Digitizing Cyrillic Manuscripts for the Historical Dictionary of the Serbian Language Using Handwritten Text Recognition Technology*

Оцифровка кириллических рукописей для исторического словаря сербского языка с использованием технологии распознавания рукописного текста

Vladimir Polomac
University of Kragujevac,
Kragujevac, Serbia

Marina Kurešević
Isidora Bjelaković
Aleksandra Colić Jovanović
Sanja Petrović
University of Novi Sad,
Novi Sad, Serbia

Владимир Поломац
Университет в Крагуеваце,
Сербия

Марина Курешевич
Исидора Бьелакович
Александра Цолич Йованович
Саня Петрович
Университет в Нови-Саде,
Нови-Сад, Сербия

Цитирование: *Поломац В., Курешевич М., Бьелакович И., Цолич Йованович А., Петрович С.* Оцифровка кириллических рукописей для исторического словаря сербского языка с использованием технологии распознавания рукописного текста // Slověne. 2023. Vol. 12, № 1. С. 295–316.
Citation: Polomac V., Kurešević M., Bjelaković I., Colić Jovanović A., Petrović S. (2023) Digitizing Cyrillic Manuscripts for the Historical Dictionary of the Serbian Language Using Handwritten Text Recognition Technology. *Slověne*, Vol. 12, № 1, p. 295–316.
DOI: 10.31168/2305-6754.2023.1.08

Abstract

The paper explores the possibilities of using information technologies based on the principles of machine learning and artificial intelligence in the process of digitizing Cyrillic manuscripts for the purposes of creating a historical dictionary of the Serbian language. Empirical research is based on the use of the *Transkribus* software platform in the creation of a model for automatic text recognition of the manuscripts by Gavril Stefanović Venclović, the most significant and prolific Serbian cultural enthusiast of the 18th century, whose extensive manuscript legacy in Serbian vernacular represents the most significant primary source for the historical dictionary of the Serbian language of this period. Following the results of conducted research, it can be concluded that the process of digitizing Cyrillic manuscripts for the purposes of creating a historical dictionary of the Serbian language can be significantly accelerated using Transkribus by creating specific and generic models for automatic text recognition. The advantage of automatic text recognition compared to the traditional methods is particularly reflected in the possibility of continuous improvement of the performance of specific and generic models in accordance with the progress of the transcription process and the increase in the amount of digitized text that can be used to train a new version of the model.

Keywords

Transkribus, automatic text recognition, artificial intelligence, machine learning, historical lexicography, serbian language, Gavril Stefanović Venclović

Резюме

В статье исследуются возможности использования информационных технологий, основанных на принципах машинного обучения и искусственного интеллекта, в процессе оцифровки кириллических рукописей в целях создания исторического словаря сербского языка. Эмпирическое исследование основано на использовании программной платформы *Transkribus* при создании модели автоматического распознавания текста рукописей Гаврила Стефановича Венцловича, самого значительного и плодовитого сербского культурного энтузиаста XVIII в., чье обширное рукописное наследие в сербском народном языке представляет собой наиболее значительный первоисточник исторического словаря сербского языка, относящегося к этому периоду. По результатам проведенного исследования можно сделать вывод, что процесс оцифровки кириллических рукописей в целях создания

* The paper was financed by the Ministry of Education, Science and Technological Development of the Republic of Serbia and German Academic Exchange Service (DAAD) (*project: Automatic Text Recognition of Serbian Medieval Manuscripts and Early Printed Books: Problems and Perspectives*). The previous version entitled *Serbian Written Heritage of the 18th century: Towards Automatic Text Recognition of Gavril Stefanović Venclović's Manuscripts* was presented at the 17th annual conference of the Slavic Linguistics Society (19–21st September 2022, Hokkaido University, Sapporo, Japan). The team of authors would like to express its gratitude to Academician Vasilije Krestić (manager) and Dr. Miroslav Jovanović (vice manager) of the SASA (Serbian Academy of Sciences and Arts) Archives for providing the digital copies of Venclović's manuscripts used in this paper.

исторического словаря сербского языка можно значительно ускорить с помощью Transkribus через создание определенных и генерических моделей для автоматического распознавания текста. Преимущество автоматического распознавания текста по сравнению с традиционным, в частности, выражается в возможности постоянного улучшения производительности определенных и генерических моделей в соответствии с ходом процесса транскрипции и увеличением объема оцифрованного текста, который можно использовать для обучения новой версии модели.

Ключевые слова

Transkribus, автоматическое распознавание текста, искусственный интеллект, машинное обучение, историческая лексикография, сербский язык, Гаврил Стефанович Венцлович

1. Introduction

Recent research on the use of the *Transkribus* software platform[1] for the automatic recognition of Russian and Serbian Church Slavonic Cyrillic manuscripts and printed books [Rabus 2019; Polomac, Lutovac Kaznovac 2021; Polomac 2022a; 2022b] triggered the investigation to be performed in the present article. In his pioneering study, German slavist A. Rabus [2019a] demonstrated that the first version of the automatically recognized text can be digitized with only 3–4% of misrecognized characters, using this particular software platform. Furthermore, this could be done in a significantly shorter amount of time, simultaneously reducing human and financial resources. The obtained output might later be used for further philological and linguistic research, especially after a manual correction (editing) by a competent philologist. The present paper further contributes by underlining that the models for automatic recognition have been made available to all *Transkribus* users; hence, its performance can be checked on other Slavic medieval manuscripts. A paper by V. Polomac and T. Lutovac Kaznovac [2021] examined the performance of Rabus's models for automatic recognition of Serbian medieval manuscripts written in different types of Cyrillic script. The authors concluded that the application of Rabus's

[1] The *Transkribus* software platform (https://readcoop.eu/transkribus/) represents a tool for manual and automatic reading and searching of old manuscripts and printed books, regardless of the time of creation, language or script. The key advantage of *Transkribus* compared to other similar applications is reflected in the ability of the user to create his/her own model for automatic text reading. Training a model for automatic text reading is an example of machine learning based on advanced neural networks in which the model compares photographs of manuscripts and the corresponding letters, words and lines of text in the diplomatic edition. For more information on the technological background and the way this platform works, see [Mühlberger et al. 2019; Rabus 2019a].

models yielded relatively favorable results on Serbian medieval manuscripts written in *poluustav* ('semi-majuscule Cyrillic script'), while the creation of specific models was suggested for manuscripts written in *brzopis* ('diplomatic minuscule Cyrillic script'). The current paper underscores the necessity of creating specific models for the recognition of Serbian medieval manuscripts and printed books in particular to speed up the work on current projects in historical corpus linguistics and lexicography of the Serbian language. It was precisely the creation of a model for the automatic recognition of Serbian Church Slavonic printed books that was the focus of the studies by V. Polomac [2022a; 2022b]. The most important result of the aforementioned studies relates to the creation of a publicly available generic model for the automatic recognition of Serbian Church Slavonic printed books of the 15th and 16th centuries, entitled *Dionisio 2.0*. In continuation of the research, the authors of the current study were interested in whether the *Transkribus* software platform can be used for the automatic recognition of the Serbian manuscript heritage of the 18th century, as well as to speed up the preparation of the electronic corpus for the historical dictionary of Serbian. Empirical research was conducted on the manuscripts by Gavrilo Stefanović Venclović, one of the most significant and prolific Serbian cultural forerunners of the 18th century, whose legacy, written in Serbian vernacular and the Serbian Church Slavonic language, includes more than 20 manuscripts, with around 10,000 pages in total.[2] The second chapter of the paper provides a more detailed presentation of conceptualization of the Serbian historical dictionary, especially referring to the principles of text digitization. After a brief review of Venclović's manuscripts written in Serbian vernacular, the third, and central chapter of the paper presents and discusses the results of the experiments on creating and evaluating models for the automatic recognition of the texts in question using the *Transkribus* software platform. The fourth, and final chapter, summarizes the results and perspectives for further research.

