

Образац 1 – Пријава теме докторске дисертације
(Кандидат доставља пријаву Програмском савету)
Студијски програм: **Интелигентни системи**
(назив студијског програма)

Кандидат: Милица Иконић Нешић, мастер математике

Молим да ми се одобри тема за израду докторске дисертације под насловом:

**„Квантитативна и семантичка анализа књижевних текстова на српском језику
заснована на машинском учењу и језичким моделима“**

1. Биографија кандидата

Милица Иконић Нешић је рођена 27. маја 1991. године у Београду. Основну школу је завршила у Барајеву, као носилац Вукове дипломе. Завршила је Математичку гимназију у Београду са просечном оценом 4,95. Дипломирала је на Математичком факултету Универзитета у Београду, на смеру теоријска математика и примена, 29. 9. 2018. године, са просечном оценом 9,27. Степен мастер математике стекла је 29. 9. 2020. године одбраном рада „Мартингални H_p простори”, под менторством професора др Драгољуба Кечкића (просечна оцена на мастер академским студијама 10,00). Уписала је Мултидисциплинарне докторске студије Универзитета у Београду, модул Интелигентни системи 2020/21 школске године (две године мировања због породилског одсуства). На докторским студијама је положила све испите и одбранила приступни рад у року (за три године), са просечном оценом 10,00.

Положени испити на докторским студијама

| Акроним | Назив | Оцена |
|---------|---|-------|
| ИСТЕ4 | Рачунарска визија | 10 |
| ИСВИ7 | Машинско учење | 10 |
| ИСПР1 | Семантички Web | 10 |
| ИСТЕ2 | Процесирање природног језика | 10 |
| ИСВИ3 | Интелигентно претраживање | 10 |
| ИС001 | Методе и технике вештачке интелигенције | 10 |
| ИСПР4 | Екстракција информација из текста | 10 |
| ИСПР3 | Интелигентна анализа података | 10 |
| ИСПР6 | Интелигентни едукативни системи | 10 |
| ИС011 | Израда и одбрана Приступног рада за докторску дисертацију | 10 |

| Предмет | Оцена | Шифра предмета | ЕСПБ |
|---|-------|----------------|------|
| Методе и технике вештачке интелигенције | 10 | ИС001 | 12 |
| Интелигентна анализа података | 10 | ИСПР3 | 9 |
| Процесирање природног језика | 10 | ИСТЕ2 | 10 |
| Екстракција информација из текста | 10 | ИСПР4 | 11 |
| Интелигентни едукативни системи | 10 | ИСПР6 | 11 |
| Машинско учење | 10 | ИСВИ7 | 11 |

| | | | |
|---|------|-------|-----|
| Семантички Веб | 10 | ИСПР1 | 9 |
| Интелигентно претраживање | 10 | ИСВИЗ | 9 |
| Рачунарска визија | 10 | ИСТЕ4 | 10 |
| Израда и одбрана Приступног рада за докторску дисертацију | 10 | ИС010 | 30 |
| Просечна оцена/Укупно | 10.0 | | 122 |

Од 2018. до 2020. године радила као сарадник у настави на Математичком факултету Универзитета у Београду за предмете основних академских студија *Анализа 1* на Катедри за анализу и *Геометрија 2* на Катедри за геометрију. Запослена је на Филолошком факултету Универзитета у Београду од фебруара 2021. године као асистент на Катедри за библиотекарство и информатику за предмете *Информатички практикум 2*, *Информатички практикум 3*, *Дигитални текст 1*, *Дигитални текст 2*, *Структура информација 1*, *Структура информација 2*, *Језичке технологије 1*, *Језичке технологије 2*, *Мултимедијални документи*, *Проналажење информација* (основне академске студије), *Напредне методе у проналажењу информација*, *Напредне језичке технологије*, *Дигитална хуманистика и библиотеке - ресурси и алати*, *Технологије семантичког веба* (мастер академске студије).

У периоду од 1. марта 2023. до 1. маја 2024. године користила је породилско одсуство и одсуство са рада ради неге детета.

У својству истраживача, ангажована је до сада:

1. на националном научном истраживачком пројекту *TESLA - Text Embeddings – Serbian Language Applications*, који финансира Фонд за науку Републике Србије, #7276, почев од 1. 5. 2024. године, <https://tesla.rgf.bg.ac.rs/>;
2. у COST акцији „Distant Reading for European Literary History“ (COST Action CA16204), 2021–2024, <https://www.distant-reading.net/>;
3. на међународном истраживачком пројекту „It-Sr-NER: CLARIN Compatible NER and Geoparsing Web Services for Parallel Texts: Case Study Italian and Serbian“, од 1. 6. до 30. 9. 2022. године, <https://github.com/jerteh/It-Sr-NER>;
4. у COST акцији „Global Network on Large-Scale, Cross-domain and Multilingual Open Knowledge Graphs“ CA23147.

Такође, члан је и:

1. Уредничког одбора часописа *Journal of Natural Language Processing*, <https://www.cambridge.org/core/journals/natural-language-engineering>;
2. Друштва за језичке ресурсе и технологије - JePTeX (<https://jerteh.rs>).
3. Организационог одбора међународне конференције RANLP 2021, Бугарска, <https://ranlp.org/ranlp2021/contacts.php>;
4. Организационог одбора међународне конференције South Slavic Languages in the Digital Environment - JuDig, Филолошки факултет, Београд, 2024 и 2026, <https://judig.jerteh.rs/inner-page.php?p=organizing-committee>.
5. Организационог одбора међународне конференције RANLP 2025, Бугарска, <https://ranlp.org/ranlp2025/contacts.php>;
6. Организационог одбора међународне конференције LaTeLL, 2026, Фес, Мароко, <https://latell.org/2026/>

Такође, ангажована је као наставни асистент (енгл. *teaching assistant*) на међународној летњој школи „The Paradigm Shift Summer School 2026“, у организацији GPLSI истраживачке групе Универзитета у Аликантеу (Шпанија), повезане са конференцијом NLPAICS 2026, <https://summer-school.gplsi.es/teaching-assistants/>.

