Петнаест писаца и њихови дигитални отисци у бројци, слици и речи

УДК 811.163.41'322.2

САЖЕТАК: У овом раду представљамо корпус 15AUTORA сачињен од 49 дела петнаест писаца који су крајем XIX и почетком XX века писали на српском језику. Овај корпус је изведен из корпуса SrpELTeC развијеног у оквиру COST акције „Удаљено читање за историју европске књижевности". Користећи постојеће анотације (реченице, фразе на страном језику, врсте речи, леме, именовани ентитети) и додатним истраживањем програмима отвореног кода Unitex и TXM откривамо какве су дигиталне отиске оставили изабрани писци у својим делима.

КЉУЧНЕ РЕЧИ: литерарни корпус, текстометрија, корпусна лингвистика, удаљено читање, српски језик, Unitex, TXM.

РАД ПРИМЉЕН: 16. март 2026.
РАД ПРИХВАЋЕН: 01. април 2026.

Цветана Крстев
ORCID 0000-0003-3328-9392
cvetana@jerteh.rs
Друштво за језичке ресурсе и технологије – JePTeX
Београд, Србија

1. Увод

Истраживање које ће бити представљено у овом раду заснива се на корпусу ELTeC, тачније на српском делу овог корпуса SrpELTeC, који је развијен у оквиру пројекта COST Action 'Distant Reading for European Literary History' (CA16204).[1] Циљеви и резултати овог пројекта, као и изазови у изградњи српског дела корпуса представљени су у радовима (Trtovac, Milnović, and Krstev 2021; Krstev 2021a).

Корпус ELTeC је до сада већ послужио као основа за разноврсна лингвистичка, филолошка, етнолошка и информатичка истраживања. Поменућемо само нека.[2] ELTeC колекција која садржи 12 потпуних потколекција (свака са по 100 романа) представља одличну основу

1. European Literary Text Collection (ELTeC)
2. Детаљнија листа

за разноврсна вишејезична и компаративна истраживања. Начин наслањавања романа у време које корпус покрива (Patras et al. 2021) и преиспитивање рачунарским методама уобичајене тезе у књижевној историји да је унутрашњи живот ликова постао централна преокупација књижевног модернизма (Radak et al. 2024) само су неке од обрађиваних тема. Byszuk et al. (2020) баве се детектовањем директног говора у романима из девет потколекција ELTeC-а коришћењем метода заснованих на правилима и трансформерској архитектури. Утврђивање ауторства коришћењем стилометријских метода заснованих на паралелним уграђивањима (енг. embeddings) и методама дубоког учења је у фокусу рада (Škorić et al. 2022).

Истраживања која су се фокусирала на српски језик и користила потколекцију SrpELTeC такође су бројна. SrpELTeC је послужио као основа за утврђивање репертоара реторичких фигура поређења у старим литерарним текстовима и његово тестирање на савременим текстовима исте врсте (Krstev 2021b). Vitas (2022) показује да пажљиво одабран литерарни корпус може послужити као основа за изучавање различитих аспеката приватног живота у одређеном периоду. Nešić et al. (2022) представљају резултате препознавања, обележавања и повезивања с базама знања основних класа именованих ентитета. У раду (Stanković, Košprdić, et al. 2022) аутори представљају резултате анализе осећања и ставова у српским романима деветнаестог века и с почетка двадесетог века, док је поређење метода за моделирање тема илустровано на примеру исте колекције текстова у (Mihajlov et al. 2024).

2. Корпус и коришћене методе и алати

Основна српска потколекција (SrpELTeC) садржи 100 романа оригинално написаних на српском језику, објављених први пут у периоду 1840–1920, а који садрже најмање 10.000 речи. Избор дела укључених у SrpELTeC је балансиран, колико је то било могуће. Тај захтев постављен је за све језичке потколекције развијене у оквиру COST акције: равномерна покривеност временског периода (тиче се године првог издања), равномерна заступљеност романа различите дужине, подједнака заступљеност мушких и женских аутора и подједнака заступљеност канонских и мање познатих дела. Коначно, потколекција би требало да садржи 9–11 аутора представљених са по три дела, док сви остали аутори треба да буду постављени с једним делом. Додатна српска потколекција (SrpELTeC-ext) садржи још 20 романа који задовољавају

исте услове. Од ових двеју потколекција сачињен је корпус SrpELTeC-108 који садржи све романе из потколекције SrpELTeC и још осам романа из потколекције SrpELTeC-ext.

У основној потколекцији SrpELTeC једанаест аутора је представљено са по три дела, док је један аутор представљен с пет. Допуњена потколекција SrpELTeC-108 садржи дела петнаест аутора који су представљени са три или више дела. То су: Јаков Игњатовић (5), Владан Ђорђевић (3), Ђура Јакшић (3), Милан Ђ. Милићевић (3+2), Лазар Комарчић (3), Драга Гавриловић (3), Јанко Веселиновић (3), Пера Тодоровић (3), Чедомиљ Мијатовић (2+1), Стеван Сремац (3), Светолик Ранковић (3), Борисав Станковић (2+1), Јелена Димитријевић (3), Светозар Ћоровић (3), Милутин Ускоковић (2+1). Списак ових дела је дат у Додатку 4. Дела ових петнаест аутора сакупљена у поткорпус 15AUTORA биће основа нашег дигиталног истраживања.

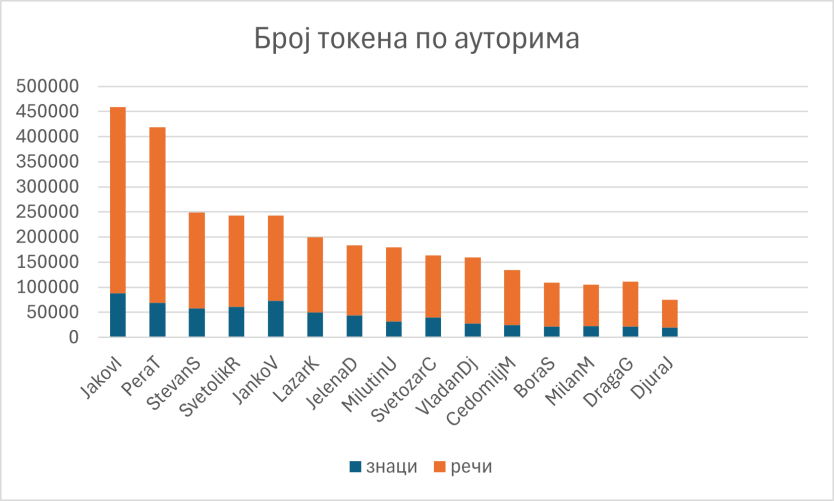
Према договору постигнутом у оквиру COST акције поткорпус SrpELTeC је опремљен у складу са TEI препорукама.[3] То значи да су обележена поглавља, пасуси, истакнути делови текста (курзив, шпационирање и сл.), делови текста на страним језицима, називи ауторских дела и фусноте. Обележавање су обавили читаоци-волонтери у току исправљања текста добијеног сканирањем и аутоматским препознавањем карактера оригиналног штампаног издања. Осим тога обављен је и низ аутоматских обележавања. Обележене су реченице, сви токени су анотирани врстом речи и лемом, а обележено је и седам категорија именованих ентитета: особе и њихове професије или позиције, локације, организације, догађаји, демоними и уметничка и остала ауторска дела (ова последња класа обележена је ручно). Више о коришћеним алатима и резултатима обележавања може се прочитати у (Stanković, Krstev, et al. 2022).

Резултати представљени у овом раду добијени су коришћењем постојећих анотација и додатним претраживањем и прорачунавањем коришћењем два програма отвореног кода. Први је Unitex (Paumier, Nakamura, and Voyatzi 2009) чији се рад заснива на електронским речницима и могућности дефинисања сложених упита и трансформација коришћењем коначних трансдуктора.[4] Овај софтвер смо користили са електронским речницима за српски језик и низом општих и наменски развијених трансдуктора (Krstev 2008) за добијање резултата представљених у одељцима 3.1, 3.2, 3.4, 3.9, 3.10, 3.12. Други софтвер

3. TEI препоруке.
4. Unitex – пакет за обраду корпуса

је TXM који је намењен разноврсним статистичким прорачунима и њиховом визуелном приказу (Pincemin, Heiden, and Mazuet 2022) (коришћен у одељцима 3.3, 3.5, 3.6, 3.7, 3.8, 3.12, 3.13).[5] Ниједан од алата који аутоматски обавља неки задатак над текстом не ради то „савршено“, увек су присутне у одређеној мери грешке и пропусти. У овом раду се нећемо тиме бавити, а успешност обављања одређених задатака може се у највећем броју случајева наћи у цитираној литератури.

3. Дигитални отисци



Слика 1. Дужине корпуса писаца мерено бројем токена: интерпункцијских и специјалних знакова и речи.

