segment from the folio 3r and the corresponding automatically recognized text.



Picture 4: *Festal Menaion* (1538)
and the *Dionisio 0.3.* model

Here also the largest number of errors is due to incorrect recognition of accents: instead of є̏си 2, 8, вьнє 3, примаше 7, моу́дре сvмеѡ́нє 10, the model reads є̀си 2, 8, вь нє 3, при́ма́ше 7, моу́дресме ѿнє 10. Several errors are connected with the space between words: instead of вьнє 3, и́ зрьцало 4/5, моу́дре сvмеѡ́нє 10, сла́ва 10, the model incorrectly renders вь нє 3, и́зрьцало 4/5, сла́ва– 10, моу́дресме ѿнє 10. The pajerak mark was not recognized in the example Въз҆лю́биль 1 (the model reads Възлю́биль 1). Inability to recognize the punctuation mark ⁘— 1, 10 (the model renders ⁘ 1, 10) was expected, since the model had no opportunity to see this mark during the training process. Errors in recognizing the letters occur in two examples: since the model did not have an opportunity to see the specific form of the letter ꙋ during the training process, it renders поло́училь 8 (instead of полꙋ̑чиль 8); the letter v is absent from the example моу́дре сvмеѡ́нє 10 (the model reads моу́дресме ѿнє 10).

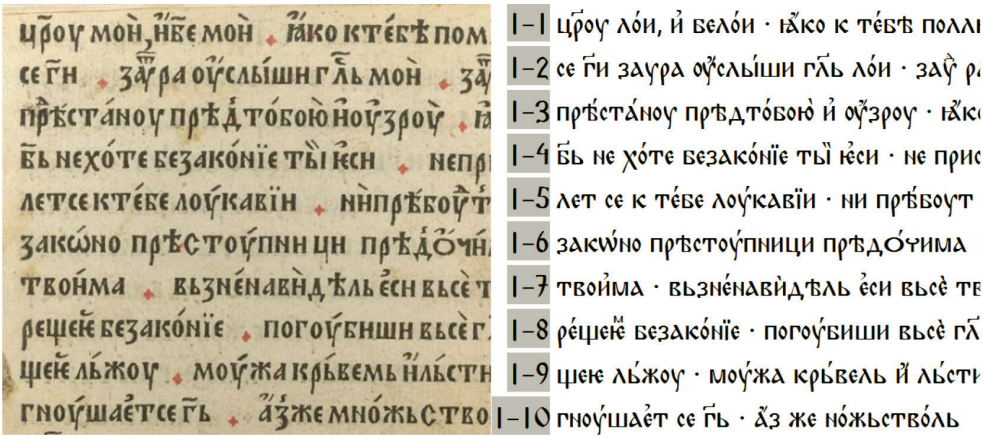
A slightly higher error rate than the one found in the two previous books was registered in *Prayer Book (Miscellany for Travellers)* from 1536. To illustrate the success of the model we can use the comparative representation of a part of the folio 49r and the corresponding automatically recognized text (Picture 5).



Picture 5: *Prayer Book (Miscellany for Travellers)* (1536) and the *Dionisio 0.3.* model

