Текстометријска и компаративна анализа двојезичног корпуса часописа *Инфотека*

УДК

САЖЕТАК: У раду се представља текстометријска и компаративна анализа двојезичног корпуса научног часописа *Инфотека*, који обухвата паралелне текстове на српском и енглеском језику. Корпус је екстрахован из дигиталне библиотеке *Библиша*, при чему су метаподаци повезани са Википодацима. Посебна пажња посвећена је анализи српског и енглеског подкорпуса применом текстометријских метода, укључујући фреквенцијску анализу, анализу кључних речи, колокација и тематских образаца. Након појединачних анализа, спроведена је компаративна анализа са циљем идентификовања разлика и сличности у лексичким карактеристикама два језика. Резултати показују да, иако су текстови преводни еквиваленти, постоје одређене разлике у дистрибуцији термина, што указује на утицај језичких и научних конвенција. Рад доприноси развоју методологије за анализу двојезичних корпуса у домену дигиталних хуманистичких наука и језичких технологија.

КЉУЧНЕ РЕЧИ: двојезични корпус, текстометрија, Инфотека, паралелни текстови, компаративна анализа, Википодаци, дигиталне библиотеке

РАД ПРИМЉЕН: 04. септембар 2020.
РАД ПРИХВАЋЕН: 25. новембар 2020.

Ранка Станковић
ORCID 0000-0001-5123-6273
ranka.stankovic@rgf.bg.ac.rs

Универзитет у Београду
Рударско-геолошки факултет
Београд, Србија

1. Увод

Двојезични корпуси имају значајну улогу у савременој лингвистици и дигиталним хуманистичким наукама, јер омогућавају систематско

поређење језичких структура, лексике и језичких образаца у различитим језицима. Као паралелни скупови текстова, они представљају драгоцен ресурс за проучавање преводилачких стратегија, терминологије и језичких варијација. У контексту језичких технологија, двојезични корпуси служе као основа за развој и евалуацију алата као што су машинско превођење, аутоматска анотација и моделирање језика. Истовремено, у дигиталним хуманистичким истраживањима, они омогућавају дубље разумевање културних и научних токова кроз вишејезичне изворе, повезујући квантитативне методе са интерпретативним приступима.

Часопис Инфотека представља релевантан и поуздан извор података за анализу у области библиотекарства, информационих наука и језичких технологија. Као двојезични научни часопис, он обухвата радове на српском и енглеском језику, што омогућава формирање паралелног корпуса погодног за компаративна истраживања. Структурисани метаподаци и доступност пуног текста чланака додатно доприносе његовој употребљивости у текстометријским и корпусним анализама, чинећи га значајним ресурсом у оквиру дигиталних хуманистичких истраживања.

Stanković et al. (2012) представили су систем *Библиша*, алат намењен унапређењу претраге великих колекција TMX (Translation Memory eXchange) докумената добијених паралелизацијом двојезичних текстова из вишејезичних дигиталних библиотека. Његове могућности су тестиране на скупу TMX докумената генерисаних из двојезичних чланака часописа Инфотека. Континуирани развој система је донео нове могућности платформе, која је постала репозиторијум паралелних текстова различитих колекција, при чему треба нагласити да је Инфотека прва колекција постављена на Библишу и уједно колекција која се редовно ажурира новим садржајем. Наиме, по изласку онлине верзије броја на OJC платформи[1], чланци се паралелизују, описују метаподацима, уносе у Библишу и у базу знања Википодаци (Stanković and Davidović 2021; Андоновски 2026).

Текстометријски приступ се већ дуго примењује као ефикасна метода за анализу корпуса у различитим областима друштвено-хуманистичких наука. Комбиновањем лексикометријских и статистичких метода са савременим корпусним технологијама, текстометрија омогућава нелинеарно, квантитативно и квалитативно проучавање дигиталних текстуалних ресурса.

1. http://infoteka.bg.ac.rs/ojs/

Софтвер TXM (Heiden 2010) представља платформу за текстометрију која омогућава различите статистичке прорачуне и графички приказ резултата[2] (Pincemin, Heiden, and Mazuet 2022).

Корпус SrpELTeC послужио је као полигон за различита текстометријска истраживања (Jaćimović 2019; Krstev 2021; Stanković, Krstev, and Vitas 2024), а након успешних резултата на књижевним текстовима, фокус се помера на анализу научних текстова. У овом раду се илуструју могућности текстометријског приступа у оквиру програмског окружења TXM, анализом двојезичног корпуса часописа Инфотека. Посебна пажња посвећена је издвајању и поређењу српског и енглеског подкорпуса, као и примени различитих текстометријских метода, укључујући анализу фреквенције, специфичности, колокација и временске прогресије. Добијени резултати додатно су интерпретирани кроз одговарајуће визуелизације, чиме се омогућава јасније сагледавање развојних трендова и тематске структуре корпуса.

2. Развој корпуса

Дигитални објекти у оквиру система *Библиша* описани су структурисаним метаподацима који су прилагођени представљању научних часописа и њихових садржаја (Stanković et al. 2012). Ови метаподаци обухватају податке о појединачним бројевима часописа, као и о чланцима који их чине. За бројеве часописа бележе се основни библиографски подаци, као што су идентификациони број, годиште, број, месец и година издавања. На нивоу чланка, метаподаци су организовани двојезично, на српском и енглеском језику, и укључују информације о ауторима, наслову, врсти рада, распону страна, сажетку, кључним речима и доступности пуног текста, уз могућност евидентирања података о преводиоцу. Овако структурисани метаподаци омогућавају систематично представљање, претрагу и анализу двојезичног корпуса. (Stanković, Obradović, and Trtovac 2012; Stanković et al. 2016)

База *Библиша* имплементирана је у оквиру Mongo[3] окружења, што омогућава флексибилно управљање подацима и њихову организацију. Систем подржава извоз података у различите формате, као што су TXT и XML, уз могућност избора нивоа детаљности.

2. TXM – платформа за текстометрију
3. https://www.mongodb.com/

За потребе ове анализе издвојен је подскуп метаподатака који обухвата идентификатор чланка, број часописа, годину издавања, наслов и податке о првом аутору. Поред тога, коришћене су и текстуалне верзије чланака на српском и енглеском језику, што омогућава њихову појединачну и упоредну анализу.