[2] All his writings represent autographs—in some of them he left his name in the preface, afterword or inscription, while others were attributed to him based on paleographic and orthographic analysis and, possibly, language or illumination. Venclović's Serbian Church Slavonic manuscript fund is somewhat more extensive (about 6,200 pages) than the one in Serbian vernacular (more details here in point 3.1), and it consists mainly of manuscripts for liturgical purposes. In them, Venclović appears primarily as a copyist and illuminator, and less often as an editor or translator [Павић 1972: 98]. The largest part of Serbian Church Slavonic manuscripts is preserved in the SASA Archives (see [Стојановић 1901: 19–21, 34–36, 38–39, 102–120]) and in the Szentendre Archives (see [Синдик et al. 1991: 107–117, 120–121]).

2. On the Historical Dictionary of the Serbian Language and Principles of Digitization

A project entitled the *Dictionary of the Serbian Language from the 12th to the 18th Century* was established in 2013[3] as part of the activities of the Department of Language and Literature of Matica srpska. Since the Serbian historical lexicography still, to a large extent, falls behind the lexicography of the Slavic world,[4] creation of such a dictionary represents one of the primary goals of Serbian diachronic investigations. Materials from the oldest preserved manuscripts in Serbian (end of the 12th century) until the beginning of the pre-standard phase in the development of the Serbian literary language (end of the 18th century) represent the corpus for the dictionary.[5] Simultaneously, the upper time limit refers to the period from which the oldest corpus for the *Dictionary of Serbo-Croatian Literary and Vernacular Language* [РСАНУ] dates, thus ensuring the continuity in the lexicographic processing of the corpora in Serbian. The corpus was divided into primary, secondary and tertiary for several reasons. Namely, Serbian vernacular functioned as a complementary and functionally marked member of a dichotomy during the period of diglossia (12th–18th century),[6] i.e. polyglossia of the 18th century.[7] Moreover, it is a well-known fact that the boundaries between the literary languages and vernacular are often not sharply delineated, yet blurred and fuzzy [Радовановић 2015; Курешевић 2016], which should likewise be taken into consideration. Thus, the primary corpus consists of texts written in Serbian vernacular, which will be lexicographically processed using total excerption. The secondary corpus is comprised of texts in which the presence of both Serbian vernacular and literary language/s was attested. These sources will be processed selectively: only the Serbian lexis not recorded in the primary sources[8] will be excerpted. Tertiary sources

[3] The project leader is academician Jasmina Grković-Major (a full member of SASA), and the project team gathers language historians (experts on the history of Serbian vernacular and literary language idioms) from the Republic of Serbia, the Republic of Srpska and the Republic of Montenegro.

[4] The only historical dictionary based on Serbian linguistic material is [Даничић 1863–1864], which contains not only Serbian but Serbian Church Slavonic corpora, as well.

[5] We should use this occasion to emphasize that there is an open possibility for Serbian words recorded in earlier sources to be processed lexicographically [Грковић-Мејџор 2021: 14].

[6] For more information about diglossia in Serbian medieval literacy see in [Грковић-Мејџор 2007: 443–459].

[7] For more information about polyossia in Serbian literacy in the 18th century see: [Суботић 2004].

[8] More on the secondary corpus of the Serbian historical dictionary cf.: [Цветковић Теофиловић 2021; Јовић 2021].

include editions to be taken into account only after we make sure the original or its transcription has not been preserved and can be regarded as a valuable historical and linguistic resource [Грковић-Мејџор 2021: 19]. In order to ensure reliability in the lexicographic processing of the material, the intention of the team of authors is to perform the excerption exclusively from the original, that is, from its digital copy. In addition to defining the theoretical concept of the dictionary [Грковић-Мејџор 2021], theoretical and methodological solutions for various issues of its microstructure have been proposed in the previous work done so far on the project (cf. [Савић, Милановић 2021; Курешевић 2021; Павловић 2021; Грковић-Мејџор, Бјелаковић 2021; Бјелаковић 2021]), as well as theoretical principles of digitization [Курешевић et al. 2021]. As part of the practical work on the project, comprehensive registers of corpora for the dictionary have been created thus far. Furthermore, digital copies of most sources have been acquired and manual digitization has already begun.

Since the materials for the historical dictionary of Serbian were written in different types of Cyrillic and Latin script, the basic principle of digitization was to standardize the paleographic variants of the letters according to the corresponding solutions in the Civil script. For these purposes, the *BeogradPro* font was used, since it contains all modern Cyrillic and Latin characters, as well as additional characters for specific old Cyrillic and Latin graphemes. The following principles were used for Cyrillic corpora digitization: 1) punctuation is transferred according to the original, 2) abbreviations are transferred without resolving, with a *titlo* mark and/or with the superscript letter written in the exponent where it belongs in the word structure, 3) types of letters are generalized, and those that have an orthographic function are retained (e.g. the letters е and є are retained, then о, ѻ and ѡ, and the letter *ђерv* is transferred with the letter *ћ*), 4) when it comes to the superscript characters, a *titlo* mark is transferred (as a mark for an abbreviated word or as a mark for the numerical value of a letter) and a *pajerak* mark in its original place, 5) ligature grapheme connections are resolved, and only traditional ligatures are retained: ѕ, щ, ꙗ, ѥ, ю, ѿ and ѽ, 6) phonetic clusters with proclitics and enclitics are separated, 7) the spelling of compounds, as well as certain adverbs and conjunctions created by grammaticalization is standardized in favor of using a hyphen in texts that are not consistent in their use, 8) the beginning of a line is marked with a vertical line next to which there is also a number of the row in the exponent, and the end of the sheet is marked with a double vertical line next to which the number of the sheet/page is written in the exponent (more details in [Курешевић et al. 2021]).

Bearing the scope of the chronological arc of the historical dictionary of Serbian (12th–18th century) in mind, as well as the volume of its corpus, the

process of digitization is currently the primary task in the realization of the project. How demanding this process is, especially in the context of limited human and financial resources, can be proved by the fact that from 2017 until today, the material of about 500,000 words (mostly from business and legal literacy and literary pieces) has been digitized, representing only an insignificant part of the entire corpus.[9] If this process continues in the traditional way and with this dynamic, it is more likely that it will take decades, rather than years, to complete. The inclusion of technology for automatic text recognition in the process of digitization could significantly improve and speed up the work on the creation of the historical dictionary. Choosing the *Transkribus* software platform for this endeavor is suitable for several reasons.[10] Not only is the software characterized by a fairly simple user interface, but also demanding computer tasks are performed on the server so that the user does not need special computer equipment. Additionally, starting from version *Transkribus 1.18.0.* this software platform has allowed training and recognition of textual tags (including text styles such as bold, italic, superscript, etc.) using the *Include Properties* option. The final reason is particularly important since, in accordance with the aforementioned principles of transferring material for the dictionary into electronic form, superscript letters and *titlo* marks are transferred by raising them to an exponent (Tag as a superscript).