The highest number of errors was found with the accents and spaces between words: instead of ѐго̀ 1, ѐди́номы́сльнїе 3, та́ко̀ждѐ 4/5, прогнѣ́ваю́щеѥ̀ 5, живоущеѥ̀ 5, ѐг҇да̀ 6, потре́се сѐ 8, сїна́ина 9, і҆сл҃ѥ́ва 10 the model reads ѐго̀ 1, ѐдн҇но мы́сльнїе 3, та́ко́ждѐ 4/5, прогнѣ́ваю́щеѥ̀ 5, живоущеѥ 5, ѐгда 6, потре се се 8, сїна́ина 9, і сл҃ѥва 10; instead of въдо́вїиц҃ь 1/2, ѐди́номы́сльнїе 3, вь до́мь 3, прѣ лю́д҇ми 6, потре́се сѐ 8, ѿ ли́ца 9, ѿ лица 9, і҆сл҃ѥ́ва 10 the model reads въ до́вїи ц҃ь 1/2, ѐдн҇но мы́сльнїе 3, вьдо́ль 3, прѣлю́д҇ ми 6, потре се се 8, ѿлица 9x2, і сл҃ѥва 10. In a smaller number of examples there are errors in recognizing the pajerak mark and superscript letters: instead of ѐди́номы́сльнїе 3, ѐг҇да̀ 6, прѣ лю́д҇ми 6 there are the incorrect forms ѐдн҇но мы́сльнїе 3, ѐгда 6, прѣлю́д҇ ми 6. It is interesting that in this book, like in *Prayer Book (Miscellany for Travellers)* from 1547, the model makes errors when recognizing the letter м: instead of вь до́мь 3, ми́мохо́ждаше 7 there is the incorrect вьдо́ль 3, мли́мохо́ждаше 7. The remaining errors in the recognition of letters were found with ѿ and и: instead of ѿ ли́ца 9, ѿ лица 9, ѿлоу́чиши 10 and ѐди́номы́сльнїе 3 the model incorrectly reads ѐдн҇но мы́сльнїе 3, ѿлица 9x29, ѿлоу́чиши 10.

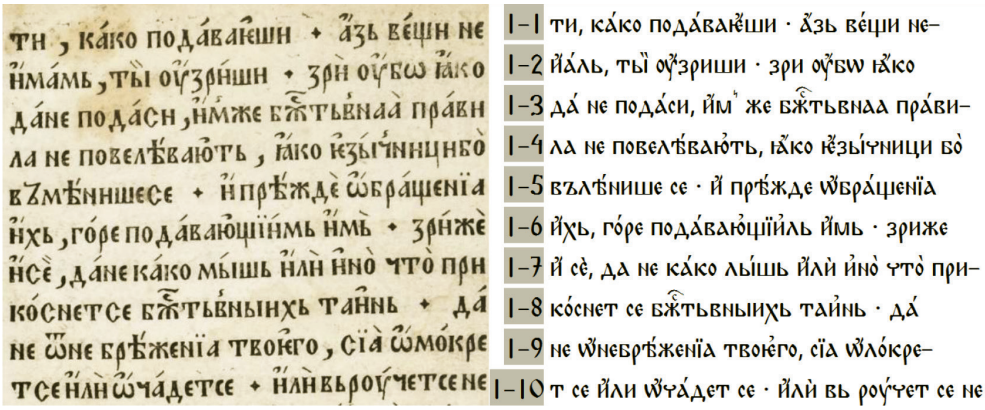
When compared to the three previous books, the performance of the *Dionisio 0.3.* model applied to *Psalter with Appendices* (1519–1520) and *Hieratikon (Liturgikon)* (1554) is worse to some extent, but taken as a whole it is quite satisfactory bearing in mind that CER of both books is smaller than 10%. The qualitative analysis showing the success of the model applied to the *Psalter with Appendices* (1519–1520) is based on the comparative representation of the image containing a segment of the 14r folio and the automatically recognized text provided in the following picture:



