Named Entity Recognition for Pre-modern Serbian: A Preliminary Study

Scientific paper DOI: 10.18485/judig.2025.1.ch17

Marija Đokić Petrović[1], *0000-0002-0568-5219*
Mihailo St. Popović[2], *0000-0002-3128-2210*
Vladimir Polomac[3], *0000-0001-5140-9330*

Abstract

Digital Humanities represent a significant advance in the research and preservation of cultural heritage, enabling the application of computational methods to the analysis of historical documents. In this context, Named Entity Recognition (NER) stands out as an important technique that allows scholars to automatically extract, classify, and organize essential information from historical sources. In this study, we focus on developing an NER model tailored to documents written in pre-modern Serbian. A NER model was trained on the historical corpus that included the Banjska Chrysobull, the third example of Dečani Chrysobull, and the collection of 13th-century charters and letters from the Dubrovnik archive. The model was evaluated on two representative documents: the Will of Miloš Belmužević from 1500 and unpublished manuscript from 1778 preserved in the archive of the Greek Orthodox Church of St. George in Vienna. The developed NER model is based on a Convolutional Neural Network architecture and trained to recognize three distinct entity types—personal names, toponyms, and demographic references. Although the model achieved an F1 score of ≈0.92 on the development set, its performance declined on the evaluation set, i.e., previously unseen

[1] School of Computing, Union University, Kneza Mihaila 6/VI, 11000 Belgrade, Serbia, mdjokicpetrovic@raf.rs

[2] Austrian Academy of Sciences, Institute for Medieval Research, Georg-Coch-Platz 2, 1010, Vienna, Austria, mihailo.popovic@oeaw.ac.at

[3] University of Kragujevac, Faculty of Philology and Arts, Liceja Kneževine Srbije 1A, 34000, Kragujevac, Serbia, v.polomac@filum.kg.ac.rs

documents. Generally, this study presents the first application of NER to medieval Serbian documents written in pre-modern Serbian, thereby contributing to historical research and the preservation of cultural heritage. The proposed approach establishes a good foundation for further refinement of the developed model.

Keywords: Named Entity Recognition; Pre-modern Serbian; Digital Humanities; Cultural Heritage

1. Introduction

In recent decades, the field of Digital Humanities (DH) has emerged as a transformative approach to the study, interpretation, and preservation of cultural heritage (Kagaba Amina, 2024). This interdisciplinary field has proven particularly valuable in the context of historical documents (Milligan, 2022). Specifically, documents written in archaic scripts benefit greatly from digital techniques that enhance accessibility and deepen scholarly understanding (Alstola & Svard, 2024).

Among the various techniques employed in DH, Named Entity Recognition (NER) plays a particularly important role. It is designed to identify and classify entities in unstructured text into predefined categories such as persons, locations, organizations, demographics, events, and others (Jehangir, Radhakrishnan, & Agarwal, 2023). In the field of history, NER serves as a powerful tool for extracting knowledge from historical documents, enabling deeper semantic analysis and supporting more comprehensive historical interpretation (Ehrmann, Hamdi, Pontes, Romanello, & Doucet, 2023).

In the context of the Serbian language, NER has been increasingly applied to modern and printed texts with the support of machine learning and transformer-based models (Marovac, Avdić, & Milošević, 2023), thus enabling more advanced information extraction from contemporary historical sources. However, considerably less attention has been directed toward historical varieties of the Serbian language. In that sense, the pre-modern Serbian, which was used in medieval administrative, legal, and ecclesiastical documents, has remained largely unexplored in computational NER research.

To address this gap, this study focuses on developing an NER model tailored to documents written in pre-modern Serbian. The model was trained on a historical corpus that includes the Charter of King Stefan Uroš II Milutin to the Monastery of St. Stefan in Banjska (Banjska Chrysobull) (Trifunović, 2011), the third example of the Charter of King Stefan Uroš III Nemanjić to the Monastery of Visoki Dečani (Dečani Chrysobull) (Ivić

& Grković, 1976), and the collection of charters and letters from the 13th-century preserved in the Dubrovnik archive (Stojanović, 1929). To evaluate the model's performance, we selected two representative documents: the Will of Miloš Belmužević (Polomac, 2025; Obradović, 2024) from 1500 and unpublished manuscript from 1778, housed in the archive of the Greek Orthodox Church of St. George in Vienna (Plöchl, 1983; Ransmayr, 2018; Turczynski, 1959; Seirinidou, 2011).