3. Creating and Evualuating Models for Automatic Text Recognition of Venclović's Manuscripts in Serbian Vernacular

3.1. Reflecting upon Venclović's Legacy Written in Serbian Vernacular

Venclović's manuscripts in Serbian vernacular were selected for investigating the possibility of including the *Transkribus* software platform in the process of the historical dictionary corpus digitization, as they represent the most extensive (about 4,400 pages) and important primary source for a dictionary of 18th century language. The advantage of automatic digitization by means of artificial intelligence and machine learning compared to traditional manual digitizing is especially evident when working with voluminous manuscripts, such as these ones. Manual digitizing requires enormous human, temporal and financial resources. It is not surprising, therefore, that even though

[9] At this moment, it is not possible to provide even an approximate estimate of the size of the corpus by the number of words.

[10] *Transkribus* is not the only software platform for automatic text recognition. At the University of Paris (Université Paris Sciences et Lettres) an open-access software platform *eScriptorium* was developed within the project *Scripta-PSL* which is currently most widely used for automatic recognition of Hebrew, Syriac and Arabic manuscripts. More on the project and platform itself see the following link https://escripta.hypotheses.org/, as well as in [Kiesling et al. 2019].

Venclović's manuscripts were discovered in the second half of the 19th century,[11] they have not yet received a complete critical edition. In Serbian vernacular, Venclović composed texts directly addressed to Orthodox believers—sermons, letters, and lessons. The choice of language is explained in several places by the need for his presentation to be understandable (according to [Павић 1972: 120–121]). This part of Venclović's written legacy includes: 1) *Поученија и слова разлика* (САНУ 94 (271), 1732); 2) *Мач духовни I* (САНУ 92 (267), 1733/34); 3) *Мач духовни II* (САНУ 93 (268), 1733/34); 4) *Великопосник* (САНУ 97 (136), 1740/41); 5) *Слова изабрана* (САНУ 101 (137), 1743); 6) *Пентикости* (САНУ 98 (272), 1743); 7) *Житија, слова и поуке* (САНУ 84 (270), 1744/45); 8) *Поученије изабраноје I* (САНУ 99 (139), 1745); 9) *Поученије изабраноје II* (САНУ 100 (269), 1746).[12]

3.2. Creation and Quantitative Evaluation of the Model

The initial methodological problem was ascribed to the fact that we lacked high-quality digital copies of any of Venclović's manuscripts written in Serbian vernacular, or transcripts that could be used to train models for automatic text recognition. By the courtesy of the SASA Archives, digital copies of the first 100 pages of the manuscript of *Слова изабрана* (САНУ (101) 137) (hereinafter abbreviated as САНУ 137) were made available to us. The choice of this manuscript was motivated by the fact that it is one of Venclović's most voluminous manuscripts in Serbian vernacular (745 pages in total), with a very neat and uniform ductus throughout. The process of creating a model for automatic text recognition started with manual digitization of the first 35 pages of the manuscript in *Transkribus*. Consequently, we obtained the minimum amount of Ground Truth data[13] (about 15,000 words) necessary for training the model.[14] During the process, we adhered to the principles of digitizing the

[11] Venclović's manuscripts reached the SASA Archives in 1870 thanks to Gavrilo Vitković, who was engaged in collecting antiquities in southern and central Hungary [Синдик et al. 1991: 3].

[12] All the mentioned books were created in the parishes of Komárno and Győr. They were described for the first time in [Стојановић 1901: 42–51, 84–171]. Based on the analysis of watermarks, M. Grozdanović-Pajić [1992] offered more precise or slightly different dates of origin for many of them, which we present in this paper. Although the degree of Venclović's originality is also questionable here, since we are talking about adaptations/translations to a considerable extent [Павић 1972: 243–246; Трифуновић 2009: 68], these manuscripts represent a very important resource in the study of the history of the Serbian language [Ивић 2014: 112–113].

[13] 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/.

[14] The minimum amount of data necessary to train a model for manuscript recognition is about 15,000 words, while training a model for recognizing printed books requires much less data (about 5,000 words) [Mühlberger et al. 2019: 959].

Cyrillic manuscripts for the historical dictionary specified here in chapter 2, except in case of compounds, certain conjunctions and adverbs. The latter were always transferred as one word (without a hyphen), since Venclović's texts belong to the epoch when the process of grammaticalization of the words in question had already ended.

The parameters and performance of the first version of the model named *Venclović 0.1.* are shown in the following table.

Table 1. Parameters and performance of the *Venclović 0.1.* model

Engine[15]	Word count on Train Set	Word count on Validation Set	Number of epochs[16]	CER on Test Set	CER on Validation Set[17]
CITlab HTR+	15 806	717	50	0.57%	6.87%

In the continuation of the transcription process, we used the *Venclović 0.1.* model for automatic digitization of the next 35 pages of the САНУ 137 manuscript. After manual correction of the automatically obtained transcripts, we had twice as much Ground Truth data necessary for training the second version of the model at our disposal. The parameters and performance of the second version of the model entitled *Venclović 0.2.* are displayed in the following table.

Table 2. Parameters and performance of the *Venclović 0.2.* model

Engine	Word count on Train Set	Word count on Validation Set	Number of epochs	CER on Test Set	CER on Validation Set
CITlab HTR+	32039	1675	50	1.39%	4.87%

We digitized the remaining 30 pages using the *Venclović 0.2.* model. After manually correcting the transcripts, we trained the *Venclović 0.3.* model, the parameters and performance of which are presented in the following table.

[15] Users of the *Transkribus* software platform have two engines for model training and automatic text recognition at their disposal: *CITlab HTR+* and *PyLaia*. Training the model on the same material using different engines yields almost identical results, which was also shown in our research. The advantage of the *PyLaia* engine is reflected only in the fact that it allows certain changes in its structure, and is thus suitable for adaptation to the specific needs of users who are familiar with the IT aspects of machine learning. For more detailed information see *Transkribus Glossary* at https://readcoop.eu/glossary/htr-plus/ and https://readcoop.eu/glossary/pylaia/.

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

[17] In all the models trained to recognize Venclović's manuscripts described in this paper, the amount of data in the validation set was 5% of the total training set.