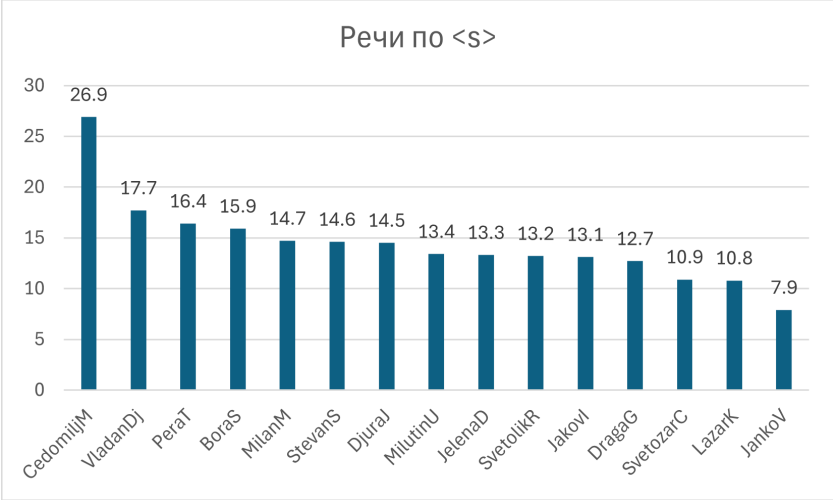
3.1 Димензија корпуса

Графикон са слике 1 показује да је у поткорпсу 15AUTORA најзаступљенији Ј. Игњатовић – његова дела чине 15,14% целог корпуса

5. TXM – платформа за текстометрију.

– док за њим следи П. Тодоровић чија дела чине 13,81% корпуса. М. Ђ. Милићевић и Ђ. Јакшић су у овом корпусу представљени најкраћим делима – 3,45%, односно 2,47%. У целом корпусу је 21,53% интерпункцијских и специјалних знакова према 78,47% речи. Највише је интерпункцијских и специјалних знакова у делима Ј. Веселиновића (30,03%) и Ђ. Јакшића (25,81%), а најмање у делима П. Тодоровића (16,56%) и В. Ђорђевића (17,23%). На основу ових података не можемо да закључимо о тенденцији неких аутора да пишу дуга или краћа дела, јер је одабир дела за SrpELTeC био вођен разноврсним критеријумима, како је напоменуто у одељку 2.

3.2 Дужина реченица

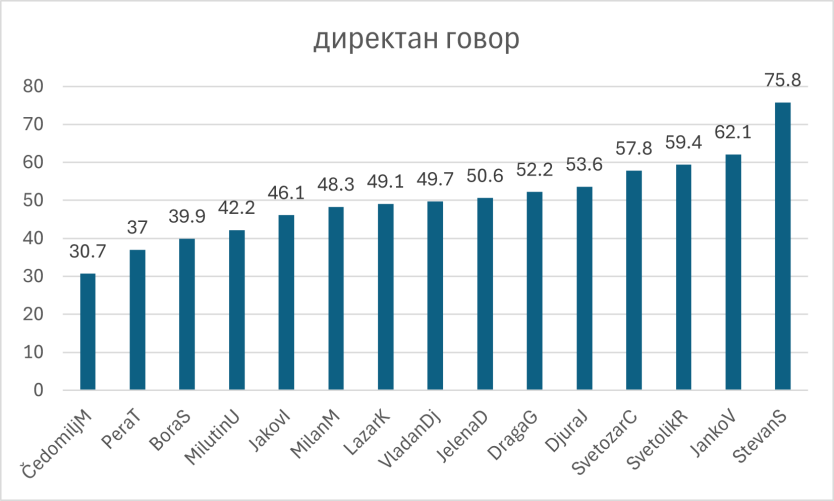


Слика 2. Дужине реченица мерено бројем речи.

Просечна дужина реченица у поткорпусу 15AUTORA је 13,4 речи по реченици. Код аутора дужина реченице варира од кратких – 7,9 речи по реченици у просеку код Ј. Веселиновића, до дугачких – 26,9 речи по реченици код Ч. Мијатовића. Ова два аутора издвајају се по дужини реченице од осталих аутора: најкраћу реченицу после Ј. Веселиновића

налазимо код Л. Комарчића (10,8 речи по реченици), а најдужу после Ч. Мијатовића код В. Ђорђевића (17,7 речи по реченици) (слика 2). Дужина реченице код М. Ускоковића је на нивоу просека поткорпуса.

3.3 Директан говор



Слика 3. Проценат пасуса који садрже директан говор по ауторима.

Речено је у одељку 1. да су већ развијани алати за препознавање и обележавање директног говора у корпусу ELTeC. Овде нас не интересује тачно препознавање директног говора у корпусу 15AUTORA већ желимо само да проценимо колико пасуса у тексту садржи директан говор. Полазимо од претпоставке да сваки пасус садржи говор само једног говорника и да директан говор отпочиње пасус што је назначено неким од интерпункцијских знакова (црта, наводници). Пример једног таквог пасуса је:

— Због̃ом, због̃ом! — вели једна другој — Лаку ноћ.

Ово свакако није најпрецизније, али нам може пружити увид у коришћење директног говора код појединачних аутора.

Уочавамо да директан говор најмање користе Ч. Мијатовић и П. Тодоровић, док убедљиво највише користе С. Сремац и Ј. Веселиновић (слика 3). Овде можемо напоменути да Ч. Мијатовић и П. Тодоровић имају најдуже реченице, Ј. Веселиновић најкраћу, а С. Сремац на нивоу просека (одељак 3.2).

3.4 Лексички параметри

У табели 1 дати су лексички параметри код писаца из корпуса 15AUTORA: број речи, број различитих речи (типова – *types*) и број различитих лема. Сви подаци у израчунати су коришћењем програма Unitex осим броја лема који је одређен програмом TXM, јер Unitex не обавља аутоматску лематизацију. Како постоје разлике у токенизацији између ових програма подаци о броју облика и лема нису директно упоредиви.[6]

Индекс коришћења лексике L израчунат је према (Smith 1973):

$$L = \frac{-\sum_X f_x \frac{X}{N} log \frac{X}{N}}{logN}$$

при чему је N дужина текста у броју речи, а X је број речи са фреквенцијом f_x. На овај начин се мери дистрибуција вокабулара у тексту и неутралише дужина текста, тј. параметри се могу упоређивати за текстове различитих дужина. Ова формула се најбоље може разумети ако претпоставимо да се неки текст састоји од једне речи која се понавља, а тада је $X = N$ и $L = 0$, док је у случају да се у тексту ниједна реч не понавља $X = 1$ и $f_x = N$ па је $L = 1$.

Слика 4 показује да учешће *hapax legomena* у тексту није у директној вези са вредношћу индекса L: процентуално учешће облика с фреквенцијом један је највеће у делима Ђ. Јакшића и М. Ђ. Милићевића (чији је удео у корпусу најмањи), док је процентуално учешће најмање у делима Ј. Игњатовића и П. Тодоровића (чији је удео у корпусу највећи). Ово сугерише да индекс L не успева да неутралише дужину текста као фактор и да би се право поређење могло добити једино коришћењем узорака исте дужине.

Уколико су речи с одређеном фреквенцијом поређане у низ од оних с фреквенцијом 1 до оних с максималном фреквенцијом, медијана

6. Програм TXM одређује врсте речи и леме коришћењем програма TreeTagger (Schmid 2013) чија је верзија за српски описана у (Utvić 2011; Stanković, Škorić, and Šandrih Todorović 2022).

аутор	речи	облици	леме	L	hapax	hapax%
JakovI	371243	38387	17545	0,8053	20958	5,65
PeraT	349721	43165	1804	0,7765	23820	6,81
JankoV	169685	23020	10953	0,7754	12989	7,65
StevanS	191094	27323	14431	0,7735	16125	8,44
SvetolikR	181467	26038	12165	0,7646	14859	8,19
BoraS	87601	14251	6806	0,7620	8492	9,69
DragaG	88733	14590	7188	0,7531	8522	9,6
JelenaD	139444	20906	10501	0,7514	11554	8,29
LazarK	149847	23812	11493	0,7504	14025	9,36
CedomiljM	110162	18312	8336	0,7396	10475	9,51
VladanDj	132100	23235	11002	0,7326	14027	10,62
SvetozarC	124390	23894	9816	0,7193	14670	11,79
DjuraJ	55579	12249	6573	0,7026	7791	14,02
MilanM	82356	17935	9459	0,6929	11200	13,6
MilutinU	147418	26354	12180	0,6836	15705	10,65

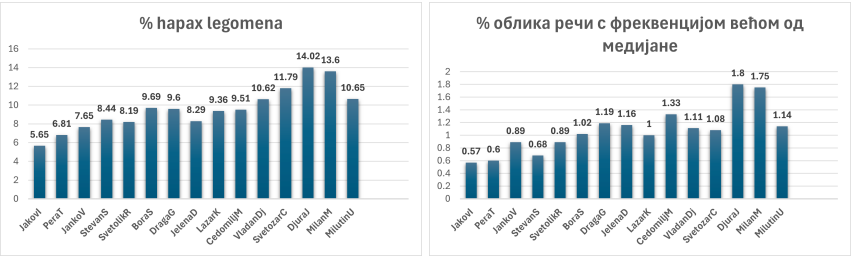
Табела 1. Лексички параметри код писаца из корпуса 15AUTORA.

представља ону фреквенцију за коју све речи са мањом фреквенцијом покривају половину текста, док речи са већом фреквенцијом покривају такође половину текста. На слици 4 уочава се да је проценат речи чија је фреквенција већа од медијане највећи опет у делима Ђ. Јакшића и М. Ђ. Милићевића, док је проценат речи чија је фреквенција већа од медијане најмањи исто код Ј. Игњатовића и П. Тодоровића. Ипак, не постоји директна веза између учешћа *hapax legomena* и речи с фреквенцијом већом од медијане.

