Serbian Early Printed Books: Towards Generic Model for Automatic Text Recognition using *Transkribus*

Vladimir Polomac*

* Serbian Language Department, Faculty of Philology and Arts, University of Kragujevac
Jovana Cvijića bb, 34 000 Kragujevac, Serbia
v.polomac@filum.kg.ac.rs

Abstract

The paper describes the process of creating and evaluating a new version of the generic model for automatic text recognition of Serbian Church Slavonic printed books within the *Transkribus* software platform, based on the principles of artificial intelligence and machine learning. The generic model *Dionisio 2.0.* was created on the materials of Serbian Church Slavonic books from various printing houses of the 15th and 16th centuries (Cetinje, Venice, Goražde, Mileševa, Gračanica, Belgrade and Mrkša's Church), and, during the evaluation of its performance, it was noticed that CER was about 2–3%. The *Dionisio 2.0.* model will be publicly available to all users of the *Transkribus* software platform in the near future.

1. Introduction

The research on creating a model for automatic text recognition of the Serbian Church Slavonic printed books from Venice using a software platform *Transkribus*,[1] presented in Polomac (2022), represents the starting point for this paper. This paper describes the process of transcription and creation of a specific model[2] for automatic text recognition of *Prayer Book (Euchologion)* printed between 1538 and 1540 in the printing house of Božidar Vuković,[3] as well as the process of creating a generic model[4] for automatic text recognition of other books printed in Venice in the printing house of Božidar Vuković and his son Vićenco.[5] The most important result of this paper is the creation of the first version of the model *Dionisio 1.0.* (named after an Italian pseudonym for Božidar Vuković – *Dionisio della Vechia*) representing the first publicly available resource for automatic reading of Serbian Church Slavonic manuscripts and printed books within the *Transkribus* software platform (cf. https://readcoop.eu/model/dionisio-1-0/).

The *Dionisio 1.0.* model structure is shown in Table 1, and its performance is displayed in Table 2.

Book	**Word count**
Prayer Book (1538–1540)	39,889
Psalter (1519–1520)	10,132
Miscellany for Travellers (1536)	10,618
Festal Menaion (1538)	10,732
Miscellany for Travellers (1547)	10,006
Hieratikon (Liturgikon) (1554)	10,196
Total	91,573

Table 1: *Dionisio 1.0.* Structure and the Amount of Training Data.

Word count	**Number of epochs**[6]	**CER**[7] **on Train set**	**CER on Validation set**
86,347	100	1.66%	2.09%

Table 2: *Dionisio 1.0* Performance.

[1] *Transkribus* (https://readcoop.eu/transkribus) represents an open-access software platform for automatic text recognition and retrieval developed as part of the READ project at the University of Innsbruck. More details about the technological background and operating system cf. Mühlberger et al. (2019).

[2] 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. For more details see Mühlberger et al. (2019) and Rabus (2019a).

[3] 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 (*Dionisius a Vetula*) and an Italian pseudonym (*Dionisio della Vecchia*) from his Serbian name and the toponym of Starčeva Gorica (at Lake Skadar), indicating his origin (Lazić, 2018). 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 (Lazić, 2020b).

[4] Unlike a specific model that is trained to recognize a single manuscript or printed book, a generic model contains material from different manuscripts or printed books. More details on the possibilities and pitfalls of training generic models can be found in Rabus (2019b).

[5] After the death of Božidar Vuković, Vićenco Vuković had reprinted several of his father's editions until 1561, and later rented his equipment to other Venetian printers. For more details about his life and work see also Pešikan (1994).

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

[7] The Character Error Rate (CER) is calculated by comparing the automatically generated text and the manually corrected version. See for more details in Transkribus Glossary https://readcoop.eu/glossary/character-error-rate-cer/.

In the continuation of the research, we aimed at examining the performance of the *Dionisio 1.0.* model on Serbian Church Slavonic books created in other printing houses, firstly in Venetian printing houses created after closing Božidar and Vićenco Vuković's printing house, and then in other old Serbian printing houses of the 15th and 16th centuries (Cetinje, Goražde, Mrkša's Church, Belgrade, Mileševa and Gračanica), thus ultimately offering a generic model for the automatic text recognition of Serbian Church Slavonic printed books as a whole.

2. Applying the *Dionisio 1.0.* Model on Books from Other Venetian Printing Houses

In the first experiment, we tested the performance of the *Dionisio 1.0.* model on several Serbian Church Slavonic books printed in Venice after closing Božidar and Vićenco Vuković's printing house: *Lenten Triodion* was printed in 1561 by Stefan of Scutari in the Camillo Zanetti's printing house, *Prayer Book* (*Miscellany for Travellers)* was printed in 1566 by Jakov of Kamena Reka, *Prayer Book (Euchologion)* was created in 1570 in the printing house of Jerolim Zagurović and *Psalter with Appendices* was printed in 1638 in the printing house of Bartol Ginammi (Pešikan, 1994). The starting hypothesis of the paper in the current experiment was that the model trained on the materials of Serbian Church Slavonic books from the printing house of Božidar and Vićenco Vuković would be useful for automatic text recognition of other Venetian editions printed using their printing equipment.

The statistical results of the experiment are shown in the following table.

Book	CER
Lenten Triodion (1561)	9.41%
Miscellany for Travellers (1566)	11.63%
Prayer Book (Euchologion) (1570)	13.67%
Psalter with Appendices (1638)	16.04%

Table 3: Application of the *Dionisio 1.0.* model on publications from other Venetian printing houses.

