

Текстометријске методе и ТХМ платформа за анализу и визуелну презентацију корпуса

УДК 811.163.41'322

Јелена Јаћимовић
jelena.jacimovic@stomf.bg.ac.rs
Универзитет у Београду
Стоматолошки факултет
Београд, Србија

САЖЕТАК: Текстометријски приступ се већ дуго примењује као корисна метода за анализу корпуса у различитим областима друштвено-хуманистичких наука. Комбинујући лексикометријска и статистичка истраживања са развијеним корпусним технологијама, текстометрија омогућава нелинеарно квантитативно и квалитативно проучавање дигиталних корпуса. У овом раду је с циљем илустровања могућности текстометријског приступа у оквиру ТХМ програмског окружења извршена анализа текуће верзије srpELTeC корпуса, уз представљање могућности визуелног приказа добијених резултата.

КЉУЧНЕ РЕЧИ: текстометрија, дигитални корпуси, српски језик, ТХМ, srpELTeC.

РАД ПРИМЉЕН: 29. јуни 2019.

РАД ПРИХВАЋЕН: 6. септембар 2019.

1. Увод

Дигитални корпуси представљају важне и практичне изворе емпиријских података неопходних за спровођење истраживања у области лингвистике и других друштвено-хуманистичких наука. Са развојем дигиталног доба и достигнућима примене језичких технологија мења се традиционални начин приступа текстовима и отварају нове могућности анализе великих количина текстуалних података применом статистичких метода. Уместо уобичајеним линераним читањем (енг. *close reading*), разумевање литературе омогућено је сакупљањем и

анализом великих количина података – читањем на даљину (енг. *distant reading*). Текстометрија се издваја као једна од дисциплина која применом својих метода омогућава другачији начин „читања“ текстова. Комбинујући лексикометријска и статистичка истраживања са развијеним корпусним технологијама, текстометрија омогућава нелинеарно квантитативно и квалитативно проучавање дигиталних корпуса.

1.1 Текстометрија

Почеци текстометријских истраживања везују се за Француску и рад Пјера Гироа (1954; 1959) и Шарла Милера (1973), који су се бавили проблемима и методама лингвистичке статистике. Методе које је Жан-Пол Банзекри развио у сарадњи са својим колегама и студентима и применио их на лингвистичке податке (Benzécri, 1973), усвојене су и примењују се и у оквиру текстометрије. Методолошку основу текстометрије такође проналазимо и у радовима Лудовика Лебарта и Андреа Салема (Lebart и Salem, 1988, 1994). Осим усвојених метода, у оквиру текстометрије развијен је и низ нових статистичких модела намењених откривању важних карактеристика текстуалних података, попут „привлачности“ која постоји између речи неког текста, линеарности и унутрашње структуре текста, интертекстуалних контраста или показатеља лексичке еволуције тј. периода које карактерише употреба одређених речи или њихово одсуство из текста. Резултати анализе текста добијени применом текстометријских метода дају синтетички, селективан и сугестиван приказ изнова организованог текста, који се сада може сагледавати кроз креирање хијерархијских листа, визуалних мапа, прегруписавања и тексту додатих информација. Хеуристичка моћ статистичких алата у анализи литерарних текстова огледа се у новом начину приступа разнородним контролисаним текстуалним подацима и новом начину „читања“ текста на основу добијених резултата у виду података који раније нису били доступни.

Текстометријски приступ подразумева да сваки текст поседује сопствену унутрашњу структуру, коју би било тешко анализирати само ручним алатима. Захваљујући рачунарским алатима и креираним хипертекстуалним везама, текстометрија на основу нумеричких показатеља истовремено пружа општи синтетички поглед на текст, али и могућност локалног сагледавања увидом у циљани контекст. Та „близина“ текста током анализе и овако уравнотежен приступ

општем и локалном у тексту отвара низ могућности за постављање херменеутичких питања и открива лингвистичку реалност која је веома важно и богато опсервационо поље. Међутим, треба имати у виду да је текстометрија процес који обезбеђује резултате и уочавање образаца и трендова, који би иначе остали скривени због велике количине података, али да тумачење добијених резултата и његова валидност зависе од стручњака и система који се користи за текстометријску анализу.

Осим текстометрије, постоје и друге дисциплине које за анализу текстуалних података користе квантитативне приступе. Проналажењем робустних метода управљања великим количинама текстова бави се и област проналажења информација (енг. *Information Retrieval*), која је, пак, усмерена на проналажење и повезивање одређених информација и докумената у којима се те информације налазе. За разлику од области проналажења информација, текстометрија се примењује на затворен, стабилан корпус текстова. Још једна област која се у одређеном смислу може поредити са текстометријом јесте латентна семантичка анализа (енг. *Latent Semantic Analysis*). И у оквиру ове области се примењују математичке методе приликом анализе текстова, али с циљем истраживања неких других поља која се налазе ван текста, попут језика и других когнитивних функција. Квантитативне методе које су познате и примењују се у текстометрији, користе се, на пример, и у области ископавања текста (енг. *Text Mining*). За разлику од ископавања текста, чија је основна идеја проналажење изузетно вредних информација које се могу екстраховати из текста, текстометријски приступ је усмерен на сам текст и откривање тенденција и језичких законитости, начина њиховог реализовања у тексту и момената када долази до њихове промене. Осим тога, предмет текстометријске анализе су добро познати корпуси, текстови и објекти. Обрада природних језика (енг. *Natural Language Processing*) такође користи статистику за изградњу система и препознавање лингвистичких јединица текста. Иако се циљеви ових двеју области разликују, може се рећи да се оне међусобно допуњују. На пример, корпуси се у обради природних језика могу користити за подешавање система намењеног препознавању одређених ентитета или изградњу правила препознавања (као што је екстракција термина или морфолошко и синтаксичко означавање), што касније у процесу текстометријске анализе може да буде од изузетног значаја. С друге стране, текстометријски приступ истраживању корпуса пружа могућност уочавања појединих ентитета или специфичности

текста, као и њиховог означавања у тексту, што може бити искоришћено приликом изградње алата намењених обради природних језика.

1.2 ТХМ програмско окружење

Квантитативни приступ истраживања текстуалних елемената корпуса није нов, али је значајно поједностављен и унапређен применом постојећих алата. Коришћење програма намењених анализи великих корпуса не води откривању нових информација о језику, већ нуди нову перспективу у сагледавању већ познатог (Hunston, 2002). Приликом анализе великих корпуса важно је омогућити корисницима постављање више различитих упита и кретање кроз текст у окружењу једноставном за рад. Постоји више програмских решења која су слободна за коришћење и имају развијен графички кориснички интерфејс (Pince-min, 2018) и која омогућавају анализу текстуалних података помоћу основних функција текстометријског приступа (као што су приказ конкорданци, рачунање *индекса специфичности*, факторска анализа међузависности, о којима ће бити више речи касније).

Заснован на текстометрији, методологији која омогућава квантитативну и квалитативну анализу текстуалних корпуса, ТХМ (Heiden et al., 2010; Heiden, 2010) је програм отвореног кода¹, веома коришћен у истраживањима различитих области друштвено-хуманистичких наука (историја, књижевност, географија, лингвистика, социологија, политичке науке). Графичко корисничко окружење ТХМ-а заснива се на коришћењу CQP² (енг. *Corpus Query Processor*) претраживача и R³ статистичког пакета. ТХМ омогућава проучавање за статистичке потребе задовољавајући обимне грађе било ког језика, велики број улазних формата текста, као и употребу различитих алата за обраду улазних текстова природних језика (нпр. аутоматских лематизатора). Овај програм такође пружа могућност изградње поткорпуса или партиција на основу метаподатака (датум, аутор, жанр итд.) или структурних јединица корпуса (текст, поглавље, пасус итд.), постављања упита (помоћу CQP претраживача) и сложеније аутоматске обраде резултата упита засноване на квантитативним методама, као и извоз резултата у табеларном или графичком облику.