2. Библиографија кандидата

(категорисано према категоризацији надлежног Министарства, објављени или прихваћени за штампу)

I. Саопштења са међународног скупа штампана у целини (M14):

Крстев, Ц., Станковић, Р., Шандрих Тодоровић, Б., & **Иконић Нешић, М.** (2023). Нове технологије за оживљавање старих текстова. У: Зборник радова Међународне научне конференције *Дигитална хуманистика и словенско културно наслеђе II*, Београд, Србија, 28–29. јун 2021. Београд: Савез славистичких друштава Србије. <https://enauka.gov.rs/handle/123456789/854220>; <http://dr.rgf.bg.ac.rs/s/repo/item/8415>.

II. Радови у часописима међународног значаја (M20)

Škorić, M., Stanković, R., **Ikonić Nešić, M.**, Byszuk, J., & Eder, M. (2022). Parallel Stylometric Document Embeddings with Deep Learning Based Language Models in Literary Authorship Attribution. *Mathematics*, 10(5), 838. <https://doi.org/10.3390/math10050838> M21a

Petalinkar, S., Stanković, R. M., & **Ikonić Nešić, M.** (2025). Comparative Analysis of Methods for Creating a Sentiment Lexicon of the Serbian WordNet. *The Electronic Library*, 43(4), 547–577. <https://doi.org/10.1108/EL-08-2024-0253> M22

Ikonić Nešić, M., Petalinkar, S., Kitanović, O., Stanković, R., & Utvić, M. (2026, in press). CNN-based Named Entity Linking: Serbian Use Case. *Poznan Studies in Contemporary Linguistics* (PSICL) M22

Stanković, R., Vučenović, T., & **Ikonić Nešić, M.** (2026, in press). Encoding, Linking, Retrieving: A Methodological Framework for Knowledge-Enriched Interview Corpora. *Computer Science and Information Systems*. M22

III. Саопштења са међународног скупа штампана у целини (M33):

- Stanković, R., Vučenović, T., Rujević, B., **Ikonić Nešić, M.**, & Škorić, M. (2026). Integrating TEI, NER/NEL, Textometry, and Linked Data for a Semantically Enriched Interview Corpus. In Proceedings of the 15th Language Resources and Evaluation Conference (LREC 2026). <https://lrec2026.info/list-of-accepted-papers/>
- Ikonić Nešić, M.**, Petalinkar, S., Stanković, R., & Mitkov, R. (2025). From zero to hero – Serbian NER from rules to LLMs. In Proceedings of the First Workshop on Comparative Performance Evaluation: From Rules to Language Models (R2LM 2025) associated with RANLP 2025 (pp. 87–96). <https://acl-bg.org/proceedings/2025/R2LM%202025/pdf/2025.r2lm-1.10.pdf>
- Stanković, R., Janković, N., Rađenović, J., & **Ikonić Nešić, M.** (2025). From LLM generation to knowledge representation: Creating and structuring the Gemini Knowledge-sr QA dataset for Serbian. In The 1st GOBLIN Workshop on Knowledge Graph Technologies. Leipzig, Germany. <https://dr.rgf.bg.ac.rs/files/original/6703fca99da6bf724e75716902d3bcb3ee53f2c2.pdf>
- Ikonić Nešić, M.**, Petalinkar, S., Škorić, M., & Stanković, R. (2024). BERT downstream task analysis: Named entity recognition in Serbian. In *Conference on Information Technology and its Applications* (pp. 333–347). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-71419-1_29
- Ikonić Nešić, M.**, Petalinkar, S., Škorić, M., Stanković, R., & Rujević, B. (2024). Advancing sentiment analysis in Serbian literature: A zero and few-shot learning approach using the Mistral model. In Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024) (pp. 58–70). <https://dr.rgf.bg.ac.rs/s/repo/item/8805>
- Ikonić Nešić, M.**, Petalinkar, S., Stanković, R., Utvić, M., & Kitanović, O. (2024). SrpCNeL: Serbian model for named entity linking. In Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) (pp. 465–473). IEEE. <https://doi.org/10.15439/2024F8827>
- Mihajlov, T., **Ikonić Nešić, M.**, Stanković, R., & Kitanović, O. (2024). Topic modeling of the srpELTeC corpus: A comparison of NMF, LDA, and BERTopic. In Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) (pp. 649–653). IEEE. <https://doi.org/10.15439/2024F1593>
- Stanković, R., **Ikonić Nešić, M.**, Perišić, O., Škorić, M., & Kitanović, O. (2024). Towards semantic interoperability: Parallel corpora as linked data incorporating named entity linking. In The 9th Workshop on Linked Data in Linguistics (LDL-2024) @ LREC-COLING-2024 (pp. 115–125). ELRA Language Resources Association. <https://aclanthology.org/2024.ldl-1.15.pdf>
- Perišić, O., Stanković, R., **Ikonić Nešić, M.**, & Škorić, M. (2023). It-Sr-NER: CLARIN compatible NER and geoparsing web services for Italian and Serbian parallel text. In Selected Papers from the CLARIN Annual Conference 2022 (pp. 99–110). Linköping University Electronic Press. <https://doi.org/10.3384/ecp198010>

Ikonić Nešić, M., Stanković, R., Schöch, C., & Škorić, M. (2022). From ELTeC text collection metadata and named entities to Linked-Data (and back). In Proceedings of the 8th Workshop on Linked Data in Linguistics (LDL-2022) within the 13th Language Resources and Evaluation Conference (pp. 7–16). <https://aclanthology.org/2022.ldl-1.2.pdf>

