Анализа породичних односа ликова корпуса SrpELTeC заснована на техникама семантичког веба и текстометрији

811.163.41’322.2

САЖЕТАК: У раду је представљена анализа породичних односа ликова у колекцији SrpELTeC српских романа објављених од 1840. до 1920. године, кроз текстометријски приступ и технике семантичког веба. Истражено је појављивање породичних односа између књижевних ликова у корпусу у односу на периоде писања романа, као и у зависности од писца дела. Узимајући у обзир преплитање књижевних праваца романтизма и реализма у периоду писања романа, испитано је у ком су правцу породични односи заступљенији, као и да ли је у колекцији број мушких ликова заступљенији од женских. Визуелизације односа међу ликовима представљене су кроз повезивање ликова са романима у википодацима и постављањем SPARQL упита, као и мрежном анализом коришћењем алата Gephi.

КЉУЧНЕ РЕЧИ: SrpELTeC, породични односи, књижевни ликови, википодаци, Gephi, SPARQL.

РАД ПРИМЉЕН: 17. јун 2024.
РАД ПРИХВАЋЕН: 22. јун 2024.

Милица Иконић Нешић
milica.ikonic.nesic@fil.bg.ac.rs
ORCID: 0000-0002-0835-8889

Универзитет у Београду
Филолошки факултет
Београд, Србија

1. Увод

Потколекција српских романа (названа SrpELTeC[1]) (Krstev and Stanković 2022, 2020; Stanković et al. 2022; Patras et al. 2020) настала

1. SrpELTeC, https://github.com/COST-ELTeC/ELTeC-srp
SrpELTeC, https://jerteh.github.io/srpELTeC/

је под окриљем COST акције CA16204 *Distant Reading for European Literary History* (Удаљено читање за европску историју књижевности). Један од најважнијих циљева ове акције био је припрема вишејезичног корпуса (названог *European Literary Text Collection – ELTeC*) који садржи по 100 романа[2] први пут објављених у периоду 1840–1920. за дванаест европских језика, који чине његове језичке потколекције[3] (Odebrecht, Burnard, and Schöch 2021). Српска потколекција романа има 5.886.528 токена и 4.769.262 речи, све речи су анотиране врстом речи и лемом (Stanković et al. 2020; Frontini et al. 2020) и информацијом о именованим ентитетима (Stanković et al. 2019), што отвара могућност примене напредних метода анализе текста, у складу са парадигмом удаљеног читања.

До сада је објављено више истраживања на SrpELTeC корпусу кроз текстометријске анализе појављивања именица и глагола у корпусу (Jaćimović 2019), стилометријске анализе у циљу препознавања аутора текста према стилу писања (Škorić et al. 2022), као и дубоко учење и обучавање модела за препознавање именованих ентитета (Šandrih Todorović et al. 2021; Ikonić Nešić et al. 2024). Појављивање реторичких фигура поређења у корпусу проучавано је кроз њихову употребу према различитим критеријумима: ауторима, периодима, полу аутора, обиму романа и популарности романа. Такође, анализирани су придеви, именице и маркери који се користе у поређењима пронађеним у SrpELTeC-у, као и ентитети на које се примењују (Krstev 2021). Витас (Vitas 2020) у свом раду, користећи 104 романа SrpELTeC корпуса, анализира различите аспекте језика хране у српском језику, навика у исхрани, као и положај жена и културне навике у српској популацији из друге половине 19. и почетка 20. века, коришћењем алата Unitex[4] помоћу система електронских морфолошких речника за српски језик, јер су у њима детаљно описани различити прехрамбени производи, јела и пића. Међутим, до сада, проучавање ликова у романима и њихови међусобни односи остали су неистражени.

Предмет овог рада је екстракција ликова и анализа породичних односа у корпусу српских романа, утврђивање зависности њиховог појављивања у односу на писца дела, као и у односу на време настанка. Користиће се напредне анализе потколекције српских романа засноване

2. Наративни текст који има више од 10.000 речи, осим романа то може да буде и новела или дужа приповетка.

3. ELTeC, https://distantreading.github.io/ELTeC/

4. Multilingual Corpus Processing Suite Unitex/GramLab

на обради метаподатака, структуре текстова и њиховог садржаја, као и на техникама семантичког веба. Оно што је недостатак оваквог приступа јесте што је потребна и контекстуална анализа, јер је, на пример, именица *жена* вишезначна. У корпусу SrpELTeC именица *жена* јавља се 5.335 пута, док се именица *супруга* јавља 132 пута, што потврђује да се именица *жена* осим за означавање женске особе користи и за означавање особе у браку. Да би се приступило таквој анализи један од начина је анализа придевских речи у функцији придевских атрибута уз именице *муж* и *жена*, као што је то представљено у раду Pantić (2018) где су придевске речи у корпусу СрпКор2013 (Utvić 2013) класификоване у 10 група, од којих је за овакво истраживање најзначајнија група која се односи на друштвени и лични статус као што су *брачни статус*: *бивши*, *будући*, *ожењен*, *удат* итд. и *емотивни статус*: *омиљен*, *вољен*, *љубљен* итд, међутим, то сада неће бити предмет овог истраживања.

Циљ овог рада је разумевање односа међу ликовима у романима корпуса SrpELTeC, кроз концепте породичних, љубавних или друштвених односа, користећи технике текстометрије, удаљеног читања и семантичког веба. Конкретно, циљ је да се одговори на следећа питања:

1. Могу ли се уочити сличности или разлике у концепту породичних односа које аутори одабраних романа показују у својим делима, кроз анализу речи попут *мајка*, *отац*, *син*, *ћерка*, *жена*, *муж*, *брат*, *сестра* као и њихових синонима, и да ли они зависе од самих аутора?
2. Узимајући у обзир да су романи писани у периоду од 1840. до 1920. године у доба преплитања књижевних праваца романтизма и реализма, а знајући да је једна од главних тема реализма управо концепт породице, да ли можемо тврдити да су породични односи заступљенији у реализму него у романтизму?
3. Родни стереотипи и недовољна заступљеност женских ликова у романима су документовани у прошлости. Да ли можемо да тврдимо да су и у корпусу SrpELTeC мушки ликови бројнији од женских?

Преглед сродних истраживања биће приказан у Одељку 2. Одељак 3. укратко приказује корпус SrpELTeC на коме је истраживање спроведено. Методологија рада приказана кроз текстометрију и повезивање ликова из романа са википодацима представљена је у Одељку 4. Резултати добијени различитим методама приказани су у Одељку 5., док је Одељак 6. посвећен закључцима изведеним из резултата.

2. Нека сродна истраживања

Историјски гледано, писана дела су првобитно проучавана из перспективе заплета као најважнијег дела приче. Међутим, модернији приступи се заснивају на ликовима (Bamman, O'Connor, and Smith 2013) и сматра се да управо они својим поступцима унапређују радњу (Min and Park 2016). Волох (Woloch 2009) посматра ликове кроз њихову позицију у односу на остале елементе радње (место, време, други ликови), тј. посматра како се ликови описују у наративу. Концепт система ликова проширује овај појам на наратив у целини, и одговара јединству простора свих ликова у причи. Више истраживача је користило овакав приступ ради проучавања и разумевања поступка којим писци граде наратив.

Поред самих ликова, истраживачи се баве интеракцијом ликова, што се сматра окосницом нарације (Cipresso and Riva 2016; Prado et al. 2016). Један од приступа анализи интеракције ликова користи графове који омогућавају представљање и проучавање система кроз интеракције његових конститутивних елемената. Мрежа ликова је граф који описује наратив представљањем ликова као чворова графа, а интеракције између њих кроз његове ивице. Морети је показао да такав приступ омогућава формалније руковање Волоховим концептима (Moretti 2011), јер Волох наглашава чињеницу да се карактерни простори морају разматрати заједно, а то је управо оно што дозвољавају графови, као релациони простор за моделирање.

Међутим, постоје и другачији приступи анализи односа међу ликовима. Сантос (Santos, Mamede, and Baptista 2010) и група истраживача, развијају приступ заснован на правилима за издвајање породичних односа из португалског наратива. Аутори развијају 99 правила укључујући нека правила пропагације и комбинују их са системима препознавања именованих ентитета и системима за разрешавање анафора. Кокинакис и Малм (Kokkinakis and Malm 2011) развијају ненадгледани приступ за издвајање међуљудских односа, укључујући и породицу. Метод се ослања на истовремене појаве имена ликова у истим реченицама. Такве истовремене појаве се идентификују као међуљудски односи на основу сличности контекста и садржаја онлајн доступних лексичко-семантичких ресурса (Borin et al. 2009; Borin and Forsberg 2010). Аутори тестирају методу на три тома шведске фикције из 19. века. Приступ издвајања породичних односа из књижевног наратива се даље развијао, укључујући технику атрибуције исказа у комбинацији

са детекцијом вокатива – експлицитних облика обраћања које користе ликови у роману (Makazhanov, Barbosa, and Kondrak 2014). Аутори користе чињеницу да поједини вокативи указују на породичне односе међу говорницима. Извучене релације се затим пропагирају коришћењем скупа правила.

У најновијим истраживањима не анализирају се само односи у породици већ и односи међу љубавним паровима као и улога жене почетком 20. века. Нека од тих истраживања анализирају главне ликове романа кроз екстракцију и анализу наратива међу њима, на пример у (Qizi Sayitqulova 2021) тако се анализира роман „Ноћ и дан" (*Kecha va Kunduz*) узбекистанског аутора А. Чулпана, док се у (Firat 2018), користе технике анализе садржаја кроз појављивање специфичних именица попут *бака*, *дека*, *љубав*, *игра* и других за анализу односа *бака*, *дека* и *унука*. Однос сестара у мађарским романима 19. века (Kucserka 2020) анализира се испитујући наратив између *сестара*.