Picture 6: *Psalter with Appendices* (1519–1520) and the *Dionisio 0.3.* model

Here too the largest number of errors is connected with the recognition of accents: instead of моѝ 1, 2, и̑ 1, ꙗ̑ко 1, оу̑зроу̀ 3, ꙗ̑ко 3, нѝ 5, прѣд᷾ о̀чи́ма 6, твои́ма 7, тво́рецеѣ 7/8, гл҃ю́цеѣ 8/9, мно́жьствомь 10, the model incorrectly reads ло́и 1, 2, и̑ 1, ꙗ̑ко 1, оу̑зроу 3, ꙗ̑ко 3, ни 5, прѣдо́чима 6, твои́ма 7, тво́ре́цеѣ 7/8, гл҃ю́цеѣ 8/9, но́жьствóль 10. The errors in the recognition of spaces between words are also frequent: instead of бе̑ моѝ 1, помл҃ю́ 1, заꙋ̑ра̀ 2, прѣд᷾ то́бою̀ 3, закѡ́нопрѣстоу́пници 6, прѣд᷾ о̀чи́ма 6, лы́стива 9, the model incorrectly reads: белóи 1, поллю́- 1, заꙋ̑ ра̀ 2, прѣдто́бою̀ 3, закѡ́но прѣстоу́пници 6, прѣдо́чима 6, лы́стива- 9. In this book also the model makes frequent errors in the recognition of the letter м: instead of моѝ 1, 2, помл҃ю́ 1, кры́вемь 9, мно́жьствомь 10 there are ло́и 1, 2, поллю́- 1, кры́вель 9, но́жьствóль 10. A higher percentage of CER in relation to previous books in connected with the higher number of errors in the recognition of superscript letters, titlo mark and pajerak mark: instead of бе̑ моѝ 1, помл҃ю́ 1, заꙋ̑ра̀ 2, гл҃ь 2, заꙋ̑ра̀ 2, пр҃ста́ноу 3, пр҃боу᷾т 5 the model incorrectly reads белóи 1, поллю́- 11, заура 2, гл҃ь 2, заꙋ̑ ра̀ 2, пр҃ста́ноу 3, пр҃боут 5; instead of прѣд᷾ то́бою̀ 3, пр҃боу᷾т 5, прѣд᷾ о̀чи́ма 6, а̑з᷾ 10, the model renders прѣдто́бою̀ 3, пр҃боут 5, прѣдо́чима 6, а̑з 10.

The qualitative analysis of the model's success applied to the *Hieratikon (Liturgikon)* (1554) from Vicenzo Vuković's printery is based on the comparative representation of the image containing a portion of the 14r folio and the automatically recognized text given in the following picture.



Picture 7: *Hieratikon (Liturgikon)* (1554) and the *Dionisio 0.3.* model

The following examples contain the observed errors related to the recognition of accents: instead of пода́ваѥ̀ши 1, оу̏зри́ши 1, зрѝ 2, бж̑тьв᷄наа̀ 3, ѥ̀зы́ч᷄ници 4, прѣ́ждѐ 5, пода́ваю̀щїи́мь 6, зри́ жѐ 6, сѐ 7, да́ 7, сїа̀ 9, и̏лѝ 10, the models incorrectly renders пода́ваѥ̀ши 1, оу̏зриши 1, зри 2, бж̑тьвнаа 3, ѥ̀зы́чници 4, прѣ́жде 5, пода́ваю̀щїи́ль 6, зриже 6, сѐ 7, да 7, сїа 9, и̏ли 10. Errors connected with the recognition of the spaces between words were recorded in the following examples: instead of не 1, зри́ жѐ 6, w̑ небрѣ́женїа 9, w̑ мо́кре- 9, вьроу́чет 10, the model incorrectly renders не- 1, зриже 6, w̑небрѣ́женїа 9, w̑ло́кре- 9, вь роу́чет 10. In this book the model makes errors in the recognition of the pajerak mark, the letter м and the letter w̑: instead of бж̑тьв᷄наа 3, ѥ̀зы́ч᷄ници 4, бж̑тьв᷄ныихь 8 there are incorrect бж̑тьвнаа 3, ѥ̀зы́чници 4, бж̑тьвныихь 8; instead of и̏ма́мь 2, въмѣ́нише 5, пода́ваю̀щїи́мь 6, мы́шь 7 there are incorrect и̏а́ль 2, възлѣ́нише 5, пода́ваю̀щїи́ль 6, лы́шь 7; instead of w̑ небрѣ́женїа 9, w̑ мо́кре- 9 we have the incorrect w̑небрѣ́женїа 9, w̑ло́кре- 9.

4. Creation and Evaluation of the Generic Model

Based on the quantitative and qualitative analyses presented in the previous chapter we can conclude that the *Dionisio 0.3.* model can also be relatively successfully used for the automatic recognition of other books from Božidar and Vicenzo Vuković's printery. In this process the model most frequently makes errors in the

recognition of accents and spaces between words, while the errors in the recognition of pajerak mark, superscript letters and the titlo mark occur less frequently. What is especially interesting is that the model had problems recognizing certain letters: most frequently the letters ѡ and ѿ. Taking these errors into account, the future process of transcribing books from Vuković printery can be enhanced by manually correcting the transcripts obtained by applying the *Dionisio 0.3.* model, and then used to train a generic model which would contain the material from various books. Our starting hypothesis is that the generic model will achieve better results when recognizing other books from Božidar and Vicenzo Vuković's printery than the special model trained only on a single book (*Prayer Book (Euchologion)* from 1538–1540).

