Именице и колокације придев–именица у српском делу паралелног корпуса часописа *Инфотека*: текстометријски приступ

Јелена Андоновски[1] [0009-0007-9764-5364],
Александра Тртовац[1] [0000-0002-0478-9980]

[1] Универзитетска библиотека "Светозар Марковић", Србија,
{andonovski|aleksandra}@unilib.rs

Научни рад DOI: 10.18485/judig.2026.2.ch12

Апстракт

Овај рад приказује фреквенцијску анализу именица и придевско-именичких колокација, као и текстометријску анализу (анализа прогресије и специфичности термина) српског дела доменског паралелног корпуса научног часописа *Инфотека*, *Infoteka-sr*. Фокус истраживања је на поређењу два периода: фазе усмерене на библиотекарство и информатику (до 2014) – поткорпус BI и фазе окренуте дигиталној хуманистици (од 2016) – поткорпус DH. За текстометријску анализу коришћен је алат TXM, а добијени резултати указују на промену тематског фокуса часописа, од традиционалних библиотечко-информационих појмова ка концептима дигиталне хуманистике, језичких ресурса и технологија. Корпус *Infoteka-sr* потврђује се као релевантан дигитални ресурс за даља истраживања у области библиотекарства и информатике и дигиталне хуманистике.

Кључне речи: Инфотека, текстометрија, библиотекарство и информатика, дигитална хуманистика, корпус, језички ресурси, фреквенцијска анализа

1. Увод

Инфотека, часопис за дигиталну хуманистику је научни часопис који је почео да излази 2000. године. За 26 година излажења прошао је кроз значајне промене које су утицале на његов профил, мењајући не само структуру и квалитет часописа

и објављених радова, већ и научноистраживачку заједницу која у њему објављује резултате својих истраживања и која је постала знатно шира него на почетку. Поседњих десет година посебно су биле изазовне за уредништво како би часопис у сваком смислу пратио технолошке промене и иновације у области научног издаваштва, али и одговорило на захтеве прописане Правилником о категоризацији и рангирању научних часописа[1] које прописује Министарство науке, технолошког развоја и иновација.

Часопис је почео да излази као *часопис за информатику и библиотекарство*, поднаслов је, затим, промењен у *часопис за библиотекарство и информатику*, али је намена остала иста: да се у њему објављују научни и стручни радови о примени информационе технологије у библиотекарству и о научном информисању уопште, о утицају информатике на савремени развој високошколских и специјалних библиотека, као и о новим знањима у информационој и библиотечкој науци. Временом, у *Инфотеци* су почели да се објављују радови из области дигиталне хуманистике те је и поднаслов часописа промењен у *часопис за дигиталну хуманистику* дефинишући га на тај начин као први и једини часопис у Републици Србији са таквим називом и једини који искључиво објављује радове из ове научне области (Andonovski, 2026, 93-94). До данас је објављено 26 годишта, 40 бројева и око 600 радова.

Због тематике радова које објављује, *Инфотека* представља релевантан и поуздан извор података за анализу у области библиотекарства, информационих наука и језичких технологија. Са друге стране, радови су у отвореном приступу и објављују се двојезично (на српском и енглеском језику) што су значајне карактеристике које су довеле до тога да часопис постане дигитални језички ресурс значајан за различите врсте истраживања и погодан за примену различитих језичких алата за српски језик које чланови Друштва за језичке ресурсе и технологије (JePTex)[2] развијају и користе у истраживањима из области рачунарске лингвистике. Тако је припремљен двојезични (енглеско-српски) паралелни корпус *INFOteka* (2026, 104-107) који је у контексту језичких технологија значајан за развој и евалуацију алата као што су машинско превођење, аутоматска

[1] Правилник о категоризацији и рангирањунаучнихчасописа
[2] Друштво за језичке ресурсе и технологије (JePTeX)

анотација и моделирање језика, а истовремено је погодан и за различита компаративна истраживања.

У области дигиталне хуманистике, овакви корпуси омогућавају дубље разумевање културних и научних токова кроз вишејезичне изворе, повезујући квантитативне методе са интерпретативним приступима (Stanković, 2026, 44), а када је реч о корпусу *INFOteka* структурисани метаподаци и доступност пуног текста чланака додатно доприносе његовој употребљивости у текстометријским и корпусним анализама, чинећи га значајним ресурсом у овој области истраживања.

2. Текстометрија

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. Овај приступ подразумева да сваки текст поседује сопствену унутрашњу структуру, коју би било тешко анализирати само ручним алатима (Jaćimović, 2019, 31).