Након екстракције, корпуси срИнфотека и енИнфотека су импортовани у програмско окружење TXM, у оквиру којег су примењени интегрисани модели за морфосинтаксичко означавање и лематизацију. За аутоматско одређивање врста речи и лема TXM се ослања на алат TreeTagger[4] (Schmid 1994), који представља један од стандардних алата за обраду корпуса и омогућава тагирање на више језика.

У случају српског језика, коришћени су модели прилагођени српском језику и усклађени са скупом етикета Universal Dependencies (UD)[5] (тагсетом), чије су обучавање и опис дати у (Utvić 2011; Stanković, Škorić, and Šandrih Todorović 2022). Овај приступ омогућава уједначено и конзистентно означавање категорија врста речи, што је посебно важно за даљу текстометријску анализу. За енглески језик коришћен је Penn Treebank тагсет[6], један од најраспрострањенијих стандарда за означавање врста речи у енглеском језику, који омогућава фину граматичку диференцијацију и широко је примењен у корпусној лингвистици и обради природног језика. Имајући у виду да су за српски и енглески језик коришћени различити тагсетови (UD и Penn Treebank), било је неопходно извршити њихово усклађивање ради омогућавања директне компаративне анализе. У том циљу, Penn Treebank ознаке су мапиране на одговарајуће категорије Universal Dependencies (UD) тагсета, који представља универзални стандард за морфосинтаксичко означавање. Процес мапирања заснивао се на постојећим конверзионим шемама и лингвистичким кореспонденцијама између два система, уз прилагођавања у случајевима када не постоји једнозначно пресликавање. Ова нормализација омогућила је уједначено посматрање врста речи у оба подкорпуса и поузданије поређење њихових граматичких карактеристика.

Поткорпуси имају по 196 радова, 30.062 упарених сегмената (углавном реченица), при чему енглески потпорпус енИнфотека има

4. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
5. https://universaldependencies.org/u/pos/
6. https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Penn-Treebank-Tagset.pdf

871 253 токена, док српски срИнфотека има 775 669 токена, што значи да је по броју токена енглески потпорпус 12% већи.

Када се броје само речи, без интерпункције тада срИнфотека има 653 215, док енИнфотека има 751 185, дакле 15% је већи енглески поткорпус. Када су у питању различити токени, срИнфотека има 68 631, а енИнфотека 40 358, што значи да српски има чак за 70% више различитих токена, понајвише због различитих флективних облика. Међутим, и број различитих лема у срИнфотека је за 16% већи, јер има 36 640, а енИнфотека 31 373, иако је енглески корпус нешто већи.

Богатство лексике се може мерити на различите начина. Једна од основних мера је Type–Token Ratio (TTR) која представља однос броја различитих речи (типова) и укупног броја токена. Како је зависна од величине корпуса, поређење се може примењивати код корпуса исте или сличне величине. За срИнфотека је овај однос 0,09, док за енИнфотека он износи 0,05, што показује веће богатство лексике у српском поткорпусу. Друга мера, Root TTR представља однос броја различитих речи и квадратног корена укупног броја токена и за срИнфотека ова мера износи 77,93, док за енИнфотека износи 43,24.

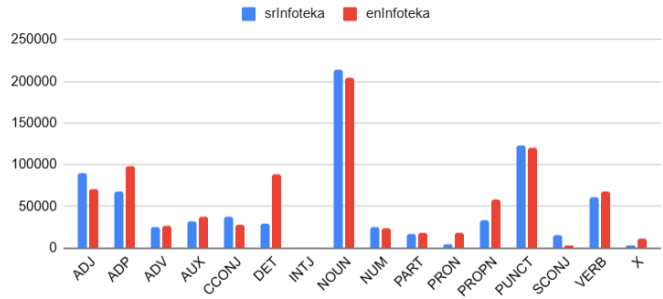
Графикон на Слици 1 приказује дистрибуцију врста речи према UD категоријама (UPOS) у српском (sr) и енглеском (en) поткорпусу, при чему су уочљиве и сличности и јасне разлике. Уочава се да је категорија именица (NOUN) доминантна у оба језика, са нешто већом фреквенцијом у српском корпусу, што указује на благо већу номиналну густину. Слично томе, интерпункција (PUNCT) је високо заступљена и релативно уједначена, што одражава структуру научног дискурса.

Значајне разлике уочавају се код одређених граматичких категорија. У енглеском корпусу изразито је већа заступљеност одредница (DET) и предлога (ADP), што је у складу са аналитичком природом енглеског језика. Насупрот томе, српски показује мању употребу ових категорија, јер се синтаксички односи често изражавају морфолошки, пре свега падежима.

Глаголи (VERB) су нешто учесталији у енглеском, што може указивати на већу експлицитност предикатске структуре, док су придеви (ADJ) и прилози (ADV) релативно уједначени у оба корпуса. Везници (CCONJ, SCONJ) показују умерене разлике, при чему је енглески склонији употреби везника.

Посебно је уочљива већа заступљеност властитих именица (PROPN) у енглеском корпусу, што може бити последица различитих конвенција у именовању или превођењу. Укупно посматрано, добијене разлике

одражавају типолошке особености српског (флективног) и енглеског (аналитичког) језика, као и специфичности научног стила у оба језика.



Слика 1. Број токена по врстама речи у српском и енглеском поткорпусу.

Може се закључити да иако корпус Инфотека представља паралелне текстове, разлике у категоријама као што су DET, VERB и ADP одражавају типолошке и стилске разлике између српског и енглеског језика.

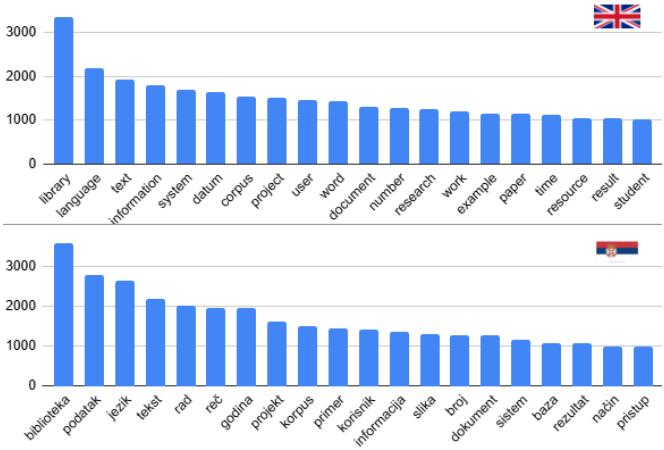
3. Фреквенцијска анализа

3.1 Фреквенције именица

Анализа фреквенције именица у паралелном корпусу Инфотека спроведена је применом одговарајућих CQL упита за енглески и српски подкорпус. У енглеском делу корпуса, коришћењем упита `[enpos="NN.*"]`, идентификовано је укупно 204 262 појављивања именица, са 16 396 различитих облика и 12 392 лема.