Ground Truth data required for the training of the generic model were acquired by manually correcting the transcripts automatically obtained by applying the special *Dionisio 0.3.* model. In line with our findings about the dependence of a model's success on the amount of data available for the training (see Table 1), along with the similar findings associated with the Church Slavonic books from the Berlin State Library (see Neumann 2021: 212), a goal was set to procure a critical amount of data consisting of around 10000 words per each printed book for the purposes of training the first version of the generic model called *Dionisio 1.0.* The structure and the mount of training data is given in Table 3.

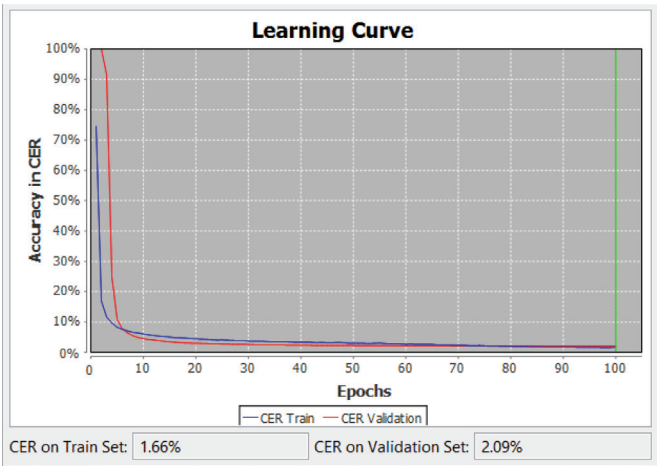
Table 3: Structure of the model and the amount of data used for the training of the *Dionisio 1.0.* generic model

Book	**Word count**	**Number of folios**
Prayer Book (Euchologion) (1538–1540)	39889	269
Psalter (1519–1520)	10132	81
Prayer Book (Miscellany for Travellers) (1536)	10618	70
Festal Menaion (1538)	10732	30
Prayer Book (Miscellany for Travellers) (1547)	10006	66
Hieratikon (Liturgikon) (1554)	10196	80
Total	**91573**	**596**

The structure and performance of the generic model *Dionisio 1.0.* after a hundred epochs have been given in Table 4, while the learning curve is provided in Picture 8.

Table 4: Structure and performance of Dionisio *1.0.* generic model

Dionisio 1.0.	**Word count**	**Number of folios**	**CER**
Training set	86347	554	1.66%
Validation set	5226	32	2.09%



Picture 8: Learning curve of the *Dionisio 1.0.* model

When comparing the performance of the special *Dionisio 0.3.* model and the generic *Dionisio 1.0.* model it is important to mention that the data from the *Prayer Book (Euchologion)* (1538) were omitted from the validation set. The comparison of the result connected with the *Dionisio 0.3.* model, which is presented in Table 2 (CER varies from 4.66% and 9.37%, depending on the book), and the result of the generic model applied to the validation set (which is 2.09%), clearly confirms the starting hypothesis that the performance of the generic model in the recognition of Serbian Church Slavonic books from Božidar and Vicenzo Vuković's printery is substantially better than the performance of the model trained only on a single book – *Prayer Book (Euchologion)* from 1538.

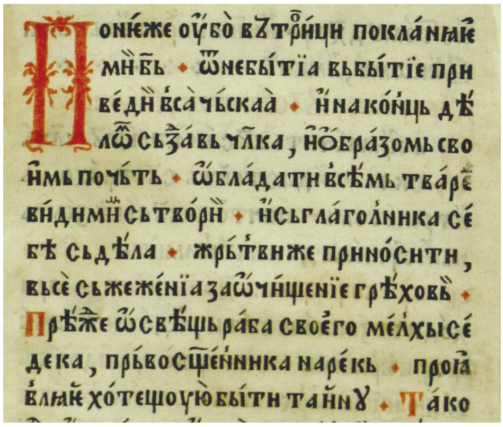
The same conclusion can be drawn based on the data from the performance comparison of the two models applied to the three books not included in the training of the generic model: *Hieratikon (Liturgikon)* (1519) and *Octoechos, Mode 5–8.* (1537) originating from Božidar Vuković's printery, and the *Prayer Book (Miscellany for Travellers)* (1560) from Vicenzo Vuković's printery.[19] A summary of the results is given in the following table:

[19] The images of these books were downloaded from the internet presentation of the National Library of Serbia: https://digitalna.nb.rs/sf/NBS/Stara_stampana_knjiga.