The unexpectedly high CER does not necessarily indicate poor performance of the *Dionisio 1.0.* model. The largest number of errors in text recognition is the result of the fact that in these books accent marks are used differently than in the books from the printing house of Božidar and Vićenco Vuković, which were used to train the *Dionisio 1.0.* model. This fact is especially evident in *Prayer Book (Euchologion)* from the printing house of Jerolim Zagurović (1570) and *Psalter with Appendices* from the printing house of Bartol Ginammi (1638) in which only *spiritus lenis* with an *oxia* over the initial vowel grapheme was used.

To illustrate this claim, we shall use a comparative presentation of a photograph of a part of sheet 2b *Prayer Book (Euchologion)* (1570) and an automatically read text using the *Dionisio 1.0.* model.

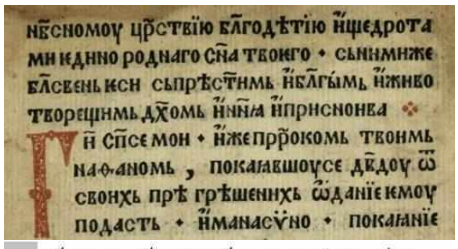


Figure 1: The Automatically Read Text of a Segment of Sheet 2b *Prayer Book (Euchologion)* from 1570.

The greatest number of errors in text recognition refers to cases in which the model outputs accent marks in accordance with the material on which it was trained, although in the text of *Prayer Book (Euchologion)* these marks were not used: so instead of щедротами 1/2, твоего 2, ними 2, бл҃свень кси 3, животворещимь 3/4, дх҃омь 4, прискноива 4, мои 5, твоимь 5, наданомь 6, своихь 7, прѣгрѣшенихь 7, емоу 7, подасть 8, манасуно 8, покаѧнїе 8 the model outputs щедро́тами 1/2, твоего́ 2, ни́ми 2, бл҃све́нь кси 3, живо́творе́щимь 3/4, дх҃о́мь 4, прискноива 4, мои́ 5, твои́мь 5, нада́номь 6, свои́хь 7, прѣгрѣше́нихь 7, е́моу 7, пода́сть 8, ма́насуно 8, покаѧнїе 8. Along with the accent marks, the model inccorectly reads a *pajerak mark* in two examples only: instead of кдинороднаго 2, покаѧвшоу 6 there is the incorrect кди́норо́дна́го 2, покаѧв҃шоу 6. In one example, instead of *oxia* there is an incorrect double *circumflex*: instead of бл҃гы́мь 3 there is the incorrect бл҃гы̑мь 3.

The same problem is exhibited by the comparative presentation of the photograph of a part of sheet 5b *Psalter with Appendices* (1638) and the automatically read text.

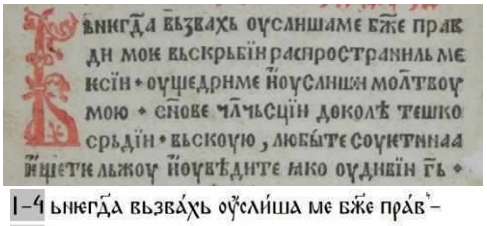
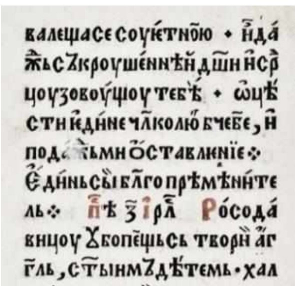


Figure 2: The Automatically Read Text of a Part of Sheet 5b *Psalter with Appendices* from 1638.

Here, too, the largest number of errors refers to cases in which the *Dionisio 1.0.* model outputs accent marks according to the patterns of their use in the Venetian books that served for its training, although in the text of Ginammi's *Psalter with Appendices* these marks were not used. Thus instead of вьзвахь 4, оуслиша 4, правди 4/5, моѥ 5, скрьбїи 5, распространиль 5, ме 5, ѥсїи 6, оущедри 6, оуслиши 6, мою 7, сн҃ове 7, до колѣ 7, тешкосрьдїи 7, вьскоую 8, любыте 8, соуѥтнаа 8, льжоу 9, оувѣдите 9, ꙗко 9, оудивїи 9 the model incorrectly outputs вьзва́хь 4, оу̑слиша 4, пра́в҆ди 4/5, моѥ́ 5, скрьбі́и 5, распростра́ниль 5, ме́– 5, ѥ́сїи 6, оу̑ще́дри 6, оу̑сли́ши 6, мою́ 7, сн҃о́ве 7, до ко́лѣ 7, те́шкосрьдїи 7, вьскоу́ю 8, лю́быте 8, соуѥ́тнаа 8, льжоу́ 9, оу̑вѣдите 9, ꙗ́ко 9, оу̑ди́вїи 9. Here, as well, the other types of errors are confirmed by isolated examples: *pajerak mark*: instead of правди 1/2 there is the incorrect пра́в҆ди 1/2; space between words: instead of ме 5 the incorrect ме́– 5; initials: instead of Вьнегд҃а 4 the incorrect ьнегд҃а 4; incorrect accent recognition: instead of и̏ 6 there is the incorrect и́ 6.

The given examples of the most common errors show that, despite the high percentage of incorrectly recognized characters, after the automatic post-correction of the transcripts which would include accent marks removal using the *Search/Replace chosen chars in transcript* option, the *Dionisio 1.0.* model can also be very efficient in recognizing Serbian Church Slavonic books created in the printing houses of Jerolim Zagurović and Bartol Ginammi during the 16th and 17th centuries.