Слика 2 приказује расподелу најфреквентнијих именица у енглеском и српском поткорпусу. У енглеском делу корпуса, идентификовано је укупно 204 262 појављивања именица, са 16 396 различитих облика и 12 392 лема. Најзаступљеније су именице *library*, *language*, *text*, *information* и *system*, што указује на доминацију термина везаних за библиотекарство, језичке технологије и обраду текста. Поред њих, често се јављају и *corpus*, *project*, *user*, *word* и *document*, који одражавају технолошки и кориснички оријентисан аспект анализираних текстова.

У српском поткорпусу забележено је 214 129 појављивања именица, са 24 996 различитих облика и 13 583 лема. Најфреквентније именице су *biblioteka*, *podatak*, *jezik*, *tekst* и *rad*, што указује на слично тематско језгро као у енглеском делу корпуса. Међу чешћим именицама налазе се и *reč*, *godina*, *projekt*, *korpus*, *primer* и *korisnik*, као и *informacija*, *slika*, *broj*, *dokument*, *sistem*, *baza*, *rezultat*, *način* и *pristup*.



Слика 2. Двадесет најфреквентнијих именица у поткорпусима.

Поређење показује висок степен поклапања у кључним појмовима два поткорпуса, што је очекивано за паралелне текстове. Ипак, уочавају се и разлике у дистрибуцији појединих лексема, при чему енглески корпус показује нешто већу концентрацију термина техничког карактера, док српски обухвата шири спектар општијих и контекстуалних именица. Ове разлике могу бити последица језичких специфичности, али и различитих преводилачких и стилских избора.

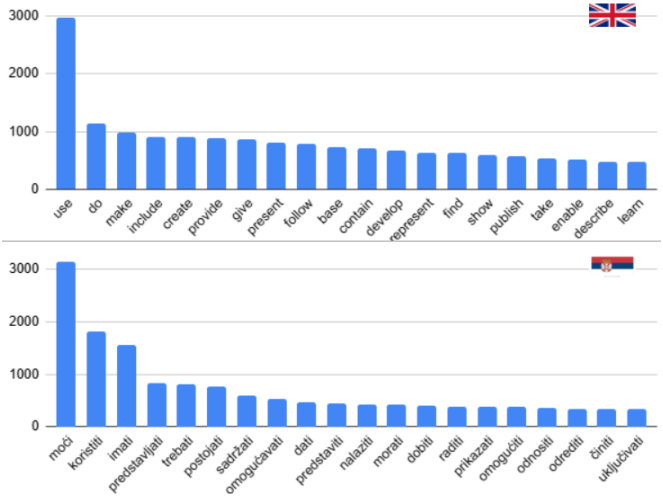
3.2 Фреквенције глагола

У српском поткорпусу, применом упита `[srpos="VERB"]`, идентификовано је укупно 61 121 појављивање глагола, са 11 272 различита облика и 3 982 леме. У енглеском поткорпусу, коришћењем упита `[enpos="VV.*"]`, пронађено је 67 891 појављивање глагола, са

5 433 различита облика и 2 328 лема. Добијени резултати указују на већу разноврсност облика у српском језику.

Слика 3 приказује расподелу најфреквентнијих глагола у енглеском и српском поткорпусу. У оба језика доминирају општи, високо фреквентни глаголи који имају широку употребу у научном дискурсу. У енглеском корпусу најзаступљенији је глагол *use*, који се јавља значајно чешће од осталих, праћен глаголима као што су *do*, *make*, *include* и *create*. Ови глаголи одражавају карактеристичан стил енглеског научног израза, са нагласком на експлицитно описивање поступака и метода.

У српском корпусу најфреквентнији је глагол *моћи*, што указује на чешћу употребу модалних конструкција и изражавање могућности. Следе глаголи *користити*, *имати*, *представљати* и *требати*, који су такође типични за академски стил.

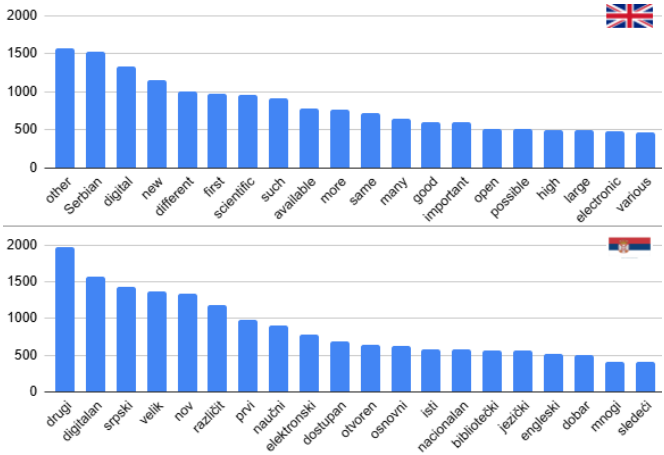


Слика 3. Двадесет најфреквентнијих глагола у поткорпусима.

Поређење указује на разлике у дискурзивним стратегијама: енглески језик тежи директнијем и акционом изражавању, док српски чешће користи модалне и описне конструкције. Ове разлике могу бити последица типолошких особености језика, али и преводилачких избора у паралелном корпусу.

3.3 Фреквенције придева

У енглеском поткорпусу, применом упита `[enpos="JJ.*"]`, идентификовано је укупно 70 582 појављивања придева, са 6 259 различитих облика и 5 400 лема. У српском поткорпусу, коришћењем упита `[srpos="ADJ"]`, пронађено је 89 363 појављивања придева, са 20 537 различитих облика и 9 153 леме. И ови резултати показују знатно већу разноврсност облика у српском језику.



Слика 4. Двадесет најфреквентнијих придева у поткорпусима.