3.5 Учесталост врста речи

Врсте речи су у корпусу 15AUTORA разврстане према препорукама пројекта *Universal Dependencies* (Marneffe et al. 2021) у 16 класа. Фреквенција појављивања ових класа у 15AUTORA је следећа:

- интерпункцијски знаци – PUNCT, $f = 567.364$ (19,3%),
- именице – NOUN, $f = 417.261$ (14,2%);
- глаголи – VERB, $f = 88.330$ (13,2%);

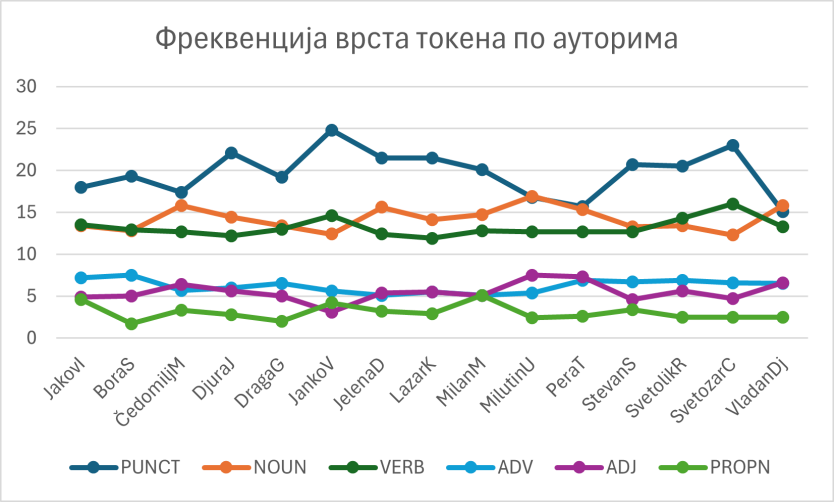


Слика 4. Проценат речи (облика) с фреквенцијом: један (лево); већом од медијане (десно). Писци су поређани по вредности индекса L, од већег ка мањем.

- предлози – ADP, $f = 188.284$ (6,4%);
- прилози – ADV, $f = 186.921$ (6,4%);
- помоћни глаголи – AUX $f = 179.460$ (6,1%);
- независни везници – CCONJ $f = 177.297$ (6,0%);
- придеви – ADJ, $f = 162.709$ (5,5%);
- заменице – PRON, $f = 147.878$ (5,0%);
- речце – PART, $f = 138.947$ (4,7%);
- детерминатори – DET, $f = 134.734$ (4,6%);
- зависни везници – SCONJ, $f = 115.004$ (3,9%);
- властите именице – PROPN, $f = 93.807$ (3,2%);
- кардинални бројеви – NUM, $f = 22.764$ (0,8%);
- узвици – INTJ, $f = 8.953$ (0,3%);
- остало (речи на страном језику, скраћенице, итд.) – X, $f = 7.969$ (0,3%).

На сликама 5 и 6 приказано је учешће ових класа код појединачних аутора. Интерпункцијски знаци, именице и глаголи се јасно издвајају у односу на остале токене: код свих аутора ове врсте токена су најучесталије. Коришћење интерпункцијских знакова доста варира у односу на просек: чак 24,8% свих токена су интерпункцијски знаци код Ј. Веселиновића, док их је код В. Ђорђевића и П. Тодоровића тек нешто више од 15%. Учешће глагола не варира значајно од аутора до аутора – најмање је 11,9% (код Л. Комарчића), највеће 16,0% (код С. Ћоровића). Веће коришћење именица уочава се код М. Ускоковића 16,9%, док је просек 14,2%.

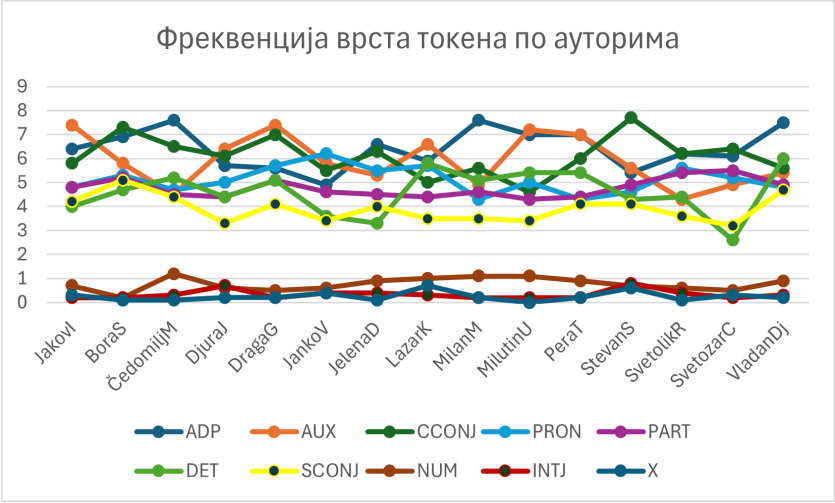
Прилози, придеви и властите именице користе се значајније мање од именица и глагола. Коришћење прилога је углавном уједначено, и креће



Слика 5. Учешће интерпункцијских знакова и значећих речи у текстовима појединачних аутора.

се око просека 6,4%, а једино искакање се уочава код Б. Станковића – 7,5%, односно Ј. Димитријевић и М. Ђ. Милићевића – 5,1%. Код Ј. Веселиновића придеви чине свега 3,1% целог текста, док је просек на целом корпусу 5,5%. Више придева у односу на просек користе М. Ускоковић 7,5% и П. Тодоровић 7,3%. Више властитих имена у односу на просек 3,2% јавља се код М. Ђ. Милићевића (5,1%) док их је најмање код Б. Станковића (1,7%).

Коришћење кардиналних бројева и узвика, као и „осталих“ речи значајно је мање у односу на остале функционалне речи и креће се од нула до 1,2%, што одговара коришћењу кардиналних бројева код Ч. Мијатовића. Коришћење свих осталих функционалних речи варира од аутора до аутора. Рецимо, коришћење именских заменица је најмање код П. Тодоровића и М. Ђ. Милићевића (4,3%) у односу на просек (5%), док их највише користи Ј. Веселиновића (6,2%). Најмање помоћних глагола користи С. Ранковић (4,3%) у односу на просек (6,1%), а највеће је учешће помоћних глагола у текстовима Ј. Игњатовића и Д. Гавриловић (7,4%).

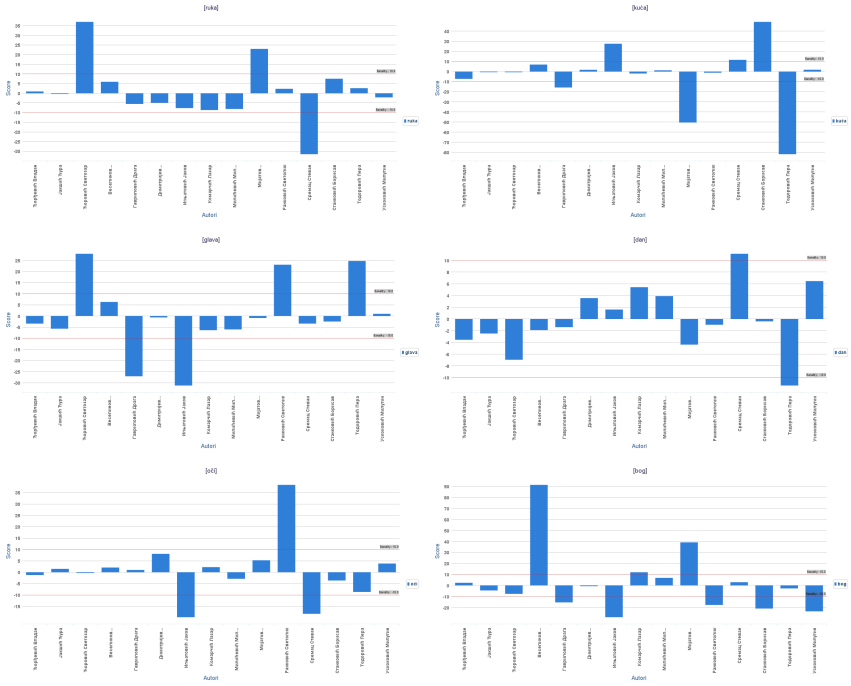


Слика 6. Учешће функционалних речи у текстовима појединачних аутора.

3.6 Именице

У корпусу 15AUTORA 20 најфреквентнијих именица су редом: *рука*, *кућа*, *глава*, *дан*, *очи*, *бог*, *чов(ј)ек*, *људи*, *пут*, *жена*, *срце*, *р(иј)еч*, *лице*, *отац*, *глас*, *година*, *мајка*, *страна*, *време*, *св(иј)ет*. Од ових 20 именица три су међу 20 најфреквентнијих код свих писаца из корпуса, то су: *рука* (на позицијама од 1 до 11), *дан* (на позицијама од 1 до 12) и *очи* (на позицијама од 1 до 15). Најређе се појављује на првих 20 позиција код појединачних писаца именица *време* (код четири аутора). Код Л. Комарчића је међу 20 најфреквентнијих именица 16 оних које су међу најфреквентнијима на целом корпусу 15AUTORA, док се код Ч. Мијатовића међу најфреквентнијима налази само 10 оних које су међу најфреквентнијима на целом корпусу.

Код већине аутора, њих 11, на првој позицији се појављује нека именица која је међу 20 најфреквентнијих на целом корпусу 15AUTORA. Издвајају се Ђ. Јакшић код кога је најфреквентнија именица *господин*, Д. Гавриловић код које је то *љубав*, док је код П. Тодоровића и Ч. Мијатовића најфреквентнија именица *кнез*. Код два аутора су на другој позицији нове именице: *чича* код Ђ. Јакшића, *поп* код С. Сремца. Од именица које су најфреквентније код појединих писаца а нема их међу



Слика 7. Специфичност употребе именица *рука*, *кућа*, *глава*, *дан*, *очи*, *бог* код 15 аутора

20 најфреквентнијих на целом корпусу најчешће се јавља *душа* (шест пута) и *земља* (пет пута). Занимљиво је да се код ауторке Д. Гавриловић осим именица *љубав*, *срце* и *душа*, јављају и именице које означавају породичне односе: *жена*, *муж*, *д(иј)ете*, *д(ј)еца* и *сестра*. Слично је и код друге ауторке Ј. Димитријевић код које се јављају *душа* и *срце* али и *мајка*, *отац*, *син*, *кћер* и *ханума*. Од именица које означавају осећања јављају се само *љубав* (код Д. Гавриловић и М. Ускоковића) и *страх* (код Б. Станковића).