The greatest number of errors in the automatic recognition of the text *Lenten Triodon* (1561) by Stefan of Scutari and *Prayer Book (Miscellany for Travellers)* (1566) by Jakov of Kamena Reka also refers to the recognition of accent marks. However, what distinguishes these books from the books from the printing houses of Jerolim Zagurović and Bartol Ginammi is that accent marks are actually used, yet in different positions compared to the books from the printing house of Božidar and Vićenco Vuković on which the *Dionisio 1.0.* model was trained. To illustrate this claim, we will first use a comparative presentation of a part of sheet 3a *Lenten Triodon* (1561) by Stefan of Scutari and the automatically read text using the *Dionisio 1.0.* model.

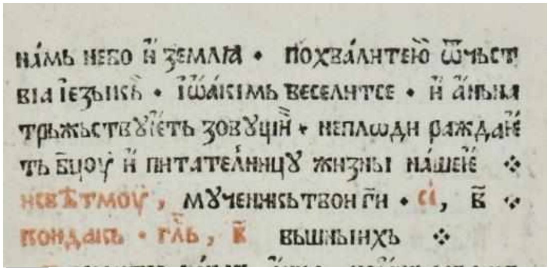


2-1 валецꙗ̀ се соуѥ́тною̀ · и да̀
2-2 жь съкроу́шен҆нѣи̇ дш҃и и ср–
2-3 цоу̑зовоу̑щоу тѣ́бѣ · ѡ̑цѣ–
2-4 сти ѥди́не чл҃колю́бче бе, и
2-5 пода́жь ми ѻ̑ста́в҆ленїе ⁘
2-6 Ѥди́нь сы̑ бл҃гопрѣмѣни́те–
2-7 ль ⁘ пѣз̃ ір҃ѧ Ро́сода́–
2-8 вницоу Ѹ҆бопе́щьсь тво́ри ӑ̑г
2-9 г҃ль, ст҃ыимъ дѣ́темь · ха́л҆–

Figure 3: The Automatically Read Text of a Part of Sheet 3a *Lenten Triodon* from 1561.

Errors in accent mark recognition: instead of валецꙗ 1, соуѥтною 2, съкроушенн҃ѣи 2, ср҃цоу 2/3, тебѣ 3, ѡ̑цѣ– 3, Ѻставлѥнїе 5, пѐщь сьтвори 8, хал– 9/10 the model incorrectly outputs ва́лецꙗ̀ 1, соуѥ́тною̀ 1, съкроу́шен҆нѣи̇ 2, срцоу̑ 2/3, тѣ́бѣ 3, ѡ̑цѣ– 3, Ѻ̑ста́в҆лѥнїе 5, пе́щьсь тво́ри 8, ха́л҆– 9/10. Errors in recognizing spaces between words are also of high frequency: instead of да– 1, цоу зовоу̑щоу 3, пѣ з҃ 7, Ѹ҆бо пѐщь сьтвори 8, ӑ̑г– 8, ст҃ыимъ дѣтемь 9 the model incorrectly outputs да́ 1, цоу̑зовоу̑щоу 3, пѣз̃ 7, Ѹ҆бопе́щьсь тво́ри 8, ӑ̑г 8, ст҃ыимъдѣ́темь 9. In a fewer number of examples, errors in recognizing *pajerak mark*, superscript letters and *titlo mark* can be found: instead of съкроушеннѣи 2, Ѻставлѥнїе 5, хал– 9/10, ср҃– 2, пѣ 7 the model incorrectly outputs съкроу́шен҆нѣи̇ 2, Ѻ̑ста́в҆лѥнїе 5, ха́л҆– 9/10, ср– 2, пѣ 7.

A comparative presentation of a part of sheet 7a *Prayer Book (Miscellany for Travellers)* from 1566 and the automatically read text using the *Dionisio 1.0.* model displays similar errors.



1-1 на́мь не бо̀ и̏ землѧ̀ · похва́литею̀ Ѿчьст–
1-2 вїа і̑е́зы́кь̏ · і̑ѡ̑а́кїмь весе́лит се · и̏ ӑ̑ньна
1-3 трь́жьствѹ̑ѥ́ть зовѹ̑щи · непло́ди раждаѥ̀
1-4 ть бцоу̑ и̏ пи́тател҆ницѹ̑ жи́зны на́шеѥ ⁘
1-5 и̏ свѣтмоу̑, мѹ̑ченикь твои г҃и · сь, в҃ ⁘
1-6 воѥ́дакь · гл҃ь, в҃ вышныихь ⁘

Figure 4: The Automatically Read Text of a Part of Sheet 7a *Prayer Book (Miscellany for Travellers)* from 1566.

Errors in recognizing accent: instead of небо 1, землѧ 1, похвалите ю̑ 1, Ѿчьствїа 1/2, і̑езыкь̏ 2, весе́лит 2, трьжьствѹ̑ѥть 3, неплоди 3, раждаѥ– 3, питател҆ницѹ̑ 4, жизны 4, на́шеѥ 4, и 5, мѹ̑ченикь 5, кондакь 6 the model incorrectly outputs не бо̀ 1, землѧ̀ 1, похва́литею̀ 1, Ѿчь́ствїа 1/2, і̑е́зы́кь̏ 2, весе́лит 2, трь́жьствѹ̑ѥ́ть 3, непло́ди 3, раждаѥ̀ 3, пи́тател҆ницѹ̑ 4, жи́зны 4, на́шеѥ 4, и̏ 5, мѹ̑че́никь 5, воѥ́дакь 6. A certain number of errors is connected to recognizing spaces between words: instead of небо 1, похвалите ю̑ 1, раждаѥ– 3, свѣт моу 5 the model incorrectly outputs не бо̀ 1, похва́литею̀ 1, раждаѥ̀ 3, свѣтмоу̑ 5. Several errors in recognizing letters may perhaps be related to poor quality of the photograph: instead of сі 5, кондакь 6 the model incorrectly outputs сь 5, воѥ́дакь 6.