3. Потколекција српских романа (SrpELTeC)

Процес креирања потколекције српских романа (SrpELTeC) као део пројекта COST акције (CA16204), трајао је од 2017. до 2021. године. У основној SrpELTeC[5] колекцији налази се 100 романа, док се у проширеној колекцији SrpELTeC-ext[6] (SrpELTeC-extended) налази 20 романа. Сва дела уврштена у било коју језичку потколекцију корпуса ELTeC задовољавају исте, унапред договорене критеријуме, а то су 1) само наративна проза (роман, новела или дужа приповетка) чија је дужина најмање 10.000 речи[7]; 2) прво издање дела треба да буде из периода од 1840. до 1920, укључујући и ове године; 3) дело треба да буде оригинално написано на српском језику, тј. језику одређене потколекције, па се преводи не узимају у обзир; 4) дело треба да буде објављено у Европи највише десет година после првог издања и 5) предност имају она дела која су у назначеном периоду објављена као књиге, а не у наставцима у серијским публикацијама.

У циљу постизања разноврсности заступљених текстова и да би се омогућила компаративна анализа потколекције, као и примена

5. SrpELTeC, https://github.com/COST-ELTeC/ELTeC-srp

6. SrpELTeC-ext, https://github.com/COST-ELTeC/ELTeC-srp-ext

7. Подразумева се да се речи броје аутоматски, на пример онако како то уради програм MS Word.

статистичких метода за анализу текстова, било је потребно да буду испуњени и додатни критеријуми приликом одабира романа, а то су 1) свака потколекција састоји се од 100 романа; 2) подједнака заступљеност женских и мушких аутора; 3) подједнака заступљеност дела са много поновљених издања и оних објављених једном или двапут; 4) равномерна покривеност периода 1840–1920; 5) равномерна заступљеност кратких и дужих дела; 6) оптимално је да 10 аутора у потколекцији буде представљено са по три дела, а остали са једним (Крстев et al. 2023). Да би равномерна покривеност периода 1840–1920. била задовољена одабрани период првог издања дела подељен је у четири периода (T1, T2, T3 и T4). Сва дела су по дужини сврстана у три групе: кратка (до 50.000 речи), средње дугачка (од 50.000 до 100.000 речи) и дугачка (више од 100.000 речи).

У српској потколекцији SrpELTeC из првог периода T1: [1840–1859] налазе се два романа, из другог T2: [1860–1879] 18 романа, из трећег T3: [1880–1899] и четвртог T4: [1900–1920] по 40 романа. У основној колекцији заступљено је 66 мушких и 4 женска аутора,[8] при чему је највише дела Јакова Игњатовића (5) (Крстев et al. 2023).

3.1 Обележавање именованих ентитета и повезивање са википодацима

Препознавање именованих ентитета (енгл. Named Entity Recognition, NER) представља идентификовање и класификацију именованих ентитета у тексту. То могу бити имена особа, њихових улога, локалитета, организација, као и препознавање нумеричких израза који укључују датум и време, монетарне вредности и слично, а они представљају један од кључних корака при екстракцији информација из текста. На нивоу целе COST акције договорено је да се у романима означава само седам категорија ентитета: **PERS** (властита имена људи стварних или фиктивних: лично име, презиме, надимак или све то у комбинацији), **ROLE** (занимања, титуле и задужења људи), **DEMO** (обележавају се становници држава, градова, региона), **ORG** (имена организација, политичких партија, образовних установа, спортских клубова, болница, библиотека, хотела, музеја, кафана и цркава), **LOC** (континенти, државе, региони, насељена места, планине, острва, пећине, реке, имена небеских тела, градске локације), **WORK** (наслови књига, драма, песама за читање или певање, музичких дела, слика, скулптура, новина),

8. SPARQL упит ка визуелизацији броја аутора, https://w.wiki/5YSy

EVENT (имена догађаја, природне катастрофе, револуције, битке, ратови, демонстрације, концерти, спортски догађаји, славе), за које је закључено да ће бити од највећег значаја за даља литерарна изучавања (Frontini et al. 2020). Систем за препознавање именованих ентитета за српски језик заснован је на ручно креираним правилима која се ослањају на свеобухватне лексичке ресурсе за српски језик (Krstev et al. 2014).

Као резултат, до сада је за 100 романа из SrpELTeC колекције завршено обележавање свих речи текста врстом речи, лемом и именованим ентитетом у складу са препорукама акције. Поред тога, урађено је обележавање именованих ентитета за још осам романа од 20 романа из проширене колекције, и они ће бити део корпуса за даљу текстометријску анализу (укупно 108 романа).

4. Методологија

У овом одељку биће представљене методе коришћене при истраживању и проналажењу одговора на постављена истраживачка питања. Да би се добио одговор на прво и друго питање приступиће се истраживању кроз текстометријски приступ, док ће се одговор на треће питање и могућности визуелизације односа међу ликовима представити кроз повезивање ликова са романима у википодацима и постављањем одговарајућих SPARQL упита (Ikonić Nešić, Stanković, and Rujević 2021; Ikonić Nešić et al. 2022).

4.1 Текстометрија

Текстометрија је рачунарски потпомогнута метода текстуалне анализе која се заснива на фреквенцијама речи (или на било ком пребројавању језичких карактеристика) повезујући статистичку обраду са референцирањем података и контекстуалним поређењем (Jaćimović 2019). Досадашње студије показале су да се текстометријска анализа показала веома успешном и корисном на различитим корпусима, од француских вести (Carrive et al. 2021), медицинских сажетака (Gledhill et al. 2019) до форензичких истраживања (Longhi 2021). У свом раду на узорку од 21 романа из корпуса старих српских романа Јаћимовић (Jaćimović 2019) је показала значајне резултате текстометријске анализе, приказујући листе учесталости појављивања лема, као и одређених врста речи, тј. именица и глагола кроз различите текстометријске

визуелизације. Узимајући у обзир све поменуто и то да је у овом моменту у оквиру проширеног ELTeC корпуса дигитализовано 120 романа, претпостављамо да ће техника текстометрије бити успешна у овом истраживању приликом приказивања како специфичности писаца, тако и романа. Један од основних алата за текстометрију је TXM[9] (Heiden 2011). TXM је програм отвореног кода који се користи у различитим истраживањима у области друштвено-хуманистичких наука. Графичко корисничко окружење TXM-а заснива се на коришћењу CQP[10] (енгл. Corpus Query Processor) претраживача и R[11] статистичког пакета. Прва и основна корпусна метода која се примењује при текстометријској анализи је израда фреквенцијских листа или листа учесталости. Поређење апсолутних фреквенција појаве језичких јединица (тачан број појављивања у корпусу) може бити корисно и даје иницијални утисак о контрасту који постоји међу деловима корпуса, али је за поређење фреквенција у деловима различитих величина неопходно извршити нормализацију, тј. изразити фреквенције заједничким фактором – релативном фреквенцијом.

Рачунање индекса специфичности на основу хипергеометријске дистрибуције у TXM окружењу показаће вероватноћу појаве језичке јединице у неком одређеном делу корпуса. TXM ће омогућити и графички приказ дистрибуције специфичности одабраних јединица. Вредности индекса специфичности веће (позитивне вредности) или мање (негативне вредности) од очекиване представљају више или мање заступљену језичку јединицу, док се резултати између -2 и +2 (вероватноћа $\geq 1\%$) сматрају статистички незначајним. На овај начин могуће је идентификовати значајно честе (позитивне кључне речи) или значајно ретке (негативне кључне речи) појаве језичке јединице у деловима у односу на цео корпус, што је корисна полазна тачка за извођење претпоставки о кључним речима текста, домену текста, ауторству текста итд.

Такође, да би се прецизније описала промена учесталости одређене речи у корпусу, у оквиру TXM-а може се користити прогресија. Постављајући образац претраге изражен CQL упитом креираће се графикон који представља „кумулативну“ криву фреквенције појављивања за речи или обрасце претраге. Није потребна нормализација, пошто нагиб директно мери и визуелно показује однос

9. TXM, https://sourceforge.net/projects/txm/
10. The IMS Open Corpus Workbench
11. R: The R Project for Statistical Computing

броја појављивања и дужине текста, што је нормализована вредност. Фреквенције речи могу да се разликују, али се нагиби могу поредити у текстовима различите дужине.

Кластер анализа у TXM алату за текстометријску анализу омогућава груписање сличних јединица (нпр. речи, лема, пасуса, поглавља или читавих докумената) на основу њихових заједничких карактеристика. У контексту анализе текстова, ово може значити груписање речи које се појављују у сличним контекстима, или груписање докумената који деле сличне тематске карактеристике.

Користећи описане приступе уз визуелизацију резултата приказаће се прогресија фреквентности појављивања именица у лематизованом корпусу од 108 романа[12] и њихових синонима: **мајка**, мати, мајка, родитељка, маћеха, мама, дада, матер, матера, маћа; **отац**, тата, бабо, поочим, очух, бабајко, ћаћа, тајо, татко; **муж**, супруг, супружник; **жена**, супруга, супружница, љуба; **ћерка**, кћи, кћер, ћер, кћерка; **син**, синак, синчић; **брат**, братић, бата, братац, браца; **сестра**, дада, сека, сеја, сестрица, секица. Фреквентност појављивања именица посматраће се у зависности од романа, али и од датума објављивања дела, па и аутора. Рачунајући индекс специфичности поменутих именица резултати ће бити приказани у зависности од аутора, као и пола аутора.

За кластер анализу изабран је поткорпус текстова чији аутори имају више од 120.000 речи у корпсу, а то су: Владимир Ђорђевић (155.706), Ивко Ћипико (121.469), Светозар Ћоровић (143.600), Јанко М. Веселиновић (225.275), Јелена Димитријевић (177.103), Јаков Игњатовић (457.434), Драгутин Ј. Илић (141.401), Лазар Комарчић (191.034), Миодраговић Јован (147.144), Чедомиљ Мијатовић (133.392), Бранислав Нушић (258.006), Светолик Ранковић (228.046), Стеван Сремац (240.862), Каменко Суботић (192.212), Пера Тодоровић (415.024), и Милутин Усковић (177.127). Оваква анализа користиће ради испитивања груписања аутора по начину приказивања породичних односа унутар дела.