Индекс *специфичности* омогућава упоређивање употребе одабраних језичких јединица (речи, фраза, конструкција) код појединих аутора. Овај индекс одражава неочекиваност појављивања одабраних језичких јединица код појединачних аутора и он се рачуна на основу: f – број појављивања одабране јединице код одабраног аутора; F – број

појављивања исте јединице у целом корпусу; t – укупан број јединица код одабраног аутора; T – укупан број јединица у целом корпусу. Да би се поредиле фреквенције појављивања неких језичких јединица у деловима корпуса (партицијама) различите величине, потребно је извршити нормализацију. Уобичајено је да се рачуна релативна фреквенција као однос апсолутне фреквенције појаве лексичке јединице у том делу корпуса и укупне дужине тог дела, што одговара *нормалној расподели*. Међутим, уочено је да појављивање језичких јединица у корпусу више одговара *хипергеометријској расподели*. Очекивање да се код једног аутора (у његовој партицији корпуса) појави f тражених лексичких јединица је:

$$Pr(X = f) = \frac{\binom{F}{f}\binom{T-F}{t-f}}{\binom{T}{t}}$$

Позитивне вредности индекса специфичности говоре да се језичка јединица појављује чешће него што је очекивано, док негативне вредности говоре да се појављује ређе. Више о индексу специфичности у (Jaćimović 2019).

Индекси специфичности представљени графички на слици 7 показују да се најфреквентније именице *рука*, *кућа*, *дан*, *глава*, *очи* и *бог* код В. Ђорђевића, Ђ. Јакшића, Ј. Димитријевић, Л. Комарчића и М. Ђ. Милићевића јављају у очекиваним границама. Код Ј. Веселиновића и М. Ускоковића само именица *бог* одступа од очекиваног, код првог се јавља знатно више, а код другог знатно мање. Код С. Ћоровића се *рука* и *глава*, а код С. Ранковића *глава* и *очи* јављају знатно изнад очекиваног, док се код Д. Гавриловић *кућа*, *глава* и *бог* јављају знатно испод очекиваног. Код Ч. Мијатовића се *рука* и *бог* јављају више од очекиваног, а *кућа* мање, код С. Сремца се *дан* јавља чешће, а *рука* и *очи* ређе од очекиваног, код Б. Станковића се *кућа* јавља изнад очекиваног а *бог* испод, док се код П. Тодоровића *глава* јавља чешће, а *кућа* и *дан* ређе. Највише одступања се уочава код Ј. Игњатовића: *кућа* се јавља изнад очекиваног, а *глава*, *очи* и *бог* испод очекиваног.

3.7 Глаголи

Најфреквентнијих 20 глагола у корпусу 15AUTORA су редом: *моћи*, *рећи*, *знати*, *вид(ј)ети*, *имати*, *доћи*, *казати*, *ићи*, *гледати*, *говорити*, *мислити*, *чути*, *морати*, *почети*, *дати*, *требати*, *узети*, *остати*, *немати*, *отићи*. Од ових 20 глагола пет су међу 20 најфреквентнијих

код свих писаца из корпуса: *моћи* (на позицијама од 1 до 5), *знати* (на позицијама од 1 до 6), *вид(ј)ети* (на позицијама од 1 до 11), *имати* (на позицијама од 2 до 11) и *доћи* (на позицијама од 2 до 14). Глагол *рећи* је међу најфреквентнијих 20 код свих писаца (на позицијама од 1 до 9) осим код Б. Станковића. Код Ј. Игњатовића су нафреквентнији сви глаголи са листе најфреквентнијих на целом корпусу осим глагола *гледати*. Најмање појављивања, тачно 16, глагола са ове листе се јавља у листама најфреквентнијих код Ј. Веселиновића, Ч. Мијатовића, Б. Станковића, М. Ускоковића и С. Ћоровића.

Код свих аутора је најфреквентнији глагол неки из листе најфреквентнијих на целом корпусу, осим код С. Сремца код кога је на врху листе *велим*, који је на целом корпусу на 21 позицији најфреквентнијих. Од глагола који су најфреквентнији код појединих писаца а нема их међу 20 најфреквентнијих на целом корпусу најчешће се јавља *погледати* (осам пута) и *одговорити*, *питати* и *стати* (пет пута). Уочава се да у листи најфреквентнијих глагола на целом корпусу и на листама најфреквентнијих код појединих писаца налазе глаголи изведени префиксацијом – *гледати* и *погледати*, *говорити* и *одговорити*, *питати*, *упитати* и *запитати* – као и видски парњаци: *осетити* и *осећати*, *стати* и *стајати*, *доћи* и *долазити*.

Глагола осећања нема међу двадесет најфреквентнијих на целом корпусу. Глагол *волети* је међу најфреквентнијим код Д. Гавриловић (на позицији 10), Ј. Димитријевић (на позицији 9) – једина два женска аутора – и М. Ускоковића (на позицији 5), а код Д. Гавриловић се јавља још и *љубити* (на позицији 11). Глагол *осећати* је међу најфреквентнијим код Б. Станковића (на позицији 12) и М. Ускоковића (на позицији 13), а код Б. Станковића се јавља и *осетити* (на позицији 19). Код Ј. Димитријевић се јавља још и *плакати* (на позицији 20).

3.8 Придеви

Двадесет придева који се најчешће јављају у корпусу 15AUTORA су редом: *други*, *велик*, *стар*, *добар*, *л(иј)еп*, *млад*, *цео/цијел*, *први*, *црн*, *нов*, *мали*, *пун*, *бео/бијел*, *жив*, *исти*, *српски*, *по(с/ш)(л/љ)едњи*, *турски*, *сре(ћ/т)ан*, *читав*. Код одређивања најфреквентнијих придева код појединачних аутора искључили смо присвојне придеве од властитих имена, нпр. *Кочин* и *Даринчин*. Тако смо добили да су пет од ових придева међу 20 најфреквентнијих код свих аутора: *други*, *велик*, *стар*, *л(иј)еп* и *млад*, док су три међу најфреквентнијим код свих аутора сем

једног: *добар* (код Б. Станковића на позицији 27), *први* (код Ђ. Јакшића на позицији 25) и *мали* (код П. Тодоровића на позицији 25). Придев *читав* се налази међу најфреквентнијих 20 код само три аутора, али зато код С. Ћоровића на позицији 3 (за разлику од целог корпуса где је на позицији 20).

Код свих аутора је најфреквентнији придев неки из листе најфреквентнијих на целом корпусу, а најчешће је то придев *други* – код 11 аутора – који је најфреквентнији и на целом корпусу. Од придева који су најфреквентнији код појединих писаца а нема их међу 20 најфреквентнијих на целом корпусу најчешће се јавља *дуг* (шест пута) и *страшан*, *тежак*, *весео* и *женски* (пет пута). Придеви *црн* и *бео/бијел* су међу 20 најфреквентнијих на целом корпусу, а оба се јављају и код већине аутора, њих девет. Код Ј. Игњатовића и Д. Гавриловић се на листи 20 најфреквентнијих не јавља ниједан од ових придева, код П. Тодоровића, С. Ранковића и С. Ћоровића се јавља само *црн*, а код С. Сремца само *бео/бијел*.

Међу 20 најфреквентнијих у корпусу 15AUTORA јављају се демоними *српски* (на позицији 16) и *турски* (на позицији 18). Оба придева јављају се међу 20 најфреквентнијих само код Ј. Игњатовића и П. Тодоровића (уз *балкански* на позицији 20). Само *српски* се јавља на листи најфреквентнијих код М. Ђ. Милићевића, Л. Комарчића и Ч. Мијатовића, а само *турски* код Ј. Веселиновића и Ј. Димитријевић, код које се јављају још и демоними *француски* и *европски*.

Осим придева *црн* и *бео/бијел*, друге боје се не појављују у листама најфреквентнијих ни код једног аутора, осим код Ч. Мијатовића код кога се јављају *црвен* и *зелен*, а осим тога још и *сребрн* и *златан* који се такође могу односити на боју (видети одељак 3.9). Придеви који се јављају у листама најфреквентнијих код неких аутора, а који нису најфреквентнији у целом корпусу, могу указивати на специфичност стила или теме аутора. Такви придеви су, на пример:

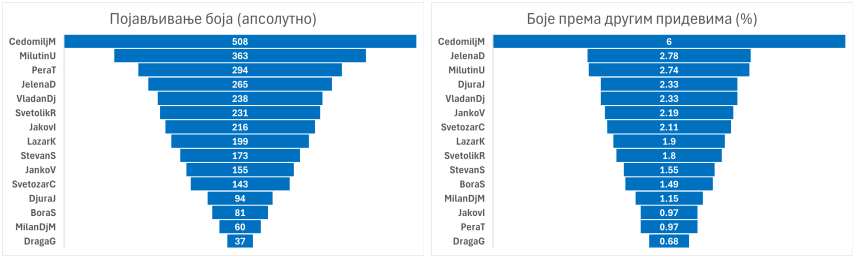
- *страшан*, *силан*, *диван*, *сјајан* и *красан* код В. Ђорђевића,
- *сирот*, *блед*, *гладан*, *сам*, *несрећан* и *тужан* код Ђ. Јакшића,
- *сирот*, *богат*, *ваљан*, *чист*, *весео* и *поштен* код Д. Гавриловић и
- *широк*, *голем*, *дуг*, *танак*, *ситан*, *дебео* и *јак* код С. Ћоровића.

3.9 Придеви који означавају боје

Посебан семантички маркер који се у електронским речницима српског језика користи за придеве који означавају боје омогућава нам

да установимо какво је коришћење ових придева у делима 15 аутора. У корпусу 15AUTORA јавља се 3.057 придева који означавају боје и то 109 различитих, ако искључимо придеве *црн* и *бео/бијел*. Придеви који се најчешће користе су редом: *црвен*, *зелен*, *златан*, *плав*, *румен*, *сребрн*, *жут*, *сив*, *модар*, *ружичаст*, *смеђ*, *сур*, *беличаст*, *жућкаст*, *риђ*, *зеленкаст*, *љубичаст*, *црвенкаст*, *плаветан*, *плавичаст*. У овом случају нисмо раздвајали значења, тако да *златан* и *сребрн* могу бити и градивни придеви.