The illustrated examples of the most frequent errors in *Lenten Triodon* (1561) and *Prayer Book (Miscellany for Travellers)* (1566) show that the *Dionisio 1.0.* model can be used for obtaining transcripts that can, after appropriate manual correction, be used for creating specific models for automatic text recognition of the aforementioned two books.

3. Applying the *Dionisio 1.0.* Model on Books from Other Serbian Printing Houses of the 15th and 16th Centuries

In the second experiment, the performance of the *Dionisio 1.0.* model was tested on selected books from other printing houses of the 15th and 16th centuries (Cetinje, Goražde, Gračanica, Mileševa, Belgrade and Mrkša'a Church). During the research, we started from the hypothesis that the model trained on the material of books from the Venetian printing house Vuković will be useful for books from other printing houses, since there are not many orthographic variations in Serbian Early Printed Books as there are in medieval manuscripts.

The results of the experiment are shown in the following table.

Book **(Printed House, Year)**	**CER**
Octoechos, mode 1–4 (Cetinje, 1495)	8.24%
Psalter with Appendices (Goražde, 1519)	6.44%
Octoechos, mode 5–8 (Gračanica, 1539)	11.11%
Prayer Book (Euchologion) (Mileševa, 1546)	5.43%
Tetraevangelion (Belgrade, 1552)	11,28%
Tetraevangelion (Mrkša's Church, 1562)	12.06%

Table 4: Application of the *Dionisio 1.0.* model on publications from other printing houses in the 15th and 16th centuries.

Based on the previous table, it can be concluded that the *Dionisio 1.0.* model achieved the best results in the automatic recognition of the text of *Prayer Book (Euchologion)* (1546) from the printing house of the Mileševa monastery and *Psalter with Appendices* (1521) from the Goražde printing house. These results can be explained by the fact that *Prayer Book (Euchologion)* (1546) had been printed in Mileševa with the same typographic characters as *Psalter with Appendices* (1521) from Božidar Vuković's printing house, as well as by the fact that *Psalter with Appendices* (1519) was printed in Goražde using the typographic equipment imported from Venice (Lazić, 2020a).[8]

To illustrate the efficiency of the *Dionisio 1.0.* model we may firstly use the comparative presentation of the photograph of a part of sheet 5b *Prayer Book (Euchologion)* (1546) from the printing house of the Mileševa monastery and the automatically read text in Figure 5.

In this book, as well, the greatest number of errors refers to accent marks recognition: instead of имєни 4, истинын 4, ѥдинородьнааго 4/5, ст҃аго 5, и ме 7, сподобльшаго 7, ѽноу 10 the *Dionisio 1.0.* incorrectly outputs и҆мє́нѝ 4, и҆сти́ный 4, ѥ҆ди́норо́днаа҄го 4/5, ст҃а́го 5, и҆ м͡е́ 7, сподо́бльшаго 7, ѽноу 10. Other errors are fewer in number and relate to recogizing initials, spaces between words and *pajerak mark*: instead of ѿ и҆мє́ни 4, подь 9 и дрѣвнꙋ҆ю 10 the model incorrectly reads и҆мє́нѝ 4, по дь 9 и дрѣв͑нꙋ҆ю 10.

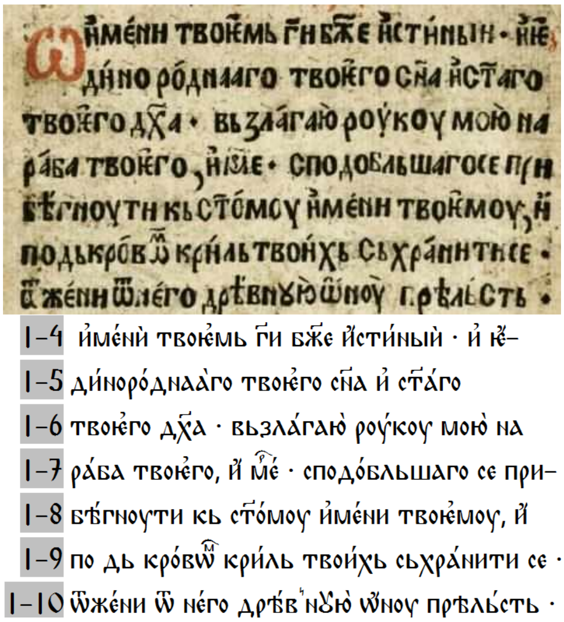


Figure 5: The Automatically Read Text of a Part of Sheet 5b *Prayer Book (Euchologion)* from 1546.

Similar errors are indicated by the comparative illustration of the photograph of a part of sheet 35a *Psalter with Appendices* (1519) from the Goražde printing house and the automatically read text using the *Dionisio 1.0.* model.