¹ Доступне су десктоп инсталације за Windows, Linux и Mac OS X, као и веб платформа.

² CQP (на вебу)

³ R (на вебу)

Дигитални корпуси које је могуће анализирати у оквиру овог окружења јесу текстуални корпуси засновани на писаним текстовима, затим корпуси транскрипата (синхронизовани са изворним аудио или видео снимком) и паралелни корпуси. ТХМ омогућава увоз текстова кодираних у складу са одређеним конвенцијама, попут текстова у препорученом кодном распореду UTF-8 (TXT формат) или XML⁴ (енг. *eXtensible Markup Language*) докумената, који могу бити кодирани у складу са TEI⁵ (енг. *Text Encoding Initiative*) упутствима (XML или XML-TEI формат). На тај начин је у оквиру ТХМ-а могуће изабрати ниво репрезентације текстова корпуса који је мање или више богат, те је стога и мање или више захтеван за припрему. Основна репрезентација чистог текста нуди неке основне могућности анализе, док текстови кодирани у XML-TEI формату, захваљујући богатијој репрезентацији текста и описаној унутрашњој структури, могу да се посматрају са далеко више аспеката (Lavrentiev et al., 2013).

На основу изворних текстова и формата у којима су сачувани, ТХМ приликом увоза генерише XML-ТХМ приказ тј. прилагођену верзију TEI модела података, која ће се користити као основни модел за спровођење свих анализа. Оваква репрезентација корпуса подразумева постојање одређених јединица корпуса, и то: текстуалних, структурних и лексичких јединица. Текстуалне јединице су **текстови** који чине корпус (нпр. књиге, чланци, интервјуи итд.) и које могу имати своје атрибуте, односно метаподатке (нпр. аутор, наслов, датум, жанр итд.). Затим, сваки од текстова може да има одређену структуру тј. може да садржи одређене унутрашње **структурне јединице** (нпр. поглавља, пасусе, управни говор и сл.) које могу имати одређена својства, односно атрибуте (као што су наслов, број итд.). На последњем нивоу су дефинисане **лексичке јединице**, јер је сваки текст састављен од низа речи које могу имати одређена својства, као што су графички облик, лема, граматичка категорија итд. Метаподаци се, као и сви XML елементи текста, у оквиру ТХМ-а посматрају као структурне јединице. На тај начин од репрезентације текста зависе и могућности његове анализе. Анотације којима је текст/корпус опремљен, затим структурне и лексичке јединице користе се за креирање поткорпуса, дељење корпуса на партиције ради њиховог међусобног поређења и за претраживање текстова. Стога је, са становишта корпусног истраживања, потребно

⁴ XML (на вебу)

⁵ TEI, (на вебу)

што детаљније дефинисати јединице на основу којих би могла да се врши анализа. За сваку текстуалну јединицу корпуса се у оквиру ТХМ-а изграђује облик текста у HTML формату, на основу ког је у сваком кораку анализе могућ повратак тексту.

Иако ТХМ модули за увоз текстова омогућавају аутоматско морфолошко и синтаксичко етикетирање, као и лематизацију текстова помоћу програма TreeTagger (Schmid, 1994), могуће је претходно обрадити корпусне податке и ван платформе помоћу других алата намењених обради природних језика. Не постоји стандардна репрезентација резултата ових алата, већ се примењују само стандарди који се користе у пракси. Са становишта ТХМ-а, резултати примене алата обраде природних језика виђени су као анотације додате XML-TEI репрезентацији текста.

Осим изградње XML-ТХМ формата, врши се конверзија и генерисање CWB⁶ (енг. *Corpus Workbench*) формата, који се користи за претрагу корпуса помоћу упита изражених CQP синтаксом. Опис CWB окружења и регуларних израза које користи корпусни упитни језик (енг. *Corpus Query Language*, скр. CQL) може се наћи код (Utvić, 2014; Evert и Hardie, 2011).

ТХМ окружење је алат који, пре свега, омогућава спровођење квалитативне анализе текстова генерисањем фреквенцијских листа, конкорданци или кретањем кроз HTML издање текста. Било која комбинација својстава дефинисаних јединица текста може да се употреби за постављање упита и приказ контекста у којима се те јединице појављују. С друге стране, статистички модели примењени на пребројавање својстава лексичких јединица омогућавају квантитативну анализу, односно анализу њихове дистрибуције по корпусима (факторска анализа, кластер анализа), њихову изузетно високу или ниску заступљеност у појединим партицијама (анализа специфичности), или анализу привлачности која постоји између појединих речи (анализа заједничког појављивања). Резултат сваке анализе може да се изведе у табеларном или графичком облику ради даљег истраживања и уређивања помоћу неког другог алата.

Циљ овога рада је представљање текуће верзије srpELTeC корпуса⁷ и илустрација могућности текстометријског приступа и примене ТХМ алата за анализу и визуелни приказ резултата. Сprovedена анализа

⁶ CWB (на вебу)

⁷ srpELTeC (на вебу)

корпуса српског романа с краја 19. и почетка 20. века требало би да истакне потенцијал текстометрије и приближи је истраживачима различитих научних области који се баве корпусном анализом.

2. Методологија текстометријске анализе **srpELTeC** корпуса

2.1 **srpELTeC** корпус

Корпус коришћен за потребе овог рада назван је **srpELTeC** корпус. Основни мотив израде овог корпуса јесте његово укључивање у вишејезичну збирку европских књижевних текстова (енг. *European Literary Text Collection*), која би требало да садржи 100 романа за сваки од језика укључених у COST акцију *Читање на даљину за историју европске књижевности* (енг. *Distant Reading for European Literary History*), којима су истекла ауторска права, а који су објављени у периоду 1850–1920. године.⁸

За разлику од многих других европских језика који су укључени у ову акцију, српски корпус се производи од самог почетка. Већина српских романа из овог периода није дигитализована на одговарајући начин или није дигитализована уопште, посебно због тога што је до првих издања многих романа било тешко доћи. Српска књижевност, а посебно роман, датог периода се ни у ком случају не може поредити по обиму са књижевном „продукцијом“ великих европских језика, као што су француски, енглески или немачки језик, те је и сам процес одабира романа и проналажења штампаних примерака био изузетно захтеван. Процес њихове трансформације у машински читљив облик подразумевао је скенирање и оптичко препознавање карактера (енг. *Optical Character Recognition*, скр. OCR). Грешке настале током оптичког препознавања

⁸ ELTeC збирка би требало да садржи прва издања књижевних текстова (романа) из различитих периода и писаних на неколико језика. Да би текст био укључен у неку од ELTeC подзбирки треба да је први пут објављен у европској земљи између 1850. и 1920. године као књига, чији текст садржи минимум 10.000 речи. Остали критеријуми за избор романа односе се, пре свега, на пол аутора и канон. У саставу сваке ELTeC подзбирке требало би да буде барем 10% до 50% текстова које су писали аутори женског пола, као и по минимум 30% веома престижних (канонизованих, који имају више издања) романа, али и оних који нису познати и утицајни (који немају ни једно или имају смо једно поновљено издање).

карактера аутоматски су кориговане помоћу специјализованог алата заснованог на српским морфолошким реченицима (Krstev, 2008). За ручну корекцију преосталих грешака и означавање текстова основним структурним јединицама ангажован је велики број волонтера⁹. У овој фази вршена је и припрема метаподатака који ће се употребити за каснију анализу корпуса, у складу са захтевима COST акције.