Слика 4 приказује расподелу најфреквентнијих придева у енглеском и српском поткорпусу. У енглеском делу доминирају општи и високо фреквентни придеви као што су *other*, *Serbian*, *digital*, *new* и *different*, који имају широку примену у научном дискурсу. Посебно се издвајају *Serbian* и *digital*, што указује на тематску усмереност корпуса ка српском језику, дигиталним ресурсима и поступцима. Међу учесталим придевима налазе се и *scientific*, *such*, *available*, *same* и *many*, док *important*, *possible*, *large* и *various* одражавају тенденцију ка генерализацији и вредносном оцењивању у академском стилу.

Српски поткорпус показује сличан образац, са доминацијом придева *други*, *дигиталан*, *српски*, *велики* и *нов*, при чему *дигиталан* и *српски* такође указују на тематску оријентацију корпуса. Међу чешћим придевима издвајају се *различит*, *први*, *научни*, *електронски* и

доступан, који доприносе прецизнијем опису и класификацији појмова. Доменску и стилску специфичност додатно обликују придеви *отворен*, *основни*, *исти*, *националан*, *библиотечки* и *језички*, док *добар*, *многи* и *следећи* указују на елементе вредновања и организацију текста.

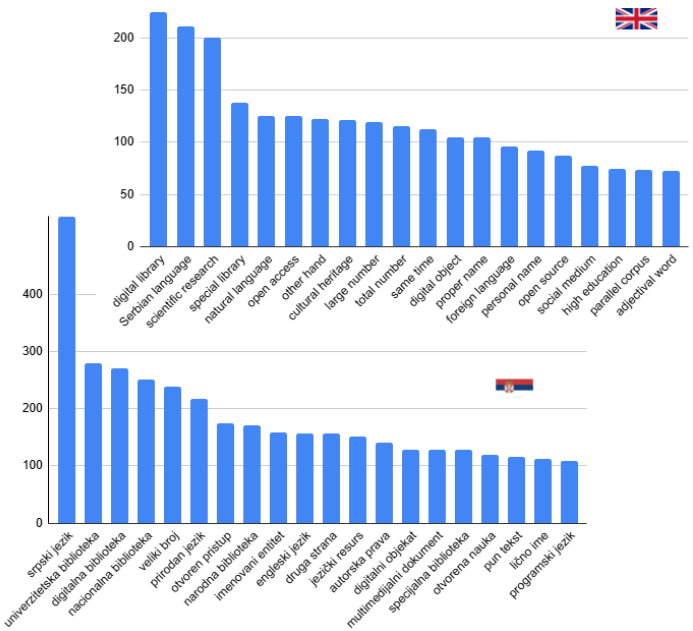
Укупно посматрано, оба поткорпуса карактерише комбинација општих и доменски релевантних придева, што је типично за научни дискурс, уз висок степен тематске подударности између енглеског и српског језика. Оваква расподела потврђује да придеви у корпусу углавном имају функцију квалификације и спецификације појмова.

3.4 Фреквенције спојева придева и именице

Међу вишечланим терминима, најчешћи спој представља придев за којим следи именица, који има значајну улогу у прецизирању и спецификацији појмова. У енглеском поткорпусу, анализом придевско-именичких колокација применом упита `[enpos="JJ.*"][enpos="NN.*"]`, идентификовано је укупно 47 813 појављивања. Ове колокације обухватају 26 731 различит облик, односно 23 901 лему. Добијени резултати указују на високу разноврсност придевско-именичких спојева, што одражава богатство номиналних синтагми у енглеском научном дискурсу.

Слика 5 (горе) приказује расподелу најфреквентнијих доменских појмова у анализираном корпусу. Уочава се да доминирају термини као што су *digital library*, *Serbian language* и *scientific research*, што указује на јасну тематску оријентацију корпуса ка дигиталним библиотекама, језичким ресурсима и научним истраживањима. Значајну заступљеност имају и појмови *natural language*, *open access* и *cultural heritage*, који одражавају шири контекст примене језичких технологија у области дигиталне хуманистике. Међу учесталим појмовима налазе се и *metadata*, *digital object*, *personal name* и *foreign language*, што указује на важност структурирања података и мултијезичности у корпусу. Присуство термина као што су *parallel corpus*, *social medium* и *high education* додатно потврђује интердисциплинарни карактер анализираних текстова.

Слика 5 (доле) приказује расподелу најфреквентнијих придевско-именичких колокација у српском поткорпусу. Уочава се доминација комбинација као што су *српски језик*, *универзитетска библиотека*, *дигитална библиотека* и *национална библиотека*, што указује на јасну тематску оријентацију корпуса ка језику, библиотекарству и дигиталним ресурсима.

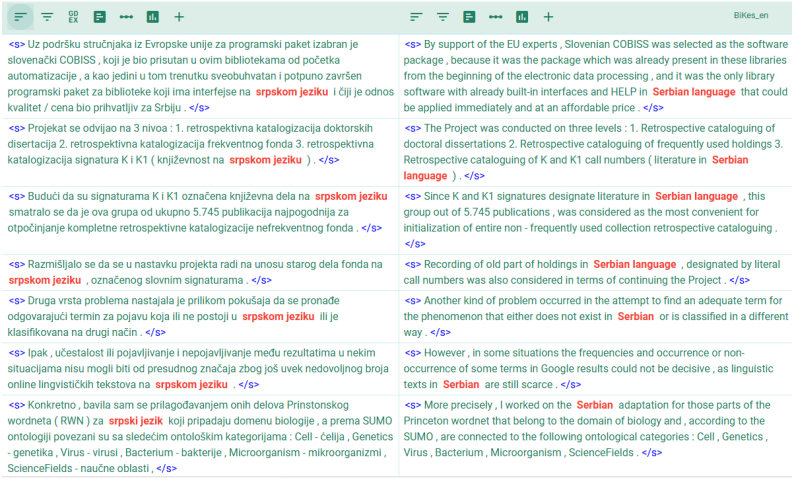


Слика 5. Двадесет најфреквентнијих именица са придевом који претходи.