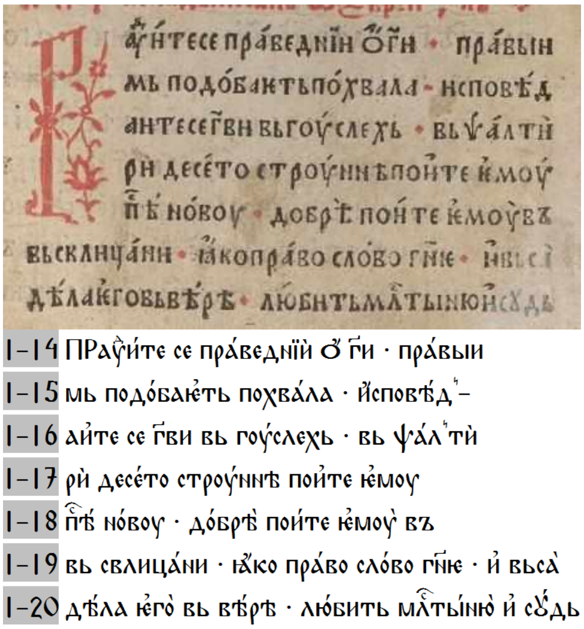


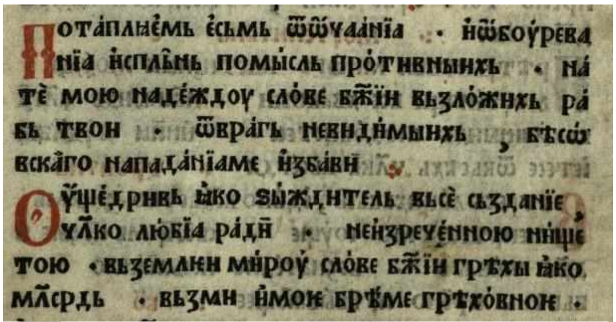
Figure 6: The Automatically Read Text of a Part of Sheet 35a *Psalter with Appendices* from 1519.

[8] Scholars likewise claim that *Psalter with Appendices* (1519) and *Prayer Book (Euchologion)* (1544) from Goražde printing house could have been printed in Venice, as well, which corresponds to the widespread practice of the time to place a counterfeit place of printing on the colophonies of editions (Lazić, 2020a).

The previous illustration demonstrates how the *Dionisio 1.0.* model makes the most frequent errors while recognizing accent marks: instead of пра́ведні́и 14, подо́баеть 15, по́хвала 15, исповѣ́даите се 15/16, ѱа́лті́ри 16/17, ѥ́моу 17, добрѣ́ 18, ѥ́го 19, ма́тыню 19, сѫ́дь 19 the model incorrectly outputs пра́веднїй 14, подо́баѥ́ть 15, похва́ла 15, и́сповѣ́да҆ите се 15/16, ѱа́л҆ті́ри 16/17, ѥ́моу 17, до́брѣ 18, ѥ́го́ 19, ма́тыню́ 19, сѫ́дь 19. The other errors pertain to recognizing spaces between words, *pajerak mark* and initials: instead of пра́вьи- 14, ѱа́лти- 16, десе́тостроу́ннѣ 17 the model incorrectly reads: пра́вьи 14, ѱа́л҆ти 16, десе́то строу́ннѣ 17; instead of исповѣ́даите се 15/16, ѱа́лти- 16 there is the incorrect и́сповѣ́д҆аи́те се 15/16, ѱа́л҆ти 16; instead of Ра́ꙋ́ите се 14 there is the incorrect ПРа́ꙋ́ите се 14. There is merely one example of an incorrectly recognized letter: instead of въсклица́ни 19 the model incorrectly reads вь свлица́ни 19.

The *Dionisio 1.0.* model also shows a similar performance during the automatic recognition of the text of the oldest printed Serbian Church Slavonic book – *Octoechos, mode 1–4* (1495) from the Cetinje printing house. The percentage of unrecognized characters is somewhat higher than in the previous two books due to poor photo quality and issues with recognizing certain letters and punctuation marks.

To illustrate the efficiency of the model, we will use a comparative presentation of a part of sheet 33b and the automatically read text in the following figure.

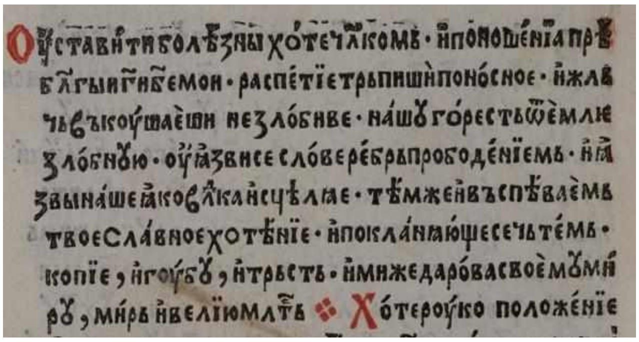


1-8 Пота́плꙗѥ́мь ѥ́сьмь ѿ ѿчаа́нїа · и ѿбоу́рева
1-9 нїа и́спльнь помы́сль про́тивнꙑихь · на
1-10 тѐ моѭ на де́ждоу сло́ве бж҃їи вьзло́жихь ра́-
1-11 вь твои · ѿ вра́гь невиди́мꙑихь · ⁘ вѣсѡ́
1-12 в҆ска́го напада́нїа ме и́зба́ви
1-13 Оу́щедривь ꙗ́ко быжди́тель вьсѐ сьзда́нїе ·
1-14 чл҃колю́бїа ра́ди · неи́зреченною ни́щѐ
1-15 тоѭ · вь зе́млѥи ми́роу сло́ве бж҃їи грѣхы ꙗ́ко
1-16 мл҃содь · вь́зми и́ моѥ́ вре́ме грѣхо́вноѥ ·

Figure 7: The Automatically Read Text of a Part of Sheet 33b *Octoechos, mode 1–4* from 1495.