Добро је познато да би приликом креирања корпуса требало водити рачуна о његовој величини, репрезентативности и балансираности (O'Keeffe и McCarthy, 2010). Припрема више прозних дела објављених на српском језику у периоду 1850-1920. године је у току, па је у srpELTeC корпус укључено за сада 21 дело које је дигитализовано до тренутка писања овог рада (табела 1). Разлог избора ових дела, дакле, није ни естетске нити тематске природе. Имајући у виду чињеницу да srpELTeC корпус тренутно не обухвата све романе објављене у датом периоду, не може се рећи да је репрезентативан, као ни балансиран јер су, на пример, већину дела која су укључена писали аутори мушког пола. Ипак, циљ овога рада је приказ спровођења текстометријске анализе помоћу ТХМ алата, за коју креирани корпус српске књижевности с краја 19. и почетка 20. века може да буде добар извор. Осим тога, овај специјализовани корпус садржи колекцију текстова изузетног значаја који нису пример савременог језика и у којој се, осим добро познатих аутора и њихових дела, налазе и они чији рад доноси зачетке модерне романескне структуре или, пак, они о којима је у историјама српске књижевности мало писано. На пример, у корпус су укључени романи Милутина Ускоковића, кога критичари и историчари књижевности сматрају зачетником београдског тј. урбаног стила, али и роман Драгомира Шишковића, аутора о коме је мало тога познато. Део srpELTeC корпуса чини и роман *Једна угашена звезда* Лазара Комарчића, који је први научно-фантастични роман у српској књижевности, као и роман *Бабадевојка* Драге Гавриловић, ауторке која је прва написала роман у српском патријархалном друштву тог времена. Текућа верзија srpELTeC корпуса доступна је у оквиру ELTeC колекције.¹⁰

⁹ Волонтери (на вебу)

¹⁰ srpELTeC (на вебу)

Аутор	Дело	Година	Дужина (w)
Гавриловић, Драга	Бабадевојка	1887	23.858
Гавриловић, Андра	Прве жртве	1893	44.929
Костић, Тадија	Господа сељаци	1896	39.349
Мијатовић, Чедомиљ	Рајко од Расине	1892	50.305
	Иконија, везирова мајка	1891	28.332
Милићевић, Милан	Десет пара	1881	12.365
	Јурумуса и Фатима	1879	21.947
Станковић, Борисав	Увела ружа	1899	12.748
	Покојникова жена	1902	12.701
Шишковић, Драгомир	Један од многих - роман из престоничког живота	1920	21.676
Ускоковић, Милутин	Потрошене речи	1911	14.580
	Дошљаци	1910	97.467
	Чедомир Илић	1914	65.073
Димитријевић, Јелена	Нове	1912	116.782
Илић, Драгутин	Хаци Ђера	1904	65.554
Јанковић, Милица	Калуђер из Русије*	1919	8.279
Комарчић, Лазар	Драгоцене огрлица	1880	65.160
	Једна угашена звезда	1902	58.334
	Просиоци	1905	28.327
Нушић, Бранислав	Општинско дете	1902	77.994
Секулић, Исидора	Ђакон Богородичине цркве	1919	62.414

* Неће ући у srgELTeC корпус због дужине

Табела 1. Прозна дела укључена у srgELTeC корпус коришћен за текстометријску анализу

Текстови srgELTeC корпуса сачувани су у XML формату, у складу са ТЕI смерницама, који је веома користан за каснију анализу. Заглавље ТЕI документа садржи библиографске податке о електронској и изворној верзији дела, као и податке о особама одговорним за поједине фазе креирања и ажурирања електронске верзије. Текстови су структурно анотирани тј. садрже информацију о логичкој структури текста. Осим прописаних ТЕI елемената и атрибута за структурну анотацију текста (заглавље, текст, тело текста, јединица текста -

поглавље, наслов и поднаслови, пасус, цитат, речи или реченице које не припадају језику текста, већ неком другом језику и сл.), кроз метаподатке у облику CSV документа унете су и додатне информације о полу аутора и врсти дела (нпр. роман, приповетка или кратка проза). Дефинисани метаподаци су у ТХМ окружењу посматрани као нови структурни елемент **text**, представљен следећим атрибутима: **author**, **title**, **date**, **gender** и **type**. На тај начин је за припремљене текстове одређена расподела по аутору, наслову, години објављивања, полу аутора и врсти дела. Метаподаци су у ТХМ окружењу коришћени за поделу корпуса на партиције, креирање поткорпуса, као и за претраживање текстова.

Ради анализе **sprELTeC** корпуса, колекција текстова у XML формату увезена је у ТХМ окружење помоћу XML-TEI Zero + CSV модула. Етикетирање текстова **sprELTeC** корпуса спроведено је помоћу програма **TreeTagger** и модела развијеног за српски језик (Utvić, 2011). Приликом увоза корпуса извршена је аутоматска сегментација, токенизација, лематизација и етикетирање врстом речи (енг. *Part of Speech tagger*, скр. **Pos**). Резултати примене овог програма за етикетирање су у оквиру ТХМ окружења посматрани као лексичке јединице, и то: **n** (бројчана ознака позиције појавног облика речи у корпусу), **srlemma** (лема придружена токену аутоматском анотацијом помоћу **TreeTagger** програма), **srpos** (врста речи придружена токену аутоматском анотацијом помоћу **TreeTagger** програма) и **word** (конкретна реализација токена у тексту) (пример 1).

Пример 1. ... из нашега **друштвеног** живота ... (део текста из романа Лазара Комарчића *Једна угашена звезда*)

n: 52.726

srlemma: друштвен

srpos: A

word: друштвеног

На основу репрезентације текстова у XML-TEI формату модул за увоз генерише XML-TXM приказ, који ТХМ софтвер користи као основни модел за представљање анотација корпуса. Поред изградње XML-TXM формата, извршена је и конверзија и генерисање CWB формата, коришћеног за претрагу корпуса помоћу упита изражених CQP синтаксом. Методологија анализе корпуса коју ТХМ окружење омогућава биће описана у следећем делу.

2.2 Текстометријске методе у ТХМ окружењу

Основни параметар законитости унутар једног корпуса јесте фреквентност, која означава колико се пута нека језичка јединица појављује у одређеном корпусном контексту (Dobrić, 2009). Подаци о фреквентности служе као основа за спровођење различитих статистичких анализа, обезбеђујући емпиријску основу за извођење теорија о некој језичкој појави.

Прва и основна корпусна метода која се примењује 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 процес поређења фреквенција у деловима различитих величина неопходно извршити нормализацију, односно изразити фреквенције заједничким фактором – релативном фреквенцијом. Очекивано би било да се релативна фреквенција рачуна као количник апсолутне фреквенције језичке јединице и укупног броја јединица у делу корпуса. Овако израчуната средња вредност представља математичко очекивање за нормалну (Гаусову) расподелу вероватноће. Међутим, појављивања језичких јединица у неком делу корпуса није нужно у складу са нормалном дистрибуцијом. Пјер Лафон (Lafon, 1984) је уочио да је вероватноћа појављивања језичких јединица у складу са хипергеометријском (негативном биномном) расподелом. Вероватноћа да ће се језичка јединица A , која је део вокабулара корпуса V , појавити f пута у делу корпуса p дужине t , узимајући у обзир укупан број појављивања те јединице F у целом корпусу дужине T , предложено у (Lafon, 1980), израчунава се формулом:

$$Prob_{specif}(card\{A \in V | A \in p\} = f) = \frac{C_F^f \times C_{T-F}^{t-f}}{C_T^t}, \text{ где је}$$

$$C_n^k = \frac{n!}{k!(n-k)!}$$

$$n! = 1 \times 2 \times 3 \times \dots \times (n-1) \times n$$

Рачунање *индекса специфичности* на основу хипергеометријске дистрибуције у ТХМ окружењу показује вероватноћу појаве језичке јединице у неком одређеном делу корпуса. ТХМ омогућава 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. На овај начин могуће је идентификовати

значајно честе (позитивне кључне речи) или значајно ретке (негативне кључне речи) појаве језичких јединица делова у односу на цео корпус, што је корисна полазна тачка за извођење претпоставки о кључним речима текста, домену текста, ауторству текста итд.