У текстометријским истраживањима углавном се користи софтвер TXM[3] (Heiden, 2010), програм отвореног кода који омогућава различите статистичке прорачуне и графички приказ резултата (Pincemin et al., 2022) анализирајући обимну грађу на најразличитијим језицима, велики број улазних формата текста, као и употребу различитих алата за обраду улазних текстова природних језика (нпр. аутоматских лематизатора). 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. Такође, програм 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 (енг. Corpus Query Processor) претраживача[4]) и сложеније аутоматске обраде резултата упита засноване на

[3] Textometry // TXM
[4] Corpus Query Processor

квантитативним методама, као и извоз резултата у табеларном или графичком облику (2019, 33-35). На основу података о фреквентности ТХМ омогућава визуелни приказ учесталости употребе одређених језичких појава кроз читав корпус, односно његове делове. На овај начин једноставно је уочити у којим деловима корпуса и са којом учесталошћу се одређене речи или изрази користе.

3. Повезани радови

У Србији је као полигон за различита текстометријска истраживања послужио корпус SrpELTeC[5] (2019); (Krstev et al., 2020), (Krstev, 2021); (Krstev, 2026); (Stanković et al., 2024).

У раду (2019) приказани су први резултати текстометријског приступа у оквиру програмског окружења ТХМ на овом корпусу. Током истраживања добијени су следећи резултати:

1. опште информације о корпусу и фреквентности - колико токена је заступљено и који од њих су најфреквентнији, колико лема је у корпусу употребљено, затим приказ различитих речи из SrpELTeC корпуса, њихов тачан број појављивања у корпусу и ранг листа учесталости, а издвојене су и најучесталије корпусне речи;

2. учесталост појављивања одређених врста речи (именице, придеви, глаголи и прилози) и њихова заступљеност у целом корпусу, поткорпусима текстови одређеног аутора, текстови одређене партиције односно текстови аутора мушког пола, текстови аутора женског пола, текстови из одређеног временског раздобља као и индекс специфичности у односу на цео корпус.

На основу података о учесталости употребе врсте речи спроведена је факторска међузависност и кластер анализа матрице[6].

[5] SrpELTeC је корпус књижевних текстова (поткорпус корпуса ELTeC) на српском језику развијен током европског пројекта COST акције (CA16204) „Удаљено читање за историју европске књижевности“ (Distant Reading for European Literary History). Пројекат је реализован у периоду од 2017. до 2021. године (Trtovac et al., 2021), а основни корпус садржи 100 романа из периода 1840-1920. године.

[6] Кластер анализа је хијерархијско груписање тј. уређеност релација које постоје између текстова аутора и лема коришћених у њима.

Рад (2020) анализира реторичке фигуре поређења у облику придева током припреме корпуса SrpELTeC. Учитавањем аноторианог корпуса у алат TXM урађена је квантитативна и квалитативна анализа препознатих реторичких фигура. Утврђена су најфреквентнија поређења и њихове компоненте, као и предмет и основа поређења одређене фигура. На основу учесталости поређења у целом корпусу и његовим појединим деловима, као и коефицијента специфичности, идентификована су значајно честа и значајно ретка појављивања придевских поређења у односу на целокупан корпус, као и специфични начини употребе придевских и именских елемената поређења. Истраживање је допуњено када је пројекат завршен и корпус комплетиран, а резултати истраживања приказани су у раду (2021). Тако је SrpELTeC је послужио као основа за утврђивање репертоара реторичких фигура поређења у старим литерарним текстовима и његово тестирање на савременим текстовима исте врсте.

Поред поменутих истраживања, на корпусу SrpELTeC примењен је текстометријски приступ у програмском окружењу TXM за анализу типичних улоге женских ликова у романима обухваћеним корпусом, као и њихова употреба у делима различитих аутора током периода који корпус обухвата (Stanković et al., 2024). Анализа је извршена захваљујући анотацији именованих ентитета спроведеној током пројекта којом су, између осталог, обухваћене професије, позиције и титуле (Stanković et al., 2021) женских и мушких ликова у романима. Будући да алат TXM омогућава укључивање структурних анотација приликом постављања упита, анализирано је 100 најфреквентнијих позиција, након чега су детаљније размотране улоге женских ликова. На основу графичког приказа прогресије праћене су промене у њиховим улогама током периода обухваћеног корпусом SrpELTeC. Поред временске заступљености женских улога, анализирана је и њихова специфичност у поткорпусима романа мушких и женских аутора појединачно.

Још једна текстометријска анализа корпуса SrpELTeC урађена је на издвојеном поткорпусу од 49 романа 15 писаца (2026). Коришћењем алата TXM одређен је број коришћених лема у корпусу, фреквенција речи, врсте токена према ауторима, фреквенција именица и специфичност употребе одређених

именица, глагола, придева уопште, придева који означавају боје, улога ликова у романима и специфичност употребе неких именица које означавају улоге. Такође је спроведена факторска анализа одабраног корпуса, којом су идентификовани обрасци и односи присутни у поткорпусу. На тај начин омогућено је сагледавање дигиталних трагова садржаних у делима, као и начина на које се они, применом текстометријских метода, могу открити и анализирати.