Table 5: Performance of *Dionisio 0.3.* and *Dionisio 1.0* models

Book	***Dionisio 0.3.*** **(CER)**	***Dionisio 1.0.*** **(CER)**
Hieratikon (Liturgikon) (1519)	9.35%	2.15%
Octoechos, Mode 5–8 (1537)	9.24%	4.50%
Prayer Book (Miscellany for Travellers) (1560)	8.29%	2.47%

To illustrate the success of the generic model, in the next picture we will use the parallel representation of the image containing the portion of the folio 100b of *Hieratikon (Liturgikon)* (1519) and the relevant automatically recognized text.



1-1 Поніе́же оу҃бо̀ въ тр҃и́ци покла́нмаѣ–
1-2 мй҃ бь҃ · ѿ невы́тїа вь бы́тїе при–
1-3 ве́ди в҃ сѧ҃чьскаа̀ · и҆ на ко́н҃ць дѣ–
1-4 лѡ҃ съ҃зда́вь чл҃ка, и҆ о҆бра́зомь сво–
1-5 и́мь почьть · ѡ҃бла́дати вс҃ѣмь тва́рѣ
1-6 ви́димй҃ сьтво́рй҃ · и҆ сьгла́гол҃ника сѣ–
1-7 вѣ сьдѣ́ла · жрь҃тв҃и же прино́сити,
1-8 вьсѣ̀ сьже́же́нїа за ѡ҃чи́щенїе грѣ́ховь҃ ·
1-9 Прѣ́же ѡ҃ св҃ѣць ра́ба своего́ мел҃хы сѣ–
1-10 дека, прьвосщ҃е́н҃ника на ре́кь · проѣ–
1-11 в҃лаѣ хо́тещоую̀ бы́ти таи́нь҃ · Та́ко

Picture 9: *Hieratikon (Liturgikon)* (1519) and the *Dionisio 1.0.* model

A small number of errors that were observed in the automatic recognition most frequently occured with the accent marks and the spaces between words: instead of приве́дй҃ 2/3, сьже́же́нїа 8, таи́нь҃ 11 the model renders приве́ди 2/3, сьже́же́нїа 8, таи́нь҃ 11; instead of ѡ҃ св҃ѣць 9, на ре́кь 10, the model renders ѡ҃ св҃ѣць 9, на ре́кь 1. We have also recorded an instance of the incorrectly recognized pajerak mark and an instance of an incorrect superscript letter: instead of ме́л҃хысе́дека 9/10 and тр҃и́ци 1, the model renders ме́лхы се́де́ка 9/10 and тр҃и́ци 1.