Stanković, R., Košprdić, M., **Ikonić Nešić, M.**, & Radović, T. (2022). Sentiment analysis of Serbian old novels. In Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data (SALLD-2) (pp. 31–38). <https://aclanthology.org/2022.salld-1.6.pdf>

Stanković, R., Krstev, C., Šandrih Todorović, B., Vitas, D., Škorić, M., & **Ikonić Nešić, M.** (2022). Distant reading in digital humanities: Case study on the Serbian part of the ELTeC collection. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022) (pp. 3337–3345). European Language Resources Association. <https://aclanthology.org/2022.lrec-1.356.pdf>

Šandrih Todorović, B., Krstev, C., Stanković, R., & **Ikonić Nešić, M.** (2021). Serbian NER&beyond: The archaic and the modern intertwined. In Deep Learning Natural Language Processing Methods and Applications – Proc. of the Int. Conf. Recent Advances in Natural Language Processing (RANLP 2021) (pp. 1252–1260). <https://dr.rgf.bg.ac.rs/s/repo/item/5139>

IV. Саопштење са међународног скупа штампано у изводу (M34):

Krstev, C., Stanković, R., Marković, A., & **Ikonić Nešić, M.** (2025). Progress in SR-ELEXIS semantic annotation: Focusing on multiword expressions, named entities, and sense repository. UniDive 3rd General Meeting, Hungarian Research Centre for Linguistics, Budapest, Hungary, 29–30 January 2025. European Cooperation in Science and Technology. <https://doi.org/10.5281/zenodo.14845204>

Stanković, R., Chiarcos, C., & **Ikonić Nešić, M.** (2024) Leveraging Linked Data NIF, and CONLL-U for Enhanced Annotation in Sentence Aligned Parallel Corpora. *UniDive 2nd general meeting*, University of Naples L'Orientale, Italy. <https://doi.org/10.5281/zenodo.11208944>

V. Радови у националном часопису (M53):

Ikonić Nešić, M. (2026). Analysis of family relationships of characters in the srpELTeC corpus based on semantic web techniques and textometry. *Infotheca — Journal for Digital Humanities*, 25(1-2), 7–40.
https://doi.org/10.18485/infotheca.2025.25.1_2.1 https://infoteka.bg.ac.rs/ojs/index.php/InfoTek/article/view/2025.25.1_2.1_en

Perišić, O., Stanković, R., **Ikonić Nešić, M.**, & Škorić, M. (2023). It-Sr-NER: Web Services for Recognizing and Linking Named Entities in Text and Displaying Them on a Web Map. *Infotheca — Journal for Digital Humanities*, 23(1), 61-77.
<https://doi.org/10.18485/infotheca.2023.23.1.3>
<https://dr.rgf.bg.ac.rs/s/repo/item/7790>

Ikonić Nešić, M., Stanković, R. & Rujević, B. (2021). Serbian ELTeC Sub-Collection in Wikidata. *Infotheca — Journal for Digital Humanities*, 21(2).
<https://doi.org/10.18485/infotheca.2021.21.2.4>
<https://repff.fil.bg.ac.rs/handle/123456789/1428>

VI. Саопштења са скупа националног значаја штампана у изводу (M64)

Stanković, R., Vučenović, T., & **Ikonić Nešić, M.** (2025). From unstructured interviews to queryable knowledge: An AI pipeline for the “Digitalne Ikone” corpus using TEI, NER, NEL, and RAG. Book of Abstracts, Artificial Intelligence Conference, Belgrade, Serbia, 9–10 October 2025, pp. 151–152.
https://www.mi.sanu.ac.rs/~ai_conf/2025/AI_Conference_Book_of_Abstracts.pdf

Ikonić Nešić, M., Petalinkar, S., & Stanković, R. (2025). Towards linguistic completeness in knowledge graphs: Generating Serbian inflections with language models. Book of Abstracts, Artificial Intelligence Conference, Belgrade, Serbia, 9–10 October 2025, pp. 129–130. https://www.mi.sanu.ac.rs/~ai_conf/2025/AI_Conference_Book_of_Abstracts.pdf

Stanković, R., Petalinkar, S., & **Ikonić Nešić, M.** (2026). "From Names to Places: Serbian Entity Linking Through the Lens of GIS". PROCEEDINGS OF THE 1ST ReLDI CONFERENCE ON LANGUAGE SCIENCE AND TECHNOLOGY, 25-26 September 2025.
<https://doi.org/10.5281/zenodo.18488312> M34

Petalinkar, S., **Ikonić Nešić, M.**, Stanković, R., & Škorić, M. (2024). *Comparison of Zero and Few-Shot Learning Approaches Using LLMs for Sentiment Analysis in Serbian Literature*. In: **Book of Abstracts of the Artificial Intelligence Conference** (pp. 55–56). Belgrade: Serbian Academy of Sciences and Arts (SASA).
https://www.mi.sanu.ac.rs/~ai_conf/previous_editions/2024/AI_Conference_Book_of_Abstracts.pdf

Ikonić Nešić, M., Petalinkar, S., & Stanković, R. (2024). *Development and Evaluation of Named Entity Linking Models for Serbian Language with Wikidata Integration*. In: **Book of Abstracts of the Artificial Intelligence Conference** (pp. 47–48). Belgrade: Serbian Academy of Sciences and Arts (SASA).
https://www.mi.sanu.ac.rs/~ai_conf/previous_editions/2024/AI_Conference_Book_of_Abstracts.pdf

Ikonić Nešić, M., & Utvić, M. (2024). *Overview of the Tesla-Ner-Nel-Gold Dataset: Showcase*

on Serbian-English Parallel Corpus. In *Book of Abstracts of the International Conference South Slavic Languages in the Digital Environment (JuDig)*, 21–23 November 2024, Belgrade, Serbia (p. 57). <https://judig.jerteh.rs/2024/judig-book-of-abstracts.pdf>