За дубљу анализу односа ликова, коришћењем CQL упита, унутар окружења TXM претраживаће се релације попут *мајка од*, *отац од*, *син од* итд. у облику уређене (X,ρ,Y) тројке, где X и Y означавају аргументе релације, а ρ означава саму релацију, дакле,

X и Y у релацији $\rho \Leftrightarrow$ постоји породична релација између X и Y

12. Serbian ELTeC Corpus TXM Edition (108 NER), https://live.european-language-grid.eu/catalogue/corpus/23621

Претрага ће служити за издвајање релевантног текста, из ког се затим могу екстраховати информације потребне за извођење следећих релација, попут:

$$(X, \text{мајка/отац од}, Y) \Rightarrow (Y, \text{син/ћерка од}, X)$$
$$(X, \text{син/ћерка од}, Y) \Rightarrow (Y, \text{мајка/отац од}, X)$$
$$(A, \text{мајка/отац од}, B)\ \&\ (B, \text{брат/сестра од}, C) \Rightarrow (A, \text{мајка/отац од}, C)$$

при чему ће се захтевати да у поменутим релацијама X и Y заправо буду именовани ентитети означени етикетом <PERS> која у овом случају означава лична имена особа. Дакле, у окружењу TXM постављајући CQL упите кроз корпус претраживаће се односи у облику

$$(\text{<PERS>X</PERS>}\ \rho?\ \text{<PERS>Y</PERS>})$$

где ρ означава релацију између двеју личности, нпр. *мајка од*, *отац од*, *брат од*, итд, док ? означава да се између релације и друге особе може наћи још један токен. Овакав приступ ће се даље користити како би се повезивали ликови у википодацима и прецизније приказали односи међу њима. Пример CQL упита и односа екстрахованих из текста помоћу њега представљен је у Табели 1.

Табела 1: Пример екстракције релације из текста

CQL упит	**екстрахован текст**	**екстрахован однос**
`<pers>[]*</pers> []{0,1} [srlemma = "majka|otac|kći |ćerka|sin|muž |žena"][]{0,1} <pers>[]*</pers>`	Anđe i oca joj Bogosava Stevanovića; Arslan-Nuri-pašina sina Murad; Darinka ćerka Sretena Cvetića; Rosa žena Milutinova; Velju sina gazda Damnjana Smiljanića;	(**Bogosav Stevanović**, otac od, **Anđa**); (**Arslan-Nuri-paša**, otac od, **Murad**); (**Sreten Cvetić**, otac od, **Darinka**); (**Milutin**, muž od, **Rosa**); (**Damnjan Smiljanić**, otac od, **Velja**);

4.2 Повезивање ликова из романа са википодацима

Википодаци (Wikidata[13]) – отворена база знања у којој корисник може креирати нову ставку и изменити постојећу. Уношења романа

13. Wikidata, https://www.wikidata.org

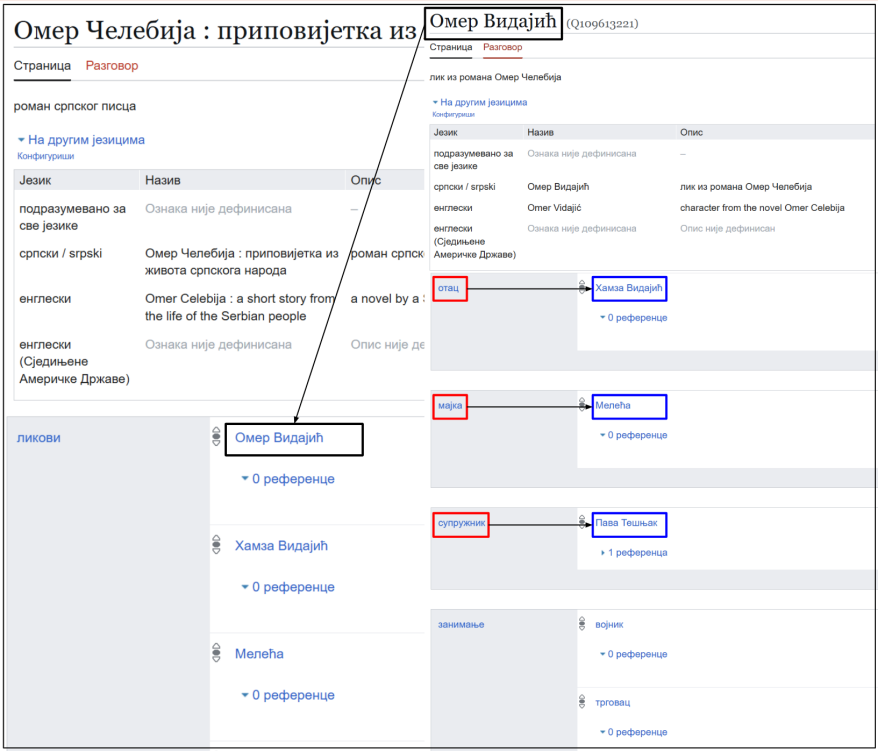
SrpELTeC колекције у википодатке као и првог, дигиталног, штампаног или ELTeC издања сваког од романа аутоматизовано је синергијом алата *OpenRefine*[14] и *QuickStatements*.[15] Уношење ликова и места радње је обављано ручно јер се подаци о ликовима нису могли екстраховати из метаподатака, већ је на основу фреквенције појављивања именованих ентитета класе PERS и класе LOC направљена селекција ликова и места радње према броју појављивања у романима, где су изабрани они са највећим фреквенцијама појављивања (Ikonić Nešić, Stanković, and Rujević 2021; Ikonić Nešić et al. 2022). Користећи SPARQL упите може се утврдити да тренутно у википодацима има 978 (https://w.wiki/5Yuz) унетих ликова романа, при чему је највећи број из романа „Ђурађ Бранковић: историческиј роман“. У википодацима налази се 100 романа (https://w.wiki/5Ydd) основне SrpELTeC колекције и 20 романа (https://w.wiki/HX8p) из проширене колекције SrpELTeC-extended. У раду аутора колекције Крстев и Станковић дат је списак свих романа у овим колекцијама (Krstev and Stanković 2022).

Сви креирани ликови су повезани са романима у википодацима користећи својство ликови (`P674`). Могућности претраживања у окружењу TXM омогућиле су нам да екстрахујемо информације о ликовима и закључимо какви су њихови односи са другим ликовима постављањем CQL упита представљених у Табели 1. Након тога, сваки лик је у википодацима повезан са другим ликовима својствима *отац* (`P22`), *мајка* (`P25`), *супружник* (`P26`), *дете* (`P40`). Пример уноса лика *Омера Видајић* (`Q109613221`) из романа „Омер Челебија: приповијетка из живота српскога народа“ и његово повезивање са осталим ликовима приказано је на Слици 1, где се може уочити да је **отац** Хамза Видајић (`Q109654569`), **мајка** Мелећа (`Q109654771`) и **супружник** Пава Тешњак (`Q109659439`). На основу оваквог уноса могуће је претраживати односе међу ликовима користећи SPARQL упите.

4.3 Мрежна анализа ликова

Након што су главни ликови ручно унети у википодатке, да би се испитале могућности које нуди мрежна анализа ликова у откривању односа ликова у романима, коришћен је алат INCEpTION[16] за повезивање ентитета у тексту са одговарајућим ставкама у

14. OpenRefine, https://openrefine.org
15. QuickStatements, https://quickstatements.toolforge.org/
16. INCEpTION, https://inception-project.github.io



Слика 1: Пример повезивања ликова у роману „Омер Челебија: приповијетка из живота српскога народа“

отвореној бази знања Википодаци, а након тога алат отвореног кода Gephi[17] (Bastian, Heymann, and Jacomy 2009) за визуелизацију и за мрежну анализу ликова на примеру одломка романа „Нечиста крв“ Борисава Станковића. INCEpTION је веб-окружење које омогућава интерактивно означавање текста и повезивање ентитета са спољашњим базама знања у које спадају и википодаци. Поступак анотирања састоји се у идентификовању помињања ентитета у тексту и његовом повезивању са википодацима. Повезивање текста са ставком (класом или инстанцом) почиње бирањем распона текста ентитета иза чега следи тражење ставке повезивања коришћењем идентификатора за аутоматско

17. Gephi, https://gephi.org/

претраживање и повезивање са ставкама са википодацима (Ikonić Nešić, Stanković, and Rujević 2021; Ikonić Nešić et al. 2022).

У википодацима се тренутно налази 13 ликова из романа „Нечиста крв“ (https://w.wiki/ApXd). Након завршеног повезивања ликова из романа са ставкама википодатака, анотирани делови романа су експортовани као датотеке са подацима раздвојеним табулатором (*tsv*) сагласно формату WebAnno TSV 3.3.[18]

Следећи корак је био припрема улазне датотеке за алат Gephi. У примеру анализиране мреже коришћени су Force Atlas layout алгоритам у Gephi окружењу, као и Network Diameter и Modularity статистике. Ове статистике пружају увиде у структуру мреже, која ће бити приказана у Одељку 5. у анализи резултата истраживања чиме се постиже боље разумевање динамике односа између ликова. Force Atlas је погодан избор за мрежну анализу ликова у романима јер омогућава визуелизацију која открива структуру и динамику међу ликовима. У романима, ликови често формирају групе или кластере на основу својих интеракција. На пример, ликови који често комуницирају или су део исте сцене могу формирати кластере. Force Atlas распоређује ликове који су међусобно повезани ближе један другом у мрежи, чиме се олакшава визуелно уочавање група или кластера ликова који су често у интеракцији. Овај алгоритам такође позиционира ликове који имају интензивну интеракцију са другим ликовима, што су често главни јунаци или битнији споредни ликови. На тај начин су централни ликови јасно истакнути, док они мање значајни који не остварују много веза са другим ликовима, заузимају место на ободу мреже. За потребе овог истраживања узимајући у обзир број чворова (мањи од сто) Force Atlas је пригодан алгоритам, док је за већи број чворова потребно користити Force Atlas 2 алгоритам који поред тога што се примењује на већи број чворова, омогућава флексибилно подешавање, тако да корисник може прилагодити визуализацију према специфичним потребама анализе, било да се фокусира на одређене ликове, сцене или теме (Moretti 2011).