Након успешних резултата на књижевним текстовима, овакав вид истраживања примењен је и у анализи научних текстова односно на двојезичном паралелном корпусу научног часописа *Инфотека* (2026). У овом истраживању посебна пажња посвећена је издвајању и поређењу српског и енглеског поткорпуса, као и примени различитих текстометријских метода, укључујући фреквенцијску анализу именица, глагола, придева и облика придев-именица, затим, анализу прогресије, специфичности тема према периодима, поређење прогресије и специфичности, као и моделирање тема над енглеским поткорпусом часописа. Поред анализе појединачних корпуса (српског и енглеског), спроведена је и компаративна анализа са циљем идентификовања разлика и сличности у лексичким карактеристикама ова два језика, а добијени резултати додатно су интерпретирани кроз одговарајуће визуелизације, чиме се омогућава јасније сагледавање развојних трендова и тематске структуре корпуса, али и самог часописа.

Сва претходно наведена истраживања представљала су основу и подстицај за ово истраживање у коме се приказује текстометријска анализа српског дела корпуса *INFOteka* са посебним акцентом на фреквенцијској анализи именица и вишечланих термина облика придев-именица у два доменска поткорпуса који обухватају следеће периоде: од 2007. до 2014. када је часопис био часопис за библиотекарство и информатику, и од 2016. када је часопис за дигиталну хуманистику. Поред фреквенције именица и колокација придев-именица (одељак 6.) дата је и временска прогресија (одељак 7.1.) као и специфичност доменске терминологије (7.2.). За разлику од рада (2026) у овом истраживању није урађена компаративна анализа са енглеским делом корпуса, као ни поређење прогресије и специфичности, и за разлику од осталих

приказаних истраживања није урађена факторска анализа што је предвиђено за нека наредна истраживања.

4. Корпус *INFOteka*

Корпус *INFOteka* смештен је у дигиталну библиотеку *Библиша* и то је први корпус који је постао њен саставни део и једини који се континуирано допуњава новим текстовима. Од самог почетка на њему се примењују различити алати за обраду природних језика, најчешће они који се односе на проналажење и екстракцију информација и терминологије из доменских корпуса. О резултатима тих истраживања више се може сазнати у радовима (Stanković et al., 2012a); (Stanković et al., 2012b); (Тртовац, 2017).

Систем *Библиша* креиран је ради унапређења могућности претраживања вишејезичних дигиталних библиотека електронских часописа (2012a), али је временом допуњен и другим врстама корпуса (Андоновски, 2021); (Moderc et al., 2023). Акценат *Библише* је на вишејезичној претрази пуног текста паралелних колекција постављањем комплексних упита уз могућност њиховог морфолошког и семантичког проширења позивањем различитих вишејезичних лексичких и терминолошких ресурса (2012a); 2012b; (Stanković et al., 2015).

Дигитални објекти у оквиру *Библише* описани су структурисаним метаподацима који су прилагођени представљању научних часописа и њихових садржаја (2012a). Ови метаподаци обухватају податке о појединачним бројевима часописа (идентификациони број, годиште, број, месец и година издавања), као и о чланцима који их чине и уносе се упоредо на српском и енглеском језику (информације о ауторима (име и презиме, имејл адреса, афилијација), наслову, врсти рада, распону страна, сажетку, кључним речима и доступности пуног текста, уз могућност евидентирања података о преводиоцу) што омогућава систематично представљање, претрагу и анализу двојезичног корпуса (Stanković et al., 2016). Временом, поред

ових метаподатака уз сваки чланак стоји и његов QID[7] из базе Википодаци.

Википодаци, вишејезична база знања општег типа, омогућава преузимање и чување структурираних података на више језика у интероперабилном машински читљивом формату, те омогућава и постављање комплексних упита који као резултат дају информације које не могу да се добију претрагом других база података, као што је на пример COBISS каталог[8] или Библиша. Такође, у зависности од добијених резултата они се интерпретирају кроз различите визуализације. Тако је и садржај корпуса *INFOteka* постао део ове базе (Stanković and Davidović, 2021) са намером да се са допуном корпуса креирају упоредо и Википодаци. Осим тога, чињеница да су Википодаци део облака Отворени повезани подаци (Linked Open Data – LOD) указује на то да је и *Инфотека*, а самим тим и српски језик, део овог великог екосистема на вебу, што додатно повећава њену видљивост, али и видљивост српског језика.