Milinković, M., & **Ikončić Nešić, M.** (2024). *Named Entities in the Digital Corpus of Spatial Plans/Planning*. In *Book of Abstracts of the International Conference South Slavic Languages in the Digital Environment (JuDig)*, 21–23 November 2024, Belgrade, Serbia (pp. 33–34). <https://judig.jerteh.rs/2024/judig-book-of-abstracts.pdf>

3. Предмет и циљеви докторске дисертације

3.1. Предмет докторске дисертације

Последњих година област обраде природног језика (енгл. *Natural Language Processing*, скр. NLP) бележи интензиван развој, пре свега захваљујући напретку у машинском учењу, дубоком учењу и развоју великих језичких модела (Grattafiori et al. 2024; Achiam et al. 2023; Devlin et al. 2019; Škorić 2024; Vreš et al. 2024). Ови приступи омогућавају аутоматску анализу, интерпретацију и обраду великих количина неструктурисаног текста, што их чини значајним за бројне примене, у претраживању и екстракцији информација, анализи текстуалних збирки, дигиталној хуманистици и другим областима у којима је неопходна обрада сложених језичких података. Ипак, већина доступних ресурса и система развијана је првенствено за језике са богатим ресурсима, док српски језик упркос великим напорима заједнице и даље има мањак јавно доступних аотираних (ручно верификованих) корпуса, отворених модела и семантичких ресурса за задатке машинског разумевања. Ова докторска дисертација ублажава те недостатке развојем модела за препознавање именованих ентитета, њихову класификацију и повезивање са базама знања. Поред тога, предложени приступ интегрише ове компоненте са текстометријом, анализом сентимента и емоција, као и моделирањем тема, специфично за књижевни корпус, уз проверу преносивости оквира на корпуре другачијег жанра.

Аутоматска анализа књижевних текстова на српском језику подразумева квантитативну и семантичку анализу текста. Квантитативна, односно текстометријска анализа, обухвата основну обраду текста и израчунавање показатеља за описивање лексичких и структурних својстава корпуса, као што су број токена и типова, дужина речи, реченица и пасуса, расподела фреквенција лексичких јединица и сличне мере. Такви показатељи представљају важну основу за даљу примену напреднијих метода обраде природног језика и машинског учења (Јаџић 2019). Под семантичком анализом текста подразумевају се поступци који омогућавају аутоматско издвајање, повезивање и тумачење значења релевантних елемената текста. У том смислу, семантичка анализа обухвата препознавање именованих ентитета (енгл. *Named Entity Recognition*, скр. NER), њихово повезивање са базама знања (енгл. *Named Entity Linking*, скр. NEL), анализу сентимента и емоција, као и екстракцију семантичких релација међу ентитетима. Интеграција ових комплементарних приступа омогућава да се са установљених лексичких и структурних параметара пређе на дубинску интерпретацију текста, чиме се поставља методолошки оквир за свеобухватно дигитално моделовање књижевног дела.

Препознавање именованих ентитета, као и повезивање са базама знања, представља један од важних предуслова за прецизније аутоматско разумевање текста и подршку другим NLP задацима. За српски језик, један од првих система у овој области био је *SrpNER*, заснован на ручно дефинисаним правилима и ослоњен на лексичке ресурсе (Krstev 2008;

Krstev et al. 2014; Krstev et al. 2018), првенствено намењен обради новинских текстова. У обради српских романа, корпуса *SrpELTeC* (Krstev and Stanković 2021, Stanković et al. 2024), први пут објављени у периоду 1840–1920, показало се да адаптирана верзија система *SrpNER*, прилагођена специфичностима књижевног текста, даје боље резултате од тадашњих модела заснованих на машинском учењу, попут *SrpCANNER* (Stanković et al. 2021; Šandrih Todorović et al. 2021). Даљи развој трансформерских архитектура поставио је нове циљеве у обради природног језика, али је истовремено нагласио потребу за проширењем постојећих аотираних скупова текстова (Stanković et al. 2021). Док је задатак препознавање именованих ентитета имао одређену традицију развоја за српски језик, задатак повезивања именованих ентитета са базама знања до сада није био предмет систематског истраживања, нити је за ову намену постојао адекватан језички ресурс.

Поред препознавања и повезивања именованих ентитета, у овом раду се обрађују и друге компоненте семантичке обраде релевантне за књижевне текстове. У домену анализе сентимента посебан значај имају истраживања која показују да и за српски језик постоје одређена решења за аутоматско одређивање поларитета текста (Batanović and Nikolić 2017; Mladenović et al. 2015; Stanković et al. 2022a). Са становишта анализа тема, моделирање тема представља важан поступак за откривање доминантних тематских структура у већим корпусима текста, од класичних статистичких модела, као што је LDA (енгл. *Latent Dirichlet Allocation*, скр. LDA), до новијих метода заснованих на векторским репрезентацијама текста и трансформерским моделима, као што је BERTopic (Blei, Ng, and Jordan 2003; Grootendorst 2022; Chu et al. 2022). Екстракција семантичких релација у књижевним текстовима често је усмерена на идентификацију и анализу односа међу ликовима, који се сматрају једним од кључних елемената наративне структуре (Cipresso and Riva 2016; Prado et al. 2016; Radak et al. 2024). Неки од приступа заснивају се на представљању ових односа кроз графове ликова, у којима се ликови моделирају као чворови, а њихове интеракције као гране графа, што омогућава формалније проучавање система ликова у наративу (Moretti 2011; Lee and Yeung 2012). Поред тога, развијени су и приступи који аутоматски издвајају конкретне типове односа, као што су породични односи, коришћењем комбинације правила, система за препознавање именованих ентитета и техника за разрешавање анафоре (Santos et al. 2010; Makazhanov et al. 2014; Manzoor 2022). На општем плану, екстракција семантичких релација развијала се од класичних приступа заснованих на ручно дефинисаним правилима и обележјима (Zhou et al. 2005; Baker, Fillmore, and Lowe 2003), преко прегледа техника заснованих на машинском учењу (Zhang, Chen, and Liu 2017), до новијих приступа који користе велике језичке моделе у режиму без претходних примера у задацима за језике с ограниченим ресурсима (Han, Liang, and Wang 2025).