За приказ мрежне анализе ликова узет је одломак романа „Нечиста крв“ од 3,912 реченица, у којима се налазило 1,195 именованих ентитета класе PERS, од којих је Софка обележена у приближно 50% случајева. Основне карактеристике мреже која је анализирана на овом одломку романа су: *број чворова* 10, *број грана* 17 (свака грана представља појављивање ликова у истој реченици) и *тип графа*: неусмерен (*undi-*

18. WebAnno TSV 3.3, https://inception-project.github.io/releases/26.8/docs/user-guide.html#sect_formats_webannotsv3

rected), а за рангирање изабран је *модул* `weighted degree`, при чему пондерисани тежински степен `weighted` као параметар приказује колико пута су се два лика поменула у истој реченици.

У овом истраживању направљен је интерфејс који од улазне *tsv* датотеке генерише *gexf* датотеку која се користи као улаз у алат Gephi приказану на Слици 2 на којој се може уочити да Софка (`Q109693861`) и Магда (`Q109746715`) имају највише заједничких појављивања у истим реченицама, чак 16.

```
</meta>
<graph defaultedgetype="undirected" idtype="string" type="static">
<nodes count="10">
<node id="Q109748906" label="Tone"/>
<node id="Q109748839" label="Tomča"/>
<node id="Q109693861" label="Sofka"/>
<node id="Q109748924" label="Ahmet"/>
<node id="Q109747507" label="Arsa"/>
<node id="Q109748862" label="Simka"/>
<node id="Q109746715" label="Magda"/>
<node id="Q109747662" label="Mita"/>
<node id="Q109748881" label="Todora"/>
<node id="Q109747266" label="Marko"/>
</nodes>
<edges count="17">
<edge id="0" source="Q109693861" target="Q109747266" weight="2.0"/>
<edge id="1" source="Q109747662" target="Q109747266"/>
<edge id="2" source="Q109748839" target="Q109693861"/>
<edge id="3" source="Q109748839" target="Q109748924"/>
<edge id="4" source="Q109748906" target="Q109693861"/>
<edge id="5" source="Q109748839" target="Q109747662"/>
<edge id="6" source="Q109747507" target="Q109747266"/>
<edge id="7" source="Q109693861" target="Q109747507"/>
<edge id="8" source="Q109693861" target="Q109748881" weight="2.0"/>
<edge id="9" source="Q109693861" target="Q109748924"/>
<edge id="10" source="Q109748906" target="Q109746715"/>
<edge id="11" source="Q109693861" target="Q109747662"/>
<edge id="12" source="Q109746715" target="Q109747662"/>
<edge id="13" source="Q109747662" target="Q109747507"/>
<edge id="14" source="Q109748924" target="Q109747266" weight="3.0"/>
<edge id="15" source="Q109693861" target="Q109746715" weight="16.0"/>
<edge id="16" source="Q109693861" target="Q109748862" weight="3.0"/>
</edges>
</graph>
</gexf>
```

Слика 2: Улазна *gexf* датотека романа „Нечиста крв“ (`SRP19101`)

5. Резултати и дискусија резултата

Рачунајући релативне фреквенције укупног појављивања лема *мајка, отац, син, ћерка, муж, жена, брат, сестра* и њихових синонима

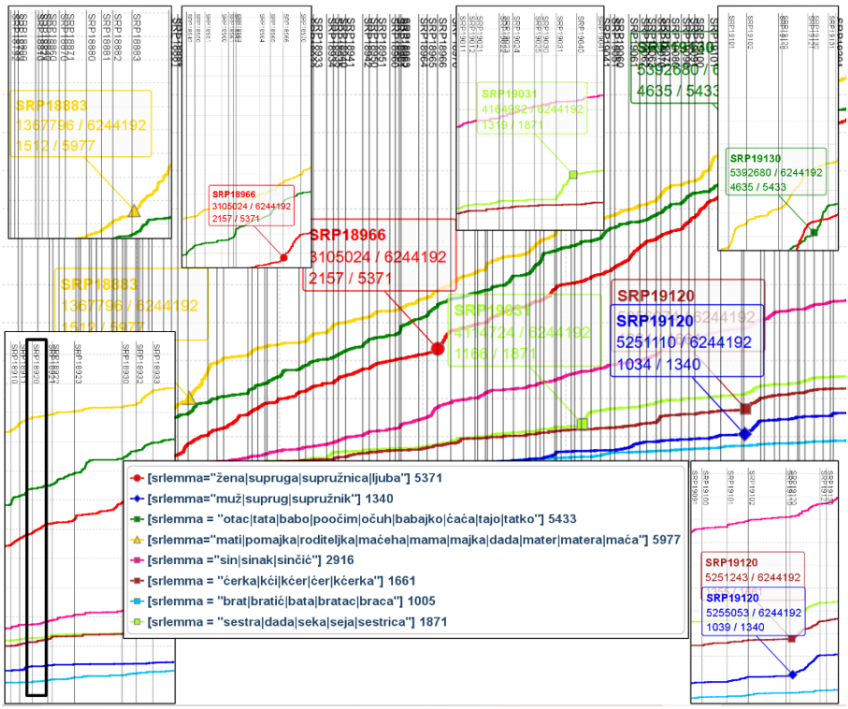
унутар корпуса показало је да приповетка „Глава шећера“ (SRP18751), Милована Глишића, има најмању релативну фреквенцију овог скупа лема 0,000553, док роман „Нове“ Јелене Димитријевић (SRP19120) има највећу релативну фреквенцију 0,013066 (Слика 3). Ови резултати потврђују нашу претпоставку да релативне фреквенције појављивања одабраних лема могу донекле указивати да ли је тема породице заступљена у роману. Релативне фреквенције у приповетки „Глава шећера“ указују да породица као тема није у њој заступљена, што је у складу са садржајем приповетке која представља положај сељака у односу на политичку сцену тог доба. Са друге стране, Јелена Димитријевић за тему романа бира трагичну судбину жене описујући њен положај у породици у којој се вековима одржавају стари обичаји. Ради прегледности резултата на Слици 3 приказане су само леме *мајка, отац, син, ћерка, муж, жена, брат, сестра* без њихових синонима.

Посматрајући дијаграм прогресије (Слика 4), за леме *мајка*, *отац*,

srlemma	žen	otac	majk	sin	sestra	muž	brat	ćerka	sum	words	rel. freq
SRP19120	475	170	288	113	117	183	19	5	1539	117785	0.013066
SRP19022	26	4	52	2	5	34	2	2	157	12702	0.012360
SRP18962	3	86	12	14	0	0	1	0	156	13379	0.011660
SRP19070	43	33	79	77	23	11	2	3	349	33291	0.010483
SRP18890	73	49	35	26	94	75	5	22	581	63341	0.009173
SRP18991	7	4	28	2	1	3	0	3	112	12748	0.008786
SRP19012	29	1	25	12	3	8	5	3	215	26027	0.008261
SRP18620	17	24	9	8	16	5	2	2	209	26669	0.007837
SRP18691	42	71	5	21	29	7	14	1	291	37610	0.007737
SRP18965	51	135	9	125	23	2	34	1	496	71130	0.006973
SRP18910	41	58	38	18	6	3	0	2	198	28452	0.006959
SRP18750	26	109	5	53	8	9	8	3	420	61167	0.006866
SRP18810	31	21	14	1	5	4	0	2	85	12380	0.006866
SRP19130	28	247	51	45	15	7	10	20	487	71023	0.006857
SRP18870	21	5	48	10	9	8	3	23	162	23889	0.006781
SRP18883	105	145	55	54	49	56	36	13	944	139670	0.006759
SRP18966	243	76	22	55	29	33	62	7	622	93865	0.006627
SRP19101	83	99	6	47	13	45	4	0	549	83104	0.006606
SRP19010	44	112	93	5	2	3	3	0	273	42991	0.006350

Слика 3: Првих 20 романа сортираних према релативним фреквенцијама укупног појављивања лема *мајка, отац, син, ћерка, муж, жена* и њихових синонима унутар корпуса у опадајућем редоследу

брат, *сестра*, *син*, *ћерка*, *муж* и *жена*, као и њихове синониме, у делу SRP18883 („Патница“) Јакова Игњатовића, уочава се скок за лему *мајка* и нешто мање за лему *отац*, што иде у прилог томе да се дело бави



Слика 4: Дијаграм прогресије лема *мајка, отац, муж, жена, ћерка, син, сестра и брат* и њихових синонима унутар корпуса.