У апсолутном износу највише придева који означавају боје јавља се код Ч. Мијатовића (508) и највише различитих (35). Најмање ових придева користи Д. Гавриловић (37), а и најмање различитих (13, као и М. Ђ. Милићевић) (слика 8, лево). У односу на све придеве које аутор користи, процентуално се највише њих односи на боје код Ч. Мијатовића, а најмање код Д. Гавриловић (слика 8, десно).



Слика 8. Учешће придева који означавају боје код појединачних аутора.

Међу 109 различитих придева налази се 17 оних који означавају да је нешто помало одређене боје или нагиње тој боји, и то у више варијанти: *беличаст*, *жу(т/ћ)каст*, *загаситоруменкаст*, *зеленкаст*, *златкаст*, *мораст*, *мркајаст*, *отворенозеленкаст*, *плаветникаст*, *плавичаст*, *плав(к/ч)аст*, *плавуш(к)аст*, *пурпураст*, *сив(к)аст*, *сребрнкаст*, *црвенкаст*, *црнкаст*.

У корпусу 15AUTORA идентификовано је 11 придева турског порекла који означавају боје : *алев*, *ален*, *алов* (отворено црвен), *ђувез* (тамноцрвен), *карпус* (боје лубенице), *крмезли* (врста црвене), *мавен* (модроплав), *мор* (љубичаст), *пембе* (ружичаст), *скрлатан* и *скрлетан* (љубичастоцрвен). Занимљиво је да се придев *алев* и његове варијанте јавља у 15AUTORA 12 пута, а од тога само два пута у изразу *алева*

паприка код М. Ускоковића у *Чедомиру Илићу* и код Ђ. Јакшића у *Чича Тими*. Остала појављивања се сва односе на одевне предмете – чарапе, шалваре, јаглак, либаде и фес.[7] Непроменљивих придева, који су по правилу страног порекла, јавља се 10 у више варијанти: *бра(о/у)н*, *виолет*, *грао*, *ђувез*, *зејтини* (боје маслиновог уља), *карпус*, *крмезли*, *мор*, *пембе*, *роз(а/е)*.

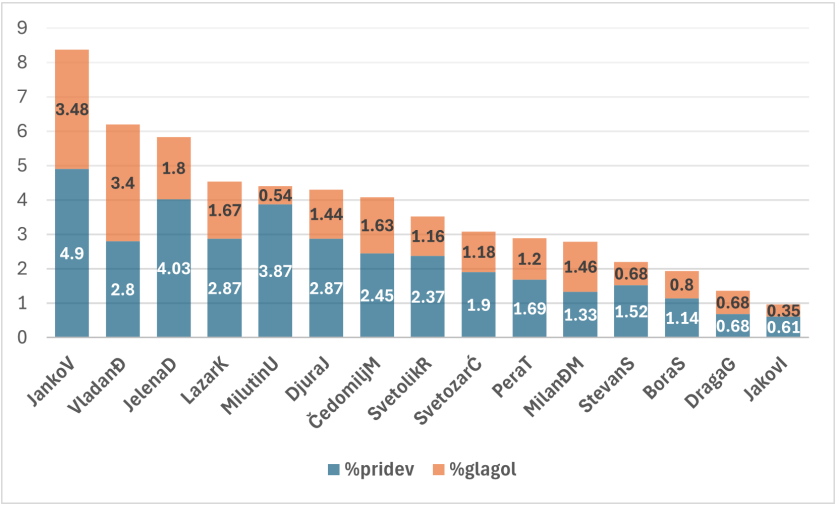
Међу идентификованим придевима који означавају боје је и 31 сложени придев од којих њих 10 означавају комбинацију боја: *беличастосребрн*, *жућкастозелен*, *зеленкастоплаветан*, *зеленожут*, *плавобео*, *плавосив*, *сивоплав*, *црвенкастосив*, *црножут* и *чивитастомодар* (индиго модар). Осталих 19 представљају модификацију боје: *бледо-* (*жут*, *зелен*, *љубичаст*, *плав*, *румен*), *загаситоруменкаст*, *затвореножут*, *мрко-* (*жут*, *сив*), *мутноплав*, *отворено-* (*зеленкаст*, *љубичаст*), *полу-* (*зелен*, *златан*), *снежнобео*, *тамно-* (*зелен*, *плав*, *сив*, *црвен*). У преостала два случаја боја се јавља као модификатор: *жућкастоблед* и *руменосјајан*.

3.10 Стилска фигура поређења

У радовима (Krstev 2021b) и (Krstev, Stanković, and Marković 2023) описали смо алат за препознавање и анотирање стилских фигура поређења за српски, засноване на опису ових стилских фигура, електронским речницима за српски и коначним трансдукторима. Овај алат нам је омогућио да детектујемо у корпусу 15AUTORA 837 стилских фигура поређења, од тога 521 придевских и 315 глаголских.

Међу 301 различитом придевском стилском фигуром поређења најфреквентнија је *бео као снег* са 30 појављивања, а за њом следе: *блед као смрт* са 25 појављивања, *хладан као лед* са 15, *блед као крпа* (10), *бео као млеко* (9), *млад као капља* (8), *црвен као крв* (7) и *добар као дан*, *жут као восак*, *љут као рис*, *мек као памук* сви са шест појављивања. У корпусу се 223 придевске стилске фигуре поређења појављују само једанпут. Међу овим стилским фигурама основа поређења су најчешће придеви *бео* и *црн* који се јављају у 12 различитих стилских фигура, а следе придеви *блед*, *црвен* и *румен* који се јављају у 10 и *чист* у девет стилских фигура. Именице *анђео*, *јагње*, *птица* и *сунце* јављају се као извор поређења у пет, док се именице *ватра*, *дан*, *жеравица*, *земља*, *небо* и *смрт* јављају све у по четири различите стилске фигуре.

7. Значајна разлика у односу на коришћење данас: у корпусу SrpKor21 придев *алев* се 976 пута јавља као модификатор именице и то је увек *паприка*.

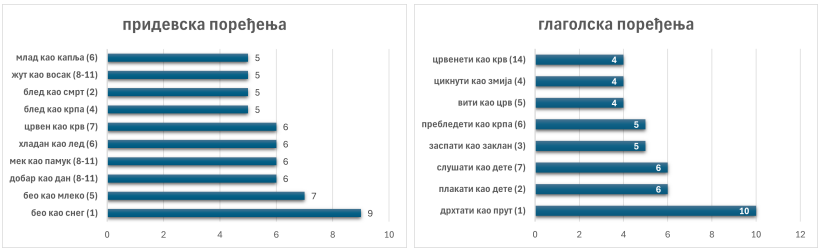


Слика 9. Релативна фреквенција придевских и глаголских стилских фигура поређења код појединачних аутора.

Међу 164 различите глаголске стилске фигура поређења најфреквентнија је *дрхтати као прут* са 20 појављивања, а за њом следе: *плакати као дете* са 13 појављивања, *заспати као заклан* и *цикнути као змија* са 8, *вити (се) као црв* и *пребледети као крпа* са 7, *слушати као дете* (6) и *смејати (се) као луд* (5). У корпусу се 99 глаголских стилских фигура поређења појављује само једанпут. Као основа поређења, глагол *стајати* се јавља у пет различитих стилских фигура, следе *живети* и *изгледати* који се јављају у четири, а затим *вити се*, *заспати*, *јурнути*, *падати*, *плакати*, *сејати*, *скочити* и *спавати* сви у три различите стилске фигуре. Именица *дете* је извор поређења у девет, именице *змија* и *црв* у шест, а *пут* и *стрела* у пет различитих глаголски фигура.

Највећи број стилски фигура у апсолутном броју – укупно 142 – детектовано је у делима Ј. Веселиновића, али је релативна фреквенција[8] такође највећа код овог аутора, као што се види на слици 9. Најмање стилских фигура поређења је уочено код Д. Гавриловић, само 12, док је

8. Релативна фреквенција је рачуната као број стилских фигура на 10,000 речи текста.



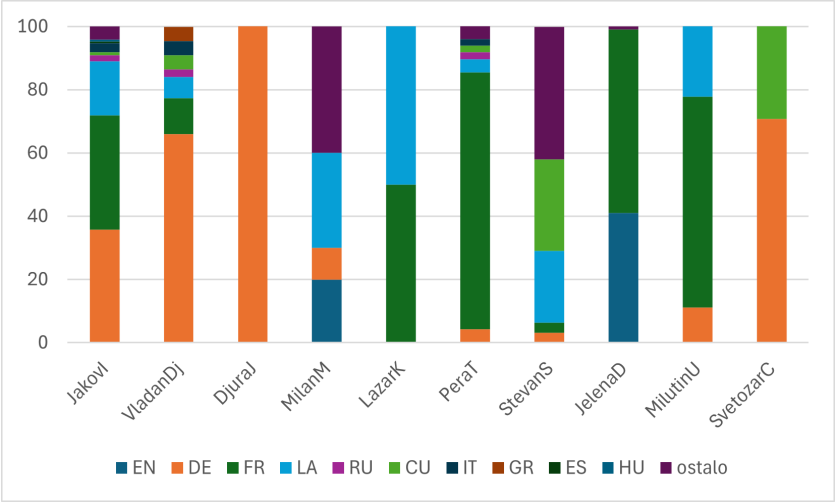
Слика 10. Стилске фигуре поређења које се јављају у делима највише аутора (у заградама је дат ранг тих израза по апсолутној фреквенцији у корпусу).