Предмет ове докторске дисертације је моделовање и квантификовање разноврсних појава које доприносе семантичкој анализи и екстракцији информација из књижевних текстова српског језика. У овом истраживању семантичка анализа обухвата препознавање сентимента у тексту (позитивно, негативно, неутрално), емоција (радост, поверење, страх, туга, изненађеност, гађење, љутња и ишчекивање) (Šošić et al. 2026), моделовање тема, као и посебан фокус на препознавању и повезивању именованих ентитета са базама знања. Истраживање се заснива на примени савремених модела машинског учења, пре свега трансформерских архитектура и великих језичких модела, уз повезивање са структурираним изворима знања и формалним моделима репрезентације података. Посебна пажња биће посвећена примени ових поступака на књижевне корпуре српског језика, као и питањима интероперабилности са постојећим ресурсима, базама знања и графовима повезаних података (енгл. *Linked Open Data*, скр. LOD) (Stanković et al. 2023), што кулминира развојем истраживачке платформе за свеобухватну текстометријску и квантитативну анализу књижевних дела опремљених богатим семантичким аотацијама.

Иако је тежиште истраживања на корпусу srpELTeC као примарном књижевном корпусу, кључни делови методолошког оквира, пре свега TEI кодирање, текстометријска анализа, препознавање и повезивање именованих ентитета, и NIF (енгл. NLP Interchange Format, скр. NIF) серијализација, биће паралелно примењени и на корпусу другог жанра. Тиме се не само квантитативно процењује преносивост модела за препознавање и повезивање именованих ентитета са књижевног на други жанр, већ се потврђује методолошка целовитост и поновна употребљивост предложеног оквира на различитим типовима текстова на српском језику (Stanković et al. 2026; Krstev et al. 2025).

3.2. Циљеви докторске дисертације

Општи циљ докторске дисертације јесте развој методологије за квантитативну и семантичку анализу књижевних текстова на српском језику применом језичких ресурса и модела машинског учења, са посебним фокусом на развој и примену модела за препознавање и повезивање именованих ентитета. Тако постављена методологија треба да обезбеди теоријску и техничку основу за идентификацију, структурисање, квантитативно описивање и семантичко повезивање језичких појава у књижевним текстовима, као и да допринесе ширем развоју технологија обраде природног језика за српски језик.

Специфични циљеви укључују:

- Изградњу и проширење језичких ресурса за српски језик, то јест, проширење и анотацију корпуса са значајним уделом књижевних текстова за потребе задатака препознавања и повезивања именованих ентитета.
- Развој, обучавање и фино подешавање модела заснованих на трансформерским архитектурама за задатке препознавања именованих ентитета и повезивања именованих ентитета са базама знања.
- Примена језичких модела у задатку повезивања именованих ентитета са отвореном базом знања Википодаци (енгл. Wikidata) (Vrandečić and Krötzsch 2014).
- Развој поступака квантитативне, односно текстометријске анализе, у циљу описивања лексичких и структурних својстава књижевних корпуса и припреме података за даље аналитичке поступке;
- Примену лексичких ресурса и језичких модела у задацима анализе сентимента, емоција и моделирања тема у књижевним текстовима на српском језику.
- Аутоматску екстракцију семантичких релација у тексту, укључујући релације релевантне за анализу књижевних садржаја, као што су породични односи и друге семантички значајне повезаности међу ентитетима.
- Развој процедура за репрезентацију резултата језичке анализе у NIF, ради стандардизоване RDF репрезентације текста, анотација и њихове интеграције са графовима знања и повезаним лингвистичким подацима (енгл. *Linguistic Linked Open Data*, скр. LLOD).

- Евалуацију појединачних компоненти и интегрисаног система кроз квантитативну и квалитативну анализу, са циљем утврђивања домета, ограничења и применљивости предложеног приступа у анализи књижевних текстова на српском језику.
- Интеграцију развијених ресурса, анотација и резултата анализе у јединствену истраживачку платформу ради њиховог представљања, визуализације и даље употребе, користећи корпус srpELTeC као пример студије случаја.
- Евалуацију преносивости развијених модела са примарног књижевног домена на текстове другог типа, ради потврде шире применљивости методолошког оквира на различите регистре српског језика.

Крајњи циљ истраживања јесте развој ефикасног, проширеног и методолошки утемељеног система који интегрише више компоненти за обраду природног језика и омогућава сложенију аутоматску анализу и визуализацију екстрахованих информација и семантичке анализе књижевних текстова на српском језику.

4. Хипотезе

Основна хипотеза докторске дисертације јесте да интеграција савремених метода машинског учења, језичких модела и структурираних извора знања унапређује прецизност и поузданост аутоматске идентификације, повезивања и анализе семантички релевантних појава у корпусима књижевних текстова на српском језику, чиме се ствара основа за напредније облике аутоматске квантитативне и семантичке анализе књижевности.

Из ове основне хипотезе произилазе следеће посебне хипотезе:

- Примена савремених модела заснованих на трансформерским архитектурама омогућава значајно већу тачност у задатку препознавања именованих ентитета у односу на системе засноване на конволуционим неуронским мрежама.
- Повезивање именованих ентитета са отвореним базама знања, као што су Википодаци, омогућава обogaћивање текстуалне анализе структурираним семантичким информацијама и подржава прецизнију интерпретацију односа међу ентитетима у књижевним текстовима.
- Комбиновање језичких ресурса, лексикона и великих језичких модела може допринети успешнијем решавању задатака анализе сентимента, емоција, моделирања тема и екстракције релација у корпусима српског језика.
- Примена квантитативних приступа заснованих на текстометрији (Stanković et al. 2022b) омогућава уочавање стилских, тематских и семантичких образаца у књижевним корпусима које није лако уочити традиционалним интерпретативним читањем.
- Употреба формализованих и међусобно повезаних репрезентација података, укључујући LLOD и NIF, доприноси интероперабилности језичких ресурса за српски језик и олакшава њихово укључивање у шире екосистеме повезаних података и знања.
- Модел за препознавање и повезивање именованих ентитета обучени претежно на књижевном корпусу, као и предложени методолошки оквир, могу се успешно

применити и на сродне жанрове.