породицом и положајем жене у друштву. Највећи скок за лему *жена* (и њене синонима *супружница*, *љуба*) *муж* и *ћерка* уочава се у SRP19120 („Нове“) Јелене Димитријевић, док се пре тога мањи скок уочава за лему *жена* у роману SRP18966 („Назарени“) Јаше Томића. У оба романа говори се о сукобу традиције и нових вредности, у чијем се центру налази породица, при чему је у случају Јелене Димитријевић то првенствено жена. У делу SRP19031 („Две сестре или Самоубиство једне шваље: слика из београдског живота“) Боже Савића уочава се скок за лему *сестра*, што и одговара у потпуности теми романа. Посматрајући роман (SRP19130) „Калуђер и хајдук: приповетка о последњим данима Србије у XV веку“ Стојана Новаковића, може се уочити скок леме *отац*. С обзиром на то да је роман историјски и главна тема нису породични односи, анализирањем конкорданци и контекста у ком се помиње лема *отац*, уочено је да у овом роману лема *отац* представља титулу

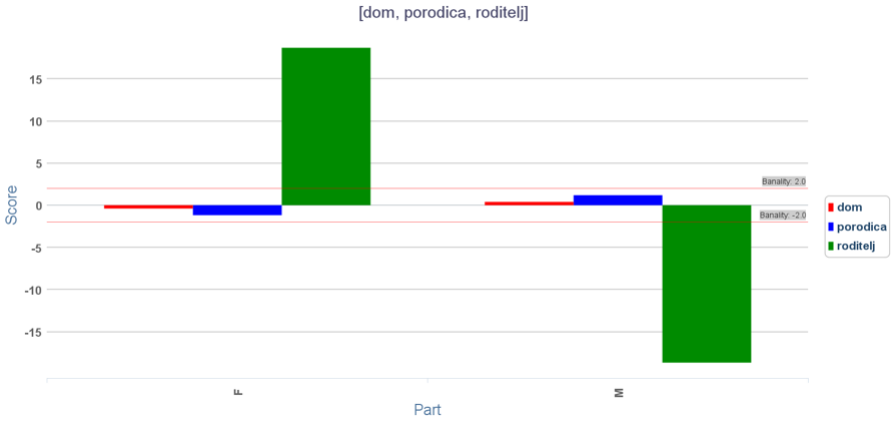
свештеног лица, у овом случају калуђера, који је и један од главних ликова романа. У SRP18923 („Београдске тајне“) Пере Тодоровића, све леме стагнирају, што и потврђује да је роман писан у реалистичком стилу, са елементима социјалне драме и авантуристичке прозе, без породице као централне теме.

Да бисмо ближе описали како појављивање лема зависи од аутора, приказаћемо најпре индексе специфичности употребе лема у зависности од пола аутора. Резултати индекса специфичности у односу на пол аутора показују да је индекс специфичности за лему отац незначајан како код женских, тако и код мушких аутора, док индекс специфичност за леме *ћерка* (>110), *мајка* (>60), *син* (>10), *муж* (>100), *жена* (>7) указује да су поменуте леме заступљеније код женских аутора него код мушких (Слика 5). Да би се дубље испитао концепт породице у романима у односу на пол аутора, била би потребна додатна квалитативна анализа контекста; овде ћемо испитати још само специфичност употреба лема *дом, породица* и *родитељ* у зависности од пола аутора.

Слика 5: Специфичности употребе лема *ћерка, мајка, муж, отац, син, жена* према полу аутора (лево женски, десно мушки аутори).

На Слици 6 представљен је хистограм специфичности употребе лема *дом*, *породица* и *родитељ* у зависности од пола аутора. Може се јасно уочити да је лема родитељ специфична за женске ауторе, док је разлика специфичности лема *дом* и *породица* статистички незначајне. Овакви резултати донекле потврђују да се женски аутори фокусирају

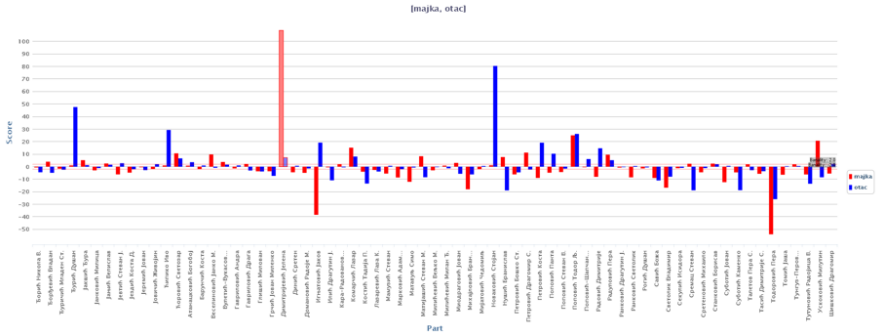
на индивидуалне улоге унутар породице, као што су улога мајке, жене, ћерке или оца, сина. То може значити да дубље истражују специфичне односе између родитеља и деце, док мушки аутори указују на шири, можда традиционалнији поглед на породичне односе и институцију породице као целине, или на важност дома као физичког и симболичког простора. Ово такође може указивати на родне разлике у приступу темама породице и дома, где жене више говоре о емоционалним и персоналним искуствима, док мушкарци више обрађују породицу као друштвену институцију.



Слика 6: Специфичности употребе лема *дом, породица, родитељ* према полу аутора (лево женски, десно мушки аутори).

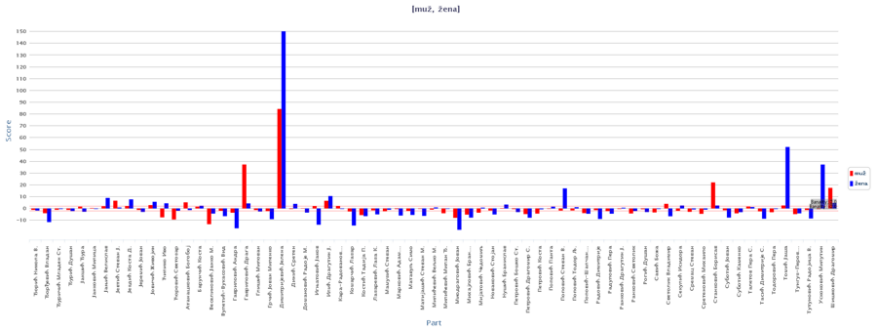
Даљом анализом заступљеност лема по ауторима, посматраћемо парове лема који описују одређени породични однос по ауторима. На Слици 7 представљене су специфичности заступљености лема *мајка* и *отац* по ауторима. Резултати указују да је лема *мајка* са индексом специфичности већим од 100, веома заступљена код Јелене Димитријевић, што се може објаснити чињеницом да се њени романи баве пре свега женама, док мала специфичност леме *отац* (<10) потврђује да у овом роману однос мајке и оца није главна тема. Сразмерна специфичност лема мајка и отац код аутора Тодора Љ. Поповића, који у овом корпусу има само један роман „Гила: новела из сеоског живота“, указује на породичне односе мајке и оца који и јесу једна од тема овог романа. Специфичност Пере Тодоровића (<-40, односно <-20), јасно потврђује да у романима овог писца мајка и отац

и њихов однос нису централна тема, што је у складу са историјским жанром дела овог аутора.



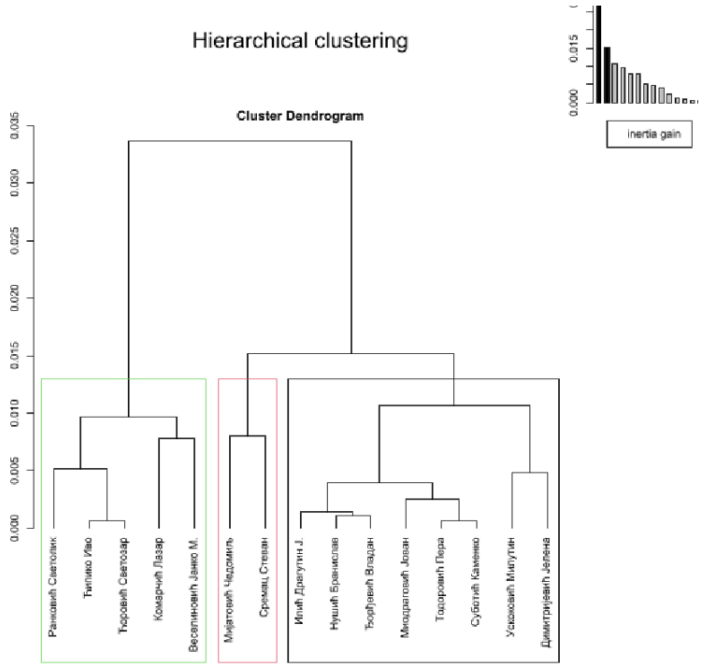
Слика 7: Специфичност употребе лема мајка и отац у SrpELTeC корпусу по ауторима.

Међутим, ако посматрамо леме жена и муж (Слика 8), висока специфичност код аутора Јелене Димитријевић још једном потврђује да је у њеним романима однос жене и мужа (садашњег, будућег или жељеног) једна од главних тема, док специфичност мања од -10 код Андре Гавриловића указује да у његовим романима жена није у фокусу, што и потврђује историјска тематика његових романа „Деспотова властела : роман из српске прошлости“ и „Прве жртве: приповетка из српске прошлости“.



Слика 8: Специфичност употребе лема *муж* и *жена* у SrpELTeC корпусу по ауторима.

Кластер анализа приказана на Слици 9 дели изабране ауторе у три групе, које се могу издвојити по начину на који су породични односи заступљени у њиховим делима. Писци као што су Иван Ћипико, Светолик Ранковић, Лазар Комарчић, Јанко Веселиновић и Светозар Ћоровић могу се груписати по неколико кључних аспеката у вези с темом породице у њиховим делима. Сви ови писци често истражују поремећаје и кризе унутар традиционалних породичних структура. Приказују како друштвене и економске промене утичу на породичне односе и разарају старе обрасце и вредности. Чедомиљ Мијатовић и Стеван Сремац користе породичне односе као средство за критичко осветљавање друштвених и социјалних проблема. Њихова дела често истражују како породични односи одражавају и утичу на шире друштвене структуре и норме. Јелена Димитријевић, Милутин Ускоковић, Бранислав Нушић, Драгутин Илић, Пера Тодоровић, Владан Ђорђевић и Јован Миодраговић истражују породичне конфликте, сукобе генерација и међусобне односе унутар породице. Њихова дела често



Слика 9: Кластер анализа породичних односа у односу на ауторе.

садрже драматичне приказе сукоба и несугласица међу члановима породице. Можемо закључити да је кластер анализа успела да издвоји ауторе тако да су у истим кластерима они који на сличан начин приступају теми породице. Међутим, с обзиром на то да је примењена аутоматски генерисана кластеризација, валидност ове кластеризације је нешто што би требало додатно да се истражи.