5. Concluding Remarks

The conducted research has shown that by using the *Transkribus* software platform extraordinarily efficient models for the automatic recognition of Serbian Church Slavonic books can be created. By using an example of *Prayer Book (Euchologion)*, printed in Venice in Božidar Vuković's (1538–1540) printery, we described the process of transcription and creation of a special model for the automatic recognition of individual Serbian Church Slavonic printed books. The efficiency of different versions of the special *Dionisio (0.1–0.3.)* model (named after Božidar Vuković's Italian pseudonym – *Dionisio della Vechia*) depends on the number of transcripts used for training. By using the transcripts with only around 4000 words, an efficient model can be created, recognizing approximately 95% of characters. If the number of manuscripts used for the training is increased, this leads to the improvement of the model and reduces the rate of the unrecognized characters to 1–2%. The paper has shown that the model trained for the automatic recognition of individual books can efficiently be used for the automatic recognition of other books belonging to the same printing tradition or the ones originating from the same printery. Thus, by using the special *Dionisio 0.3.* model, we obtained transcripts (around 10000 words per book) of other Serbian Church Slavonic books printed in Venice in Božidar and Vicenzo Vuković's printery: *Psalter* (1519–1520), *Prayer Book (Miscellany for Travellers)* (1536), *Festal Menaion* (1538), *Prayer Book (Miscellany for Travellers)* (1547) and *Hieratikon (Liturgikon)* (1554). After the manual correction, the transcripts were used to create the first version of the *Dionisio 1.0.* generic model, which manifested extraordinary results in the automatic recognition of the Venetian editions. The Character Error Rate (CER) ranges from 2–5%, depending on the book, and the errors are most frequently connected with the failure to recognize accent marks and blanks between words. The *Dionisio 1.0.* generic model represents the first publicly available model for the automatic recognition of the Serbian Church Slavonic Cyrillic script operating within the *Transkribus* software platform (see https://readcoop.eu/model/dionisio-1-0/). When this model is applied, the platform users can automatically obtain transcripts of Serbian Church Slavonic books from Božidar and Vicenzo Vuković's printery, which, after manual correction, can further be used to create digital editions and electronic corpora, as well as for different philological and linguistic investigations. In this way, by using artificial intelligence and machine learning, the *Transkribus* software platform enables us to perform mass digitization of Serbian Church Slavonic printed books, thus allowing for the philological and linguistic investigations of the Serbian Church Slavonic language to be based on larger and more representative samples. In the next stage of the research, the performance of the *Dionisio 1.0.* model will be tested on Serbian Church Slavonic books printed in other old Serbian printeries

(Cetinje, Goražde, Mrkšina Crkva, Belgrade, Mileševa, Gračanica, etc.), which should in the near future lead to the creation of a generic model for the automatic recognition of Serbian Church Slavonic printed books in their totality.

REFERENCES

Burlacu and Rabus 2021: Burlacu Constanţa and Rabus Achim. "Digitising (Romanian) Cyrillic using Transkribus: new perspectives." *Diacronia* 14 (2021), 1–9.

Grbić et al. 1994: Грбић, Душица, Катарина Минчић-Обрадовић и Катица Шкорић . *Ћирилицом штампане књиге 15–17. века Библиотеке Матице српске.* Нови Сад: Матица српска, 1994.

Lazić 2013: Лазић, Мирослав. "Између патриотизма, побожности и трговине: мотиви издавачке делатности Божидара Вуковића." *Археографски прилози* 35 (2013), 49–94.

Lazić 2018a: Лазић, Мирослав. "Од Божидара Вуковића до Дионизија дела Векије: идентитет и псеудоним у култури раног модерног доба." У *Scala Paradisi.* Ур. Анатолий А. Турилов и др. *Scala Paradisi*. Београд: САНУ, 2018, 165–185.

Lazić 2018b: Lazić, Miroslav. "Venice and editions of Early Serbian Printed Books." *Thesaurismata* 48 (2018), 161–192.

Lazić 2020a: Лазић, Мирослав. "Инкунабуле и палеотипи: српскословенске штампане књиге од краја 15. до средине 17 века." У *Осам векова аутокефалије Српске православне цркве*. 2. Ур. Владислав Пузовић, Владан Таталовић. Београд: Свети архијерејски Синод Српске православне цркве – Православни богословски факултет у Београду, 2020, 325–344.

Lazić 2020b: Lazić, Miroslav. "Between an Imaginary and Historical Figure: Božidar Vuković`s Professional Identity." *Ricerche Slavistiche* 63 (2020), 141–156.

Mühlberger et al. 2019: Mühlberger, Günther, L. Seaward, M. Terras, S. Oliveira Ares, V. Bosch, M. Bryan, S. Colluto, H. Déjean, M. Diem, S. Fiel, B. Gatos, A. Greinoecker, T. Grüning, G. Hackl, V. Haukkovaara, G. Heyer, L. Hirvonen, T. Hodel, M. Jokinen, P. Kahle, M. Kallio, F. Kaplan, F. Kleber, R. Labahn, M. Lang, S. Laube, G. Leifert, G. Louloudis, R. McNicholl, J. Meunier, J. Michael, E. Mühlbauer, N. Philipp, I. Pratikakis, J. Puigcerver Pérez, H. Putz, G. Retsinas, V. Romero, R. Sablatnig, J. Sánchez, P. Schofield, G. Sfikas, C. Sieber, N. Stamatopoulos, T. Strauss, T. Terbul, A. Toselli, B. Ulreich, M. Villegas, E. Vidal, J. Walcher, M. Wiedermann, H. Wurster and K. Zagoris. "Transforming scholarship in the archives through handwrittwn text recognition." *Journal of Documentation* 5 (75) (2019), 954–976.