5. План рада

Истраживање у оквиру предложене докторске дисертације извршиће се у три фазе и састојаће се од следећих корака:

Фаза припреме

- (i) Изучавање релевантне научне литературе из области обраде природног језика, дигиталне хуманистике и квантитативне анализе књижевних текстова.
- (ii) Анализа савремених приступа у задацима препознавања именованих ентитета, повезивања ентитета са базама знања, анализе сентимента, емоција, екстракције релација и моделирања тема.
- (iii) Преглед постојећих језичких ресурса за српски језик, укључујући постојеће корпусе, лексиконе и векторске репрезентације речи и реченица.
- (iv) Анализа постојећих метода текстометријске анализе корпуса.
- (v) Дефинисање анотационе шеме, критеријума за избор корпуса, као и корпуса за студију случаја и методолошког оквира истраживања.

Фаза развоја

- (vi) Анотација корпуса текстова на српском језику као ресурса за задатке препознавања и повезивања именованих ентитета.
- (vii) Развој, обучавање и фино подешавање модела за препознавање именованих ентитета.
- (viii) Развој система за повезивање именованих ентитета (пре свега локација) са базама знања, у овом случају Википодаци.
- (ix) Испитивање примена савремених великих језичких модела у задацима машинског разумевања текста.
- (x) Развој поступака за квантитативну, односно текстометријску анализу корпуса;
- (xi) Развој компоненте за семантичку анализу у мери неопходној за проверу шире применљивости предложеног оквира.
- (xii) Развој процедура за конвертовање резултата у NIF формат и публикавање добијених анотација у том формату на SPARQL приступној тачки.
- (xiii) Пројектовање и имплементација истраживачког портала, укључујући развој модула за приказ корпуса, анотација и резултата анализе, као и подршку за два режима рада: преглед постојећих анотација и креирање нових анотација.

Финална фаза

- (xiv) Евалуација развијених модела и метода применом одговарајућих

- квантитативних метрика за семантичку анализу.
- (xv) Квалитативна анализа добијених резултата и њихова интерпретација у контексту анализе књижевних текстова.
 - (xvi) Компаративна анализа различитих приступа, укључујући класичне статистичке методе, вештачке неуронске моделе и велике језичке моделе;
 - (xvii) Квантитативно профилисање језичких и семантичких појава у корпусима, текстометријским приступом.
 - (xviii) Дискусија резултата и формулисање закључака о могућностима и ограничењима примене савремених метода обраде природног језика у анализи књижевних текстова на српском језику.
 - (xix) Интеграција развијених језичких ресурса, анотираних корпуса, модела и резултата евалуације у имплементирани истраживачки портал, његово тестирање и постављање у продукционо окружење за даљу истраживачку употребу.
 - (xx) Евалуација развијених модела за препознавање и повезивање именованих ентитета на тексту другог жанра као секундарном скупу за валидацију, упоредна анализа резултата у односу на референтни корпус srpELTeC, и формулисање закључака о преносивости предложеног методолошког оквира на сродне жанрове српског језика.
 - (xxi) Формулација закључака о дoметима, ограничењима и могућностима даљег развоја.

6. Методе које се користе у истраживању

Више метода ће се користити за решавање постављених проблема у овом раду:

I. У почетној фази истраживања биће спроведена систематична анализа релевантне научне литературе, постојећих језичких ресурса и доступних софтверских алата из области обраде природног језика и рачунарске лингвистике за српски језик, са фокусом на књижевне дела. Истовремено ће се применити аналитички и компаративни приступи ради упоређивања различитих метода коришћених у задацима семантичке анализе текста.

II. У другој фази истраживања биће примењене методе рачунарске лингвистике у процесу анотације корпуса књижевних текстова на српском језику. Ови поступци биће интегрисани са техникама машинског и дубоког учења ради развоја, обучавања и финог подешавања језичких модела намењених анализи текста. Биће развијени и обучени модели засновани на трансформерским архитектурама за задатке препознавања именованих ентитета и повезивања ентитета са базама знања, циљано за локације, где ће се испитати и примена великих језичких модела, посебно у контексту обраде падежних облика именованих ентитета. У задацима анализе сентимента примењиваће се комбинација лексикона, као и коришћење језичких модела. За анализу тема у књижевним корпусима примењиваће се методе моделирања тема, укључујући традиционалне приступе, као и векторске репрезентације. Задатак екстракције релација биће приказан коришћењем технологија семантичког веба, текстометрије, уз евалуацију решења која се ослањају на велике језичке моделе и на систем коначних трансдуктора развијен у алату Unitex (Maurel and Krstev 2024; Krstev 2008).

III. У финалној фази евалуација развијених модела вршиће се комбинацијом аутоматске и ручне евалуације, уз квантитативну и квалитативну анализу добијених резултата. За задатке препознавања и повезивања именованих ентитета користиће се стандардне метрике класификације, као што су прецизност, одзив и F1 мера. Применом статистичких метода и компаративне анализе утврђиваће се разлике у перформансама различитих приступа и формулисати закључци о њиховој применљивости у квантитативној анализи књижевних текстова. Поред евалуације на основном корпусу `srpELTeC`, биће спроведена и евалуација на корпусу другог жанра за проверу преносивости развијених модела на други регистар. Развијени ресурси, модели и резултати анализе биће интегрисани у истраживачки портал који ће служити као окружење за обједињено представљање и даљу употребу резултата истраживања.