5. Корпус истраживања

Истраживање које се у овом раду приказује фокусирано је на српски део паралелног корпуса *INFOteka* (*Infoteka-sr*). За разлику од рада (2026) у којем је на нивоу целог корпуса анализирана фреквенција лексичких јединица уопште (именица, глагола и придева) и вишечланих термина придев-именица, као и анализа прогресије и специфичности тема по периодима и дата компаративна анализа српског и енглеског поткорпуса, у овом раду фокус је на фреквенцији именица и вишечланих термина облика придев-именица из домена библиотекарства и информатике, рачунарске лингвистике и дигиталне хуманистике. За потребе истраживања корпус *Infoteka-sr* експортован је из Библише и подељен на два поткорпуса: поткорпус из периода од 2007. до 2014. године када је часопис носио назив *часопис за*

[7] QID је трајни јединствени идентификатор (енгл. permanent identifier) који се аутоматски додељује приликом креирања записа у бази Википодаци и не може се касније мењати. Идентификатор је позитиван цео број са префиксом великог слова Q и повезан је са насловом и описом ентитета. Може имати више псеудонима и одређен број изјава којима се представљају својства и и вредности ентитета који се описује.
[8] COBISS

библиотекарство и инфроматику (поткорпус BI), и поткорпус од 2016. до 2026. од када часопис носи назив *часопис за дигитлану хуманистику* (поткорпус DH). Издвојен је подскуп метаподатака који обухвата идентификатор чланка, број часописа, годину издавања, наслов и податке о првом аутору, а коришћене су текстуалне верзије чланака на српском. Поткорпус BI има 97 паралелизованих чланака, 16.516 упарених сегмената и око 410.000 речи, док поткорпус DH има 98 паралелизованих чланака, 13.546 упарених сегмената и преко 420.000 речи. На Слици 1 приказана је величина оба поткорпуса према броју речи.

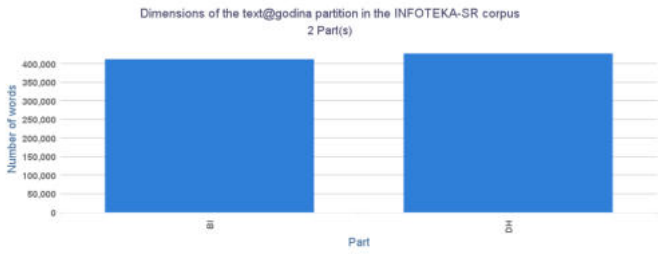

Figure 1: Број речи у поткорпуима BI и DH.

Након екстракције поткорпуси су учитани у 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[9] (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)[10] (тагсетом), чије су обучавање и опис дати у (Utvić, 2011); (Stanković et al., 2021). Овај приступ омогућава уједначено и конзистентно означавање категорија врста речи, што је посебно важно за даљу текстометријску анализу.

6. Фреквенцијска анализа

6.1. Фреквенције именица

Анализа фреквенције именица у корпусу *Infoteka-sr* спроведена је применом одговарајућих CQL упита. За анализу су

[9] TreeTagger - a part-of-speech tagger for many languages
[10] Universal POS tags

одабрани упити “biblio.”, “infor.”, “korpus.”, “leksikon|rečn.”, “digitaliz.”, “metapod.”:

[srlemma=“biblio.*”], [srlemma=“infor.*”], [srlemma=“korpus.*”], [srlemma=“leksikon|rečn.*”], [srlemma=“digitaliz.*”], [srlemma=“metapod.*”]

Добијени резултати анализирани су на нивоу целог корпуса *Infoteka-sr* и поткорпуса BI и DH појединачно. У Табели 1 приказане су најфреквентније именице које садрже леме из постављених упита и њихов број појављивања. Увидом у добијене резултате запажа се да највећи број именица у корпусу садржи лему “biblio”. У складу са тим најфреквентнија је “библиотека” са укупно 3569 појављивања, од чега у поткорпусу BI 2437, а у DH 1132. Затим, “библиотекар” са 592 појављивања у корпусу, од чега у BI 342, а у DH 250. После ње је “библиотекарство” са 257 појављивања у корпусу, у поткорпусима 184 (BI) и 73 (DH). Именица “библиографија” се јавља 123 пута, од чега у поткорпусу BI 52, а у DH 71. Поред поменутих, анализирана је и фреквенција именице “еБиблиотеке” и утврђено је да се у целом корпусу појављује 23 пута и то само у поткорпусу DH, док се у поткорпусу BI ова именица уопште не појављује.

После именица које садрже лему “biblio” најфреквентније су именице које садрже лему “infor” и њих има 2222. Ту се издваја именица “информација” са 1342 појављивања од чега у поткорпусу BI 909, а у DH 433. Следи именица “информатика” са 180 појављивања у целом корпусу, док се у поткорпусу BI јавља 118, а у DH 62 пута. Поред ових заступљена је и именица “информатичар” - у целокупном корпусу 19, од чега у поткорпусу BI има 11, а у DH 8 појављивања.