In this book, too, the largest number of errors in the automatic text recognition occurs with accent marks: instead of ѥ́сьмь 8, и́спльнь 9, на́ 9, моѭ 10, твои 11, вѣсѡ́вска́го 11/12, и́зба́ви 12, ꙗ́ко 13, сьзда́нїе 13, неи́зрече́нною 14, ни́щетою 14/15, зе́млѥи 15, ꙗ́ко 15, вьзми 16, и́ 16 the *Dionisio 1.0.* model incorrectly reads: ѥ́сь́мь 8, и́спльнь 9, на 9, моѭ́ 10, твои́ 11, вѣсѡ́ в҆ска́го 11/12, и́зба́ви 12, ꙗко 13, сьзда́нїе 13, неи́зрече́нною 14, ни́щѐтоѭ́ 14/15, зе́млѥи 15, ꙗ́ко 15, вь́зми 16, и́ 16. The issues with recognizing spaces between words and *pajerak mark* can be illustrated by the following examples: instead of ѿбоу́рева- 8, наде́ждоу 10, вѣсѡ́- 11 there is the incorrect ѿбоу́рева 8, на де́ждоу 10, вѣсѡ́ 11; instead of вѣсѡ́вска́го 11/12 there is the incorrect вѣсѡ́ в҆ска́го 11/12. In this book, as we have already mentioned, the *Dionisio 1.0.* model likewise incorrectly recognizes certain letters and punctuation marks: instead of ѿ 8, ѕыжди́тель 13, мл҃срдь 16 there is the incorrect ѿ 8, бы́ждитель 13, мл҃содь 16; instead of невиди́мꙑихь, 11, и́зба́ви :·12 there is the incorrect невиди́мꙑихь · ⁘ 11 и҆зба́ви :·12.

In the rest of the books listed in Table 4, (*Octoechos, mode 5–8* (1539) from Gračanica, *Tetraevangelion* (1552) from Belgrade and *Tetraevangelion* (1562) from Mrkša's Church), the CER is slightly higher, around 11–12%. The categories in which the *Dionisio 1.0.* model outputs errors are mostly the same in all three books, so we will only take a comparative presentation of a part of sheet 27b *Octoechos, mode 5–8* (1539) from Gračanica and the automatically read text as an illustration.



1-1 Оу́ста́вити бо́лѣзны҆ хо́те чл҃комь · и́ попоше́нїа прѣ-
1-2 бл҃гыи г҃и б҃е мои · распе́тїе трь́пиши поно́сноѐ · и́ жл҃в
1-3 чьвькоу́шаѥ́ши незло́биве · на́шꙋ го́ресть ѿе́млѥ
1-4 зло́бнꙋѭ · оу́ꙗ́зви се сло́ве ре́брь прободе́нїемь · и ꙗ-
1-5 звы на́ше ꙗко вл҃ка и́сцѣлꙗѥ̀ · тѣ́м҆же и́ вьспѣ́ваѥ́мь
1-6 твоѐ сла́вноѐхотѣ́нїе · и́ покла́нꙗюще се чь те́мь ·
1-7 копїе, и́ гоу́бꙋ, и́ трьсть · и́ ми́же даро́ва своѐ мл҃ми-
1-8 рꙋ, мирь и́ ве́лїю мл҃ть ⁘ Хо́те роу́ко положе́нїе

Figure 8: The Automatically Read Text of a Part of Sheet 27b *Octoechos, mode 5–8* from 1539.

The greatest number of errors is related to the recognition of accent marks: instead of бо́лѣзны 1, и́ 1, 2, 5, 6, 8, мои 2, трьпиши́ 2, поно́сноѐ 2, чь вькоу́шаѥ́ши 3, ѿе́млѥ 3, прободе́нїемь 4, ꙗ́звы 4/5, ꙗ́ко 5, и́сцѣлꙗѥ 5, вьспѣ́ваѥ́мь 5, твоѥ 6, сла́вноѥ хотѣ́нїе 6, покла́нꙗю́ще се 6, и́миже 7, своѥ́мꙋ ми- 7, ве́лїю 8 the *Dionisio 1.0.* model incorrectly outputs бо́лѣзны҆ 1, и́ 1, 2, 5, 6, 8, мои́ 2, трь́пиши́ 2, поно́сное 2, чьвькоу́шаѥ́ши 3, ѿе́млѥ̀ 3, прободе́нїемь 4, ꙗ́звы 4/5, ꙗко

5, їсцѣлꙗѣ 5, вьспѣваѣмь 5, твоѣ 6, славноѣхотѣнїе 6, поклањающе се 6, ѝ миже 7, своѣ мѣми- 7, вєлїю 8. Recognizing spaces between words represents the problematic issue in a multitude of cases: instead of жль- 2, ъ вькоушаѣши 3, славное хотѣнїе 6, ътємь 6, ѝмиже 7, своѣмѣ ми- 7, роукоположєнїе 8 the model incorrectly outputs жл҃в 2, ъвькоушаѣши 3, славноѣхотѣнїе 6, ъ тємь 6, ѝ миже 7, своѣ мѣми- 7, роуко положєнїе 8. The other errors pertain to the recognition of superscript letters and *pajerak mark*, as well as regular letters in a few examples: instead of жль- 2, мр̂ть 8 the model outputs жл҃в 2, мл̂ть 8; instead of тѣмже 5 there is the incorrect тѣм᷾же 5; instead of поношєнїа 1, жль- 2, ѿѣмлю 3 the model reads попошєнїа 1, жл҃в 2, ѡ̑ѣмлю̑ 3.

The quantitative and qualitative analysis conducted in this chapter demonstrates that the *Dionisio 1.0.* recognizes the text of the Serbian Church Slavonic books created in other printing houses of the 15th and 16th centuries with varying degrees of success. The quantitative analysis shows that the lowest CER was recorded in books from Mileševa and Goražde printing houses, which is expected considering the fact that these books were printed using the typographic printing equipment from Venice. An acceptable CER was noted during the recognition of *Octoechos, mode 1–4* (1494) from the Cetinje printing house, while this percentage exhibited in books from other printing houses (Belgrade, Gračanica, Mrkša's Church) underscores the need for training a new version of the generic model with improved performance. The qualitative analysis showed that the *Dionisio 1.0.* model usually makes errors when recognizing accent marks, but also when recognizing spaces between words. The errors in recognizing superscript letters, *pajerak mark*, initials and regular letters are far less common.