Међу учесталим колокацијама налазе се и *велики број*, *природан језик*, *отворен приступ*, *народна библиотека* и *именовани ентитет*, које одражавају кључне концепте из области обраде природног језика и управљања информацијама. Присуство колокација као што су *енглески језик*, *друга страна*, *језички ресурс* и *ауторска права* указује на мултијезички и правни контекст анализираних текстова. Даље, колокације *дигитални објекат*, *мултимедијски документ*, *специјална библиотека*, *отворена наука*, *пун текст*, *лично име* и *програмски језик* потврђују интердисциплинарни карактер корпуса. Укупно посматрано, расподела придевско-именичких спојева показује да српски поткорпус карактерише богата и тематски кохерентна употреба номиналних синтагми, које имају кључну улогу у прецизном именовању и спецификацији појмова у научном дискурсу.

Значајна разлика у фреквенцијама се уочава код термина *српски језик* (536) и преводног еквивалента *Serbian language* (211). Разлог је што је у енглеском тексту *language* обично изостављен (подразумева се)

што се може видети на Слици 6 која приказује конкорданце паралелног корпуса Инфотека са платформе Noske[7] – NoSketch Engine (Kilgarriff et al. 2014).



Слика 6. Конкорданце поткорпуса Инфотека на платформи Noske.

У енглеском језику се често користи структура именица–именица тамо где се у српском јавља придевско-именичка конструкција. То показује случај *универзитетске библиотеке*, која је друга у српском а нема је у енглеском, јер се тамо реализује као именица–именица (*university library*). Слична појава уочава се и у другим примерима: *језичке технологије* – *language technology*, *матерњи језик* – *mother tongue*, *здравствена заштита* – *health care*, *образовни процес* – *education process*.

Укупно посматрано, расподела појмова показује да корпус обједињује теме из области језичких технологија, библиотекарства, дигиталних ресурса и образовања, уз наглашену улогу обраде природног језика и управљања дигиталним садржајем.

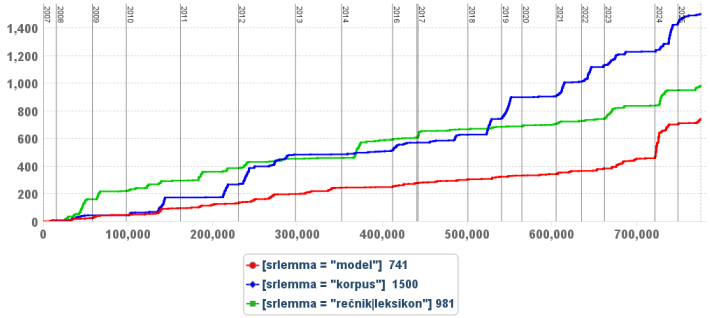
7. https://noske.jerteh.rs/ – инстанца система NoSketch Engine коју одржава Друштво за језичке ресурсе и технологије JePTex

4. Текстометријска анализа

4.1 Прогресија

Прогресија у ТХМ-у представља приказ расподеле и промене учесталости изабраних језичких јединица дуж текста или кроз задати корпус, омогућавајући увид у њихову динамику и развој кроз време или структуру документа. Слика 7 приказује временску прогресију појављивања појмова *модел*, *корпус* и *речник/лексикон* у српском поткорпусу. Уочава се да појам *корпус* (плава линија) има највећу укупну учесталост и показује континуиран раст током целог периода, са израженијим повећањем око 2012., 2019. и од 2024. наовамо, што указује на несмањено интересовање за корпусе и њихово коришћење у истраживањима. Појам *речник/лексикон* (зелена линија) бележи стабилан и умерен раст, са нешто ранијим порастом у односу на остале посматране термине, и интензиван раст 2024. године што одражава традиционално снажно присуство лексичких тема. Ипак, његов раст је постепенији и мање динамичан у односу на *корпус*. С друге стране, појам *модел* (црвена линија) показује спорији раст у ранијем периоду, али са уочљивим убрзањем у каснијим фазама, што може бити повезано са развојем савремених метода машинског учења и језичких модела.

Укупно посматрано, прогресија указује на померање фокуса ка корпусно заснованим и моделским приступима, што одражава шире трендове у области језичких технологија.



Слика 7. Прогресија појмова везаних за језичке ресурсе.

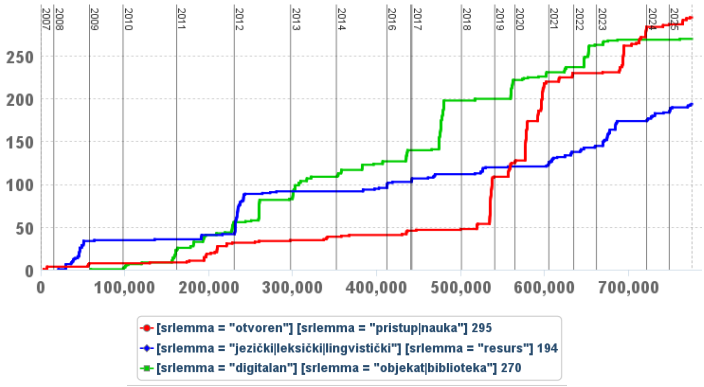
Слика 8 приказује временску прогресију појављивања одабраних концепата у српском поткорпусу, илуструјући развој тематских области

током периода обухваћеног корпусом. Посматране су три групе појмова: *отворен приступ/наука*, *језички/лексички/лингвистички ресурс* и *дигиталан објекат/библиотека*.

Уочава се да дигитални објекти и библиотеке (зелена линија) показују релативно стабилан раст у периоду 2011–2023, са наглим скоком у 2017. години, што одражава постепену дигитализацију и развој инфраструктуре дигиталних библиотека. Тематски број часописа је 2023. године посвећен систему еНаука, јавно доступном информационом систему који је почео са радом у првом кварталу 2023. године.

Помињање језичких ресурса (плава линија) показују нагли раст током 2008., 2012. и 2023. године, и раст након тога, што указује на јачање интересовања за језичке технологије и ресурсе у научном дискурсу последњих година. С друге стране, отворена наука (црвена линија) први раст има 2011. и након тога интензивно тек од 2018. Овакав тренд одражава ширење иницијатива отворене науке и све већу релевантност политике отвореног приступа у научној заједници.