Следеће анализаране именице су оне које садрже лему “korpus”. Ова лема има 1609 појављивања, а најфреквентнија је именица “korpus” која се у *Infoteka-sr* јавља 1501 пут, од чега у поткорпусу BI 514, а у DH 987.

Анализа упита “leksikon|rečn” показује да се именице са овом лемом појављују 1052 пута. Најфреквентније су: “лексикон” која се у целом корпусу појављује 108 пута, од чега у поткорпусу BI 53, а у DH 55, и именица “речник” која у целом корпусу има 871 појављивање, од чега у поткурпусу BI 539, а у DH 332.

На Слици 2 приказан је графикон фреквенције претходно анализираних лема у целом корпусу *Infoteka-sr*.

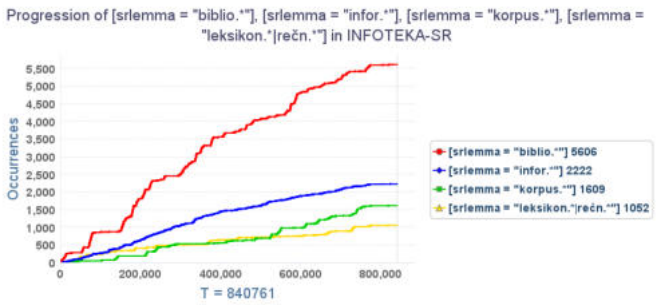

Figure 2: Фреквенција лема “bilio”,“infor”, “korpus”, “leksikon|rečn” у корпусу *Infoteka-sr*.

Table 1: Фреквенција именица које садрже леме “bilio”, “infor”, “korpus”, “leksikon|rečn”.

Именице	**Фреквенција у Infoteka-sr**	**Фреквенција у BI**	**Фреквенција у DH**
Библиотека	3.569	2.437	1132
Библиотекар	592	342	250
Библиотекарство	257	184	73
Библиографија	123	52	71
еБиблиотека	23	нема	23
Информација	1342	909	433
Информатика	180	118	162
Информатичар	19	11	8
Корпус	1501	514	987
Лексикон	108	53	55
Речник	871	539	332

Поред претходно поменутих анализирана је и фреквенција упита “digitaliz.” и “metapod.” за које је приказана и временска прогресија у одељку 7.1.:

[srlemma=“digitaliz.*”] [srlemma=“metapod.*”]

Лема “digitaliz” се појављује 607 пута у целом корпусу, а именице са највећом фреквенцијом које садрже ову лему су “дигитализација” и “дигитализовање”. Именица „дигитализација“ се у целом корпусу јавља 360 пута, од чега у поткорпусу BI 222, а у поткорпусу DH 138, док се именица “дигитализовање”

у целокупном корпусу јавља 14 пута, под чега у поткорпусу BI броји 12 појављивања, а у DH само 2. Разлог за то може бити чињеница да се овај израз ређе користи у последњих десетак година као последица његове стилске обележености и постепене замене стандардизованим изразом "процес дигитализације".

Лема "metapod" се у целом корпусу појављује 672 пута, а најфреквентнија именица је "метаподатак" са 665 појављивања у целом корпусу, од чега у поткорпусу BI 222, а у DH 443.

6.2. Фреквенција придевско-именичких колокација

У српском језику су продуктивне колокације састављене од придева и именице па се то одражава и у доменском корпусу *Infoteka-sr*. Најфреквентнија придевско-именичка колокација у целом корпусу је "српски језик" са 537 појављивања, од чега се у поткорпусу BI појављује 194 пута и друга је по фреквенцији, док је у DH најфреквентнија са 343 појављивања. Друга колокација по фреквенцији у целом корпусу је "универзитетска библиотека" која се јавља 280 пута и то у поткорпусу BI 95 где је на осмом месту по заступљености, док се у DH налази на другом месту са 185 појављивања. Трећа по заступљености у целом корпусу је колокација "дигитална библиотека" са 270 појављивања и у оба поткорпуса налази се на трећем месту са 127 појављивања у поткорпусу BI и 143 у DH.

Поред три најфреквентније у целом корпусу анализирали смо још неке колокације карактеристичне за домене "библиотекарство и информатика" и "дигитална хуманистика" и њихова фреквенција приказана је у Табели 2. Из Табеле видимо да се колокација "национална библиотека" у целом корпусу појављује 250 пута и на четвртом је месту по заступљености. Ова колокација је у поткорпусу BI најфреквентнија са 238 појављивања, али се у DH појављује само 12 пута и налази се чак на 212. месту. Колокација "природни језик" се у целом корпусу појављује 216 пута и налази се на шестој позицији, док је у поткорпусу BI на четвртом месту са 120 појављивања, а у DH на седмом месту са 96 појављивања. Посебно је анализиран упит [srlemma="dig.*"][srlemma="hum.*"]. Придевско-именичке колокације са лемама у основи из овог упита има укупно 127 у целом корпусу *Infoteka-sr*. На основу добијених резултата анализирана је придевско-именичка колокација "дигитална