The developed NER model, employing a Convolutional Neural Network (CNN) architecture and trained to recognize three distinct entity types—personal names, toponyms, and demographic references—achieved an F1 score of approximately 0.92 on the development set. Its performance, however, declined when applied to previously unseen documents, reflecting the linguistic and orthographic variability inherent in historical sources. Despite these challenges, this study represents the first application of NER for pre-modern Serbian, thereby contributing to both historical scholarship and the preservation of cultural heritage. The presented approach provides a robust foundation for further methodological improvements of the developed NER model.

The remainder of this paper is structured as follows. Section 2 provides an overview of related work. Section 3 presents the development of the NER model and the evaluation strategy. Section 4 discusses the findings. Finally, Section 5 offers concluding remarks and outlines directions for future research.

2. Related Work

NER is a fundamental task in Natural Language Processing (NLP) that involves identifying and categorizing entities in unstructured text into predefined classes such as personal names, organizations, locations, time expressions, quantities, and other domain-specific types (Jehangir, Radhakrishnan, & Agarwal, 2023). Although NER has been an established NLP task since the 1990s, its application to the Serbian language remained limited until the early 2010s, when more systematic and domain-specific research efforts began to emerge (Marovac, Avdić, & Milošević, 2023).

Early efforts in Serbian-language NER include the system developed by (Vitas & Pavlović-Lažetić, 2008), which combined morphological and lexical analysis with dictionary-based resources to recognize personal names and geographic entities. Authors (Ljubešić, Stupar, Jurić, & Agić) developed a machine learning-based NER system for Croatian and Slovene languages, closely related to Serbian, using conditional random fields. The model was trained on annotated web and news corpora (e.g., SETimes, Vjesnik) and incorporated linguistic and distributional similarity features derived from large unannotated monolingual corpora. Krstev, Obradović, Utvić, & Vitas (2014) enhanced an earlier rule-based approach (Krstev, 1997), incorporating transducers and thesauri to identify personal and geopolitical entities. This approach was later utilized to construct a gold-standard dataset of news articles annotated with personal names, which served as training data for machine learning models such as Stanford NER and spaCy (Šandrih, Krstev, & Stanković, 2019). In the context of cultural heritage, (Tanasijević, 2019) developed a system for annotating cultural heritage documents with metadata by recognizing entities like years, person names, and document topics. Authors in (Todorović Šandrih, Krstev, Stanković, & Nešić Ikonić, 2021), developed a NER model (SrpCNNER) based on a CNN architecture, trained on a corpus of Serbian novels written between 1840 and 1920. The model was designed to recognize seven named entity categories —persons, locations, organizations, roles, events, demos, and art works—and achieved an F1 score of approximately 91% on the test dataset. Authors in (Nešić Ikonić, Petalinkar, Stanković, & Utvić, 2024) introduced the SrpCNNeL model, which builds upon the NER backbone SrpCNNER2 by incorporating an entity-linking layer designed to align recognized entities with entries in the Wikidata knowledge base. The SrpCNNER2 model is trained using the spaCy Python module, employing the same model architecture as SrpCNNER (Todorović Šandrih, Krstev, Stanković, & Nešić Ikonić, 2021) across a diverse dataset containing sentences from Serbian novels (1840-1920), legal documents, as well as sentences generated from the Wikidata knowledge base and the Leximirka lexical database.

More recently, transformer-based models have brought significant advancements to NER in South Slavic languages. For example, (Ljubešić & Lauc, 2021) introduced a multilingual transformer model BERTić, pre-trained on 8 billion tokens from web-crawled corpora in Bosnian, Croatian, Montenegrin, and Serbian. This model was fine-tuned on several domain-specific datasets, including SETimes.SR (Batanovic, Ljubešić, & Samaradžić, 2018), ReLDI-sr (Ljubešić, Erjavec, Batanović, Miličević, & Samardžić, 2025), and news article corpora, demonstrating strong performance in modern NER tasks. Authors in (Ikonić Nešić, Petalinkar, Škorić, & Stanković, 2024) conducted a comparison of different architectures and techniques for preparing NER models via integrating BERT with spaCy. They trained models to recognize seven entity types—persons, locations, organizations, professions, events, demonyms, and artworks—using a set of Serbian novels (1840-1920), modern newspaper articles, and knowledge-base–derived sentences from the Wikidata knowledge base and Leximirka lexical database.

Despite the progress of NER for the contemporary Serbian language, the application of such methods to its historical varieties remains limited. This study presents the first attempt to develop and apply an NER model tailored to such script. By doing so, this paper addresses a critical gap in the digital processing of Serbian medieval textual heritage and contributes to the broader objective of enhancing access to historical content for both scholars and the wider public.

3. NER Model for Pre-modern Serbian

In this section, we present the training process of the NER model[4] for pre-modern Serbian, followed by a detailed evaluation of its performance.