Table 3. Parameters and performance of the *Venclović 0.3.* model

Engine	Word count on Train Set	Word count on Validation Set	Number of epochs	CER on Test Set	CER on Validation Set
CITlab HTR+	46118	2421	50	2.04%	4.49%

The quantitative indicators of the models for the automatic recognition of Venclović's manuscripts can be rated as exceptional, since it was already in the second version of the model *Venclović 0.2.* that the percentage of incorrectly recognized characters fell below 5%.[18] In other words, this means that the model can be trained to automatically recognize the rest of the manuscript with 95% accuracy only on the basis of one tenth of the manuscript. The progress in the quantitative performance of the model is more pronounced between its first and second versions—cf. CER on Validation Set for model *Venclović 0.1.* and *Venclović 0.2.* The *Venclović 0.3.* model shows that each subsequent version of the model exhibits minimal improvement in quantitative performance with the new training material. Unfortunately, as we did not possess digital recordings of the rest of the manuscripts, we were not able to continue the process of automatic recognition and model enhancement. However, even based on this experiment, as well as on the experience with training models for automatic recognition of Serbian Church Slavonic printed books [Polomac 2022a; 2022b], we can assertively assume that further refinement of the model could lead to the percentage of misrecognized characters dropping even lower. Nevertheless, insisting on reducing the percentage of incorrectly recognized characters to an even lower percentage does not contribute much in the practical sense, since the text obtained by automatic recognition must be edited by a competent philologist anyhow[19].

All three versions of Venclović's manuscripts recognition model were trained in a fifty-epoch process. The dependency of the training results expressed by the percentage of incorrectly recognized characters and the number of epochs for training the model can be shown for each model using the learning curve. A typical learning curve can be seen in the example of the *Venclović 0.3.* model in Figure 1.

The learning curve demonstrates that, in the process of machine learning, the model achieves the most significant progress during the first few epochs

[18] According to [Mühlberger et al. 2019: 962] it can be considered exceptional if the percentage of incorrectly recognized characters during automatic manuscript recognition is less than 5%. In the case of printed books, this percentage can be lower and amount to about 1–2%. Cf. our results on the material of Serbian Church Slavonic printed books in [Polomac 2022a; 2022b].

[19] In the paper by J. Besters-Dilger and A. Rabus [2021] a very interesting thesis was presented stating that a large amount of material obtained by automatic text recognition and tagging can be used for quantitative linguistic research even without the manual correction of the text.

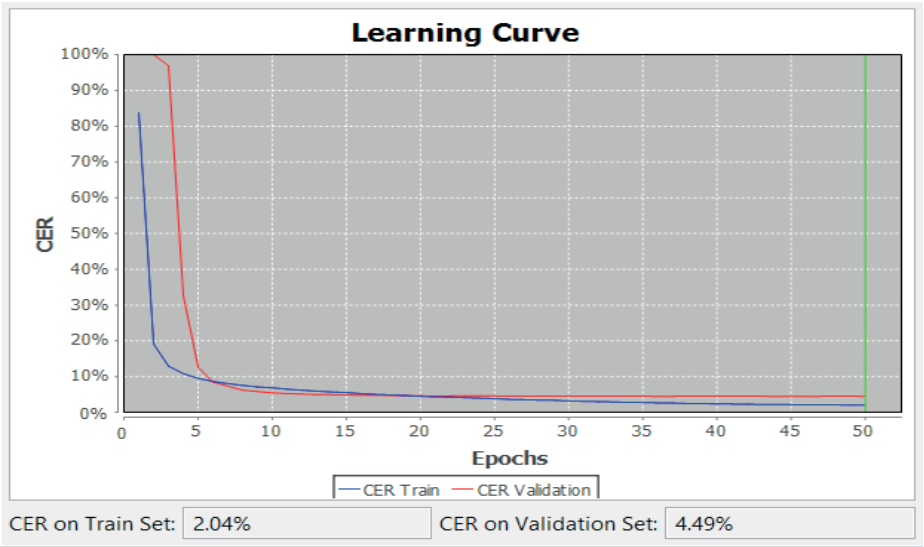


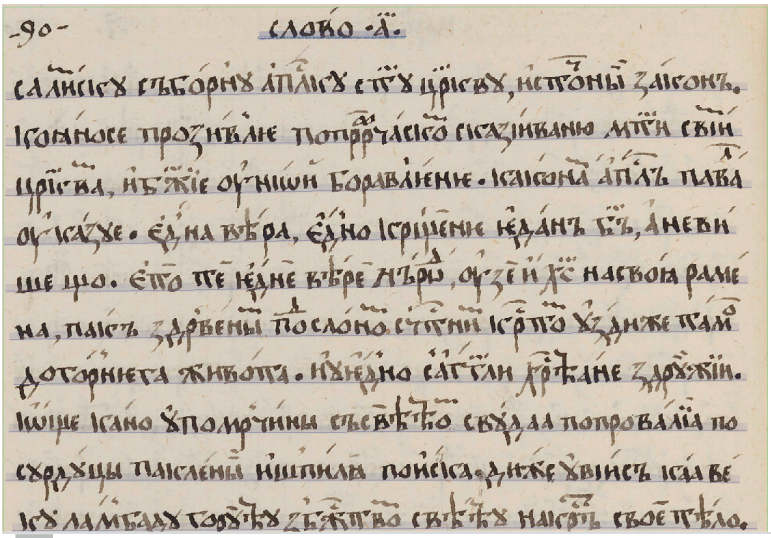
Figure 1. The learning curve of the *Venclović 0.3.* model

of training. Subsequently, the percentage of incorrectly recognized characters stabilizes very quickly at a certain level (only after ten epochs). By the end of the training process, it only slightly decreases, which means that increasing the number of epochs would not necessarily lead to a lower percentage of misrecognized characters.

3.3. The Qualitative Analysis of the *Venclović 0.3.* Model

Previous research (cf. only [Rabus 2019b: 13]) showed that the percentage of incorrectly recognized characters was not always a realistic indicator of model quality. Considering that during the automatic statistical calculation of the percentage of incorrectly recognized characters all interventions in the text are taken into account (e.g. insertion, deletion or replacement of characters, including spaces and punctuation marks),[20] qualitative indicators of the model's success are often better than quantitative ones. For the qualitative analysis of the *Venclović 0.3.* model a comparative display of sheet 90b САНУ 137 was used along with the automatically digitized text, which is presented in the following figure.

[20] For more precise data on the method of calculating the percentage of incorrectly recognized characters see Transkribus Glossary at https://readcoop.eu/glossary/character-error-rate-cer/.



1-1 слово . а .
1-2 салимскꙋ съборꙗнꙋ апслкꙋ сѵ̑ꙋ цр̑квꙋ, источны законъ .
1-3 коꙗано се прозивꙗлѥ по пророчаском сказиваню мѵ̑ти свим
1-4 цр̑ква, и бж̑ѥ оу нѣми боравꙗлѥнѥ . како нам апслъ пава-
1-5 оуказꙋе . єдꙗна вѣра, єдꙗно крꙗщенѥ ѥданъ бъ̑, а не ви-
1-6 ше що . єто те ѥдꙗне вѣре нърѡд, оузе и хс̑ . на своꙗ раме-
1-7 на, пакъ з дрꙗвеным подслоном с чс̑тним крс̑то ꙋздиже тамо
1-8 до горꙗнѥга живота . и ꙋѥдꙗно сꙗ аг̑ли хрс̑ћане здрꙋжїи .
1-9 ѡще кано ꙋ помрꙗчины съ свѣћом свꙋдаа попровалїа по-
1-10 сꙋрꙗдꙋцы паклены и шꙗпилꙗ поиска . диже ꙋвиисъ каа ве-
1-11 кꙋ ламꙗбадꙋ горꙋћꙋ зꙗ бж̑ством свѣћꙋ на крхъ свое тѣло .