хуманистика" и утврђено је да се у целом корпусу појављује 53 пута, од тога у поткорпусу BI 40, а у DH само 13. Остало су различити граматички облици овог појма, као и појам "дигитални хуманиста". У складу са овим подацима ако се погледа временска прогресија (Слика 3) придевско-именичких колакација са овим лемама у основи види се да је њихов раст интензиван само током 2013. и 2014. године. Из тог разлога анализиране су неке колокације које се односе ресурсе и технологије које се у дигиталној хуманистици примењују: "рачунарска лингвистика" - у целом корпусу појављује 56 пута, од тога у поткорпусу BI 35, а у DH 21 пут; затим, "језичке технологије" - у *Infoteka-sr* појављује се 106 пута, од тога у BI 75 пута, у DH 31; и "језички ресурс" - у *Infoteka-sr* појављује се 150 пута, од тога у BI 76 пута, а у DH 74 пута.

Из овога се може закључити да је највероватније појам "дигитална хуманистика" за кратак временски период посто устаљен и усвојен и да се после тога објављени радови нису бавили појмом као таквим већ истраживањима који се у оквиру дигиталне хуманистике спроводе, технологијама и ресурсима које се користе и добијеним резултатима.

Са друге стране, од 2016. године промењен је и поднаслов часописа у "часопис за дигиталну хуманистику", а самим тим и профил часописа што све иде у прилог претходно постигнутим резултатима истраживања.

Детаљном анализом утврђено је да су бројне колокације састављене од придева "библиотечки" и именице која стоји уз њега фреквентније у поткорпусу BI у односу на поткорпус DH. Такве колокације су на пример: "међубиблиотечка позајмица", "библиотечка грађа", "библиотечка делатност", "библиотечки фонд", "библиотечки материјал", "библиотечко пословање", "библиотечко-информациона делатност", "библиотечки систем", "библиотечке услуге".

Већа фреквенција колокација са придевом "библиотечки" у поткорпусу BI указује на доминацију тема из традиционалног библиотекарства и информационе науке. Ове колокације односе се на кључне аспекте библиотечког пословања, управљања збиркама и организације рада библиотека. Њихова учесталост потврђује да су институционална и стручна питања

Table 2: Фреквенција неких колокација придев-именица у корпусу *Infoteka-sr*

Колокација придев-именица	**Фреквенција у Infoteka-sr**	**Фреквенција у BI**	**Фреквенција у DH**
српски језик	537	194	343
универзитетска библиотека	280	95	185
дигитална библиотека	270	127	143
национална библиотека	250	238	12
природни језик	216	120	96
дигитална хуманистика	53	40	13
рачунарска лингвистика	56	35	21
језичке технологије	106	75	31
језички ресурс	150	76	74

библиотекарства била у средишту интересовања аутора у овом периоду на шта и сам поднаслов часописа из тог периода указује.

С друге стране, њихова мања заступљеност у поткорпусу DH одражава померање тематског фокуса ка дигиталним ресурсима, метаподацима, корпусима и технологијама. Ова разлика представља јасан показатељ трансформације тематског профила часописа *Инфотека* и преласка са библиотекарско-информационе ка дигитално-хуманистичкој оријентацији и у скалду са тим и промену поднаслова часописа.

7. Текстометријска анализа

7.1. 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 (2026, 53). Слика 3 приказује временску прогресију појављивања именица које садрже леме “digitaliz” и “metapod”, и придевско-именичке колокације које садрже леме “dig”“hum” у корпусу *Infoteka-sr*.

Запажа се да именице са лемом “metapod” (зелена линија) имају већу укупну учесталост од именица са лемом “digitaliz” и показује спорији раст у периоду од средине 2011. до средине 2017. од када креће њихово изражено повећање све до 2024. године након чега се уочава стабилност. Са друге стране, именице са лемом “digitaliz” (плава линија) показују континуирани раст у периоду од краја 2010. године до краја 2022. године са мањим периодима стагнације, након чега се уочава стабилност. За разлику од претходно анализираних лема, раст придевско-именичке колокације са лемама “dig”“hum” (црвена линија) уочава се само током 2013. и 2014. године након чега следи стабилност до данас.

Укупно посматрано прогресија показује да лема “metapod“ представља једно од кључних тематских упоришта дигиталне хуманистике. Повећање фреквенције ове леме, нарочито до 2024. године, сведочи о њеној све израженијој актуелности и о интензивирању истраживачког интересовања у најновијем временском периоду.

Са друге стране, прогресија колокација са лемама “dig”“hum” показује изразиту стабилност од 2016. што може да буде показатељ да је појам “дигитална хуманистика” постао устаљен за јако кратак временски период и да се акценат померио на ресурсе и технологије које се примењују у области дигиталне хуманистике.