7. Мултидисциплинарност теме

Мултидисциплинарност предложене докторске дисертације огледа се у интеграцији вештачке интелигенције, лингвистике, рачунарства, статистике и дигиталне хуманистике ради свеобухватне анализе књижевних текстова на српском језику. Истраживање комбинује корпусну и рачунарску лингвистику са савременим методама машинског и дубоког учења, омогућавајући постављање смерница за семантичку анализу текста. Статистички приступи и технике квантитативне анализе омогућавају моделовање тема, процену перформанси модела и идентификацију језичких феномена који нису очигледни применом традиционалне интерпретативне анализе. Истовремено, развијени језички модели, алати и представљање ресурса у стандардним форматима (као што је NIF) доприносе развоју језичких технологија за српски језик и пружају практичну примену у системима за аутоматску обраду и разумевање текста. Оваква синергија дисциплина обезбеђује да рад има и научни, и примењени, и технолошки значај.

8. Очекивани научни допринос докторске дисертације

Очекивани научни допринос докторске дисертације огледа се у више међусобно повезаних резултата. Пре свега, истраживање доприноси развоју језичких ресурса за српски језик, кроз проширење и анотацију корпуса текстова намењених задацима препознавања и повезивања именованих ентитета. Такви ресурси представљају важну основу за даља истраживања и развој модела у области обраде природног језика за српски језик. У том контексту, истраживање ће директно допринети националном научном пројекту ТЕСЛА (Text Embeddings – Serbian Language Applications) (<https://tesla.rgf.bg.ac.rs/>) кроз креирање анутираних скупова података, чиме се значајно проширује опсег примене савремених језичких технологија за српски језик на специфичне врсте неструктурисаних текстова.

Други значајан допринос односи се на развој и евалуацију модела за препознавање именованих ентитета и њихово повезивање са базама знања, прилагођених специфичностима српског језика и књижевних текстова. Дисертација ће понудити нове научне увиде у ефикасност векторских репрезентација развијених у оквиру пројекта ТЕСЛА, кроз процесе њихове доменске адаптације и финог подешавања за потребе обраде морфолошки богатих језика, конкретно језика српских романа. Развијени методолошки оквир за обраду књижевних текстова, који обухвата TEI кодирање, текстометријску анализу, препознавање и повезивање именованих ентитета, као и серијализацију у формату повезаних података, биће паралелно примењен и на корпусу другог жанра. Тиме се не само квантитативно процењује преносивост модела на сродне жанрове, већ се потврђује

методолошка целovitost и поновна употребљивост предложеног оквира на различитим типовима текстова на српском језику.

Поред тога, кроз интеграцију више нивоа аутоматске анализе текста, укључујући анализу сентимента, емоција, моделирање тема и екстракцију семантичких релација, биће предложен шири оквир за рачунарски потпомогнуту анализу књижевних текстова на српском језику. Технички аспект научног доприноса огледа се у формалној репрезентацији резултата у стандардизованим форматима, чиме се осигурава њихова интероперабилност са глобалним екосистемима повезаних података и графова знања.

На крају, дисертација ће пружити важан методолошки мост између области обраде природног језика, интелигентних система и дигиталне хуманистике. Обједињавањем развијених ресурса, анотација, квантитативно и семантички обogaћене анализе књижевних текстова на српском језику и резултата анализе у оквиру јединствене истраживачке платформе, обезбеђује се дугорочна доступност и поновна употребљивост научног рада, чиме се постављају темељи за будућа мултидисциплинарна истраживања српске књижевности и језика.

9. Библиографски подаци релевантни за докторску дисертацију

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & Zoph, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Baker, C. F., Fillmore, C. J., & Lowe, J. B. (2003). The structure of the FrameNet database. *International Journal of Lexicography*, 16(3), 281–296.
- Batanović, V., & Nikolić, B. (2017). Sentiment classification of documents in Serbian: The effects of morphological normalization and word embeddings. *Telfor Journal*, 9(2), 104–109.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Chu, K. E., Keikhosrokiani, P., & Asl, M. P. (2022). A topic modeling and sentiment analysis model for detection and visualization of themes in literary texts. *Pertanika Journal of Science & Technology*, 30(4), 2535–2561.
- Cipresso, P., & Riva, G. (2016). Computational psychometrics meets Hollywood: The complexity in emotional storytelling. *Frontiers in Psychology*, 7, 227145.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., ... & Meta. (2024). The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.
- Han, P., Liang, G., & Wang, Y. (2025). A zero-shot framework for low-resource relation extraction via distant supervision and large language models. *Electronics*, 14(3), 593.
- Jaćimović, J. (2019). Textometric methods and the TXM platform for corpus analysis and visual presentation. *Infotheca — Journal for Digital Humanities*, 19(1), 30–54. <https://doi.org/10.18485/infotheca.2019.19.1.2>
- Krstev, C. (2008). *Processing of Serbian – Automata, text and electronic dictionaries*. Belgrade: Faculty of Philology.
- Krstev, C., Maurel, D., & Vitas, D. (2018). Serbian language integration in Prolexbase multilingual dictionary. *Infotheca - Journal for Digital Humanities*, 18(2), 29–52. <https://doi.org/10.18485/infotheca.2018.18.2.2>