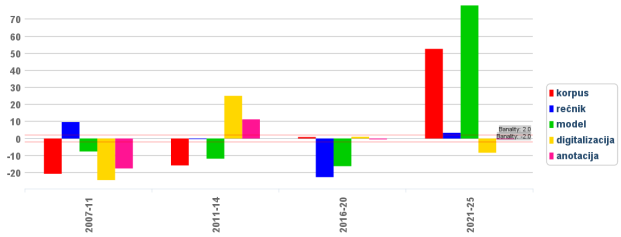
Укупно посматрано, прогресија указује на јасан развој инфраструктурних тема, језичких ресурса и отворене науке, при чему свака од посматраних категорија одражава различите фазе развоја дигиталног научног окружења.



Слика 8. Прогресија изабраних истраживачких тема.

4.2 Специфичности тема по периодима

Специфичност (енгл. *specificity*) у TXM-у представља статистичку меру која показује у којој мери је одређена језичка јединица прекомерно или недовољно заступљена у посматраном подскупу у односу на референтни корпус, омогућавајући идентификацију карактеристичних појмова за дати контекст или период (Heiden 2010). Слика 9 приказује специфичности одабраних појмова (*корпус*, *речник*, *модел*, *дигитализација*, *анотација*) за четири временска периода (од 2007.–2025. године), при чему су вредности изражене у односу на просек корпуса.



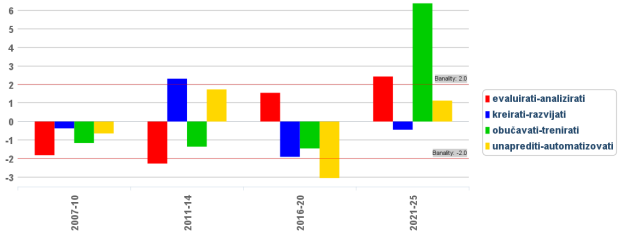
Слика 9. Специфичност изабраних истраживачких тема.

У првом периоду (2007–2011) уочава се да се речници помињу више него што је просек на нивоу целог корпуса, док остали термини, посебно *дигитализација* и *корпус*, имају негативну специчност што указује на њихову релативно малу заступљеност у том периоду. У периоду 2011–2014. долази до пораста специфичности *дигитализације* и *анотације*, што указује на јачање интересовања за дигиталне ресурсе и обраду текста. Период 2016–2020. карактерише релативна стабилизација, са вредностима блиским просеку за већину појмова, уз негативну специфичност *речника* и *модела*. За овај период, изразито позитивну специфичност имају на пример придеви *нормативан*, *терминолошки*, *отворен*, као и именице *репозиторијум* и *синтагма*.

Најизраженије промене уочавају се у периоду 2021–2025, где *модел* и *корпус* бележе снажан пораст специфичности, што указује на доминацију креирања и коришћења корпуса и језичких модела. Уочимо да су ова два термина тек у последњем периоду добила позитивну специфичност. Истовремено, *дигитализација* губи на значају, што

сугерише да ова тема постаје подразумевана инфраструктура, а не централни предмет истраживања.

Слика 10 приказује промену специфичности група глагола који описују истраживачке активности кроз четири временска периода. У раном периоду (2007–2010) све групе имају негативну или ниску специфичност, што указује на њихову слабу заступљеност. У периоду 2011–2014 долази до пораста специфичности групе *креирати–развијати*, као и умереног раста *унапредити–аутоматизовати*, док остале групе остају мање изражене. У периоду 2016–2020 уочава се пад специфичности већине група, посебно *унапредити–аутоматизовати*, што указује на привремено смањење њихове релевантности. Најзначајније промене јављају се у периоду 2021–2025, где група *обучавати–тренирати* бележи снажан пораст специфичности, док *евалуирати–анализирати* такође расте, што указује на јачање описивања језичких модела и аналитичких приступа. Ови резултати одражавају прелаз од развојних и инфраструктурних активности ка методама заснованим на машинском учењу и евалуацији.



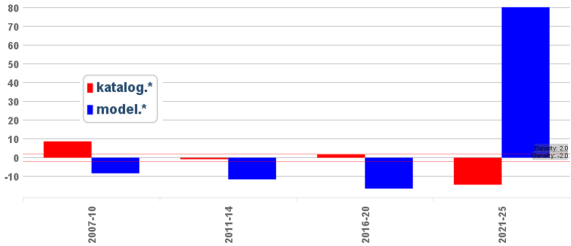
Слика 10. Специфичност изабраних истраживачких активности.

Укупно посматрано, резултати показују јасан развојни ток: од ране фазе усмерене на лексикографију, преко фазе дигитализације и анотације, ка савременом периоду у коме доминирају корпусни и моделски приступи.

4.3 Поређење прогресије и специфичности

Поређење прогресије и специфичности појмова *модел.** и *каталог.** на истом CQL упиту указује на јасну промену тематског фокуса кроз време.

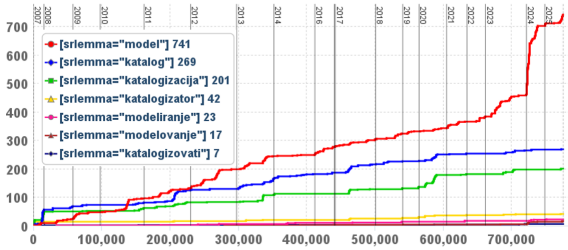
Анализа специфичности показује да је појам *каталог.** у раном периоду (2007–2010) позитивно специфичан, што указује на његову релативну доминацију у односу на просек корпуса (слика 11). У наредним периодима његова специфичност опада и остаје на нивоу просека, са падом у последњем периоду (2021–2025), што сугерише да овај појам губи централну улогу у дискурсу. Насупрот томе, *модел.** показује негативну специфичност у ранијим фазама, али у последњем периоду бележи изразито висок пораст, што указује на његову снажну тематску релевантност у савременом корпусу и савременим истраживањима.



Слика 11. Специфичност изабраних истраживачких активности.

Овај налаз је у складу са анализом прогресије, која показује да *модел* има релативно умерен и постепен раст до око 2023. године, након чега следи нагло повећање учесталости (слика 12). С друге стране, *каталог* и сродни појмови (*каталогизација*, *каталогизатор*) показују стабилан, али умерен раст без значајних скокова до 2021. а после тога такорећи стагнира (што значи да се заправо скоро и не појављује), што указује на њихову континуирану, али све мање доминантну улогу.