Још једна стандардна метода текстометрије, осим идентификације карактеристичних јединица помоћу специфичних фреквенција, јесте факторска анализа међузависности (Benzécri, 1973). Принципи факторске анализе међузависности, развијени у оквиру француске школе „Analyse des Données”, коришћени су за анализу корпуса, али и многих других врста података (Beaudouin, 2016). Ова статистичка техника омогућава приказ и преглед скупа података, односно међузависности које постоје између делова корпуса и језичких јединица, у дводимензионалном графичком облику.

Почетна идеја била је утврђивање образаца међусобних релација двају скупова елемената забележених у табеларном облику. На примеру текстуалних корпуса, ако се у колонама табеле налазе текстови, а у редовима речи, на пресеку колона и редова налазе се показатељи присутности, односно учесталости појављивања речи у тексту (нпр. фреквенција појављивања речи, специфична фреквенција појављивања речи). Помоћу алгоритама анализе података могућа је синтеза информација садржаних у матрицама. Факторска анализа тежи реорганизацији матрица тако да садрже максималну количину информација. Другим речима, основна идеја спровођења факторске анализе међузависности јесте упрошћавање сложеног скупа тј. облака података и проналажење начина да се што више информација представи у простору мањих димензија. Да би ово било могуће, прво се рачуна центар гравитације облака, око кога се мери распршеност, односно дисперзија облака. У следећем кораку се конструишу факторске равни, односно главне осе дисперзије. Тачке су пројектоване на ове равни, а њихове координате на овим осама су фактори. На плану дефинисаном помоћу прве две осе може да се добије најбоља пројекција облака, која у највећој мери умањује губљење информација. Основни циљ је визуелни приказ удаљености између атрибута, односно од насумичне дистрибуције.

Факторска анализа међузависности често је комбинована са кластер анализом, односно хијерархијским класификовањем које се заснива на координатама елемената факторских оса. Ова метода класификације служи да се идентификују хомогене подгрупе текстова и речи. Примењена упоредо са факторском анализом, кластер анализа

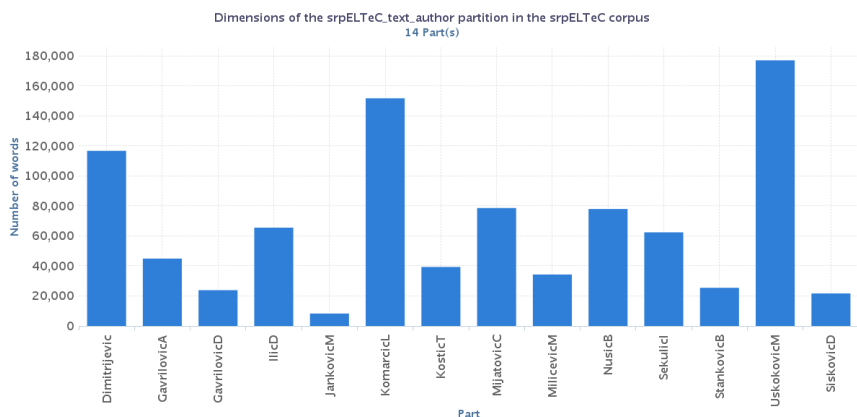
омогућава боље разумевање података и поједностављује њихову интерпретацију.

3. Резултати текстометријске анализе

3.1 Опште информације о корпусу и фреквентности

Посматрани *srpELTeC* корпус састоји се од текстова који укупно садрже 935.902 речи. У погледу заступљености текстова одређеног аутора, најобимније делове корпуса чине текстови Милутина Ускоковића, Лазара Комарчића и Јелене Димитријевић, док партиција у којој се налази роман Милице Јанковић садржи само 8.279 речи (слика 1). Резултати говоре да је у овом корпусу, који има 78.542 токена, употребљено 32.604 леме. Најфреквентнијих 20 токена, који носе мало семантичких информација, чини готово 30% укупног броја конкретних реализација у корпусу, док хапакса тј. токена који се појављују само једном у корпусу има 42.004, што је у односу на укупан број токена нешто више од 50%. У табели 2 приказане су различите речи *srpELTeC* корпуса (*word*), њихов тачан број појављивања у корпусу (апсолутна фреквенција *F*) и ранг на листи учесталости сортираној по опадајућој фреквенцији (*rank*). Издвојене најучесталије корпусне речи, као и у случају Корпуса савременог српског језика (*SrpKor*) (Utvić, 2014), јесу функцијске речи из затворених класа речи попут предлога, везника, помоћних глагола, заменица итд.

На основу додељених морфолошких етикета (енг. *PoS tags*) генерисана је листа учесталости појављивања одређених врста речи. У оквиру табеле 3 приказане су вредности атрибута *srpos*, њихова апсолутна фреквенција (*F*) и ранг на листи учесталости сортираној по опадајућој фреквенцији (*rank*). Осим именица и глагола које доминирају у текстовима *srpELTeC* корпуса, по високој заступљености издвајају се и заменице као морфолошка категорија чију би сложеност употребе и експресивне могућности било занимљиво истражити помоћу ТХМ алата. Имајући у виду чињеницу да је у српском језику употреба личне заменице произвољна јер дати глаголски облик и сам указује на лице вршиоца радње, што карактерише неутрални стил изражавања, висока фреквенција употребе заменица својствена је стилски специфичним књижевно-уметничким текстовима, попут текстова *srpELTeC* корпуса (Katnić-Bakaršić, 1999).