Испитивање породичних односа унутар корпуса посматраних према књижевним правцима реализму и романтизму, у циљу добијања одговора на 2. истраживачко питање, посматраће се хронолошки кроз специфичности лема *мајка*, *отац*, *син*, *ћерка*, *жена*, *муж*, *брат*, *сестра* по временским периодима по којима је корпус подељен у поткорпусе, T1: [1840–1859], T2: [1860–1879], T3: [1880–1899] и T4: [1900–1920] (Слика 10).

Слика 10: Специфичност лема према периодима.

Романтизам као правац српске књижевности развија се од 1848. до 1870. године. Оснивач српског лирског романтизма је Бранко Радичевић (1824–1853), док су се у прози највише истакли Богобој Атанацковић и најзначајнији путописац Љубомир Ненадовић (1826–1895). У српској књижевности реализам се као уметнички покрет јавља почетком 1870-их, док су 1880-те период када Глишић пише своје приче, објављују се новеле Лазе Лазаревића, као и дела Симе Матавуља и Јанка Веселиновића. Стеван Сремац и Светолик Ранковић објавили су најзначајнија дела 1890-их година. Развој реализма траје све до почетка 20. века (Deretić 1983). Посматрајући колекцију српских романа

написаних у периоду од 1840. до 1920. године, било је тешко повући јасну границу између ових књижевних праваца. На основу претходног можемо рећи да у периоду T1 доминира романтизам, док је период T2 испреплетан како романтизмом тако и реализмом. Периоди T3 и T4 посвећени су реализму, при чему је реализам као правац достигао врхунац у стваралаштву у периоду T4.

Табела 2 представља фреквенције појављивања лема f_{Tn} по периодима Tn, где је $n \in \{1, 2, 3, 4\}$, док S_{Tn} представља индекс специфичности за лему у периоду Tn. Посматрајмо индексе специфичности (Табела 2) и хистограмски приказ (Слика 10) лема *ћерка*, *мајка*, *муж*, *отац*, *син*, *жена*, *брат* и *сестра* подељених према периодима на које је корпус подељен. Индекси специфичности (Табела 2)

Табела 2: Индекси специфичности лема *мајка*, *ћерка*, *отац*, *муж*, син, брат и сестра по периодима

лема	Σ_F	f_{T1}	S_{T1}	f_{T2}	S_{T2}	f_{T3}	S_{T3}	f_{T4}	S_{T4}
жена	5.108	64	-12,9	353	-34,1	2.040	-5,9	2.651	**46,3**
отац	4.870	267	**23,9**	604	0.6	1.970	-4,3	2.029	-0.4
мајка	3.637	55	- 6.5	320	-10,0	1.242	-28,4	2.020	**60,8**
син	2.366	111	**6,9**	223	-4,7	965	-2,0	1.067	3,0
муж	1.293	33	-0,4	119	-3,3	495	-3,7	646	**8,5**
ћерка	433	7	-1,1	48	-0,5	208	**1,6**	170	-0,8
брат	789	47	**5,9**	112	1.4	358	0,9	272	-4,9
сестра	1.542	10	**-8,9**	198	0,7	627	-1,6	707	3,0

указују да су леме *жена*, *мајка*, *муж* специфичније за период T4 у односу на друге периоде, што се може уочити и на Слици 10. Значајна заступљеност и леме *син*, у одређеној мери потврђује да су породични односи високо заступљени у периоду краја реализма, више него на његовом почетку, ако се посматра период T3. За лему *отац*, која се издваја од осталих по специфичности, уочавамо да је у периоду T1 веома специфична +23,9 (Слика 10, Табела 2). Овакав резултат може се објаснити тиме што за период T1 у корпусу постоје само два романа што доводи до пристрасности резултата: од укупног број појављивања леме *отац* у целом корпусу (267) у периоду T1 само у роману Јакова Игњатовића, „Ђурађ Бранковић: историческиј роман“, ова реч се појављује 200 пута, а од тога значајан број пута у значењу „свештено, духовно лице“.

Међутим, истраживање односа између ликова захтевало је другачији приступ, због чега су коришћени подаци о романима у википодацима.

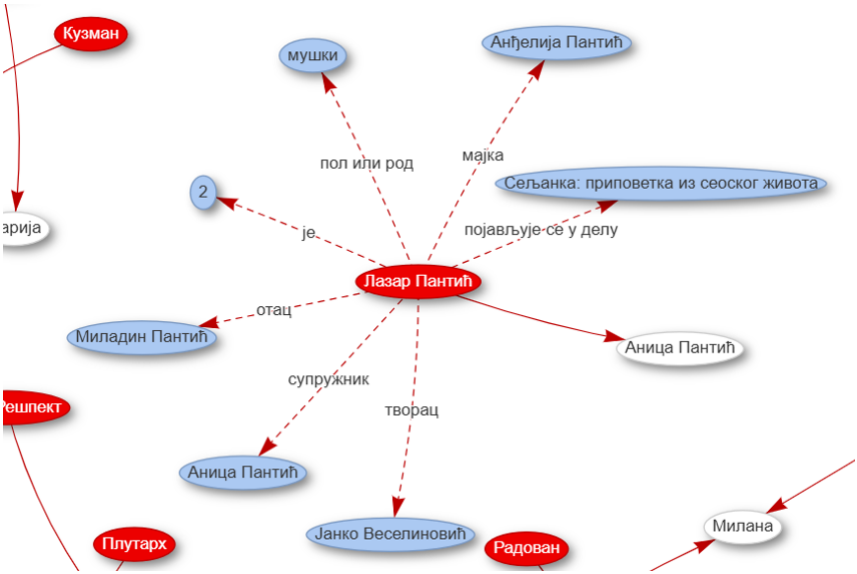
Сваки роман у википодацима представља једну RDF (Resource Description Framework)[19] структуру, а једна таква структура приказана је на примеру романа „Ивкова слава“, Стевана Сремца.

```
wd:Q109336082      # Ивкова слава
      wdt:P31      wd:Q7725634;   # је књижевно дело
      wdt:P50      wd:Q559989;    # аутор Стеван Сремац
      wdt:P407     wd:Q9299;      # језик_дела српски
      wdt:P840     wd:Q3711;      # место радње Београд
      wdt:P840     wd:Q129259;    # место радње Ниш
      wdt:P1433    wd:Q106927517; # објављено_у ELTeC
      wdt:P1476    Ивкова слава : приповетка   # наслов
      wdt:P2408     wd:Q212829       # радња се дешава у време
Ђурђевдана
   # ликови у роману
      wdt:P674      wd:Q109378039
      wdt:P674      wd:Q109378481
      wdt:P674      wd:Q109554049
      …
   Q559989        # Стеван Сремац
      wdt:P31      wd:Q5;       # је човек
      wdt:P21      wd:Q6581097;   # пол је мушки
      wdt:P19      wd:Q571136;    # место рођења Сента
```

С обзиром на приказану структуру могуће је коришћењем Wikidata Query Service[20] приказати различите визуелизације односа ликова у романима. Предност анализе односа ликова помоћу википодатака у томе је што можемо веома брзо да приступимо подацима и добијемо тачну информацију о међусобним односима појединих ликова.

Користећи упите SPARQL могу се визуелно приказати односи ликова који су унети у википодатаке. Тако се упитом SPARQL може приказати визуелни приказ односа свих супружника унутар основне и проширене колекције романа (https://w.wiki/AqMU), а део те мреже је издвојен на Слици 11. Кликом на Лазар Пантић (`Q110271637`) могу се приказати и додатна својства и везе тог лика (испрекидане црвене стрелице), на пример, ко му је отац (*отац* је Миладин Пантић (`Q110271589`)), ко му је мајка (*мајка је* Анђелија Пантић (`Q110271431`)), и у ком делу се појављује (*појављује се у делу* „Сељанка: приповетка из сеоског живота“ (`Q109336237`)). Својство *је* (`P31`) показује на број 2, што овде значи да

19. RDF, https://w3c.github.io/rdf-primer/spec/
20. Wikidata Query Service, https://query.wikidata.org/



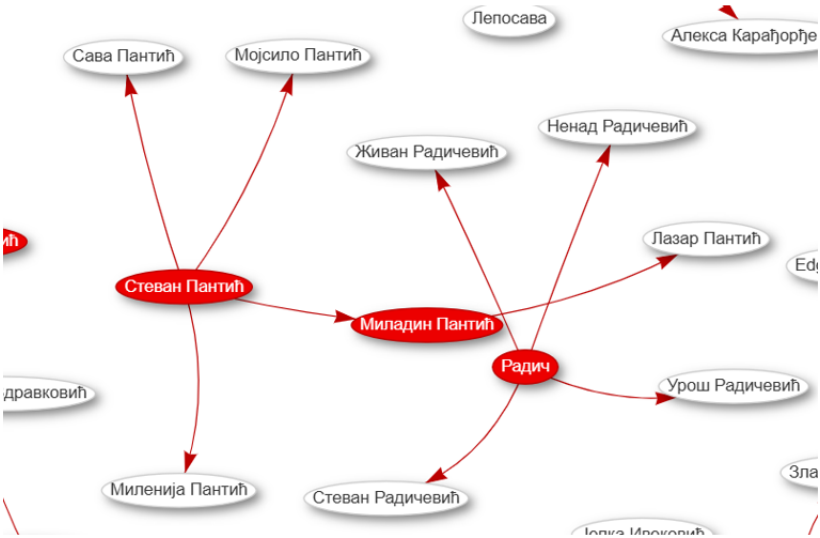
Слика 11: Граф односа муж–жена унутар колекције; црвеном бојом обојени су мужеви док пуна црвена стрелица указују на њихове жене.

је Лазар Пантић повезан са две изјаве: *измишљени човек* (`Q15632617`) и *књижевна личност* (`Q3658341`). За сада има 121 оваква релација муж–жена за ликове SrpELTeC колекције.