4. Creation and evaluation of the generic model *Dionisio 2.0.*

When creating a new version of the model, we started from the transcripts of Serbian Church Slavonic books listed in Table 4 obtained using the *Dionisio 1.0.* model. By means of the manual correction of the transcripts, the Ground Truth[9] data was obtained for training the generic model *Dionisio 2.0.* In accordance with our findings on the interdependence of model success and the amount of training data (Polomac, 2022), as well as similar findings for Church Slavonic books from the Berlin State Library (Neumann, 2021), the goal was set to provide a critical mass of at least 10000 words for each printed book in order to train the generic model *Dionisio 2.0.* While training the generic model *Dionisio 2.0.* we used the Ground Truth data prepared for the *Dionisio 1.0.* model (see Table 1 here), as well as the new Ground Truth data from Serbian Church Slavonic books printed in other printing houses of the 15th and 16th centuries listed in the following table.

***Book* (Printed House, Year)**	**Word count**
Octoechos, mode 1–4 (Cetinje, 1495)	15,667
Psalter with Appendices (Goražde, 1519)	16,445
Octoechos, *mode 5–8* (Gračanica, 1539)	15,179
Prayer Book (Euchologion) (Mileševa, 1546)	15,003
Tetraevangelion (Belgrade, 1552)	15,333
Tetraevangelion (Mrkša's Church, 1562)	15,733

Table 5: The *Dionisio 2.0.* model – Ground Truth data from other printing houses of the 15th and 16th centuries.

The performance of the generic model *Dionisio 2.0.* is shown in the following table.

Word count	**Number of epochs**	**CER on Train set**	**CER on Validation set**
176,481	200	2.03%	2.44%

Table 6: Performance of the generic model *Dionisio 2.0.*

In order to compare the performance of the two models, we tested them on ten sheets from *Psalter with Appendices* (1495) from the Cetinje printing house and *Hieraticon* (1521) from the Goražde printing house, the latter two representing Serbian Church Slavonic books that did not form the material for training the model. The results of the experiments are shown in the following table.

***Book* (Printed House, Year)**	***Dionisio 1.0.* CER**	***Dionisio 2.0.* CER**
Psalter with Appendices (Cetinje, 1495)	5.71%	1.50%
Hieraticon (Goražde, 1521)	9.38%	4.61%

Table 7: Comparing the Performance of the Two Models on Books from Cetinje and Goražde Printing Houses.

As can clearly be seen from the previous table, the *Dionisio 2.0.* model displays significantly better results compared to the *Dionisio 1.0.* model. To illustrate the exceptional efficiency of the *Dionisio 2.0.* model we provide a comparative presentation of a part of sheet 3b *Psalter with Appendices* (1495) from Cetinje printing house and the automatically read text in the figure 9.

As we can see in the figure, the *Dionisio 2.0.* model erros only in a few examples in which the *spiritus lenis* and *perispomena* are insufficiently clearly differentiated: instead of ѭнїи 8, поіѽть 9, ѝстиноѭ 10, накаꙁоуіѽть 11 the model incorrectly outputs ѭнїи 8, поіѽть 9, ѝстиноѭ 10, накаꙁоуіѽть 11. There is a single example of the model mixing *spiritus lenis* and *oxia*: instead of ѿи́де 13 there is the incorrect ѿи́де 13. The space between words was also

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

incorrect in one example solely: instead of мнѣти 9 there is the incorrect мнѣ ти 9. In the other examples on the shown part of sheet 3b the *Dionisio 2.0.* model regularly recognizes letters, spaces between words, *titlo* and accent marks. The exceptional efficiency of the *Dionisio 2.0.* model in recognizing *Psalter with Appendices* (1495) from the Cetinje printing house, especially compared to *Hieraticon* (1521) from the Goražde printing house, has resulted from the fact that there are no superscipt letters in *Psalter with Appendices* (1495), while accent marks are given in expected positions.

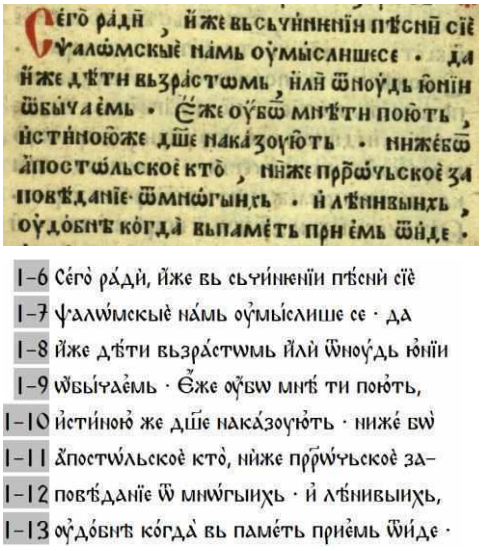


Figure 9: The Automatically Read Text of a Part of Sheet 3b *Psalter with Appendices* (1495).

On the other hand, superscript letters, as well as accent marks, found frequently in unexpected positions, are present in *Hieraticon* (1521) from the Goražde printing house, which definitely affects a somewhat less efficient CER in this book. To illustrate the aforementioned, we shall use the comparative presentation of a part of sheet 9b and the automatically read text in the following figure.