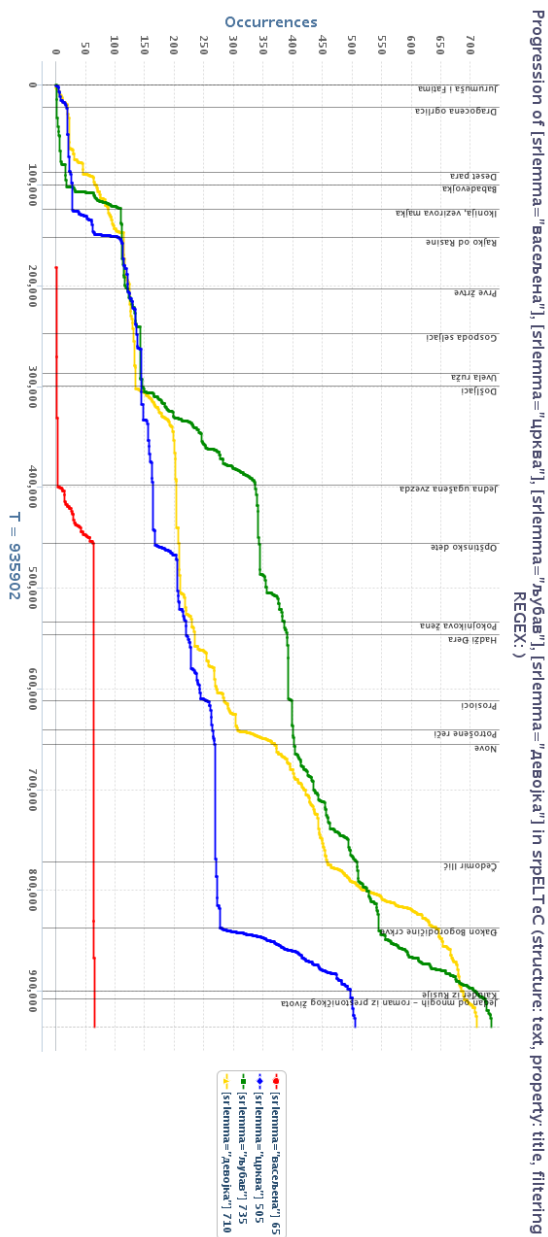


Слика 1. Димензије партиција srpELTeC корпуса креираних на основу ауторстава

На основу података о фреквентности ТХМ омогућава визуелни приказ учесталости употребе одређених језичких појава кроз читав корпус, односно његове делове креиране на основу постојећих структурних јединица. На овај начин једноставно је уочити у којим деловима корпуса и са којом учесталошћу се одређене речи или изрази користе. Илустративан приказ учесталости и позиције коришћења лема *васељена*, *црква*, *љубав* и *девојка* у делима читавог srpELTeC корпуса дат је на слици 2. На пример, лема *љубав* најчешће се спомиње у романима *Бабадевојка*, *Дошљаци* и *Закон Богородичине цркве*, у којима љубав и јесте један од доминантних мотива, док значајну употребу леме *васељена* уочавамо искључиво у првом српском научно-фантастичном роману *Једна угашена звезда* Лазара Комарчића. Лему *црква* Нушић највише употребљава на почетку свог дела *Општинско дете*, за разлику од Исидоре Секулић која је равномерно помиње кроз читав роман *Закон Богородичине цркве*.

3.2 Индекс специфичности

Подаци о учесталости појављивања именица, придева, глагола и прилога у srpELTeC корпусу приказана је у табели 3. Њихова



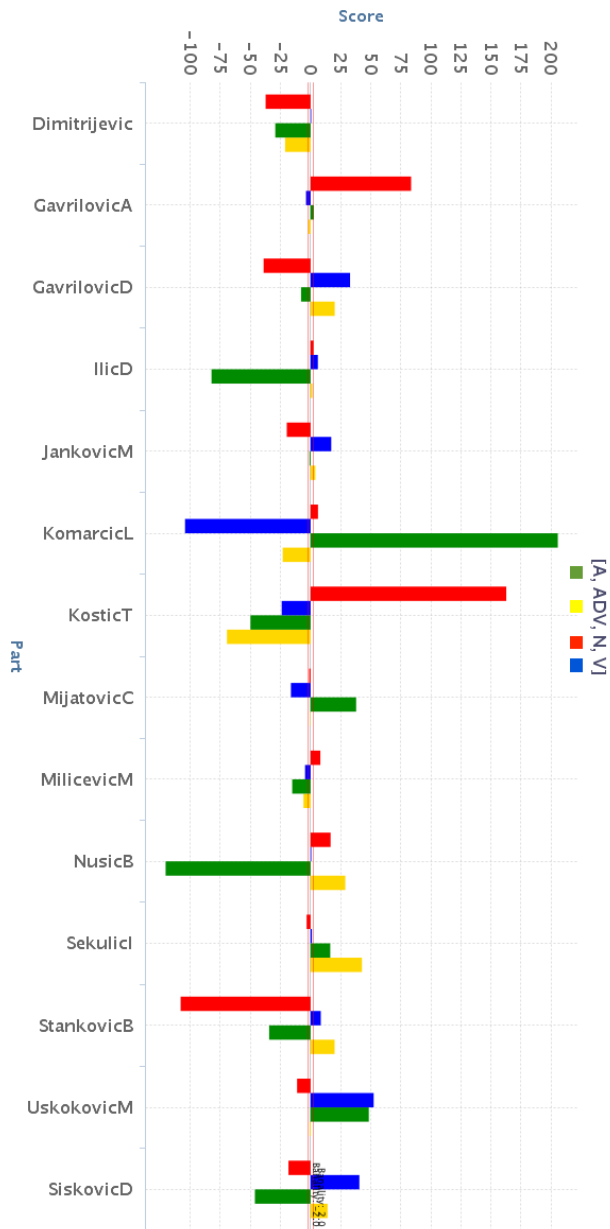
Слика 2. Графички приказ употребе појединих лема у текстовима srBELTeC корпуса

srpELTeC			SrpKor		
rank	word	F	rank	word	F
1	и	28.545	1	и	4.330.865
2	је	25.422	2	је	4.103.542
3	се	21.128	3	у	3.513.009
4	да	18.932	4	да	3.261.285
5	у	14.932	5	се	2.107.336
6	на	9.233	6	на	1.751.270
7	не	6.721	7	за	1.381.402
8	а	5.642	8	су	1.258.361
9	од	5.234	9	од	919.922
10	што	5.011	10	са	779.469
11	су	4.935	11	а	740.476
12	као	4.857	12	који	650.144
13	за	4.460	13	не	612.218
14	па	4.253	14	о	517.105
15	то	4.132	15	ће	505.643

Табела 2. Првих 15 редова листе учесталости srpELTeC и SrpKor корпуса

заступљеност у текстовима одређеног аутора (f), упоредо са *индексом специфичности* (S) дате врсте речи у односу на цео корпус, видљива је из табеле 4. Дистрибуција учесталости ових врста речи у srpEL-TeC корпусу подељеном на партиције на основу ауторства илустрована је графиком на слици 3. Резултати показују да су придеви изузетно специфични за текстове Лазара Комарчића ($S_A=205,6$), док Тадија Костић и Андра Гавриловић користе именице много чешће него остали аутори чија су дела укључена у овај корпус (респективно $S_N=162,7$ и $S_N=83,6$). С друге стране, глаголи су код Лазара Комарчића, у односу на њихов степен употребе у целом корпусу, далеко мање коришћени, што је исказано високом негативном вредношћу *индекса специфичности* ($S_V=-104,4$). Специфично ниску фреквенцију употребе придева уочавамо у романима Бранислава Нушића и Драгутина Илића ($S_A=-120,5$ и $S_A=-82,4$), док су именице, имајући у виду њихов степен употребе у осталим деловима корпуса, далеко мање заступљене у приповеткама Борисава Станковића ($S_N=-108$).

Слика 3. Специфичност употребе именица (N), глагола (V), придева (A) и прилога (ADV) у srpELTeC корпусу по ауторима



rank	srpos	F
1	N (именица)	192.865
2	V (глагол)	173.994
3	PUNCT (знак интерпункције)	103.824
4	PRO (заменица)	97.239
5	CONJ (везник)	90.248
6	A (придев)	64.242
7	PREP (предлог)	62.584
8	ADV (прилог)	50.628
9	SENT (ознака краја реченице)	49.184
10	PAR (речца или партикула)	36.307
11	NUM (број)	9.208
12	UNDEF (неодређено)	2.272
13	? (остало)	2.033
14	INT (узвик)	680
15	ABB (скраћеница)	527
16	RN (римски број)	46
17	PREF (префикс)	21

Табела 3. Листа учесталости могућих вредности позиционог атрибута `srpos` `srpELTeC` корпуса

3.3 Факторска анализа међузависности и кластер анализа

Ради поједностављења приказа и боље видљивости добијених резултата спроведене факторске анализе међузависности, `srpELTeC` корпус подељен је на само четири партиције (што је минималан број делова корпуса над којима може да се спроводи факторска анализа међузависности) на основу података о полу аутора и веку у коме је објављено његово дело. Тако у овом случају разликујемо следеће партиције: *f19* – дела аутора женског пола, објављена у 19. веку; *f20* – дела аутора женског пола, објављена у 20. веку; *m19* – дела аутора мушког пола, објављена у 19. веку; и *m20* – дела аутора мушког пола, објављена у 20. веку. Величина текстова корпуса у зависности од пола аутора и века у коме је дело објављено приказана је на слици 4, где се јасно види да је партиција која садржи дела ауторки из 19. века значајно мања у односу на остале делове.