3.1. Training

We trained our model on the training corpus, which includes Banjska Chrysobull, Dečani Chrysobull and the collection of 13th-century charters and letters preserved in the Dubrovnik archive. We should mention that the Banjska Chrysobull, along with the Dečani Chrysobull, are the richest medieval sources of personal and geographical names (Grković, 1983; Loma, 2013).

The training corpus was first manually annotated, including entities like personal names (PERS), locations (LOC), and demographic references (DEMO).

To prepare the corpus for training, we first segmented the texts into sentences, resulting in a total of 6715 sentences (38900 words), of which 4112 contain named entities. Table 1 shows the distribution of entities in the corpus.

Table 1. Distribution of Entities in the Training Corpus

Entity	Total	Unique
PERS	7562	2490
LOC	1525	1275
DEMO	85	54

[4] https://github.com/marijadjokic/ner_pre_modern_serbian

We then created annotations list as tuples sentence-list of entities. An example of such tuple is: „семоу оубо сн҃ь бж҃и прѣⷣстоѥ ѡ десноую ѡц҃а ꙗвии се не ꙗко работ҅нь нь ꙗко влⷣка и г҃ь всѣхь долоу влѣкоущоую ѥго пл҅ть бестрⷭ҇т҅ноу творе и приводе кь ѡц҃оу и сего паче моисеꙗ прославлꙗѥ не бо ꙗкоже моиси зад҅нꙗа бж҃иꙗ видѣ вь камени“, „entities“: [[171, 177, „PERSON“], [201, 206, „PERSON“]].

Afterwards, spaCy v3.7.5 allows us to specify a custom CNN architecture within a simple text file. Using the quick-start widget[5], we set up the default settings. For our model, the language was Serbian, containing only the *ner* component. The model was trained on a CPU. We made specific adjustments to the default configuration:

- [components.tok2vec.model.encode]: The token-to-vector layer size was increased from 96 to 300 (the maximum recommended value).
- [components.tok2vec.model.encode]: The architecture used is HashEmbedCNN, with input and output width set to 300, 4 convolutional layers, 2,000 rows in hash embedding tables, a window size of 1 token for concatenation during convolutions, and no pretrained static vectors.

We then employed a Python library to randomly partition the loaded annotated data into two separate sets: a training set and a testing set, where 20% of the data is allocated to the test set, while the remaining 80% is used for training. This stratification ensures that the model can be trained on the majority of the data and subsequently evaluated on a held-out subset, providing an unbiased estimate of its performance.

Model training ended up after 15 epochs having 92.08%, 93.11% and 91.08% F1 score, precision (P), and recall (R), on the development set, respectively. The number of epochs is automatically generated.

3.2. Evaluation

Our evaluation dataset comprises two historical documents previously mentioned: the Will of Miloš Belmužević from the 1500s (Figure 1) and the document from the Greek Orthodox Church of St. George in Vienna from 1778 (Figure 2).