Figure 2. САНУ 137, part of sheet 90b and the automatically digitized text

A comparison of the photo and the automatically digitized text demonstrates that the *Venclović 0.3.* model failed to recognize the superscript textual tag, which, in accordance with the principles of digitizing corpora for the historical dictionary of Serbian, introduces a superscript letter. The correction of the automatically cleared text was performed in two steps: in the first version, all the errors of character recognition were corrected, followed by the correction of the superscript textual tag that marks superscript letters in the second step. The percentage of misrecognized characters was identical in both versions (2.69% in total). The latter indicates that, not only did *Transkribus* fail to recognize the superscript textual tag, but also did not take its corrections into account when calculating misrecognized characters. This fact sheds a different

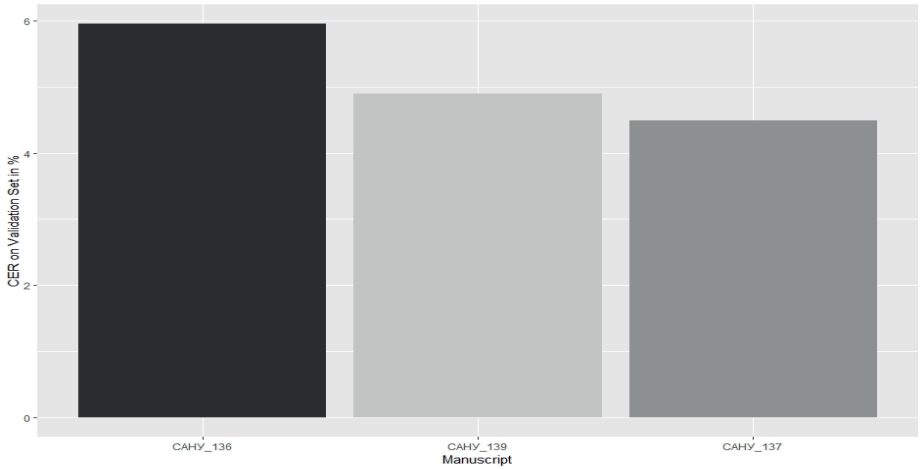
light on the quantitative indicators of the model presented in Tables 1–3. The percentage of incorrectly recognized characters on the validation set does not include the correction of the superscript textual tag. This actually means that the quantitative indicators of the model are slightly worse because of the specific way of marking superscript letters when digitizing the dictionary material.

If we return to our example of a part of sheet 90b and analyze individual errors in automatic text recognition, we can conclude that, most frequently, the *Venclović 0.3.* model makes mistakes in recognizing superscript letters and a *titlo* mark: so instead of сали^{м}скѫ 2, ц$\overset{р}{}$ква^{м} 4, пава^{л} 4, кр$\overset{ст}{}$то^{м} 7, кр^{ст}ъ 11 the model incorrectly reads сали$\overset{м}{}$скѫ 2, ц$\overset{р}{}$ква 4, пава- 4, кр$\overset{ст}{}$то 7, крхъ 11. In two examples, errors were recorded in the recognition of spaces between words: instead of по провалїа 8, по сѫр^{ь}дѫцы 8/9 the model incorrectly reads попровалїа 8, посѫр^{ь}дѫцы 8/9. Other errors refer to superscript letters and *titlo* marks that are recognized but not raised to a superscript: instead of ап^{с}лкѫ 2, исто^{ч}ны 2, пр^{о}р^{о}часком 3, свои^{м} 3, на^{м} 4, ап^{с}лъ 4, нърѡ^{д} 6, др^{ь}вены^{м} 7, по^{д}слоно^{м} 7, ч^{с}тни^{м} 7, там^{о} 7, хр^{с}ѣане 8, свѣѣо^{м} 9, бж^{с}твом 10 the model reads ап$\overset{с}{}$лкѫ 2, исто$\overset{ч}{}$ны 2, пр$\overset{о}{}$р$\overset{о}{}$часком 3, свiим 3, нам 4, ап$\overset{с}{}$лъ 4, нърѡд 6, др^{ь}веным 7, подслоном 7, ч$\overset{с}{}$тним 7, там$\overset{о}{}$ 7, хр$\overset{с}{}$ѣане 8, свѣѣом 9, бж$\overset{с}{}$твом 10. Taking the aforementioned errors into consideration, it seems that the *Venclović 0.3.* model can be rated as excellent in the qualitative sense, as well. We hope that the problem of not recognizing textual tags will be solved in the future by technical improvement of *Transkribus*. However, even if it stays the same, the process of digitizing texts for the historical dictionary of the Serbian language will be accelerated significantly.

3.4. Application of the *Venclović 0.3.* model on other manuscripts in Serbian vernacular

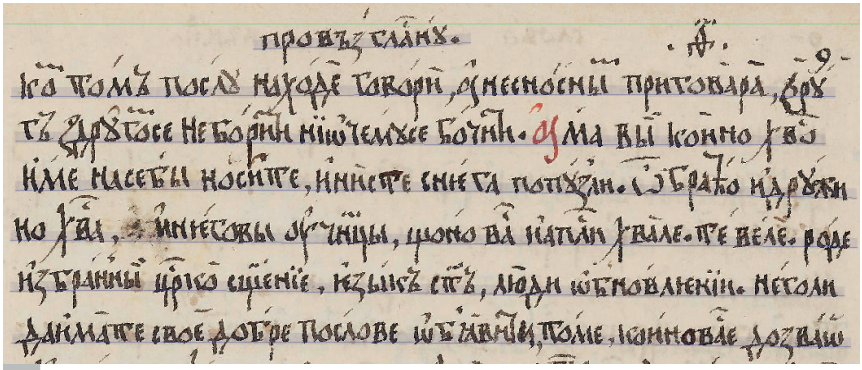
In continuation of the research, we hypothesized that the *Venclović 0.3.* model, trained on САНУ 137 material, will be able to successfully automatically recognize Venclović's other manuscripts in Serbian vernacular. In order to test the hypothesis, we created an experiment in which we used the *Venclović 0.3.* model applied to the first ten pages of the manuscript *Великопосник* (САНУ 97 (136) from the year 1740/41) (hereinafter САНУ 136) and *Поученија изабрана I* (САНУ 99 (139) from 1745) (hereinafter САНУ 139). These two manuscripts were chosen for the experiment because they were written in the same style as САНУ 137, as well as because their high-quality digital recordings already existed in the SASA Archives.

As can be seen in Graph 1, the quantitative performance of the *Venclović 0.3.* model on the manuscripts САНУ 136 and 139 is negligibly lower than on САНУ 137. Excluding the superscript letters and *titlo* marks in the superscript, the percentage of misrecognized characters on САНУ 136 is 5.95%, and on САНУ 139 is even lower—4.90%.



Graph 1. Application of the *Venclović 0.3.* model on manuscripts САНУ 136 and САНУ 139

For the qualitative evaluation of the result, a comparative view of a part of sheet 9a of the САНУ 136 manuscript is presented in the following figure.