7.2. Специфичност тема према поткорпусима

Специфичност (енгл. specificity) у систему TXM представља статистичку меру која показује степен прекомерне или недовољне заступљености одређене језичке јединице у посматраном подскупу у односу на њену заступљеност у референтном корпусу. На тај начин омогућава се идентификовање карактеристичних појмова и језичких јединица за одређени контекст или период (2010); (2026, 21). Вредност специфичности заснована на хипергеометријској расподели, показује вероватноћу појављивања лексичке јединице

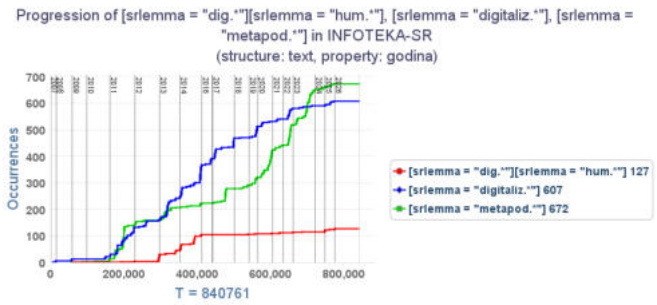

Figure 3: Прогресија лема “digitaliz” и “metapod” и придевско-именичке колокације са лемама “dig”“hum”

у одређеном делу корпуса, а програм TXM омогућава графички приказ расподеле специфичности за изабране јединице (2024, 55).

Слика 4 приказује специфичност одређених појмова у корпусу *Infoteka-sr* подељену према поткорпусима BI и DH. У поткорпусу BI изразито негативну специфичност има именица “роман” за разлику од поткорпуса DH. Разлог за то је што се у поткорпусу DH, у оквиру бројних радова и тематског броја часописа Vol. 21, No. 2(2021)[11] анализира корпус SrpElTeC који углавном садржи романе српске књижевности (Trtovac et al., 2021).

Поред именице “роман” посебно су анализиране и именице “информација” и „софтвер” које имају изразито позитивну специфичност у поткорпусу BI, посебно именица „софтвер”, док у поткорпусу DH имају негативну специфичност.

Разлог за ову појаву вероватно лежи у чињеници да је процес аутоматизације библиотека у нашој средини у великој мери заокружен око 2010. године, након чега се фокус постепено помера ка другачијем терминолошком оквиру. У том контексту, појам “информација” у последњој деценији све више уступа место широј и прецизније структурираној употреби термина “подаци”, што се огледа у различитим савременим формама као што су “метаподаци”, “истраживачки подаци”, “административни метаподаци” и сл.

[11] Vol. 21, No. 2(2021)

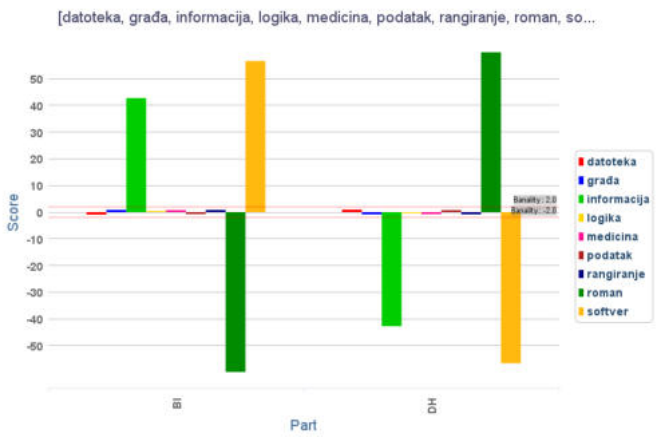

Figure 4: Специфичност одређених именица у поткорпусима BI и DH

8. Закључак

Резултати спроведене текстометријске анализе потврђују да корпус часописа *Инфотека* представља поуздан извор за праћење развоја терминологије и тематских усмерења у областима библиотекарства, информатике и дигиталне хуманистике. Анализа фреквенције именица и придевско-именичких колокација, као и прогресије и специфичности термина показала је јасну промену у тематском профилу часописа током посматраног периода. Док су у поткорпусу BI доминирали појмови карактеристични за библиотекарство и информатику, у поткорпусу DH уочава се све израженије присуство термина повезаних са дигиталном хуманистиком – "језички ресурс", "рачунарска лингвистика", "језичке технологије". На тај начин, промене у структури и садржају корпуса одражавају и ширу трансформацију научног и професионалног поља којем је часопис посвећен.