Комбиновано посматрање ова два аспекта омогућава јасније тумачење: док прогресија показује апсолутни раст употребе појмова, анализа специфичности открива њихов релативни значај у оквиру корпуса. У том смислу, иако *каталог.** остаје присутан током читавог периода, његов релативни значај опада услед експлозивног раста појма *модел.**. Овај тренд одражава шири помак од традиционалних библиотечких и каталогизационих тема ка савременим приступима заснованим на моделима и машинском учењу.



Слика 12. Специфичност изабраних истраживачких активности.

5. Моделирање тема

Моделирање тема спроведено је над енглеским поткорпусом часописа, при чему је издвојено шест тематских целина које одражавају различите аспекте анализираног корпуса (табела 1). Поступак моделирања тема заснован на LDA (Latent Dirichlet Allocation) обухвата неколико кључних корака: претпроцесирање текста, које укључује чишћење, токенизацију и лематизацију, као и уклањање стопречи. На тако припремљеним подацима тренира се LDA модел ради издвајања латентних тема у корпусу.

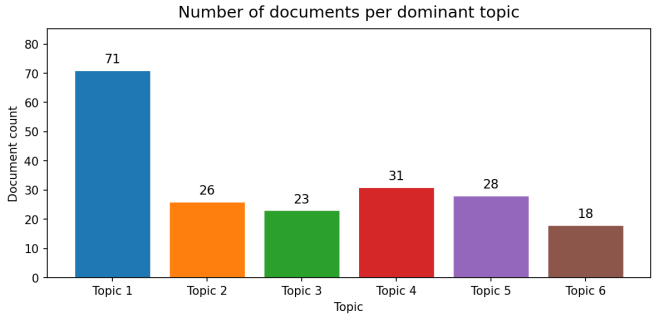
Прва тема обухвата појмове везане за библиотечке системе, научна истраживања и дигиталну инфраструктуру, док је друга усмерена на корпусну лингвистику и језичке ресурсе, укључујући обраду текста и превођење. Трећа тема повезана је са лексикографским аспектима, као што су речници, облици речи и примери употребе. Четврта тема односи се на дигиталне библиотеке и софтверске пројекте, укључујући платформе као што је *Europeana*. Пета тема обухвата образовни контекст, са фокусом на студенте, наставу и учење у дигиталном окружењу. Шеста тема повезује корпусне и језичке анализе са библиографским и ауторским аспектима, укључујући цитирање и структуру текста. Добијене теме указују на интердисциплинарни карактер корпуса, а самим тим и часописа, који обједињује библиотекарство, језичке технологије, образовање и дигиталне хуманистичке науке.

Добијени резултати се потом анализирају и интерпретирају уз помоћ различитих визуелизација, као што су интерактивни прикази тема, облаци речи и графикони расподеле тема по документима. Слика 13

Тема	Кључне речи
Тема 1	library, system, data, research, scientific, user, university, digital, national, science
Тема 2	corpus, language, word, text, resource, serbian, model, used, document, translation
Тема 3	name, language, serbian, word, dictionary, example, used, form, figure, text
Тема 4	digital, library, project, software, material, europeana, data, computer, user, page
Тема 5	student, text, language, document, course, learning, school, web, programming, figure
Тема 6	word, language, serbian, citation, used, data, corpus, author, noun, number

Табела 1. Теме издвојене из поткорпуса енИнфотека.

приказује број радова по темама, при чему је за сваки рад аутоматски одређена по једна доминантна тема помоћу LDA модела.

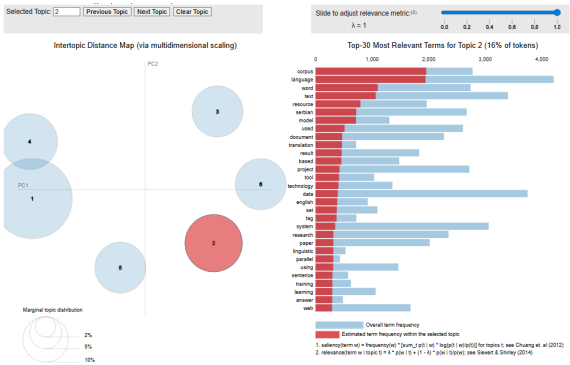


Слика 13. Дистрибуција радова по темама.

Слика 14 приказује интерактивну визуелизацију резултата моделирања тема помоћу LDA модела pyLDAvis (Sievert and Shirley 2014). Лева страна представља мапу међутематских удаљености (Intertopic Distance Map), добијену применом мултидимензионалног скалирања, где сваки круг одговара једној теми. Величина круга указује на релативну заступљеност теме у корпусу, док растојање

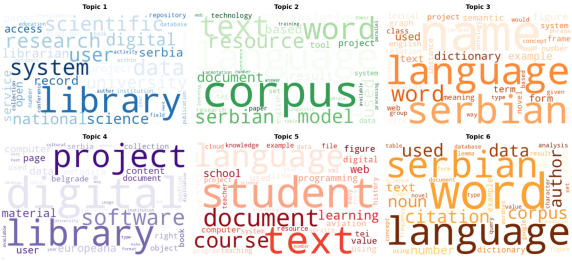
између кругова одражава степен сличности између тема. Уочава се да су поједине теме блиске и делимично преклапајуће, док су друге јасније издвојене.

Десна страна приказује најрелевантније термине за изабрану тему (Тема 2), која обухвата око 16% укупног корпуса. Међу кључним терминима издвајају се *corpus*, *language*, *word*, *text*, *resource*, *serbian* и *model*, што указује да је ова тема повезана са корпусном лингвистиком и језичким ресурсима. Плаве траке представљају укупну учесталост термина у корпусу, док црвене траке означавају њихову релевантност у оквиру конкретне теме (Sievert and Shirley 2014). Наиме, црвене траке не представљају само апсолутну учесталост речи, већ њену релевантност у оквиру изабране теме, односно процењену учесталост те речи у документима који припадају тој теми. Другим речима, оне показују у којој мери је дата реч карактеристична управо за ту тему, а не за корпус у целини. Нека реч може бити карактеристична за више тема, али са различитом релевантношћу. Оваква визуелизација омогућава лакше тумачење структура тема и њихових карактеристичних појмова.



Слика 14. Интерактивно истраживање тема у корпусу.

Слика 15 приказује облаке речи за шест идентификованих тема, што потврђује тематску разноврсност и интердисциплинарни карактер корпуса.