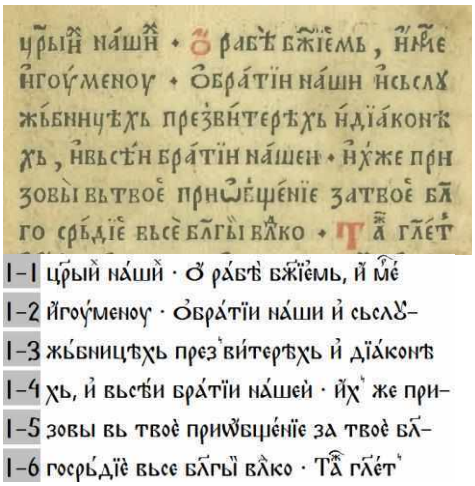


Figure 10: The Automatically Read Text of a Part of Sheet 9b *Hieraticon* (1521).

The previous illustration points to the fact that the *Dionisio 2.0.* model makes errors almost exclusively during accent marks recognition. Thus, instead of рабѣ 1, бжїѣмь 1, мѣ 1, йгоу́меноу 2, и́ 3, на́шеи 4, йх҃ 4, призовы 4/5, твоѥ̀ 5x2, приѡбще́нїе 5, бл҃госрьдїѥ̀ 5/6, вьсѣбл҃гы̀ 6 the model incorrectly reads ра́бѣ 1, бжїѣмь 1, мѣ́ 1, йгоу́меноу 2, и́ 3, на́шеѝ 4, йх҃ 4, призовы 4/5, твоѥ̀ 5x2, приѡбще́нїе 5, бл҃госрьдїѥ̀ 5/6, вьсе бл҃гы̀ 6. Along with the aforementioned errors, there are a few examples of incorrect recognition of spaces between words: instead of о̀ бра́тїи 2, сь слѫ-2, дїа́конѣ- 3 вьсѣбл҃гы̀ 6 the model reads о̀бра́тїи 2, сьслѫ- 2, дїа́конѣ 3 вьсе бл҃гы̀ 6.

5. Concluding Remarks

The research showed how the *Transkribus* software platform, based on the principles of machine learning and artificial intelligence, could be used to create efficient models for automatic text recognition of Serbian Church Slavonic printed books from the end of the 15th to the middle of the 17th century. Having in mind the limitations of the *Dionisio 1.0.* model in the automatic recognition of the text of the Serbian Church Slavonic books printed outside Venice, the paper describes the process of creating a generic model *Dionisio 2.0.*, capable of recognizing Serbian Church Slavonic printed books as a whole. The generic model *Dionisio 2.0.* was trained on the material of the Serbian Church Slavonic books printed in various Serbian printing houses of the 15th and 16th centuries: Cetinje, Venice, Goražde, Gračanica, Mileševa, Belgrade and Mrkša's Church. The quantitative analysis of the performance of this model showed that it could be used to automatically obtain transcripts with a minimum percentage of incorrectly recognized characters (about 2-3%). Most frequently, CER depends on the quality of the photo of the book, the frequency of use of accent marks and superscripts, as well as the correct use of accent marks in the appropriate positions. Using the *Dionisio 2.0.* model transcripts of Serbian Church Slavonic printed books can be obtained automatically, which, after being edited by a competent philologist, can be used for further philological and linguistic research, primarily for creating searchable digital editions of books, as well as electronic corpora, thus creating opportunities for diachronic research of Serbian early modern literacy on a large quantity of data. In the near future, the generic model *Dionisio 2.0.* will become publicly available to all users of the *Transkribus* software platform, which will enable further improvement of its performance, which could ultimately lead to the creation of a generic model for automatic text recognition of Church Slavonic printed books as a whole.

6. Acknowledgment

The research conducted in the paper was financed by the Ministry of Education, Science and Technological Development of the Republic of Serbia, contract no. 451-03-68/2022-14/ 200198, as well as by the German Academic Exchange Service (DAAD) within the project *Automatic Text Recognition of Serbian Medieval*

Manuscripts and Early Printed Books: Problems and Perspectives.

7. References

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

Miroslav Lazić. 2018. Od Božidara Vukovića do Dionizija dela Vekije: identitet i pseudonim u kulturi ranog modernog doba. In: Anatolij A. Turilov et al., eds., *Scala Paradisi*, pages 165–185, SANU, Beograd.

Miroslav Lazić. 2020a. Inkunabule i paleotipi: srpskoslovenske štampane knjige od kraja 15. do sredine 17. veka. In: Vladislav Puzović and Vladan Tatalović, eds., *Osam vekova autokefalije Srpske pravoslavne crkve*, Vol. 2, pages 325–344. Sveti arhijerejski sinod Srpske pravoslavne crkve–Pravoslavni bogoslovski fakultet, Beograd.

Miroslav Lazić. 2020b. Between an Imaginary and Historical Figure: Božidar Vuković´s Professional Identity. *Ricerche Slavistiche*, 43:141–156.

Vladimir Neumann, 2021. Deep Mining of the Collection of Old Prints *Kirchenslavica Digital*. *Scripta & e-Scripta* 21: 207–216.

Vladimir Polomac. 2022. Serbian Early Printed Books from Venice. Creating Models for Automatic Text Recognition using *Transkribus*. *Scripta&e-Scripta*, 22 [in print].

Günther Mühlberger, 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. 2019. Transforming scholarship in the archives through handwrittwn text recognition. *Journal of Documentation*, 5 (75):954–976.

Mitar Pešikan. 1994. Leksikon srpskoslovenskog štamparstva. In: Mitar Pešikan et al., eds., *Pet vekova srpskog štamparstva 1494–1994: razdoblje srpskoslovenske štampe XV–XVII*, pages 71–218, Narodna biblioteka Srbije–Matica srpska, Beograd.

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

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