Подаци о учесталости појављивања именица, придева, глагола и прилога у `srpELTeC` корпусу подељеном на партиције на основу пола

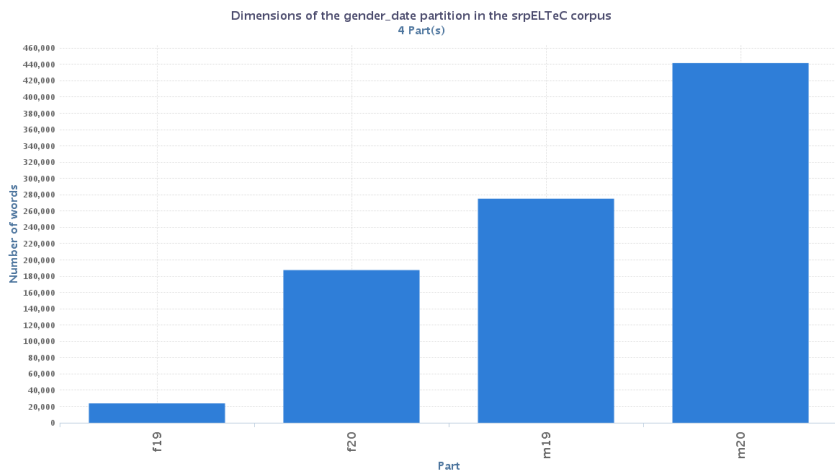
Аутор	f_N	S_N	f_V	S_V	f_A	S_A	f_{ADV}	S_{ADV}
Uskokovic M	35331	-11.2	35446	52.5	13502	48.4	9577	-0.8
Komarcic L	31869	6.2	25443	-104.4	13193	205.6	7481	-23.1
Dimitrijevic J	22330	-37.4	21938	0.5	7066	-29.3	5687	-21.2
Nusic B	16926	16.6	14675	0.6	3812	-120.5	4952	28.9
Mijatovic C	15985	-1.2	13862	-16.3	6259	37.9	4276	-0.4
Ilic D	13728	2.4	12740	6.1	3319	-82.4	3684	1.6
Sekulic I	12497	-3.3	11826	1.1	4772	16.4	4183	42.7
Gavrilovic A	10879	83.6	8125	-3.8	3215	2.7	2356	-1.7
Kostic T	10276	162.7	6606	-24.0	1983	-50.0	1411	-69.5
Milicevic M	7464	8.1	6140	-4.6	1981	-15.3	1680	-5.9
Gavrilovic D	4105	-39.0	5197	32.8	1417	-7.9	1635	20.1
Stankovic B	3867	-108.0	5126	8.5	1268	-34.4	1731	20.0
Siskovic D	3937	-18.4	4838	40.6	981	-46.3	1445	14.1
Jankovic M	1371	-19.8	1860	17.2	539	-0.9	530	4.0

Табела 4. Фреквенција употребе именица (N), глагола (V), придева (A) и прилога (ADV) по ауторима и њихов *индекс специфичности* за цео корпус

аутора и века у коме је дело објављено приказани су у табели 5. За сваку врсту речи дат је укупан број појављивања у целом корпусу (F), укупан број појављивања у текстовима одређене партиције (f) и *индекс специфичности* (S) дате врсте речи у односу на цео корпус.

Специфична употреба врста речи својствена ауторима мушког или женског пола и временског раздобља (19. или 20. века) представљена је и на слици 5. У делима која су објавили аутори мушког пола у 19. веку (партиција $m19$) именице су далеко заступљеније него у осталим деловима ($S_{m19}=129, 4128$), док је употреба глагола специфичнија за део корпуса $f19$ ($S_{f19}=32, 8464$), у коме је роман Драге Гавриловић објављен 1887. године. Степен специфичности употребе глагола и прилога статистички значајно је негативан у делима које су објавили аутори мушког пола у 19. веку ($S_{m19}=-48, 6432$ за глаголе, $S_{m19}=-53, 5126$ за прилоге).

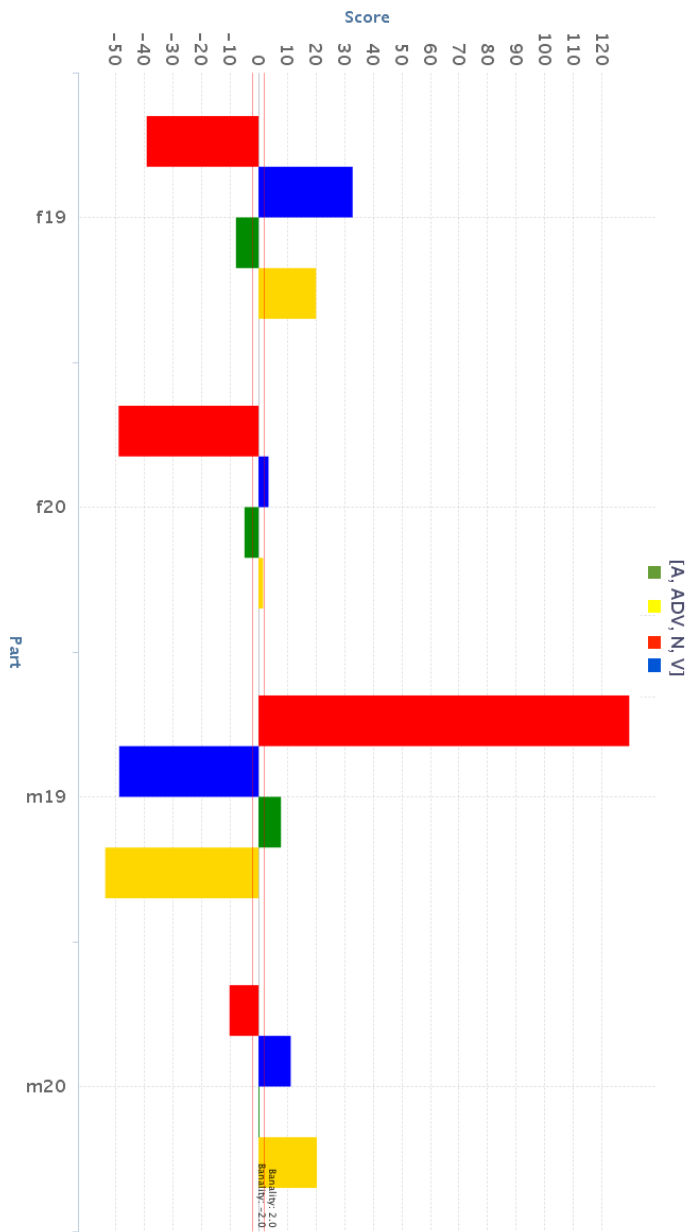
На основу података о учесталости употребе врста речи, а према χ^2 расподели, спроведена је факторска анализа међузависности, представљена у дводимензионалном графичком облику (слика 6). На добијеној факторској мапи је уочљиво да су глаголи и прилози чешће коришћени у партицијама $f19$ и $m20$, па се налазе на истој страни