Слика 15. Облаци речи по темама.

6. Дискусија

Резултати текстометријске анализе двојезичног корпуса часописа Инфотека показују висок степен тематске и лексичке подударности између српског и енглеског поткорпуса, што је очекивано имајући у виду њихову паралелну природу. На нивоу фреквенцијске анализе, уочава се да оба поткорпуса деле заједничко језгро појмова из области библиотекарства, језичких ресурса и дигиталних технологија, што потврђује тематску конзистентност корпуса.

Ипак, уочене разлике у дистрибуцији врста речи и лексичких јединица указују на утицај типолошких особености језика. Енглески језик, као аналитички, показује већу употребу функционалних речи, док српски, као флективни језик, испољава већу морфолошку разноврсност, што се огледа у већем броју различитих облика и вишем степену лексичког богатства. Анализа придевско-именичких спојева, указује на стабилност доменске терминологије у оба језика, уз присуство кључних појмова као што су *digital library*, *natural language* и *open access*, односно њихових српских еквивалената. Истовремено, уочене су разлике које произилазе из преводилачких конвенција, као што је изостављање именице *language* у енглеском језику.

Текстометријска анализа временске прогресије и специфичности указује на јасан развојни ток тематских интересовања. Рани период карактерише доминација лексикографских и библиотечких тема, затим следи фаза дигитализације и анотације, док савремени период обележава снажан пораст корпусних приступа и језичких модела. Овај тренд је посебно уочљив у анализи појмова *модел.** и *каталог.**, где се показује померање фокуса са традиционалних библиотечких активности ка савременим методама заснованим на машинском учењу.

Резултати моделирања тема додатно потврђују интердисциплинарни карактер корпуса, који обједињује области библиотекарства, језичких технологија, образовања и дигиталних хуманистичких наука. Идентификоване теме показују да корпус није тематски хомоген, већ обухвата више повезаних, али различитих истраживачких праваца.

7. Закључак

У раду је представљена текстометријска и компаративна анализа двојезичног корпуса часописа Инфотека, са циљем испитивања лексичких, структурних и тематских карактеристика српског и енглеског поткорпуса. Применом различитих метода, укључујући фреквенцијску анализу, анализу колокација, временску прогресију, специфичности и моделирање тема, добијени су резултати који омогућавају свеобухватно сагледавање корпуса.

Анализа је показала да, иако су текстови паралелни, постоје разлике условљене особинама језика. Српски поткорпус одликује већа морфолошка и лексичка разноврсност, док енглески показује већу структурну експлицитност. Истовремено, тематска анализа указује на снажан развој области језичких технологија, са померањем фокуса ка корпусима и језичким моделима у савременом периоду.

Добијени резултати потврђују да двојезични корпуси представљају вредан ресурс за лингвистичка и интердисциплинарна истраживања, омогућавајући истовремено квантитативну и квалитативну анализу. Посебан допринос рада огледа се у примени текстометријских метода на доменски специфичном корпусу, као и у интеграцији различитих аналитичких приступа.

Као правци даљег истраживања могу се издвојити проширење корпуса, као и примена других метода обраде текста и семантичке анализе, ради дубљег разумевања структуре и еволуције научног дискурса.

Захвалност

Захвалност за садржај корпуса дугујемо проф. др Цветани Крстев која је 16 година уређивала часопис Инфотека, бројним ауторима, рецензентима, преводиоцима, лекторима. За паралелизацију корпуса захваљујемо др Јелени Андоновски, др Биљани Рујевић и др Александри

Томашевић. Истраживање је подржао Фонд за науку Републике Србије кроз пројекат број 7276, „Text Embeddings – Serbian Language Applications (TESLA)“.

Литература

Heiden, Serge. 2010. “The TXM platform: Building open-source textual analysis software compatible with the TEI encoding scheme.” In *Proceedings of the 24th Pacific Asia conference on language, information and computation,* 389–398.

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.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. “The Sketch Engine: Ten Years on.” *Lexicography* 1 (1): 7–36.

Krstev, Cvetana. 2021. “White as Snow, Black as Night – Similes in Old Serbian Literary Texts.” *Infotheca - Journal for Digital Humanities* 21 (2): 119–135. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.2.6.

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.

Schmid, Helmut. 1994. “Probabilistic Part-of-Speech Tagging Using Decision Trees.” In *Proceedings of the International Conference on New Methods in Language Processing.* Manchester, UK.

Sievert, Carson, and Kenneth Shirley. 2014. “LDAvis: A Method for Visualizing and Interpreting Topics.” In *Proceedings of the workshop on interactive language learning, visualization, and interfaces,* 63–70.

Stanković, Ranka, and Lazar Davidović. 2021. "Infotheca (Q25460443) in Wikidata." *Infotheca - Journal for Digital Humanities* 21 (1): 87–98. ISSN: 2217-9461. https://doi.org/10.18485/infotheca.2021.21.1.5.

Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac, and Miloš Utvić. 2012. "A Tool for Enhanced Search of Multilingual Digital Libraries of E-journals." In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012),* edited by Nicoletta et al. Calzolari, 1710–1717. Istanbul, Turkey: European Language Resources Association (ELRA).

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, Cvetana Krstev, Duško Vitas, Nikola Vulović, and Olivera Kitanović. 2016. "Keyword-based Search on Bilingual Digital Libraries." In *International KEYSTONE Conference on Semantic Keyword-Based Search on Structured Data Sources,* 112–123. Springer.

Stanković, Ranka, Ivan Obradović, and Aleksandra Trtovac. 2012. "An Approach to Development of Bilingual Lexical Resources." In *Proceedings of the Fifth Balkan Conference in Informatics (BCI 2012),* edited by Zoran Budimac, Mirjana Ivanović, and Miloš Radovanović, 101–104. Novi Sad, Serbia: Faculty of Sciences, Department of Mathematics / Informatics.

Stanković, Ranka, Mihailo Škorić, and Branislava Šandrih Todorović. 2022. "Parallel Bidirectionally Pretrained Taggers as Feature Generators." *Applied Sciences* 12 (10): 5028. https://doi.org/10.3390/app12105028.

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

Андоновски, Јелена. 2026. "Инфотека: часопис за дигиталну хуманистику - 2000-2026 -." *Инфотека – часопис за дигиталну хуманистику* 26 (1).