релативна фреквенција коришћења најмања код Ј. Игњатовића. Код Ј. Веселиновића је највећа апсолутна фреквенција коришћења придевских (83) и глаголских (59) фигура поређења, као и њихова релативна фреквенција. Код неких аутора се уочава значајније коришћење придевских у односу на глаголске фигуре (код Ј. Димитријевић, а посебно М. Ускоковића), док В. Ђорђевић користи у својим делима више глаголских фигура.

Најфреквентнија придевска фигура поређења *бео као снег* јавља се и у делима највише аутора, њих девет, а за њом следи *бео као млеко* која се јавља код седам аутора (видети слику 10, лево). Од глаголских фигура поређења у делима десет аутора јавља се *дрхтати као прут*, а *плакати као дете* и *слушати као дете* код шест аутора (видети слику 10, десно).

3.11 Коришћење страних језика

Волонтери (акцијаши) који су читали и кориговали скениране и ишчитане текстове окруживали су XML етикетама `<foreign>` и `</foreign>` делове текста на страном језику, а као вредност атрибута `xml:lang` уносили двословну скраћеницу језика. С обзиром на то да волонтери у највећем броју случајева нису били стручњаци за српски језик, речено им је да користе скраћеницу “cu” за црквено-словенски али и за све варијанте српског које нису биле записане Вуковом азбуком. У текстовима су се понекад јављали делови текста на страном језику, рецимо турском или немачком, али записани ћирилицом и према изговору, нпр. „мајн грус“ (од немачког ‘Meine Grüße’). Сва оваква појављивања су у нашој анализи груписана под „остало“. Не може се очекивати да се неки од делова на страном језику понавља, па



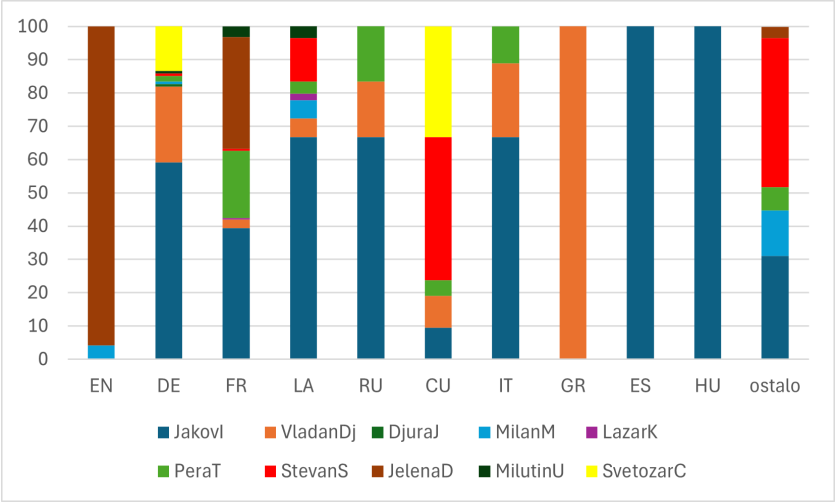
Слика 11. Коришћење различитих језика по ауторима изражено у процентима у односу на укупан број сегмената на страном језику.

ипак фраза на француском „Mon Dieu“ јавља се 17 пута (увек код Ј. Димитријевић у роману *Нове*) а „corpus delicti“ четири пута (код С. Сремца у *Поп Ћира и поп Спира* и *Ивкова слава*).

Ове етикете смо користили да анализирамо коришћење делова на страном језику у корпусу 15AUTORA. Проналазимо да од 15 аутора, њих пет није користило стране језике у својим делима. То су: Д. Гавриловић, Ј. Веселиновић, С. Ранковић, Ч. Мијатовић и Б. Станковић. На основу овога свакако не можемо да закључимо ништа о познавању или непознавању страних језика аутора, а за то је еклатантан пример Ч. Мијатовић.[9]

Остали аутори су користили делове на страном језику често (Ј. Игњатовић и Ј. Димитријевић – 291, односно 112 појављивања) или веома ретко (Ђ. Јакшић и Л. Комарчић – једно, односно два појављивања). Коришћење код осталих аутора је: П. Тодоровић 48 појављивања, В. Ђорђевић 44, С. Сремац 31, С. Ћоровић 24, М. Ђ. Милићевић 10 и М. Ускоковић 9. Слика 11 показује да су неки аутори уметали у текст делове на разним језицима – Ј. Игњатовић је користио

9. Чедомиљ Мијатовић у Википедији.



Слика 12. Коришћење језика код различитих аутора изражено у процентима у односу на број сегмената на том језику.

све језике који се појављују у корпусу 15AUTORA сем енглеског и грчког, В. Ђорђевић све сем енглеског, шпанског и мађарског, П. Тодоровић све сем енглеског, грчког, шпанског и мађарског. Ј. Димитријевић је користила само енглески и француски. Немачки језик је користило осам, а француски шест од десет аутора.

Слика 12 нам говори да је енглески језик користила само Ј. Димитријевић (изузев два погрешно анотирана сегмента код М. Ђ. Милићевића), док су немачки, латински и француски користили многи аутори. Руски језик се појављује код три аутора, али је таквих појављивања мало – само шест. Види се да су С. Сремац и Ј. Игњатовић највише прибегавали записивању страног језика српским писмом и по изговору.

3.12 Улоге људи у романима

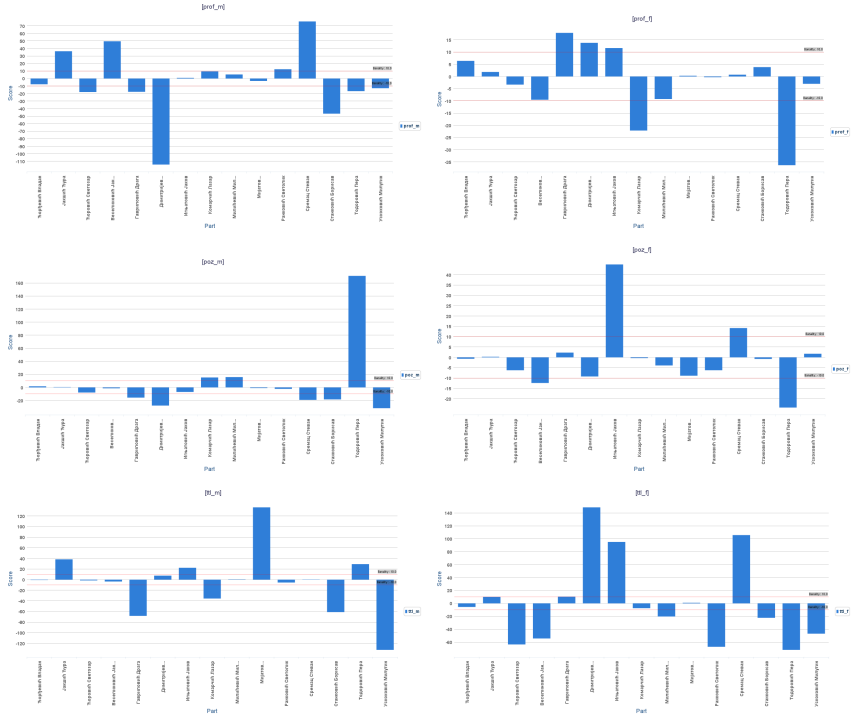
Као што је већ речено у одељку 1., у корпусу SrpELTeC, па и у његовом изводу 15AUTORA, аутоматски је обележено седам класа именованих етикета, од којих се једна односи на улоге људи који се јављају у романима (ROLE), било да је реч о стварним или

фиктивним личностима. Електронски морфолошки речници српског језика у којима су улоге људи обележене адекватним семантичким маркерима омогућили су нам да разврстамо сва појављивања улога у корпусу на професије људи, на позиције у друштву или некој организацији и на титуле, односно начине обраћања. Осим тога, речници су нам омогућили да их даље разврстамо на она која се односе на мушкарце, односно, на жене. У табели 2 дато је 20 најфреквентнијих улога у корпусу 15AUTORA разврстаних према наведеним критеријума. Из табеле се види да су професије, позиције и титуле мушкараца знатно заступљеније него исте улоге код жена, као што је показују прелиминарни резултати у (Stanković, Krstev, and Vitas 2024).

проф/м	*F*	**проф/ж**	*F*	**поз/м**	*F*	**поз/ж**	*F*	**ттл/м**	*F*	**ттл/ж**	*F*
поп	1892	слушкиња	157	везир	806	газдарица	173	кнез	2254	госпођа	1163
учитељ	676	учитељица	147	хајдук	589	попадија	159	господин	1693	гђа	498
доктор	558	калуђерица	48	газда	519	натарошевица	24	г .	1423	госпођица	293
кмет	523	кухарица	46	отац	403	ученица	18	војвода	800	фрајла	289
сељак	481	машамода	40	ђак	356	приорка	15	паша	739	ханума	244
калуђер	476	ашчика	33	министар	295	службеница	12	капетан	686	ханум	207
слуга	400	редара	31	начелник	251	надзорница	10	деспот	605	мадама	185
војник	280	измећарка	28	председник	225			игуман	587	баба	145
писар	260	куварица	28	владика	201			цар	478	госпоја	142
трговац	253	сељанка	28	хоџа	121			чича	400	царица	141
пандур	236	собарица	26	пуковник	119			султан	352	тетка	131
кочијаш	166	бабица	25	старешина	97			ага	330	госпа	125
официр	156	чочек	24	посланик	93			ђенерал	262	кнегиња	89
лекар	145	учитељка	22	намесник	77			гроф	260	пашиница	78
професор	138	куварка	21	паор	60			бој	247	фрау	75
мајстор	134	бирташица	19	ађутант	57			бег	227	фрајлица	59
адвокат	121	гувернанта	19	великаш	57			краљ	223	мадам	56
прота	119	дадиља	14	секретар	56			мула	187	милостива	55
чиновник	116	служавка	14	преседник	54			чика	152	графица	48
стражар	103	играчица	12	управитељ	53			витез	106	капетаница	37

Табела 2. Најфреквентније професије, позиције и титуле мушкараца и жена у корпусу 15AUTORA.