[5] *Quick-start spaCy3 widget, https://spacy.io/usage/training#quickstart*

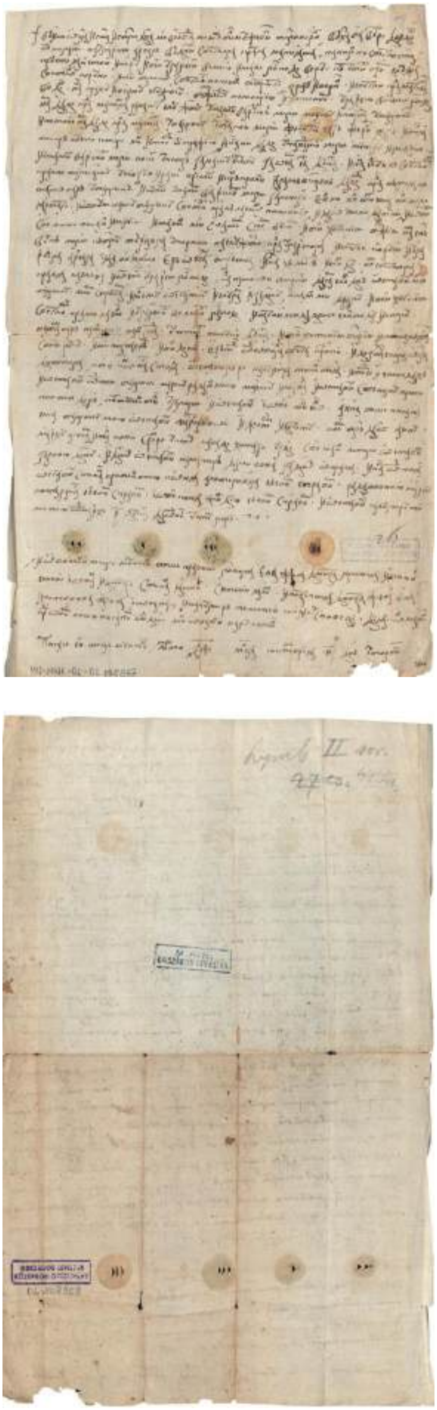
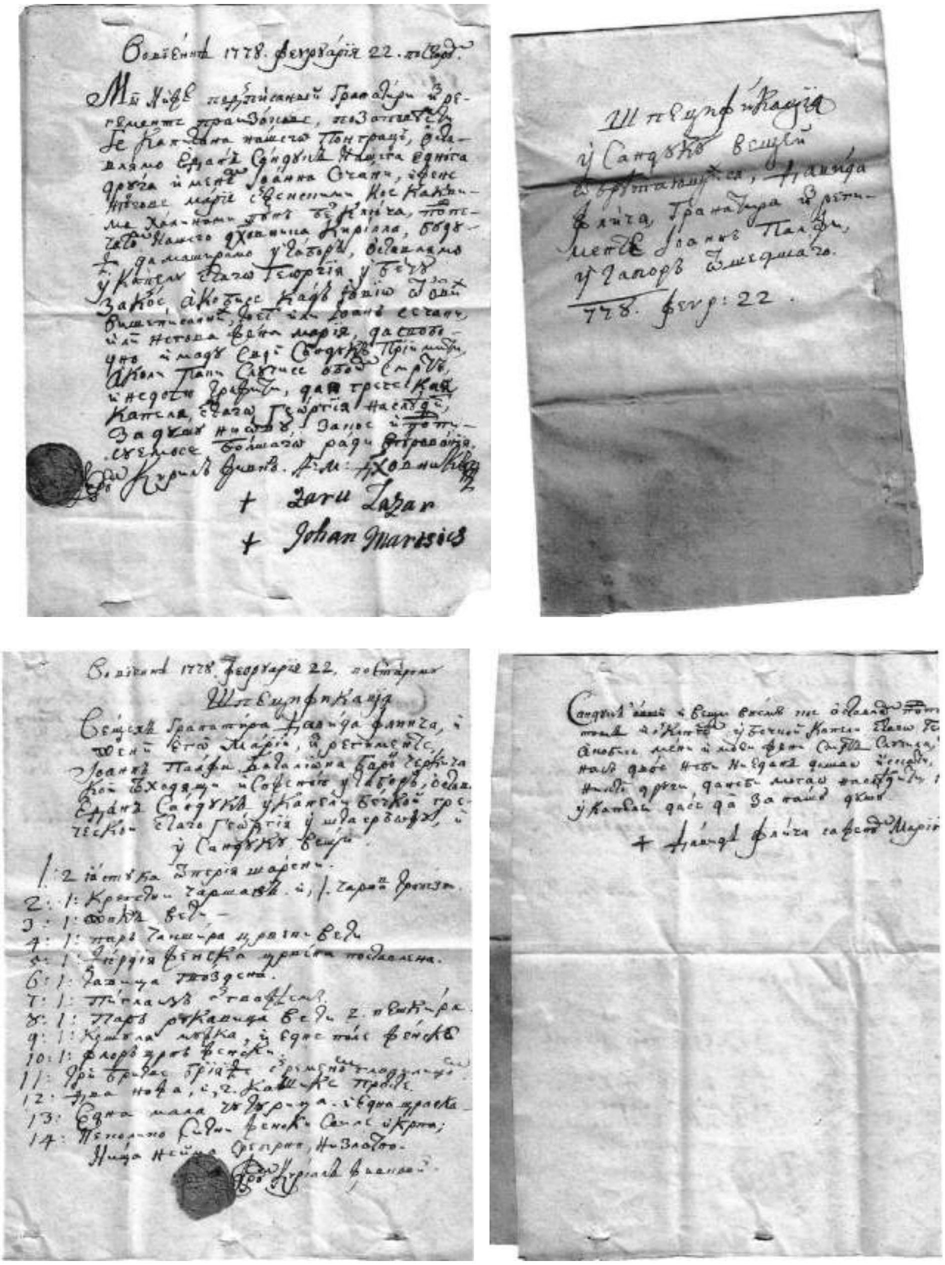


Figure 1. The Original Scanned Will of Miloš Belmužević, dated September 8, 1500. Source (Obradović, 2024)



Въ Віенни 1778. февруарія 22. по Свѣд.