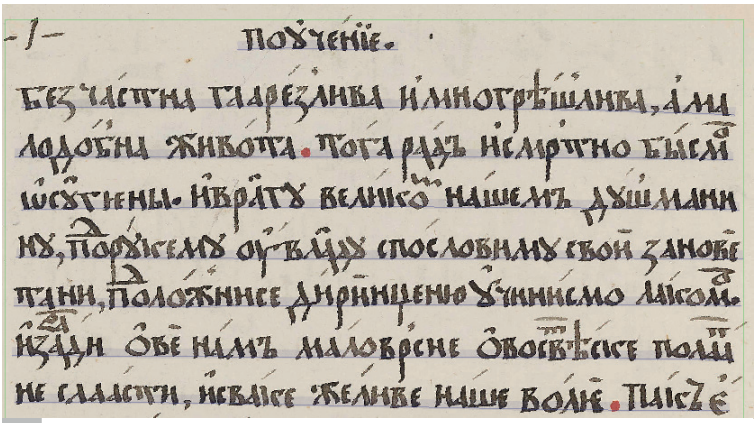


1-1 провъ зглан҃ѕ .
1-2 ком томъ посл҃ѕ находе говори, а не сносны приговара, зр҃ѕ-
1-3 гъ з др҃ѕгом се не борити нї ѡ чем҃ѕ се бочити . ама вы коино х҃сво-
1-4 име на себы носите, и ни сте с҃ нѣга попѧзли . ѡ брат҃ом и др҃ѕжи-
1-5 но х҃сва м҃нѣговым оучн̃цы, щоно ва҃с и ап҃сли хвале . те веле . роде
1-6 избран҃ны цр҃ском с҆ще҆нїе, ꙗзыкъ с҃тъ, люди ѡб҃нов҃лѣнїи . него ли
1-7 да имате свое добре послове ѡбꙗвити, томе, коино вам е дозваѡ

Figure 3. САНУ 136, part of sheet 9a and the automatically recognized text

Along with the errors in recognizing superscripts: instead of комъ 2, дрѵгомъ 3, хс̑во 3, хс̑ва 5, вас̑ 5, апс̑ли 5 there is the incorrect ком 2, дрѵгом 3, хс̑во- 3, хс̑ва 5, вас̑ 5, апс̑ли 5, the model makes errors in recognizing superscript letters and *titlo* marks, as well as spaces between words, more frequently than in the case of the manuscript САНУ 137: thus instead of провъзгласс̑нѵ 1, несноснымъ 2, браћо 4, а и нѣговы 5, црс̑ко 6, вас̑ 7, the model incorrectly reads провъ зглас̑нѵ 1, не сносны 2, браћом 4, мꙗнѣговым 5, црс̑ком 6, вам 7; thus instead of провъзгласс̑нѵ 1, несноснымъ 2, хс̑во 3, нисте 4 the model incorrectly reads провъ згла̑нѵ 1, не сносны 2, хс̑во- 3, ни сте 4. Unlike САНУ 137, errors of recognizing a *pajerak* mark are recorded here: thus instead of с нѣга попѵзꙗли 4, а и нѣговы 5, ѡбꙗновлѣнїи 6, ѡбꙗвити 7 the model incorrectly reads сꙗ нѣга попѵзли 4, мꙗнѣговым 5, ѡбꙗновꙗлѣнїи 6, ѡбꙗвити 7. Errors in recognizing letters in examples дрѵ- 2, а и нѣговы 5 (incorrect зрѵ- 2, мꙗнѣговым 5) can be explained by an illegible recording.

For the qualitative evaluation of the results of the efficiency of the model on the САНУ 139 manuscript, a comparative view of a part of sheet 1b and the automatically recognized text is displayed in the following figure.



1-1 по ѵченїе .
1-2 без̑ частна гаарез̑лива и многрѣш̑лива, ама
1-3 ло доб̑на живота . тога радъ и смр̑тном бысмо̑
1-4 ѡсѵгнены . и врагѵ велико нашемъ дѵш̑мани
1-5 нѵ, под рѵке мѵ оу владѵ с послови мѵ своим занове
1-6 тани, подлож̑ни се дирин̑ценю ѵчини смо лакомо̑ .
1-7 и задил̑ ове намъ мало вр̑сне ово свѣ̑тске полаа-
1-8 не слаасти, и сваке желиве наше волѥ . пакъ є-

Figure 4. САНУ 139, part of sheet 1b and the automatically recognized text

The previous presentation shows that the *Venclović 0.3.* model makes most frequent errors in recognizing spaces between words: thus instead of повченїе 1, безчастна 2, а малодобꙗна 2/3, дꙋшꙗманинꙋ 4/5, зановетани 5/6, ꙋчинисмо 6, маловрꙗсне 7, овосвѣ͂ске 7 the model incorrectly reads по ꙋченїе 1, безꙗ частна 2, ама ло добꙗна 2/3, дꙋшꙗманинꙋ 4/5, занове тани 5/6, ꙋчини смо 6, мало врꙗсне 7, ово свѣ͂тске 7. Digitizing superscripts represents an issue in the following examples: смрꙗтно 3, великоᵐ 4, свои 5, полаа᷆- 7, заᵖᵃди 7 the model incorrectly outputs смрꙗтном 3, велико 4, своим 5, полаа- 7, задиᷣ 7. In relation to this category and in relation to САНУ 137, there are a few examples with unrecognized superscript textual tag: instead of бысмᵒ 3, поᵈ 5, поᵈложꙗни 6, лакомᵒ 6, овосвѣ͂ске 7 the model outputs бысмо͡ 3, под 5, подложꙗни 6, лакомо͡ 6, ово свѣ͂тске 7. The errors in the recognition of the *pajerak* mark appeared in two examples only: instead of безчастна 2, дириꙗценю 6 the model incorrectly outputs безꙗ частна 2, дириꙗценю 6.

4. Concluding Remarks and Future Research Perspectives

The results of the previously presented research point to the conclusion that following the principles of artificial intelligence and machine learning, and using the *Transkribus* software platform, the digitization process of Cyrillic manuscripts can be significantly accelerated in order to create an electronic corpus for the historical dictionary of Serbian. Using the example of the Gavril Stefanović Venclović's manuscripts written in Serbian vernacular of the 18th century, the study shows that the process of transcription of voluminous manuscripts can be digitized by creating specific models for automatic text recognition. The *Venclović 0.3.* model was created on the material of 100 pages of the voluminous САНУ 137 manuscript (774 pages in total), with an acceptable percentage of misrecognized characters of about 4–5%. Using the same model, the rest of the voluminous manuscript САНУ 137 can likewise be digitized in a significantly shorter amount of time, significantly reducing human and financial resources, if, of course, complemented by a final proofreading and edition by a competent philologist. The *Venclović 0.3.* model can also be used fairly successfully for the automatic recognition of other Venclović's manuscripts in Serbian vernacular written in similar style. The percentage of incorrectly recognized characters on the САНУ 136 and 139 manuscripts was around 5–6% and is only slightly lower than the САНУ 137 manuscript on which the model was trained. The qualitative analysis of the most common errors in automatic recognition can lead to the conclusion that the most frequent problems the model has pertain to recognizing superscript letters, *titlo* marks and spaces between words. Errors in the recognition of a *pajerak* mark are much less frequent, and errors in the recognition of regular letters are found merely exceptionally.