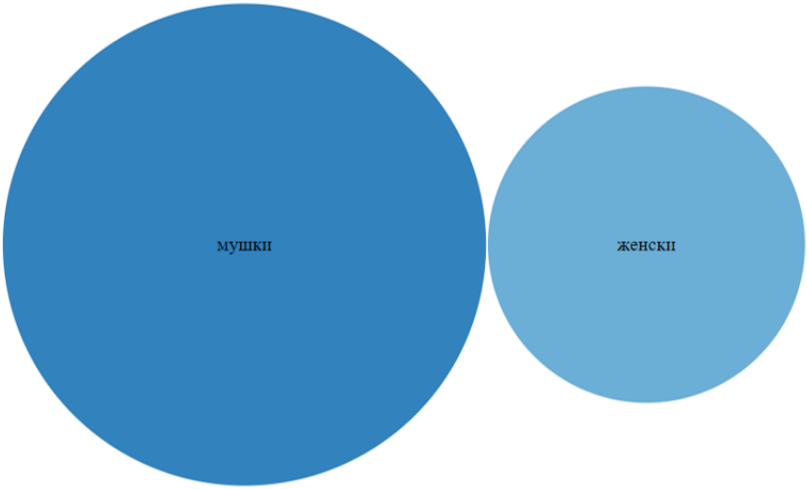
Сличним упитом може се анализирати однос дете–родитељ (https://w.wiki/AqN4). На Слици 12 је део визуелног приказа мреже добијене овим упитом. Може се уочити да је Стеван Пантић *отац* Миладина Пантића, Саве Пантића, Мојсија Пантића и Миленије Пантић, док је Миладин Пантић *отац* Лазара Пантића. Укупно за сада има 175 оваквих релација отац–дете за ликове SrpELTeC колекције.

Да бисмо добили одговор на треће истраживачко питање, користићемо SPARQL упит (https://w.wiki/AqNF) који резултате представља користећи мехурасти графикон. Величине кругова одражавају релативан однос броја ликова по полу. Резултати упита приказани на Слици 13 показују да су мушки ликови (671) заступљенији од женских (289) што потврђује претпоставку да у корпусу старих српских романа преовлађују мушки ликови.

Након текстометријске анализе и примењених техника семантичког веба, као пример могућности коришћења мрежне анализе при



Слика 12: Граф односа *отац–дете* унутар колекције; црвеном бојом је обојен чвор оца, а пуне стрелице показују његову децу.



Слика 13: Појављивање ликова у романима SrpELTeC колекције према полу.

одређивању односа ликова у роману, представљен је граф на примеру одломка романа „Нечиста крв“. Статистичка анализа мреже приказана је у Табели 3. У колони *Label* су имена ликова, док *Id* у овом случају представља QID одговарајуће ставке у википодацима.

Табела 3: Статистика мреже

Id	Label	MC	WD	E	CC	HCC	BC
Q109748906	Tone	0	2,0	2,0	0,56	0,61	0,0
Q109748839	Tomča	1	3,0	2,0	0,6	0,67	0,33
Q109693861	Sofka	0	28,0	1,0	1,0	1,0	23,67
Q109748924	Ahmet	1	5,0	2,0	0,6	0,67	0,33
Q109747507	Arsa	1	3,0	2,0	0,6	0,67	0,0
Q109748862	Simka	0	3,0	2,0	0,53	0,56	0,0
Q109746715	Magda	0	18,0	2,0	0,6	0,67	0,5
Q109747662	Mita	1	5,0	2,0	0,69	0,78	2,33
Q109748881	Todora	0	2,0	2,0	0,53	0,56	0,0
Q109747266	Marko	1	7,0	2,0	0,64	0,73	0,83

Ликови са различитим класама модуларности (енг. Modularity Class, скр. MC) вредностима припадају различитим деловима наративне мреже и подељени су у два кластера. Један кластер (MC = 0, на Слици 14 чворови зелене боје) могао би представљати групу ликова која је део одређене сцене или заплета, док други кластер (MC = 1, на слици 15 чворови плаве боје) представља другу групу ликова која је повезана са другачијим делом приче. Класа MC = 0 којој припадају ликови, Софка, Магла, Симка, Тодора, Тоне, повезује групу у којој доминира женски свет унутар романа и ликови који који обележавају радњу унутар куће. Овде се уочава и Магда која је Софкин верни пратилац, што објашњава и њихову повезаност. Софкин брат Тоне припада пасивном делу породице чији лик фигурише унутар куће, што га структурално чини ближим женским ликовима него мушкој групи, што потврђује припадност класи MC = 0. Класи MC = 1 припадају Марко, Ахмет, Мита, Томча, Арса и ову класу представља група у којој доминирају мушки ликови или ликови патријархалне моћи и трговине. Марко (WD = 7,0) и Мита (WD = 5,0) су кључне фигуре у овој групи.

Чворови представљају ликове романа, при чему је за представљање величина чвора коришћена централност посредовања (енг. Betweenness centrality, скр. BC) статистика. У анализи ликова то значи да лик који често посредује у комуникацији или интеракцији између других ликова има већу BC вредност и представљен је већим чвором

Слика 14: Граф ликова у одломку романа „Нечиста крв“.

у мрежи. Споредни ликови који не повезују различите групе или сцене представљени су мањим чворовима. Коефицијент корелације (енг. Correlation Coefficient, скр. CC) квантификује степен корелације између једног лика и других ликова у реченици и што је већи указује на већу интеракцију лика са другим ликовима у реченицама и обрнуто, што се може уочити већ код Софке (1,0), као лика који се највише појављује у анализираном одломку, али и остварује највише веза са другим ликовима, док основне централности (као што је BC) мере колико је лик централан у мрежи. Хијерархијски коефицијент кластеровања (енг. Hierarchical Clustering Coefficient, скр. HCC) даје додатне информације о томе како су ти централни ликови повезани са својим окружењем, посебно у смислу хијерархијске структуре. Већа вредност хијерархијског коефицијента кластерновања указује на то да је чвор или лик у мрежи јаче повезан са својим окружењем унутар хијерархијске структуре. Конкретно, већа HCC вредност сугерише да је чвор повезан са већим бројем суседних чворова који су такође међусобно повезани, као што је то пример код Софке (1.0). Ексцентричност (енг. Eccentricity, скр. E) је мера која се користи за описивање удаљености једног чвора од најудаљенијих чвора у мрежи. Ликови са ниском ексцентричношћу често

су кључни за причу јер су повезани са већином других ликова и играју централну улогу у друштвеној мрежи. Тако се из Табеле 3 може уочити да је Софка централни лик јер је њена вредност овог параметра најнижа (1,0). Тежински степен (енг. Weighted Degree, скр. WD) представља тежину (важност) израчунату као збир тежина (броја појављивања) свих веза повезаних са одређеним чвором (ликом у роману).

Сумирањем свих параметара из Табеле 3, статистичка анализа мреже потврђује да Софка заузима централну позицију у наративној структури посматраног одломка. Висок тежински степен (WD = 28,0) указује на њену свеприсутност и доминантан број интеракција, док вредност централности по посредовању (BC = 23,67) дефинише Софку као кључни лик који повезује све остале ликове. Њену стратешку предност додатно наглашава минимални ексцентрицитет (E = 1,0), који потврђује да је Софка непосредно повезана са свим актерима, чиме се успоставља као примарни лик у наративу. Иако има релативно мали број интеракција (WD = 5,0) Мита има другу по величини BC = 2,33. Иако би природно требало да буде центар породичног кластера (класа 0), он се појављује као спољни фактор који Софку повезује са купцима (класа 1), што потврђује његов високи BC и припадност класи ликова који су ван кућног дома. Овакав резултат такође потврђује и повезаност ефенди-Мите и Софке у релацији отац-ћерка.

Магда (WD = 18,0, али BC = 0,5) се често појављује, али скоро увек уз Софку. Њен низак BC указује на то да она нема самосталну улогу у повезивању различитих група ликова, она је „локално“ битна у Софкином окружењу, али не служи као лик који повезује различите кластере. Периферност ликова попут Тона, Арсе, Симке и Тодоре потврђена је њиховом минималном централношћу (BC = 0.0) и Ахмет и Томча (BC = 0.33), што указује на то да њихове интеракције не утичу на интеграцију шире мреже односа, већ остају ограничене на изоловане наративне делове. Ивице између ликова указују на међусобну повезаност унутар реченица романа. Што је линија дебља то је већи број заједничког појављивања повезаних ликова унутар реченица у роману. Тако се на Слици 14 може уочити да су најповезанији ликови Софка и Магда, што је и очекивано с обзиром на то да су у питању мајка и ћерка, док Софка и Арса имају само једно заједничко појављивање на шта указује и веома танка линија међу њима.

Мрежна анализа је недвосмислено потврдила породичну повезаност кроз високе вредности коефицијента груписања (CC) (Тодора мајка Софке, Мита отац Софке, Тоне брат Софкин) и модуларности (MC), који

су издвојили унутрашњи свет дома као посебан кластер. Статистички параметри су осликали породичну динамику, указујући да док висока учесталост веза (WD) доказује свакодневну блискост укућана (Магда WD = 18,0 слушкиња Софкина), високе вредност BC (Софка 23,67 и Мита 2,33) потврђује патријархалну улогу оца као кључног моста који контролише однос између породичног дома и спољашњег света трговине и моћи.

6. Закључак

Проучавајући резултате на прво истраживачко питање, закључак је да је текстометријска анализа дала очекиване резултате приликом претраживања специфичности употребе лема *мајка, отац, син, ћерка, жена, муж, сестра* и *брат* као и њихових синонима. Потврђена је зависност специфичности употребе ових лема у односу на пол аутора и да се сама специфичност поклапа са тематиком романа аутора. Даља истраживања повезана са овим питањем могу се спроводити у неколико праваца. Прво, могу се применити додатни NLP алати, за детекцију именица и глагола или фраза које се могу довести у везу са породичним односима, као и придевских атрибута који се појављују уз леме. Истраживање породичних односа унутар књижевних текстова може укључивати и истраживање њиховог појављивања у односу на одређене социолошке карактеристике писаца (пол, године, социјално порекло, образовање, професија итд.), њихове индивидуалне стилске одлике, као и према теми и типу наратива.