Мы нижеподписанныи Гранадіри Регементе [illegible], позволенїемъ Капитана нашего Понграца, оддавламо едан Сандукъ [illegible] единога друга и мене Іоанна Стани, и жене [illegible] Марїе [illegible] Кап[illegible] [illegible] попеме [illegible] нашего духовника Кирілла, будући да машћирамо у Табору, оставламо у Капели Светаго Георгїя у Бечу закос, ако би се кадъ годъ [illegible] вышеписанни [illegible] Іоанъ Стани [illegible] и негова жена Марїя, да слободно [illegible] Сандукъ [illegible] ако ли [illegible] и недоћи [illegible], да трече Капела Светаго Георгїя наследи, за [illegible] [illegible] радї [illegible]

Кирілъ [illegible] [illegible] Ходни[illegible]

+ Zaru Lazar

+ Johan Marisics

Шпецифікація у Сандуко вещей [illegible], [illegible] Гранадира Регементе Іоанне Палфи, у Табору [illegible]

1778. февр: 22.

Въ Віенни 1778. Февруарїе 22. по Свѣд.

Шпецифікація

Свещах Гранатира [illegible] Іоанна Палфи [illegible] [illegible] Марїи, и регементс [illegible] [illegible] [illegible] [illegible] у Табору, оставламо Едан Сандукъ у Капели Бечкой Греческой Светаго Георгїя у [illegible], и у Сандуку вещи.

1: 2 [illegible] [illegible] шарени.
2: 1: [illegible] [illegible] и, 1. [illegible] [illegible]
3: 1: [illegible] [illegible]
4: 1: пар [illegible] [illegible] [illegible]
5: 1: [illegible] [illegible] [illegible] [illegible]
6: 1: [illegible] [illegible]
7: 1: [illegible] [illegible]
8: 1: Пар рукавица [illegible] 2. [illegible]
9: 1: [illegible] [illegible], и едне [illegible] [illegible]
10: 1: [illegible] [illegible] [illegible]
11: [illegible] [illegible] [illegible] [illegible] [illegible] [illegible]
12: [illegible] нова, [illegible] [illegible] [illegible]
13: Една мала [illegible] и една [illegible]
14: [illegible] [illegible] [illegible] и Крсти;
Ниша нейма [illegible], низлато.

[illegible] Кирілъ [illegible]

Сандукъ [illegible] и вещи [illegible] [illegible] [illegible] у Бечкой Капели Светаго Г[illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] у Капели дасе да [illegible] [illegible]

+ [illegible] [illegible] са [illegible] Марїи

Figure 2. The Original Scanned Document from the Greek Orthodox Church of St. George in Vienna, dated February 22, 1778.

Table 2 provides the results, i.e. named entites of manual annotation of these documents.

Table 2. Distribution of Named Entities in the Evaluation Dataset

Entity	Will of Miloš Belmužević	Church Document
PERS	40	33
LOC	20	9
DEMO	2	0

We have run the previously trained model on these documents separately, and obtained the precision (P), recall (R) and F1 scores displayed in Table 3.

Table 3. Classification Report of NER Model on the Evaluation Dataset

Entity	Will of Miloš Belmužević			Church document		
	P	R	F1	P	R	F1
PERS	0.26	0.33	0.29	0.53	0.24	0.33
LOC	0.17	0.25	0.20	0.22	0.22	0.22
DEMO	0.0	0.0	0.0	0.0	0.0	0.0

Table 4 represents normalized confusion matrix, where the columns P_* (P_P, P_L, P_D) represent the predicted labels, while the rows T_* (T_P, T_L, T_D) represent the true (gold standard) labels (P refers to PERS, L refers to LOC, and D refers to DEMO). T_O refers to all tokens that do not belong to any of the annotated entity types, while P_O refers to all tokens that the model predicted as non-entity.

Table 4. Confusion Matrix on the Evaluation Dataset

Entity	Will of Miloš Belmužević				Church Document			
	P_P	P_L	P_D	P_O	P_P	P_L	P_D	P_O
T_P	13	1	0	26	8	0	0	25
T_L	3	5	0	12	0	2	0	7
T_D	0	1	0	1	0	0	0	0
T_O	34	0	0	2	7	7	0	0

4. Disscusion

During the model development, we encountered 68 errors, primarily resulting from orthographic variation, tokenization mismatch, and annotation inconsistency. A major linguistic challenge stems from pre-modern orthography, which exhibits unstable spelling conventions and variable use of diacritical marks (e.g., аⷣ, уⷪ, иⷦ, а̄, ...). These forms disrupt morphological segmentation, as the same morpheme may appear in several orthographic variants that do not align with modern normalization rules. Such discrepancies hinder both tokenization and the recognition of entity boundaries because standard NLP pipelines are trained on normalized modern orthography and lack mechanisms to account for the historical graphemic inventory of early Cyrillic scripts. Future improvements should thus include orthography-aware tokenization and morphological normalization layers capable of mapping graphemic variants to canonical lemma forms before NER training. For example, the errors can be observed in the following sentences:

[28, 31] по милости бж҃иеи ꙗ стефань влаⷣславь пишꙋ те кнеже дꙋбровьчьскии всѣи ѡпькинѣ (wrongly annotated entity влаⷣславь);

[7, 16] милʿко богдановиⷱа брат муⷪ годоѥ и драгославь а сн҃ь моу прибиль а дѣд моу братохна (wrongly annotated entity богдановиⷱ);

[17, 19] ꙋ га̄ ба̄ наш̄го ис̄ ха̄ (wrongly annotated entity ис̄ ха̄)

Following manual correction of the misaligned entity spans, the model achieved robust results on the development set. However, the model‘s performance on previously unseen documents remained limited, as indicated by the metrics in Table 3 and Table 4. A detailed study was therefore conducted. It is important to note that both the training and validation corpora were fully lower-cased.

In the case of the document from the Greek Orthodox Church, the model is significantly more effective at identifying PERS than LOC entities. It correctly detected given names such as марїе, марїя, zaru, johan, martsis, марїи, гewргїя, and гewргїя. Seven false positives were recorded, with the most notable being the word капитана (“captain”) and нѣгове (“his”), which the model misclassified as personal names. For LOC entities, two true positives were retrieved (вїеннѣ and капели бечкой), whereas seven false positives appear - including печатⷨ, сандꙋкъ and штаеръwфꙋ. The last token was rejected as a true positive solely because of the trailing comma. No DEMO-type entities occurred in this document.

In the case of the Will of Miloš Belmužević, the model exhibited a similar pattern: higher recall for personal names than for toponyms. The model correctly recognised forms such as милѿⷲ, вꙋкоⷨ, матеꙗшь, тимотеи, маркꙋ, прибеновикю, дмитра, ꙗкшикꙗ, дмитрь, степаⷩ, єла, їѡвана, and вꙋкъ. Nevertheless, thirteen false positives were annotated. The most illustrative error was the token беⷧмꙋжевиⷦ, which was mis-labelled because its non-standard characters prevented normalisation; the same mechanism explains the false assignments for маркꙋ and матеꙗ. Tokens such as маⷮреⷷ ("mother") and неговꙋ ("his") were likewise mis-tagged as names. For LOC entities, the model proposed моишꙋ, мꙋнарꙋ, єнⷪво, шаⷲварь, and кокоⷮ, but confusion between PERS and LOC entities was common. In addition, two DEMO-type entities appear as false negatives.

In general, the model performed well on the development set but still struggled with previously unseen material. Across the evaluation dataset, the model identified personal names more reliably than locations, which is expected, given the large number of PERS entities in the training data. Findings underscore the need for domain-specific preprocessing, such as Unicode normalization and character-trained tokenization, as well as stringent annotation guidelines, to improve generalization and robustness on previously unseen documents.

It should be noted that, although the current evaluation corpus is relatively small and therefore insufficient for broader statistical generalization, the results are nonetheless valuable as a proof of concept demonstrating the feasibility of automatic entity recognition in historical Serbian texts. This initial success establishes a solid foundation for subsequent linguistic and computational advancements.

5. Conclusion and Future Work

This paper represents the first end-to-end NER pipeline designed for pre-modern Serbian. By assembling and annotating a historical corpus comprising of the Banjska Chrysobull, Dečani Chrysobull and the 13th-century charters and letters from the Dubrovnik archive, we trained a SpaCy-based model that achieves F1 ≈92%. Evaluation of two previously unseen documents—the document from the Greek Orthodox Church of St. George in Vienna (1778) and the Will of Miloš Belmužević (1500)—confirmed that the system substantially outperforms manual lookup but still struggles when confronted with non-standard orthography.

Despite these challenges, we could conclude that the proposed model enables faster analysis of Serbian medieval documents, thereby advancing historical scholarship while simultaneously enriching digital humanities methodologies.

Future work will focus on expanding the training corpus, particularly to increase the number of entities related to location and demographic data to improve model performance. Additionally, we will stringent annotation guidelines and explore how transformer-based models can be utilized for the NER task in the same domain.

References

[1] Alstola, T., & Svard, S. (2024). Digital Humanities Meets Ancient Languages. U M. Nissinen, & J. Jokiranta (Urednici), *Changes in Sacred Texts and Traditions: Methodological Encounters and Debates* (str. 193-233). Society of Biblical Literature. Preuzeto sa https://cart.sbl-site.org/books/0603116P

[2] Batanovic, V., Ljubešić, N., & Samaradžić, T. (2018). SETimes. SR--a Reference Training Corpus of Serbian. *Proceedings of the Conference on Language Technologies Digital Humanities 2018 (JT-DH 2018)*, (str. 11-17).