The recognition of the superscript textual tag used to mark the superscript letter following the principles of digitization of dictionary materials is fairly problematic. Although *Transkribus* offers the possibility of training and recognizing textual tags since version 1.18.0., our research has shown that this option is still not fully applicable.[21] *Transkribus* does not read textual tags during initial recognition, yet only if there is a version of the digitized text in *Transkribus*. In neither case does it take textual tags into account when calculating misrecognized characters. Therefore, the qualitative performance of the *Venclović 0.3.* model is slightly less efficient than the percentage of incorrectly recognized characters shows, but still excellent, especially compared to traditional manual digitizing. This problem could be overcome in the near future either by further improving the technical performance of *Transkribus* or by minimally modifying the principles of digitization and, also by improving the font, so that superscript letters could be marked with special characters compliant with the Unicode standard. After solving this problem, the manually digitized material obtained so far throughout the project of digitizing the historical dictionary of Serbian could, after prior preparation, be imported into *Transkribus* and used for training specific and generic models to automatically recognize other Cyrillic manuscripts.

The advantage of automatic text recognition as compared to the traditional process is especially evident in the possibility of constant improvement of the performance of specific and generic models in accordance with the progress of the transcription process and the increase in the amount of digitized text that can be used to train a new version of the model. In order to further improve the model for automatic text recognition of Venclović's manuscripts written in Serbian vernacular, it seems necessary to completely digitize the manuscripts within the SASA Archives, and to establish cooperation with scientific and cultural institutions (SASA and Matica srpska) to become potential leaders of a particular project related to preparing and publishing the critical edition of Venclović's manuscripts in Serbian vernacular. With the development of technology for automatic text recognition, we are not only approaching the critical edition of Venclović's manuscripts, but also the possibility of creating a digital edition and a special electronic corpus.

[21] *Transkribus* software, version 1.20.0. was used for all the experiments described in the paper.

Bibliography

Primary Sources

САНУ 136
Венцловић Стефановић Г. [*Великопосник*], Архив Српске академије наука и уметности, сигнатура САНУ 97 (136).

САНУ 137
Венцловић Стефановић Г. [*Слова изабрана*], Архив Српске академије наука и уметности, Београд, сигнатура САНУ 101 (137).

САНУ 139
Венцловић Стефановић Г. [*Поучение изабраноје, први део*], Архив Српске академије наука и уметности, Београд, сигнатура САНУ 99 (139).

Literature

Besters-Dilger, Rabus 2021
Besters-Dilger J., Rabus A., Neural Morphological Tagging for Slavic: Strengths and Weaknesses, *Scripta&e-Scripta*, 2021, 21, 79–92.

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

Kiesling et al. 2019
Kiesling B., Tissot R., Stokes P., Stökl Ben Ezra D., eScriptorium: An Open Source Platform for Historical Document Analysis, *2019 International Conference on Document Analysis and Recognition Workshop (ICDARW)*. Sydney, 2019, 19–24.

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

Polomac, Lutovac Kaznovac 2021
Polomac V., Lutovac Kaznovac T., Automatic Recognition of Serbian Medieval Manuscripts by Applying the *Transkribus* Software Platform: Current State and Future Perspectives, *Зборник Матице српске за филологију и лингвистику*, 2021, LXIV/2, 7–26.

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

——— 2022b
Polomac V., Serbian Early Printed Books. Towards Generic Model for Automatic Text Recognition using *Transkribus*, D. Fišer, T. Erjavec, eds., *Proceedings of the Conference on Language Technologies and Digital Humanities*, Ljubljana, 2022b, 154–161.

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

——— 2019b
Rabus A., Training generic models for Handwritten Text Recognition using Transkribus: opportunities and pitfalls, *Proceeding of the Dark Archives Conference*, Oxford, 2019b, in print.

Васиљев 1996
Васиљев Љ., Буквар из 1717. године — дело Гаврила Стефановића Венцловића, *Зборник Матице српске за филологију и лингвистику*, 1996, 39/2, 169–184.

Бјелаковић 2021
Бјелаковић И., Предлог микроструктуре историјског речника српског језика, Ј. Грковић-Мејџор, И. Бјелаковић, М. Курешевић, ур., *Историјска лексикографија српског језика*, Нови Сад, 2021, 387–400.

Грковић-Мејџор 2007
Грковић-Мејџор Ј., *Списи из историјске лингвистике*, Сремски Карловци, Нови Сад, 2007.

——— 2021
Грковић-Мејџор Ј., Ка историјском речнику српског језика, Ј. Грковић-Мејџор, И. Бјелаковић, М. Курешевић, ур., *Историјска лексикографија српског језика*, Нови Сад, 2021, 11–24.

Грковић-Мејџор, Бјелаковић 2021
Грковић-Мејџор Ј., Бјелаковић И., Дефинисање лексичког значења у историјском речнику српског језика, Ј. Грковић-Мејџор, И. Бјелаковић, М. Курешевић, ур., *Историјска лексикографија српског језика*, Нови Сад, 2021, 367–386.

Гроздановић-Пајић 1992
Гроздановић-Пајић М., Хартија и водени знаци у Венцловићевим рукописима писаним у Коморану и Ђуру, *Сентандрејски зборник*, 1992, 2, 177–197.

Даничић 1863–1864
Даничић Ђ., *Речник из књижевних старина српских, I–III*, У Биограду, 1863–1864.

Ивић 2014
Ивић П., *Преглед историје српског језика*, Сремски Карловци, Нови Сад, 2014.

Јовић 2021
Јовић Н., Медицински списи као извор за историјски речник српског језика, Ј. Грковић-Мејџор, И. Бјелаковић, М. Курешевић, ур., *Историјска лексикографија српског језика*, Нови Сад, 2021, 185–198.

Курешевић 2016
Курешевић М., Језик *Слова Акира Премудрог* из рукописног зборника Народне библиотеке Србије бр. 53, *Јужнословенски филолог*, 2016, 72/1–2, 105–126.

——— 2021
Курешевић М., Граматичке информације у историјском речнику српског језика: полазни принципи, Ј. Грковић-Мејџор, И. Бјелаковић, М. Курешевић, ур., *Историјска лексикографија српског језика*, Нови Сад, 2021, 319–345.

Курешевић et al. 2021
Курешевић М., Лутовац Казновац Т., Цолић Јовановић А., Бајић В., Рашчитавање и пренос у електронску форму ћирилске грађе за историјски речник српског језика: недоумице и могућа решења, Ј. Грковић-Мејџор, И. Бјелаковић, М. Курешевић, ур., *Историјска лексикографија српског језика*, Нови Сад, 2021, 81–113.

Павић 1972
Павић М., *Гаврил Стефановић Венцловић*, Београд, 1972.

Павловић 2021
Павловић С., Лексикографска обрада граматичких речи у историјским речницима, Ј. Грковић-Мејџор, И. Бјелаковић, М. Курешевић, ур., *Историјска лексикографија српског језика*, Нови Сад, 2021, 345–366.

Радовановић 2015
Радовановић М., *Фази лингвистика*, Сремски Карловци, Нови Сад, 2015.

РСАНУ, 1–
Речник српскохрватског књижевног и народног језика, Београд: 1959–.