Да бисмо утврдили постоји ли специфичност употребе неких именица из шест класа улога, издвојили смо све оне који се јављају у корпусу с фреквенцијом 10 или већом, а резултати су приказани на слици 13. Видимо да се професије мушкараца (први ред лево) јављају значајније мање у односу на очекивано, тј. у односу на оно што се јавља у целом корпусу, код Ј. Димитријевић и Б. Станковића, а значајније више код С. Сремца (доприносом именице *поп* која је друга најфреквентнија именица код овог аутора, видети одељак 3.6) и Ј. Веселиновића. Професије жена



Слика 13. Специфичност појављивања професија, позиција и титула мушкараца и жена код 15 аутора.

(први ред десно) више него очекивано се јављају код Д. Гавриловић, Ј. Димитријевић и Ј. Игњатовића, а значајније мање код П. Тодоровића и Л. Комарчића. Што се тиче позиција мушкараца (средњи ред лево), изразито више од очекиваног се јављају код П. Тодоровића (*кнез* је најфреквентнија именица код овог аутора), а мање код М. Ускоковића и Ј. Димитријевић. Позиције жена (средњи ред десно) преовлађују у односу на очекивано код Ј. Игњатовића, а изразито су мање присутне код П. Тодоровића. Титуле мушкараца (доњи ред лево) значајно су више присутне од очекиваног код Ч. Мијатовића, а значајно мање код М. Ускоковића, Ј. Димитријевић и Б. Станковића. Што се тиче титула жена (доњи ред десно) значајно одступање се уочава код већине аутора: веће је код Ј. Димитријевић, Ј. Игњатовића и С. Сремца, а мање код П.

Тодоровића, С. Ранковића, С. Ћоровића, Ј. Веселиновића, Б. Станковића и М. Ђ. Милићевића.

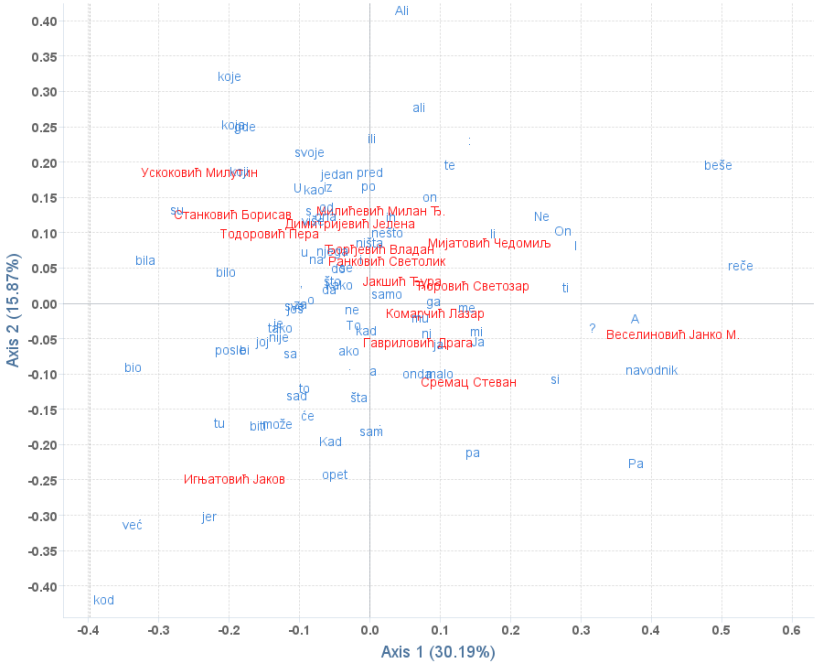
Из свега овога може се закључити да се било која улога или одређење жена користи значајно мање код П. Тодоровића, а значајно више код Ј. Игњатовића. Код Б. Станковића и М. Ускоковића ни одређења мушкараца ни жена не играју неку посебну улогу – за све категорије крећу се у оквиру очекиваног или испод – док су код В. Ђорђевића увек у оквиру очекиваног.

3.13 Факторска анализа

Анализа кореспонденције, или факторска анализа, често се користи за анализу текстуалних података. Она је посебно осмишљена за табеле контингентности, на пример, за оне које укрштају текстове и речи. Може се применити на облике речи, леме или ознаке врста речи у корпусу. У области анализе текстуалних података, факторска анализа се најчешће користи за визуелизацију преко дводимензионалних факторијалних мапа.

У програму TXM, који смо ми користили, граде се лексичке табеле које у нашем случају у колонама садрже текстове аутора, а у врстама одабране речи. На пресеку колона и врста налазе се фреквенције. Факторска анализа реорганизује ове податке с циљем визуелног приказа и бољег увида у саме податке, шта нам они говоре. Прво се рачуна центар гравитације облака података у односу на који се мери распршеност облака, а затим се одређују главне осе дисперзије у односу на које се приказују тачке, тј. њихове координате су израчунати фактори. У дводимензионалној презентацији приказане су прве две осе тј. оне које максимизују дисперзију података, а тачке у равни су из оба скупа података – колоне (речи) и врсте (аутори). Ове осе немају другу вредност, тј. тумачење, а тачке у равни треба посматрати само једне у односу на друге.

Понављајући експеримент из (Lavrentiev et al. 2021) укрстили смо ауторе са 200 најфреквентнијих токена из корпуса 15AUTORA, а резултати су приказани на слици 14. Приликом припреме података различите наводнике смо су „нормализовали“, тј. свели као један токен – *наводник*. На слици уочавамо да се Ј. Веселиновић јасно издваја од осталих аутора, а његова релативна близина са С. Сремцем и тачком *наводник* кореспондира с резултатом из одељка 3.3 где је уочено да ова два аутора највише користе директан говор. Такође се уочава да је Ј. Игњатовић јасно издвојен од осталих аутора.

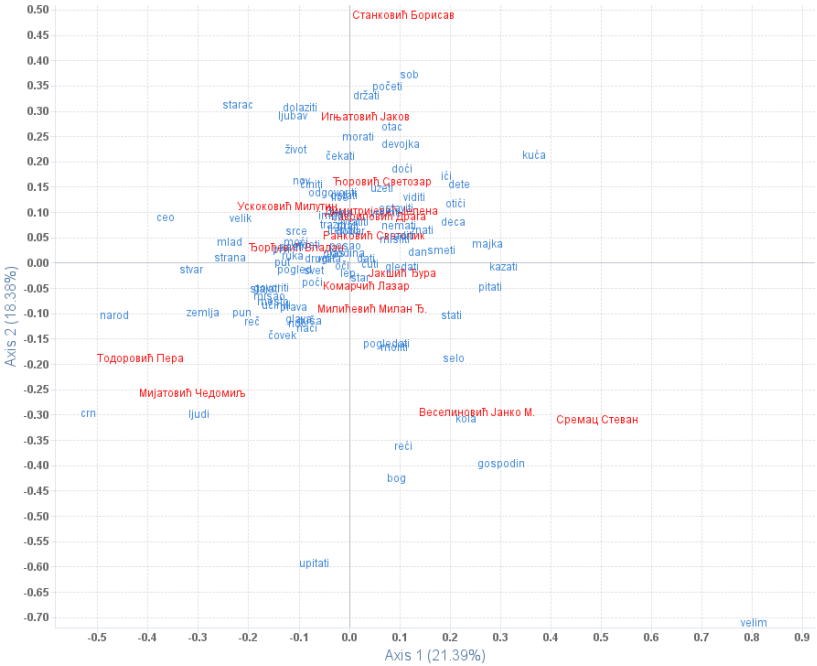


Слика 14. Факторска анализа аутора укрштених са 200 најфреквентнијих токена.

Нешто другачија слика се добија када се аутори укрсте са 100 најфреквентнијих именица, глагола и придева из 15AUTORA (слика 15): П. Тодоровић и Ч. Мијатовић су јасно издвојени од осталих аутора, као и Ј. Веселиновић и С. Сремац (њихова близина је уочена и на слици 14). М. Ускоковић, Д. Гавриловић и Ј. Димитријевић стоје веома близу на овом дијаграму – повезује их лема *жена*.

4. Закључак

Циљ овог рада није да процењује литерарне вредности изабраних писаца нити да те вредности супротстави оценама које су о њима дали књижевни критичари и сами читаоци. Намера нам је била да прикажемо какве дигиталне трагове су изабрани писци оставили у својим делима,



Слика 15. Факторска анализа аутора укрштених са 100 најфреквентнијих лема (именица, глагола и придева).

и начине како се до њих долази, што све може бити полазна основа за даља филолошка и информатичка истраживања.

У нашим даљим истраживања радићемо на проширењу овог корпуса. Већ сада је припремљен знатан број дела истих аутора, која су по врсти садржаја романи, али има и приповедака, путописа и мемоара. Нова истраживања ће тако омогућити и поређења дигиталних отисака откривених у романима према, на пример, онима у путописима. Неке прорачуне ваља поновити на балансираним узроцима, посебно она која се тичу лексичког богатства (одељак 3.4) и факторске анализе (одељак 3.13). За дубље проучавање лексике појединих писаца свакако би било корисно и разједначавање лема по значењу – на пример, да ли је именица *срце* која је друга најфреквентнија код Д. Гавриловић и В. Ђорђевића употребљена код ових аутора у истом значењу?

Представљени дигитални отисци као и многи други које тек треба открити могу послужити за аутоматско или полуатоматско класификовање изабраних дела по жанру, тону аутора и сл. Овакво класификовање би се могло даље користити за контрастивну анализу дела и језика изабраних петнаест аутора, као и за анализу дигиталних отисака према жанру и другим карактеристикама.