- Krstev, C., Obradović, I., Utvić, M., & Vitas, D. (2014). A system for named entity recognition based on local grammars. *Journal of Logic and Computation*, 24(2), 473–489. <https://doi.org/10.1093/logcom/exs079>
- Krstev, C., & Stanković, R. (2021). Novels and Authors of the Serbian ELTeC Collection. *Infotheca - Journal for Digital Humanities*, 21(2), 172–186.
- Krstev, C., Stanković, R., Marković, A., & Ikonić Nešić, M. (2025). Progress in SR-ELEXIS semantic annotation: Focusing on multiword expressions, named entities, and sense repository. In *UniDive 3rd General Meeting*, Budapest, Hungary, 29-30 January 2025. European Cooperation in Science and Technology. <https://doi.org/10.5281/zenodo.14845204>
- Lee, J., & Yeung, C. Y. (2012). Extracting networks of people and places from literary texts. *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation (PACLIC 26)*, 209–218.
- Makazhanov, A., Barbosa, D., & Kondrak, G. (2014). Extracting family relationship networks from novels. *arXiv preprint arXiv:1405.0603*.
- Manzoor, K. R. (2022). Sentiment analysis, opinion mining and topic modelling of epics and novels using machine learning techniques. *Materials Today: Proceedings*, 51, 576–584.
- Maurel, D., & Krstev, C. (2024). Unitex Getting Started.
- Mladenović, M., Mitrović, J., Krstev, C., & Vitas, D. (2015). Hybrid sentiment analysis framework for a morphologically rich language. *Journal of Intelligent Information Systems*, 46(3), 599–620. <https://doi.org/10.1007/s10844-015-0372-5>
- Moretti, F. (2011). Network theory, plot analysis. Stanford Lit Lab Pamphlet 2. <https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf>
- Prado, S. D., Dahmen, S. R., Bazzan, A. L., Carron, P. M., & Kenna, R. (2016). Temporal network analysis of literary texts. *Advances in Complex Systems*, 19(03), 1650005
- Radak, T., Burnard, L., François, P., Hilger, A., Jannidis, F., Palkó, G., Patras, R., Preminger, M., Santos, D., & Schöch, C. (2024). Towards a computational history of modernism in European literary history: Mapping the inner lives of characters in the European novel (1840–1920). *Open Research Europe*, 3, 128.
- Santos, D., Mamede, N., & Baptista, J. (2010). Extraction of family relations between entities. *InForum 2010*, 9–10.
- Stanković, R., Chiarcos, C., Utvić, M., & Kitanović, O. (2023). Towards ELTeC-LLOD: European Literary Text Collection Linguistic Linked Open Data. *LDK 2023 – 4th Conference on Language, Data and Knowledge*, 12-15 September in Vienna, Austria. <https://doi.org/10.34619/srmk-injj>
- Stanković, R., Košprdić, M., Ikonić Nešić, M., & Radović, T. (2022a). Sentiment analysis of Serbian old novels. *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data*, 31–38.
- Stanković, R., Krstev, C., & Vitas, D. (2024). SrpELTeC: A Serbian Literary Corpus for Distant Reading. *Primerjalna književnost*, 47(2).
- Stanković, R., Krstev, C., Šandrih Todorović, B., & Škorić, M. (2021). Annotation of the Serbian ELTeC Collection. *Infotheca - Journal for Digital Humanities*, 21(2), 43–59. <https://doi.org/10.18485/infotheca.2021.21.2.3>
- Stanković, R., Škorić, M., & Popović, P. (2022b). SrpELTeC on Platforms: Udaljeno čitanje, Aurora, NoSketch. *Infotheca - Journal for Digital Humanities*. <https://doi.org/10.18485/infotheca.2021.21.2.7>
- Stanković, R., Vučenović, T., Rujević, B., Ikonić Nešić, M., & Škorić, M. (2026). Integrating TEI, NER/NEL, Textometry, and Linked Data for a Semantically Enriched Interview Corpus. In *Proceedings of the 15th Language Resources and Evaluation Conference (LREC 2026)*. <https://lrec2026.info/list-of-accepted-papers/>

- Šandrih Todorović, B., Krstev, C., Stanković, R., & Ikonić Nešić, M. (2021). Serbian NER&Beyond: The archaic and the modern intertwined. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)* (pp. 1252–1260). INCOMA Ltd. https://aclanthology.org/2021_ranlp-1.141/
- Škorić, M. (2024). New language models for Serbian. *Infotheca - Journal for Digital Humanities*, 24(1).
- Šošić, M., Graovac, J., & Stanković, R. (2026). Building an emotion lexicon for Serbian using curated language resources. *Language Resources & Evaluation*, 60(9). <https://doi.org/10.1007/s10579-025-09894-5>
- Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10), 78–85. <https://doi.org/10.1145/2629489>
- Vreš, D., Božić, M., Potočnik, A., Martinčić, T., & Robnik-Šikonja, M. (2024). Generative model for less-resourced language with 1 billion parameters. *arXiv preprint arXiv:2410.06898*.
- Zhang, Q., Chen, M., & Liu, L. (2017). A review on entity relation extraction. 2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 178–183. IEEE.
- Zhou, G., Su, J., Zhang, J., & Zhang, M. (2005). Exploring various knowledge in relation extraction. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 427–434.

5. Изјава да предложеној тему кандидат није пријављивао на другој високошколској установи у земљи или иностранству

Ја, Милица Иконић Нешић, ЈМБГ 2705991715321, изјављујем под пуном моралном и материјалном одговорношћу да предложеној тему „Квантитативна и семантичка анализа књижевних текстова на српском језику заснована на машинском учењу и језичким моделима“ нисам пријављивала на другој високошколској установи у земљи или иностранству.

1. Предлог два ментора и комисије за оцену теме докторске дисертације (име, презиме, звање, институција, ужа научна област)

1. Проф. др Ранка Станковић, редовни професор, Рударско-геолошки факултет Универзитета у Београду (математика и информатика)
2. Проф. др Јелена Граовац, ванредни професор, Математички факултет Универзитета у Београду (рачунарство и информатика).

Прилози:

1. Сагласност ментора (треба да садржи име, презиме, звање, институцију и потпис)
2. Подаци о ментору

У Београду, 11.05.2026.

Подносилац молбе (потпис)



Милица Иконић Нешић