[3] Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., & Doucet, A. (2023). Named Entity Recognition and Classification on Historical Documents: A Survey. *ACM Computing Surveys, 56*(2), 1-47. doi:10.1145/3604931

[4] Grković, M. (1983). *Imena u dečanskim hrisovuljama.* Novi Sad: Filozofski fakutet u Novom Sadu, Institut za južnoslovenske jezike.

[5] Ikonić Nešić, M., Petalinkar, S., Škorić, M., & Stanković, R. (2024). BERT downstream task analysis: Named Entity Recognition in Serbian. *Conference on Information Technology and its Applications* (str. 333-347). Cham: Springer Nature Switzerland.

[6] Ivić, P., & Grković, M. (1976). *Dečanske hrisovulje.* Novi Sad: Institut za lingvistiku.

[7] Jehangir, B., Radhakrishnan, S., & Agarwal, R. (2023). A survey on named entity recognition—datasets, tools, and methodologies. *Natural Language Processing Journal, 3*, 100017. doi:10.1016/j.nlp.2023.100017

[8] Kagaba Amina, G. (2024). Digital Humanities: Using Technology to Analyze Cultural Artifacts. *IDOSR Journal of Humanities and Social Sciences, 9*(3), 1-8. doi:10.59298/IDOSRJHSS/2024/93180000

[9] Krstev, C. (1997). *Jedan prilaz informatiekom modeliranju teksta i algoritmi njegove transformacije.* Retrieved from http://elibrary.matf.bg.ac.rs/handle/123456789/4134

[10] Krstev, C., Obradović, I., Utvić, M., & Vitas, D. (2014). A System for Named Entity Eecognition Based on Local Grammar. *Journal of Logic and Computation, 24*(2), 473-489. doi:10.1093/logcom/exs079

[11] Loma, A. (2013). *Toponomija Banjske hrisovulje.* Beograd: Srpska akademija nauke i umetnosti.

[12] Ljubešić, N., & Lauc, D. (2021). Bertić–the Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. *arXiv preprint arXiv:2104.09243*.

[13] Ljubešić, N., Erjavec, T., Batanović, V., Miličević, M., & Samardžić, T. (2025). *ReLDI-NormTagNER-sr 2.1*. Preuzeto sa ReLDI CENTAR ZA JEZIČKE PODATKE: https://github.com/reldi-data

[14] Ljubešić, N., Stupar, M., Jurić, T., & Agić, Ž. (2013). Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene. *Slovenščina 2.0: empirical, applied and interdisciplinary research, 1*(5), 35-57.

[15] Marovac, U., Avdić, A., & Milošević, N. (2023). A survey of resources and methods for natural language processing of serbian language. *arXiv preprint arXiv:2304.05468*. doi:https://arxiv.org/abs/2304.05468

[16] Milligan, I. (2022). *The transformation of historical research in the digital age.* Cambridge University Press. doi:10.1017/9781009026055

[17] Nešić Ikonić, M., Petalinkar, S., Stanković, R., & Utvić, M. a. (2024). SrpCNNeL: Serbian Model for Named Entity Linking. *19th Conference on Computer Science and Intelligence Systems (FedCSIS)* (str. 465-473). IEEE.

[18] Obradović, N. (2024). Testament Miloša Belmuževića. *Mešovita građa, 45*, 35-56. doi:10.34298/IC2473035O

[19] Plöchl, W. M. (1983). Die Wiener orthodoxen Griechen. Eine Studie zur Rechts- und Kulturgeschichte der Kirchengemeinden zum Hl. Georg und zur Hl. Dreifaltigkeit und zur Errichtung der Metropolis von Austria. *Kirche und Recht*. Preuzeto sa https://ixtheo.de/Record/013109626

[20] Polomac, V. (2025). Srpski jezik u Ugraskoj krajem XV veka (na primeru testamenta vojvode Miloša Belmuževića). (M. Kovačević, Ur.) *Aktuelna pitanja istorije srpskog jezika*. Preuzeto sa u štampi

[21] Ransmayr, A. (2018). *Untertanen des Sultans oder des Kaisers: Struktur und Organisationsformen der beiden Wiener griechischen Gemeinden von den Anfängen im 18. Jahrhundert bis 1918.* Vienna : V&R unipress, Vienna University Press. Preuzeto sa https://ucrisportal.univie.ac.at/de/publications/untertanen-des-sultans-oder-des-kaisers-struktur-und-organisation

[22] Seirinidou, V. (2011). *Έλληνες στη Βιέννη (18ος – μέσο 19ου αιωνα)[=Greeks in Vienna, 18th - mid 19th C.].* Αθήνα: Ηρόδοτος.