Neumann 2021: Neumann, Vladimir. "Deep Mining of the Collection of Old Prints *Kirchenslavica Digital*." *SeS* 21 (2021), 207–216.

Pešikan 1994: Пешикан, Митар. "Лексикон српскословенског штампарства." У Митар Пешикан и др. *Пет векова српског штампарства 1494–1994: раздобље српскословенске штампе XV–XVII.* Београд: Народна библиотека Србије – Матица српска, 1994, 71–218.

Polomac and Lutovac Kaznovac 2021: Polomac, Vladimir and Tamara Lutovac Kaznovac. "Automatic Text Recognition of Serbian Medieval Manuscripts by applying the

Transkribus software platform: current state and future perspectives.“ *Зборник Матице српске за филологију и лингвистику* 64/2 (2021), 7–26.

Rabus 2019a: Rabus, Achim. “Recognizing Handwritten Text in Slavic Manuscripts: a Neural-Network Approach using Transkribus.“ *SeS* 19 (2019), 9–32.

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

Rabus 2022: Rabus, Achim. “Handwritten Text Recognition for Croatian Glagolitic.“ *Slovo* 72 (2022), 181–192.

Thompson 2021: Thompson, Walker. “Using Handwritten Text Recognition to Transcribe Historical Multilingual Lexica.“ *SeS* 21 (2021), 217–231.

Transkribus Glossary: https://readcoop.eu/glossary/<15.01.2022>

[Grbić, Dušica, Katarina Minčić-Obradović i Katica Škorić. *Ćirilicom štampane knjige 15–17. veka Biblioteke Matice srpske. Novi Sad: Matica srpska*, 1994.

Lazić, Miroslav. “Između patriotizma, pobožnosti i trgovine: motivi izdavačke delatnosti Božidara Vukovića.“ *Arheografski prilozi* 35 (2013), 49–94.

Lazić, Miroslav. “Od Božidara Vukovića do Dionizija dela Vekije: identitet i pseudonim u kulturi ranog modernog doba“. U *Scala Paradisi*. Ur. Anatolij A. Turilov i dr. *Scala Paradisi*. Beograd: SANU, 2018, 165–185.

Lazić, Miroslav. “Inkunabule i paleotipi: srpskoslovenske štampane knjige od kraja 15. do sredine 17. veka.“ U *Osam vekova autokefalije Srpske pravoslavne crkve*. 2. Ur. Vladislav Puzović, Vladan Tatalović. Beograd: Sveti arhijerejski Sinod Srpske pravoslavne crkve – Pravoslavni bogoslovski fakultet u Beogradu, 2020, 325–344.

Pešikan, Mitar. “Leksikon srpskoslovenskog štamparstva“. U Mitar Pešikan i dr. *Pet vekova srpskog štamparstva 1494–1994: razdoblje srpskoslovenske štampe XV–XVII*. Beograd: Narodna biblioteka Srbije – Matica srpska, 1994, 71–218.]

About the author…

Vladimir Polomac is an Associate Professor at the Department of the Serbian Language at the Faculty of Philology and Arts, University of Kragujevac (Serbia). For the monograph entitled Језик повеља и писама Српске деспотовине [The Language of Chaters and Letters of Serbian Despotate] he received the “Pavle and Milka Ivić” award by the Serbian Slavic Association for the best book in the field of linguistic Slavic studies in Serbia in 2016. His current scientific interests include historical (corpus) linguistics (historical dialectology and onomastics of the Serbian language), philological and linguistic research of Serbian medieval literacy especially. He has been a member of the Onomastics Committee of the Serbian Academy of Sciences and Arts since 2015.