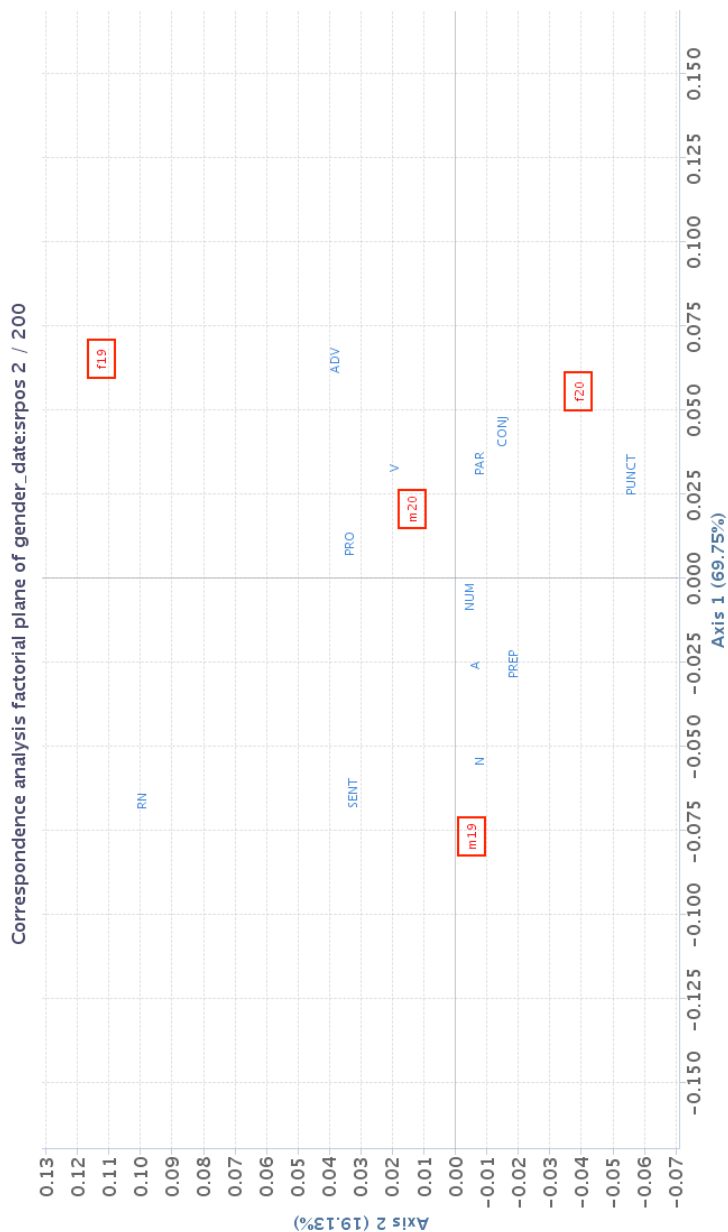
Слика 4. Димензије партиција srpELTeC корпуса креираних на основу информација о полу аутора и веку у коме је дело објављено

хоризонталне осе (*индекс специфичности* има позитивну вредност). Са супротне стране налази се партиција *m19*, код које је забележен изразито негативан *индекс специфичности* употребе глагола и прилога, али и партиција *f20*, чији је *индекс специфичности* употребе глагола и прилога позитиван, али указује на далеко мању употребу глагола и прилога у овој партицији у односу на партиције *f19* и *m20*. Из тог разлога се партиције *m19* и *f20*, које карактерише мања заступљеност глагола и прилога, налазе у супротним квадрантима у односу на вертикалну осу. Посматрајући вертикалну осу видимо да се са једне стране налази партиција *m19*, у којој је забележен изузетно висок *индекс специфичности* употребе именица, док су партиције у којима су именице далеко мање заступљене, смештене на супротну страну.

Далеко сложенији је визуелни приказ резултата спроведене факторске анализе међузависности над корпусом издељеним на партиције на основу аутора, а у погледу фреквенција коришћених лема (слика 7). Овако спроведена анализа омогућава уочавање и проучавање законитости и трендова, који иначе нису лако приметни због велике количине разноврсних података. На пример, Милутин Ускоковић и Милица Јанковић леме *живот*, *љубав*, *волети*, *мисао* и *осећати* користе



Слика 5. Специфичност употребе одређених врста речи у sPELTeC корпусу по полу аутора (m – мушког или f – женског) и временском раздобљу (19. или 20. век)



Слика 6. Резултат факторске анализе међузависности примењене над корпусом издељеним на партиције по полу аутора и датуму објављивања дела

Unit	F	f_{f19}	S_{f19}	f_{f20}	S_{f20}	f_{m19}	S_{m19}	f_{m20}	S_{m20}
N	190565	4105	-39.0398	36198	-48.8455	60820	129.4128	89442	-10.1284
V	173822	5197	32.8464	35624	3.4805	49007	-48.6432	83994	11.2368
A	63307	1417	-7.8845	12377	-4.9007	19383	7.8305	30130	0.3083
ADV	50628	1635	20.0939	10400	1.6149	13477	-53.5126	25116	20.3649

Табела 5. Фреквенција употребе и индекс специфичности именица (N), глагола (V), придева (A) и прилога (ADV) по партицијама креираним на основу пола аутора и временског раздобља

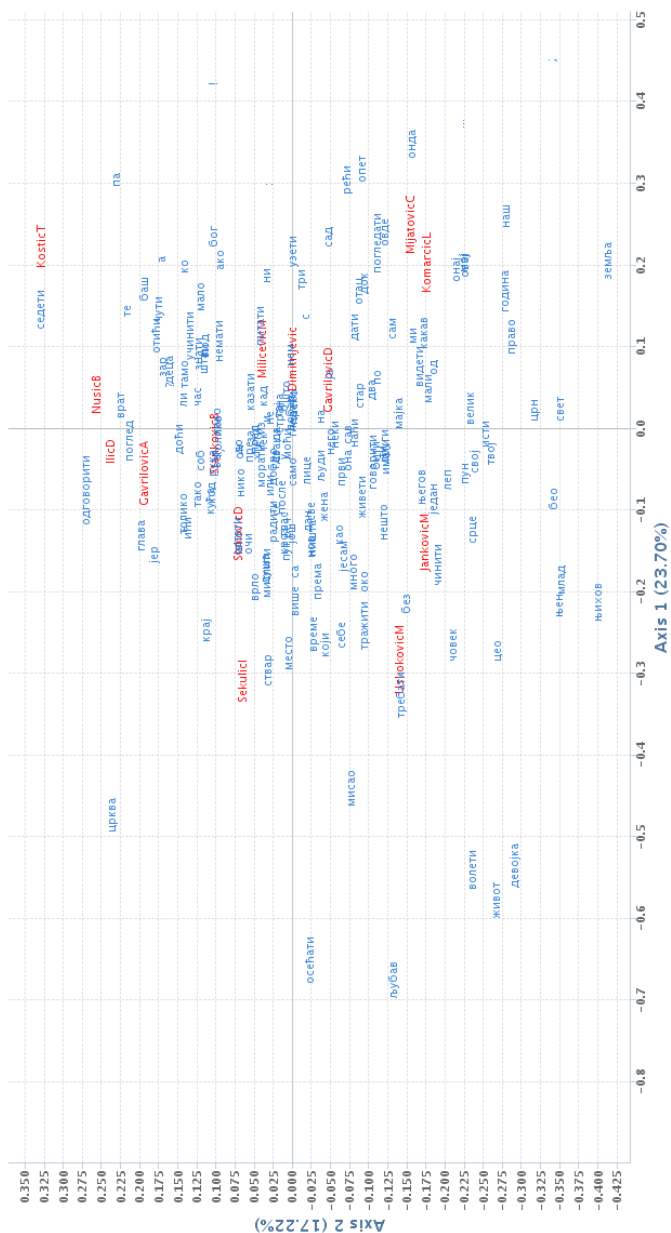
далеко чешће него остали аутори, те су смештени у исти квадрант факторске мапе. Са супротне стране хоризонталне осе, у горњем левом квадранту, налази се лема *црква*, као и ауторка Исидора Секулић, која је користи чешће у односу на остале ауторе. С друге стране, с обзиром на то да Исидора Секулић, осим леме *црква*, веома често користи и леме *љубав* и *живот*, као и аутори Ускоковић и Јанковић, поменути аутори и коришћене леме налазе се на истој страни у односу на вертикалну осу. Резултати овакве анализе свој пуни значај добијају тек након адекватног стручног тумачења, што превазилази оквире овог рада.