Указивање на већу специфичност испитиваних лема у периоду T4 у односу на друге периоде, довело је у некој мери до потврђивања тврдње да су породични односи заступљенији у периоду највећег развоја реализма него у периоду романтизма. Међутим, у будућности би се за проучавање овог истраживачког питања могли користити подаци о романима у википодацима, што ће бити могуће тек након што се сваком роману у википодацима дода и податак о припадности књижевном правцу.

Повезивање романа са википодацима и уношење ликова омогућило је пребројавање мушких и женских ликова у целој колекцији, што је показало да су родни стереотипи заступљени у SrpELTeC колекцији. Тиме је омогућена и визуелизација породичних односа за ликове и њихове односе који су унети у википодатке.

На самом крају приказана је и могућност мрежне анализе ликова у одломку роману „Нечиста крв“ која је статистичком анализом параметара мреже потврдила породичне односе међу ликовима и будућа истраживања ће засигурно ићи у смеру мрежне анализе ликова над целим корпусом.

Захвалница

Ово истраживање је финансирао Фонд за науку Републике Србије, пројекат бр. 7276, „Text Embeddings – Serbian Language Applications (TESLA)“. Аутор се захваљује професорки Цветани Крстев и професорки Ранки Станковић на уступљеном, опсежно припреманом скупу података, као и на драгоценим сугестијама које су унапредиле рад.

Литература

Bamman, David, Brendan O'Connor, and Noah A. Smith. 2013. "Learning Latent Personas of Film Characters." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,* 352–361. Association for Computational Linguistics.

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. "Gephi: an open source software for exploring and manipulating networks." In *Proceedings of the international AAAI conference on web and social media,* 3:361–362. 1.

Borin, Lars, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2009. "Thinking green: toward swedish FrameNet++." In *FrameNet Masterclass and Workshop,* 3:1–8.

Borin, Lars, and Markus Forsberg. 2010. "Beyond the synset: Swesaurus–a fuzzy Swedish wordnet." In *Workshop on Re-thinking synonymy: Semantic sameness and similarity in languages and their description. Helsinki,* 00137–7.

Carrive, Jean, Abdelkrim Beloued, Pascale Goetschel, Serge Heiden, Antoine Laurent, Pasquale Lisena, Franck Mazuet, Sylvain Meignier, Bénédicte Pincemin, Géraldine Poels, et al. 2021. "Transdisciplinary analysis of a corpus of French newsreels: The ANTRACT Project." *Digital Humanities Quarterly* 15 (1).

Cipresso, Pietro, and Giuseppe Riva. 2016. "Computational psychometrics meets Hollywood: the complexity in emotional storytelling." *Frontiers in Psychology* 7:227145. https://doi.org/10.3389/fpsyg.2016.01753.

Deretić, Jovan. 1983. *Историја Српске Књижевности.* Нолит.

Firat, Hatice. 2018. "Grandparent-Grandchild Relationships in Turkish Children's Novels." *Universal Journal of Educational Research* 6 (10): 2047–2060. https://files.eric.ed.gov/fulltext/EJ1192746.pdf.

Frontini, Francesca, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. "Named entity recognition for distant reading in ELTeC." In *CLARIN Annual Conference 2020,* 37–41.

Gledhill, Christopher, Hanna Martikainen, Alexandra Mestivier, and Maria Zimina-Poirot. 2019. "Towards a linguistic definition of 'simplified medical English': Applying textometric analysis to cochrane medical abstracts and their plain language versions." *LCM-La Collana/The Series,* 91–114.

Heiden, Serge. 2011. "The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme." In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation,* 389–398.

Ikonić Nešić, Milica, Saša Petalinkar, Mihailo Škorić, and Ranka Stanković. 2024. "BERT downstream task analysis: Named Entity Recognition in Serbian." In *Conference on Information Technology and its Applications,* 333–347. Springer.

Ikonić Nešić, Milica, Ranka Stanković, and Biljana Rujević. 2021. "Serbian ELTeC Sub-Collection in Wikidata." *Infotheca - Journal for Digital Humanities* 21 (2): 60–87. https://doi.org/10.18485/infotheca.2021.21.2.4.

Ikonić Nešić, Milica, Ranka Stanković, Christof Schöch, and Mihailo Škorić. 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 within the 13th Language Resources and Evaluation Conference,* edited by Thierry Declerck, John P. McCrae, Elena Montiel, Christian Chiarcos, and Maxim Ionov, 7–16. European Language Resources Association. https://aclanthology.org/2022.ldl-1.2/.

Jaćimović, Jelena. 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.

Kokkinakis, Dimitrios, and Mats Malm. 2011. “Character Profiling in 19th Century Fiction.” In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage,* edited by Cristina Vertan, Milena Slavcheva, Petya Osenova, and Stelios Piperidis, 70–77. Association for Computational Linguistics. https://aclanthology.org/W11-4111/.

Krstev, Cvetana. 2021. “White as Snow, Black as Night – Similes in Old Serbian Literary Texts.” *Infotheca - Journal for Digital Humanities* 21 (2): 119–135. https://doi.org/10.18485/infotheca.2021.21.2.6.

Krstev, Cvetana, Ivan Obradović, Miloš Utvić, and Duško Vitas. 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, Cvetana, and Ranka Stanković. 2020. “Old or new, we repair, adjust and alter (texts).” *Infotheca - Journal for Digital Humanities* 19 (2): 61–80. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2019.19.2.3.

Krstev, Cvetana, and Ranka Stanković. 2022. “Novels and Authors of the Serbian ELTeC Collection.” *Infotheca - Journal for Digital Humanities* 21 (2): 172–186. ISSN: 2217-9461. https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2021.21.2.13_en.

Kucserka, Zsófia. 2020. “Friends or Enemies? Sisterhood in Nineteenth-Century Hungarian Novels and Diaries.” *The Hungarian historical review: new series of Acta Historica Academiae Scientiarum Hungaricae* 9 (4): 650–666. https://doi.org/10.38145/2020.4.650.

Longhi, Julien. 2021. “Using digital humanities and linguistics to help with terrorism investigations.” *Forensic science international* 318:110564.

Makazhanov, Aibek, Denilson Barbosa, and Grzegorz Kondrak. 2014. “Extracting family relationship networks from novels.” *arXiv preprint arXiv:1405.0603.*

Min, Semi, and Juyong Park. 2016. "Network Science and Narratives: Basic Model and Application to Victor Hugo's Les Misérables." In *Complex Networks VII: Proceedings of the 7th Workshop on Complex Networks CompleNet 2016,* edited by Hocine Cherifi, Bruno Gonçalves, Ronaldo Menezes, and Roberta Sinatra, 257–265. Cham: Springer International Publishing.

Moretti, Franco. 2011. "Network theory, plot analysis."

Odebrecht, Carolin, Lou Burnard, and Christof Schöch. 2021. *European Literary Text Collection (ELTeC): April 2021 release with 14 collections of at least 50 novels.* https://zenodo.org/records/4662444.

Pantić, Marija. 2018. "Adjectival attributes with the nouns "čovek", "žena", "muškarac" and "muž"." *Infotheca - Journal for Digital Humanities* 17 (2): 66–96. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2017.17.2.4.

Patras, Roxana, Carolin Odebrecht, Ioana Galleron, Rosario Arias, J. Berenike Herrmann, Cvetana Krstev, Katja Mihurko Poniž, and Dmytro Yesypenko. 2020. *Dataset for ELTEC titles [Data set].* http://zenodo.org/records/4268669.

Prado, Sandra D., Silvio R. Dahmen, Ana LC Bazzan, Padraig Mac Carron, and Ralph Kenna. 2016. "Temporal network analysis of literary texts." *Advances in Complex Systems* 19 (03): 1650005.

Qizi Sayitqulova, Zilola Husniddin. 2021. "The Reflection of Family Relations in the Novel of A. Chulpan "Kecha Va Kunduz"." *International Journal of Linguistics, Literature and Culture* 7 (4): 188–193. https://doi.org/10.21744/ijllc.v7n4.1635.

Šandrih Todorović, Branislava, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. "Serbian NER&Beyond: The Archaic and the Modern Intertwinned." In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021),* 1252–1260. Held Online: INCOMA Ltd. https://aclanthology.org/2021.ranlp-1.141/.

Santos, Daniel, Nuno Mamede, and Jorge Baptista. 2010. "Extraction of family relations between entities." In *INForum,* 9–10.

Škorić, Mihailo, Ranka Stanković, Milica Ikonić Nešić, Joanna Byszuk, and Maciej Eder. 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.

Stanković, Ranka, Cvetana Krstev, Branislava Šandrih Todorović, Duško Vitas, Mihailo Škorić, and Milica Ikonić Nešić. 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,* 3337–3345. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.lrec-1.356/.

Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. "Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian." In *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 3954–3962. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.lrec-1.487/.

Stanković, Ranka, Diana Santos, Francesca Frontini, Tomaz Erjavec, and Carmen Brando. 2019. "Named entity recognition for distant reading in several european literatures." *DH Budapest 2019.*

Utvić, Miloš. 2013. "Izgradnja referentnog korpusa savremenog srpskog jezika." PhD diss., Univerzitet u Beogradu, Filološki fakultet. https://phaidrabg.bg.ac.rs/view/o:10061.

Vitas, Duško. 2020. "Food as Text." *Infotheca - Journal for Digital Humanities* 19 (2): 139–161. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2019.19.2.7.

Woloch, Alex. 2009. *The One vs. the Many: Minor Characters and the Space of the Protagonist in the Novel.* Princeton University Press.

Крстев, Цветана, Ранка Станковић, Бранислава Шандрих Тодоровић, and Милица Иконић Нешић. 2023. "Нове технологије за оживљавање старих текстова." In *Зборник радова Међународне научне конференције Дигитална хуманистика и словенско културно наслеђе II, Београд, 28-29 јуни 2021,* 1252–1260. Београд: Савез слависти чких друштава Србије.