[23] Stojanović, L. (1929). *Stare srpske povelje i pisma* (T. 1). Beograd: Srpska kraljevska akademija.

[24] Šandrih, B., Krstev, C., & Stanković, R. (2019). Development and Evaluation of Three Named Entity Recognition Systems for Serbian-the Case of Personal Names. U R. Mitkov, & G. Angelova (Ur.), *Proceedings of the International*

Conference on Recent Advances in Natural Language Processing (RANLP 2019) (str. 1060-1068). Varna, Bulgaria: INCOMA Ltd. doi:10.26615/978-954-452-056-4_122

[25] Tanasijević, I. (2019). Toward Automatic Tagging of Cultural Heritage Documents. *IPSI Transactions on Advanced Research, TAR, 15*(1).

[26] Todorović Šandrih, B., Krstev, C., Stanković, R., & Nešić Ikonić, M. (2021). Serbian NER&Beyond: The Archaic and the Modern Intertwinned. U R. Mitkov, & G. Angelova (Ur.), *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (str. 1252-1260). Held Online: INCOMA Ltd. Preuzeto sa https://aclanthology.org/2021.ranlp-1.141/

[27] Trifunović, Đ. (2011). Charter of King Milutin to the Banjska Monastery : The Chrysobull of St. Stephen. Službeni glasnik, Srbija. Retrieved from https://books.google.at/books?id=ZjQeMwEACAAJ

[28] Turczynski, E. (1959). Die Deutsch-Griechischen Kulturbeziehungen Bis Zur Berufung KöNig Ottos. (F. Valjavec, Ur.) *Südosteuropäische Arbeiten, 48*.

[29] Vitas, D., & Pavlović-Lažetić, G. (2008). Resources and Methods for Named Entity Recognition in Serbian. *INFOtheca - Journal of Informatics & Librarianship, 9*(1-2), 35a.

Prepoznavanje imenovanih entiteta za istorijske varijetete srpskog jezika: preliminarna studija

Marija Đokić Petrović,
Mihailo St. Popović,
Vladimir Polomac

U ovom radu predstavljen je prvi model za prepoznavanje imenovanih entiteta za istorijske varijetete srpskog jezika. Model je treniran na ručno anotiranom istorijskom korpusu koji obuhvata povelju Kralja Uroša II manastiru Banjska (Banjska hrisovulja), trećem primerku povelje Kralja Stefana Uroša III Nemanjića manastiru Visoki Dečani (Dečanska hrisovulja) i skupu povelja i pisama iz dubrovačkog arhiva iz XIII veka. Anotacije su obuhvatile tri kategorije entiteta – lična imena (PERS), toponime (LOC) i demografske reference (DEMO) – pri čemu su ukupno zabeležena 9172 entiteta u 6715 rečenica. Konkretno, 7562 entiteta tipa PERS, 1525 entiteta tipa LOC i 85 entiteta tipa DEMO. Broj anotiranih rečenica je iznosio 4112.

Za obuku modela je primenjena Convolutional Neural Network arhitektura u okviru spaCy-ja. Posle 15 epoha model je na razvojnom skupu postigao preciznost (P) 93,11%, odziv (R) 91,08 % i F1 92,08%.

Evaluacija je izvršena nad dva dokumenta: Testamentu Miloša Belmuževića iz 1500. godine i do sada neobjavljenom dokumentu koji se nalazi u arhivi grčke pravoslavne crkve Svetog Đorđa u Beču, a koji datira iz 1778. godine. Na ovom setu performanse modela su opale: F1 rezultati za PERS, LOC i DEMO entitete variraju od 0% do 33%, uz izraženo bolju tačnost za lična imena nego za toponime.

Detaljna analiza 68 uočenih grešaka prilikom procesa treniranja modela, pokazala je da su najčešći uzroci lošeg rada modela ortografska kompleksnost i retke dijakritike koje remete tokenizaciju, što dovodi do neusklađenih opsega anotacija. Rešavanjem problema pogrešno anotiranih entiteta značajno je poboljšana stabilnost treninga i metrika na razvojnom skupu, ali su problemi ostali prisutni na tekstovima koji nisu poznati modelu.

Ovaj rad doprinosi digitalnim humanističkim istraživanjima i razvoju alata za automatsku obradu srpskih srednjovekovnih izvora, postavljajući osnovu za buduće studije. Nalazi ukazuju na potrebu za proširenjem korpusa, Unicode normalizacijom, karakter-baziranom tokenizacijom, strožim smernicama za anotaciju i primenom transformer modela radi bolje generalizacije nepoznatih dokumenata.

Ključne reči: prepoznavanje imenovanih entiteta; istorijski varijeteti srpskog jezika; digitalne humanističke nauke; kulturno nasleđe.