У последњем кораку урађена је и кластер анализа матрице добијене претходно спроведеном факторском анализом међузависности. Дијаграм у форми стабла (слика 8) приказује хијерархијско груписање тј. уређеност релација које постоје између текстова аутора и лема коришћених у њима. Оваква класификација текстова омогућава боље разумевање и једноставнију интерпретацију резултата факторске анализе међузависности.

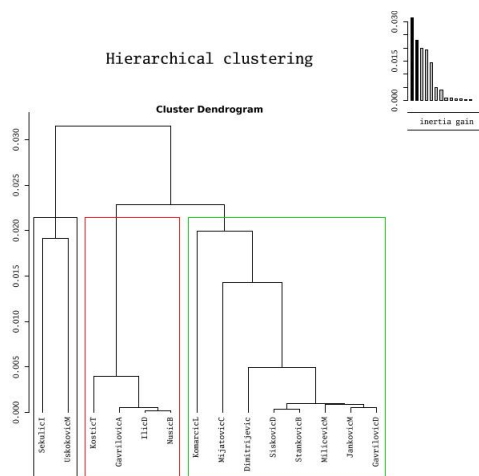
4. Закључак

У овом раду представљена је текућа верзија srpELTeC корпуса, који се састоји од прозних дела српске књижевности с краја 19. и почетка 20. века. Ради илустрације могућности текстометријског приступа, у оквиру ТХМ програмског окружења извршена је анализа srpELTeC корпуса, уз представљање могућности визуелног приказа добијених резултата. Спроведена анализа srpELTeC корпуса, односно неки сценарији примене ТХМ алата, немају друге намене до да прикажу могућности коришћења алата који указује на значај текстуалности и предлаже неке смерове

Correspondence analysis factorial plane of SRPROMAN_text_author_svi:rlemma 2 / 200



Слика 7. Резултат факторске анализе међузависности примењене над корпусом издељеним по ауторима



Слика 8. Кластер анализа спроведена над резултатима факторске анализе међузависности

анализе оних делова који се истичу и произилазе из пројекција самог корпуса.

Текстометријски приступ се већ дуго примењује и веома је корисна метода за анализу корпуса различитих области друштвено-хуманистичких наука. Законитости и закључци који произилазе из текстометријских истраживања заснивају се на квалитативној и квантитативној анализи. Квалитативна анализа омогућава постављање почетних хипотеза, које је онда могуће испитати квантитативном анализом на већем узорку. Намена примене квантитативних тј. статистичких метода је указивање на она места у тексту која се по одређеним својствима разликују и одступају. Овај другачији начин читања текстова омогућава постављање нових питања на правим местима, и то не ради давања одговора, већ ради идентификовања места која је потребно поново читати тј. додатно анализирати, што води валидном тумачењу.

Захвалност

Аутор се захваљује подршци COST акције 16204 – Distant Reading for European Literary History, која је омогућила ово истраживање и боравак аутора овог текста (STSM-CA16204-42562) у IHRIM (Institut d'Histoire des Représentations et des Idées dans les Modernités) лабораторији, École Normale Supérieure de Lyon у Француској. Аутор се посебно захваљује домаћинима Сержу Еидену и Бенедикт Пенсемин на гостопримству и корисним коментарима и сугестијама.

Литература

- Beaudouin, Valérie. "Statistical Analysis of Textual Data: Benzécri and the French School of Data Analysis." *Glottometrics* Vol. 33 (2016): 56–72
- Benzécri, Jean-Paul. *L'analyse des données*. Vol. 2, Dunod Paris, 1973
- Dobrić, Nikola. "Korpusna lingvistika kao osnovna paradigma istraživanja jezika". *Naučnostručni časopis za jezik, književnost i kulturu Philologia* Vol. 7 (2009): 47–57
- Evert, Stefan и Andrew Hardie. "Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium", 2011
- Guiraud, Pierre. *Les caractères statistiques du vocabulaire*. Presses universitaires de France, 1954
- Guiraud, Pierre. *Problèmes et méthodes de la statistique linguistique*. D. Reidel Publishing Company, 1959
- Heiden, Serge. "The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme". У *24th Pacific Asia conference on language, information and computation*, 389–398. Institute for Digital Enhancement of Cognitive Development, Waseda University, 2010
- Heiden, Serge, Jean-Philippe Magué и Bénédicte Pincemin. "TXM: Une plateforme logicielle open-source pour la textométrie-conception et développement". У *10th International Conference on the Statistical Analysis of Textual Data-JADT 2010*, Vol. 2, 1021–1032. Edizioni Universitarie di Lettere Economia Diritto, 2010
- Hunston, Susan. *Corpora in applied linguistics*. Ernst Klett Sprachen, 2002
- Katnić-Bakaršić, Marina. *Lingvistička stilistika*. Budimpešta: Open Society Institute, 1999
- Krstev, Cvetana. *Processing of Serbian – Automata, Texts and Electronic Dictionaries*. Belgrade: University of Belgrade, Faculty of Philology, 2008

- Lafon, Pierre. “Sur la variabilité de la fréquence des formes dans un corpus”. *Mots. Les langages du politique* Vol. 1, no. 1 (1980): 127–165
- Lafon, Pierre. *Dépouillements et statistiques en lexicométrie*, Vol. 24. Paris: Slatkine-Champion, 1984
- Lavrentiev, Alexei, Serge Heiden и Matthieu Decorde. “Analyzing TEI encoded texts with the TXM platform”. У *The Linked TEI: Text Encoding in the Web. TEI Conference and Members Meeting 2013*, 2013
- Lebart, Ludovic и André Salem. *Analyse statistique des données textuelles: questions ouvertes et lexicométrie*. Dunod Paris, 1988
- Lebart, Ludovic и André Salem. *Statistique textuelle*. Dunod Paris, 1994
- Muller, Charles. *Initiation au méthodes de la statistique linguistique*. Classiques Hachette, 1973
- O’Keeffe, Anne и Michael McCarthy. *The Routledge handbook of corpus linguistics*. Routledge, 2010
- Pincemin, Bénédicte. “Sept logiciels de textométrie”, , 2018, URL <https://halshs.archives-ouvertes.fr/halshs-01843695>, working paper or preprint
- Schmid, H. “TreeTagger-a language independent part-of-speech tagger”. <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/> (1994), URL <https://ci.nii.ac.jp/naid/20000989946/en/>
- Utvić, Miloš. “Izgradnja referentnog korpusa savremenog srpskog jezika”. Ph.D. thesis, Univerzitet u Beogradu, Filološki fakultet: Beograd, 2014
- Utvić, Miloš. “Annotating the corpus of contemporary Serbian”. *INFOtheca* Vol. 12, no. 2 (2011): 39–51