Савић, Милановић 2021
Савић В., Милановић А., Идентификација и формирање одредница у српском историјском речнику, Ј. Грковић-Мејџор, И. Бјелаковић, М. Курешевић, ур., *Историјска лексикографија српског језика*, Нови Сад, 2021, 277–318.

Синдик et al. 1991
Синдик Н., Гроздановић-Пајић М., Мано-Зиси К., *Опис рукописа и старих штампаних књига Библиотеке Српске православне епархије будимске у Сентандреји*, Београд, Нови Сад, 1991.

Стефановић, Јовановић 2013
Стефановић Д., Јовановић Т., *Венцловићев сентандрејски буквар: 1717*, Будимпешта, Београд, 2013.

Стојановић 1901
Стојановић Љ., *Каталог рукописа и старих штампаних књига Српске краљевске академије*. Београд, 1901.

Суботић 2004
Суботић Љ., Из историје књижевног језика: питање језика, В. Васић, ур., *Предавања из историје језика*, Нови Сад, 2004, 142–191.

Трифуновић 2009
Трифуновић Ђ., *Стара српска књижевност: основи*, Београд, 2009.

Цветковић Теофиловић 2021
Цветковић Теофиловић И., Путописи као извори за израду речника српског језика XII–XVIII века, Ј. Грковић-Мејџор, И. Бјелаковић, М. Курешевић, ур., *Историјска лексикографија српског језика*, Нови Сад, 2021, 165–184.

References

Besters-Dilger J., Rabus A., Neural Morphological Tagging for Slavic: Strengths and Weaknesses, *Scripta&e-Scripta*, 2021, 21, 79–92.

Bjelaković I., Predlog mikrostrukture istorijskog rečnika srpskog jezika, *Istorijska leksikografija srpskog jezika*, Novi Sad, 2021, 387–400.

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

Cvetković Teofilović I., Putopisi kao izvori za izradu rečnika srpskog jezika XII–XVIII veka, *Istorijska leksikografija srpskog jezika*, Novi Sad, 2021, 165–184.

Grković-Mejdžor J., *Spisi iz istorijske lingvistike*, Sremski Karlovci, Novi Sad, 2007.

Grković-Mejdžor J., Ka istorijskom rečniku srpskog jezika, *Istorijska leksikografija srpskog jezika*, Novi Sad, 2021, 11–24.

Grković-Mejdžor J., Bjelaković I., Definisanje leksičkog značenja u istorijskom rečniku srpskog jezika, *Istorijska leksikografija srpskog jezika*, Novi Sad, 2021, 367–386.

Grozdanović-Pajić M., Hartija i vodeni znaci u Venclovićevim rukopisima pisanim u Komoranu i Đuru, *Sentandrejski zbornik*, 1992, 2, 177–197.

Ivić P., *Pregled istorije srpskog jezika*, Sremski Karlovci, Novi Sad, 2014.

Jović N., Medicinski spisi kao izvor za istorijski rečnik srpskog jezika, *Istorijska leksikografija srpskog jezika*, Novi Sad, 2021, 185–198.

Kiesling B., Tissot R., Stokes P., Stökl Ben Ezra D., eScriptorium: An Open Source Platform for Historical Document Analysis, *2019 International Conference on Document Analysis and Recognition Workshop (ICDARW)*. Sydney, 2019, 19–24.

Kurešević M., The Language of the Story of the Sage Ahiquar from Serbian Manuscript No. 53 of the National Library of Serbia, *Južnoslovenski filolog*, 2016, 72/1–2, 105–126.

Kurešević M., Gramatičke informacije u istorijskom rečniku srpskog jezika: polazni principi, *Istorijska leksikografija srpskog jezika*, Novi Sad, 2021, 319–345.

Kurešević M., Lutovac Kaznovac T., Colić Jovanović A., Bajić V., Raščitavanje i prenos u elektronsku formu ćirilske građe za istorijski rečnik srpskog jezika: nedoumice i moguća rešenja, *Istorijska leksikografija srpskog jezika*, Novi Sad, 2021, 81–113.

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

Pavić M., *Gavril Stefanović Venclović*, Beograd, 1972.

Pavlović S., Leksikografska obrada gramatičkih reči u istorijskim rečnicima, *Istorijska leksikografija srpskog jezika*, Novi Sad, 2021, 345–366.

Polomac V., Lutovac Kaznovac T., Automatic Recognition of Serbian Medieval Manuscripts by Applying the *Transkribus* Software Platform: Current State and Future Perspectives, *Matica Srpska Journal of Philology and Linguistics*, 2021, LXIV/2, 7–26.

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

Polomac V., Serbian Early Printed Books. Towards Generic Model for Automatic Text Recognition using *Transkribus*, D. Fišer, T. Erjavec, eds., *Proceedings of the Conference on Language Technologies and Digital Humanities*, Ljubljana, 2022b, 154–161.

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

Radovanović M., *Fazi lingvistika*, Sremski Karlovci, Novi Sad, 2015.

Savić V., Milanović A., Identifikacija i formiranje odrednica u srpskom istorijskom rečniku, *Istorijska leksikografija srpskog jezika*, Novi Sad, 2021, 277–318.

Sindik N., Grozdanović-Pajić M., Mano-Zisi K., *Opis rukopisa i starih štampanih knjiga Biblioteke Srpske pravoslavne eparhije budimske u Sentandreji*, Beograd, Novi Sad, 1991.

Stefanović D., Jovanović T., *Venclovićev sentandrejski bukvar: 1717*, Budimpešta, Beograd, 2013.

Subotić Lj., Iz istorije književnog jezika: pitanje jezika, *Predavanja iz istorije jezika*, Novi Sad, 2004, 142–191.

Trifunović Đ., *Stara srpska književnost: osnovi*, Beograd, 2009.

Vasiljev Lj., Bukvar iz 1717. godine — delo Gavrila Stefanovića Venclovića, *Matica Srpska Journal of Philology and Linguistics*, 1996, 39/2, 169–184.

Vladimir Polomac, PhD, redovni profesor
Univerzitet u Kragujevcu
Filološko-umetnički fakultet
Jovana Cvijića bb, 34000 Kragujevac
Srbija / Serbia
v.polomac@filum.kg.ac.rs

Marina Kurešević, PhD, redovni profesor
Univerzitet u Novom Sadu
Filozofski fakultet
Zorana Đinđića 2, 21 000 Novi Sad
Srbija / Serbia
marina.kuresevic@gmail.com

Isidora Bjelaković, PhD, redovni profesor
Univerzitet u Novom Sadu
Filozofski fakultet
Zorana Đinđića 2, 21 000 Novi Sad
Srbija / Serbia
isidora.bjelakovic@gmail.com

Aleksandra Colić Jovanović, PhD, asistent sa doktoratom
Univerzitet u Novom Sadu
Filozofski fakultet
Zorana Đinđića 2, 21 000 Novi Sad
Srbija / Serbia
aleksandra.colic@ff.uns.ac.rs

Sanja Petrović, doktorand
Univerzitet u Novom Sadu
Filozofski fakultet
Zorana Đinđića 2, 21 000 Novi Sad
Srbija / Serbia
sanja.lj.petrovic@gmail.com

Received September 27, 2022