Захвалност

Аутор се захваљује сарадницима COST акције 'Distant Reading for European Literary History' (CA16204) уз чију помоћ и подстицај је формиран корпус SrpELTeC. Немерљив је допринос Универзитетске библиотеке „Светозар Марковић" у чијем фонду је пронађена и сканирана већина потребних књига, као и њених сарадника др Александре Тртовац и др Василија Милновића који су овај изузетно захтеван посао организовали. Посебно треба истаћи помоћ многих читалаца који су кориговали текстове и обавили основне анотације, а пре свега др Душка Витаса и др Ивана Обрадовића који су сами кориговали десетине књига. Без помоћи проф. др Ранке Станковић корпус не би био снабдевен свим додатним напредним анотацијама и потом понуђен корисницима на даље слободно коришћење.

Литература

Byszuk, Joanna, Michał Woźniak, Mike Kestemont, Albert Leśniak, Wojciech Łukasik, Artjoms Šeļa, and Maciej Eder. 2020. "Detecting direct speech in multilingual collection of 19th-century novels." In *Proceedings of lt4hala 2020-1st workshop on language technologies for historical and ancient languages,* 100–104.

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.

Krstev, Cvetana. 2008. *Processing of Serbian – Automata, Texts and Electronic Dictionaries.* Belgrade: University of Belgrade, Faculty of Philology.

Krstev, Cvetana. 2021a. “The Serbian Part of the ELTeC Collection Through the Magnifying Glass of Metadata.” *Infotheca – Journal for Digital Humanities* 21 (2): 26–42. https://doi.org/10.18485/infotheca.2021.21.2.2.

Krstev, Cvetana. 2021b. “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, Ranka Stanković, and Aleksandra Marković. 2023. “Multiword Expressions-Comparative Analysis Based on Aligned Corpora.” In *Book of Abstracts of the UniDive 1st general meeting, 16-17 March 2023, Paris-Saclay University, France.* COST.

Lavrentiev, Alexey M., Tatiana Yu. Sherstinova, Andrey M. Chepovskiy, and Benedict Pincemin. 2021. “Using TXM platform for research on language changes over time: The dynamics of vocabulary and punctuation in Russian Literary Texts.” *Вестник Томского государственного университета. Филология,* no. 70, 69–89.

Marneffe, Marie-Catherine de, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. “Universal Dependencies.” *Computational Linguistics* (Cambridge, MA) 47, no. 2 (June): 255–308. https://doi.org/10.1162/coli_a_00402.

Mihajlov, Teodora, Milica Ikonić Nešić, Ranka Stanković, and Olivera Kitanović. 2024. “Topic modeling of the SrpELTeC corpus: A comparison of NMF, LDA, and BERTopic.” In *2024 19th Conference on Computer Science and Intelligence Systems (FedCSIS),* 649–653. IEEE. https://doi.org/10.15439/2024F1593.

Nešić, Milica Ikonić, 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,* 7–16.

Patras, Roxana, Carolin Odebrecht, Ioana Galleron, Rosario Arias, Berenike J Herrmann, Cvetana Krstev, Katja Mihurko Poniž, and Dmytro Yesypenko. 2021. “Thresholds to the “Great Unread”: Titling Practices in Eleven ELTeC Collections.” *Interférences littéraires/Literaire interferenties* 25:163–187.

Paumier, Sébastien, Takuya Nakamura, and Stavroula Voyatzi. 2009. "Unitex, a corpus processing system with multi-lingual linguistic resources." In *eLEX2009 – Book of Abstracts – eLexicography in 21st century: New chalanges, new applications,* 173–175.

Pincemin, Bénédicte, Serge Heiden, and Franck Mazuet. 2022. "The textometric concept of active corpus." In *JADT 2022 Proceedings of the 16th International Conference on Statistical Analysis of Textual Data,* edited by Michelangelo Misuraca, Germana Scepi, and Maria Spano, II:691–698. Naples, Italy: VADISTAT - Per Simona Balbi, Univ. of Naples Federico II.

Radak, Tamara, Lou Burnard, Pieter Francois, Agnes Hilger, Fotis Jannidis, Gábor Palkó, Roxana Patras, Michael Preminger, Diana Santos, and Christof Schöch. 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.

Schmid, Helmut. 2013. "Probabilistic part-of-speech tagging using decision trees." In *New methods in language processing,* 154–164. Routledge.

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.

Smith, Raoul N. 1973. *Probabilistic Performance Models of Language.* Mouton.

Stanković, Ranka, Miloš Košprdić, Milica Ikonić Nešić, and Tijana Radović. 2022. "Sentiment Analysis of Serbian Old Novels." In *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data,* edited by Ilan Kernerman, Sara Carvalho, Carlos A. Iglesias, and Rachele Sprugnoli, 31–38. Marseille, France: European Language Resources Association, June.

Stanković, Ranka, Cvetana Krstev, Branislava Šandrih Todorović, and Mihailo Škorić. 2022. "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, Ranka, Cvetana Krstev, and Duško Vitas. 2024. "SrpELTeC: A Serbian Literary Corpus for Distant Reading." *Primerjalna književnost* 47 (2): 45–63. https://doi.org/10.3986/pkn.v47.i2.03.

Stanković, Ranka, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. "Parallel bidirectionally pretrained taggers as feature generators." *Applied Sciences* 12 (10): 5028.

Trtovac, Aleksandra, Vasilije Milnović, and Cvetana Krstev. 2021. "The Serbian Part of the ELTeC – from the Empty List to the 100 Novels Collection." *Infotheca – Journal for Digital Humanities* 21 (2): 7–25. https://doi.org/10.18485/infotheca.2021.21.2.1.

Utvić, Miloš. 2011. "Annotating the corpus of contemporary Serbian." *Infotheca: Journal of informatics and librarianship* 12 (2): 36a–47a.

Vitas, Duško. 2022. "From Onions to Champagne – Food and Drink in the SrpELTeC Corpus." *Infotheca – Journal for Digital Humanities* 21 (2): 88–118. https://doi.org/10.18485/infotheca.2021.21.2.5.

Дела укључена у корпус 15AUTORA

1. Игњатовић, Јаков (1822–1889):
   - *Ђурађ Бранковић : историческиј роман*, 1859;
   - *Једна женидба : слика из живота*, 1862;
   - *Милан Наранџић*, 1863;
   - *Васа Решпект*, 1875;
   - *Патница : роман*, 1888.
2. Ђорђевић, Владан (1844–1930):
   - *Кочина крајина : историјски роман*, 1863;
   - *Гмунденско језеро : путничка новела*, 1869;
   - *У фронт : приповетка из живота једног бившег краља*, 1913.
3. Јакшић, Ђура (1832–1878):
   - *Сељаци : приповетка из сеоског живота, из године 1857*, 1874;
   - *Сирота Банаћанка*, 1875;
   - *Чича Тима : приповетка из учитељског живота*, 1876.
4. Милићевић, Милан Ђ. (1831–1908);
   - *Потурченица Лејла : (црте из ратова за слободу)*, 1879;
   - *Јурмуса и Фатима или Турска сила сама себе једе: прича о ослобођењу шест округа 1832-1834*, 1879;
   - *Хајдуци : биљешке с пута по Рујну*, 1879;

- *Десет пара: прича из живота у вароши*, 1881;
- *Омер Челебија : приповијетка из живота српскога народа*, 1886.

5. Комарчић, Лазар (1833–1909):
   - *Драгоцена огрлица: прича у своје време*, 1880;
   - *Мој кочијаш : слике 1883 године*, 1887;
   - *Један разорен ум*, 1893.
6. Гавриловић, Драга (1854–1917):
   - *Из учитељичког живота*, 1884;
   - *Бабадевојка*, 1887;
   - *Девојачки роман*, 1889.
7. Веселиновић, Јанко М. (1862–1905);
   - *Сељанка : приповетка из сеоског живота*, 1888;
   - *Борци : роман из сеоског живота*, 1889;
   - *Хајдук Станко : историјски роман*, 1896.
8. Тодоровић, Пера (1852–1907):
   - *Силазак с престола : роман / написао Карио Амурели*, 1889;
   - *Београдске тајне : историски роман из српске прошлости с краја прошлог века!*, 1892;
   - *Смрт Карађорђева : историски роман из недавне прошлости*, 1983.
9. Мијатовић, Чедомиљ (1842–1932):
   - *Иконија везирова мајка: приповетка из XVII века*, 1891;
   - *Рајко од Расине: приповетка с краја XVII века*, 1892;
   - *Кнез Градоје од Орлова Града : приповетка из времена боја на Косову*, 1899.
10. Сремац, Стеван (1855–1906):
    - *Поп Ђира и поп Спира : приповетка*, 1894;
    - *Ивкова слава : приповетка*, 1895;
    - *Зона Замфирова : приповетка*, 1907.
11. Ранковић, Светолик (1863–1899):
    - *Горски цар : роман*, 1897;
    - *Сеоска учитељица : роман*, 1899;
    - *Порушени идеали : роман*, 1900.
12. Станковић, Борисав (1876–1927):
    - *Увела ружа*, 1899;
    - *Нечиста крв*, 1901;
    - *Покојникова жена*, 1902.
13. Димитријевић, Јелена (1862–1945):
    - *Ђул-Марикина прикажња : приповетка*, 1901;
    - *Фати-султан*, 1907;

- *Нове : роман*, 1912.

14. Ћоровић, Светозар (1875–1919):
    - *Женидба Пере Карантана*, 1905;
    - *Јарани : приповетка*, 1913;
    - *У ћелијама*, 1919.
15. Ускоковић, Милутин (1884–1915):
    - *Дошљаци : роман*, 1910;
    - *Потрошене речи*, 1911;
    - *Чедомир Илић : роман*, 1914.