Посебно је значајно што се промена поднаслова часописа непосредно одражава и на расподелу кључних термина и колокација. Резултати указују да је у периоду након 2016. године дошло до стабилизације терминологије везане за дигиталну хуманистику, при чему се тежиште истраживања постепено помера са самог дефинисања ове области ка ресурсима, методама и технологијама које се у њој примењују. Учестало појављивање термина као што су "метаподаци", "корпуси",

"језички ресурси" сведочи о све већој интеграцији рачунарских метода у хуманистичка истраживања и о растућем значају дигиталне инфраструктуре за научни рад.

Одређени резултати захтевају и додатно тумачење. Мања учесталост појма "еБиблиотека" може се објаснити чињеницом да се овај термин јавља превасходно у оквиру појединачних радова посвећених конкретним пројектима или системима, те стога није достигао ширу употребу у оквиру анализираног корпуса. Слично томе, релативно мање присуство термина "библиографија" у поткорпусу библиотекарства и информатике вероватно је условљено чињеницом да у корпус нису укључени сви бројеви часописа објављени пре 2007. године. Може се претпоставити да би укључивање тих бројева довело до веће заступљености овог појма, с обзиром на то да су библиографска питања у ранијем периоду имала значајније место у оквиру библиотечко-информационе делатности.

Свеукупно посматрано, добијени резултати потврђују вредност корпуса *Infoteteka-sr* као релевантног дигиталног ресурса за истраживања у области корпусне лингвистике, рачунарске лингвистике, библиотекарства и дигиталне хуманистике. Истовремено, они показују потенцијал текстометријских метода за праћење развоја научних дисциплина и идентификовање промена у терминолошким и тематским обрасцима током дужег временског периода.

У наредној фази истраживања планирано је проширење корпуса укључивањем додатних бројева часописа, као и поновно спровођење текстометријске анализе ради провере уочених образаца и њиховог прецизнијег тумачења. На тај начин биће могуће поузданије сагледати динамику развоја кључних појмова и додатно осветлити процес трансформације тематског и терминолошког профила часописа *Инфотека*. Додатно, компаративном анализом са енглеским делом корпуса биће могуће утврдити сличности и разлике у дистрибуцији кључних појмова како у домену библиотекарства и информатике тако и у домену дигиталне хуманистике.

Литература

Jelena Andonovski. 2026. Infotheca: Journal for Digital Humanities - 2000-2026 -. *Infotheca – Journal for Digital Humanities*, 26(1):92–123.

Serge Heiden. 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*, pages 389–398.

Jelena Jaćimović. 2019. Textometric Methods and the TXM Platform for Corpus Analysis and Visual Presentation. *Infotheca – Journal for Digital Humanities*, 19(1):30–54.

Cvetana Krstev. 2021. White as Snow, Black as Night – Similes in Old Serbian Literary Texts. *Infotheca - Journal for Digital Humanities*, 21(2):119–135.

Cvetana Krstev. 2026. Fifteen Writers and their Digital Imprints in Numbers, Images and Words. *Infotheca – Journal for Digital Humanities*, 26(1):9–42.

Cvetana Krstev, Jelena Jaćimović, and Duško Vitas. 2020. Analysis of Similes in Serbian Literary Texts (1840-1920) Using Computational Methods. In *Proceedings of the 4thInt. Conference Computational Linguistics inBulgaria (CLIB 2020)*, pages 31–41. Sofia, Bulgaria:Institute for Bulgarian Language "Prof. Lyubomir Andreychin", Bulgar-ian Academy of Sciences.

Saša Moderc, Ranka Stanković, Aleksandra Tomašević, and Mihailo Škorić. 2023. An Italian-Serbian Sentence Aligned Parallel Literary Corpus. *Review of the NCD*, 47(43):78–91.

Bénédicte Pincemin, 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*, volume II, pages 691–698, Naples, Italy. VADISTAT - Per Simona Balbi, Univ. of Naples Federico II, Vadistat Press.

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

Ranka Stanković. 2026. Textometric and Comparative Analysis of the Bilingual Corpus of the Journal *Infotheca*. *Infotheca – Journal for Digital Humanities*, 26(1):43–66.

Ranka Stanković and Lazar Davidović. 2021. Infotheca (Q25460443) in Wikidata. *Infotheca - Journal for Digital Humanities*, 21(1):87–98.

Ranka Stanković, Cvetana Krstev, Biljana Lazić, and Dalibor Vorkapić. 2015. A Bilingual Digital Library for Academic and Entrepreneurial Knowledge Management. In *Proceeding of 10th International Forum on Knowledge Asset Dynamics — IFKAD 2015: Culture, Innovation and Entrepreneurship: connecting the knowledge dots, Bari, Italy, 10-12 June 2015*, pages 1764–1777. Novi Sad: Faculty of Sciences, Department of Mathematics and Informatics.

Ranka Stanković, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac, and Miloš Utvić. 2012a. A Tool for Enhanced Search of Multilingual Digital Libraries of E-journals. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1710—1717. Istanbul: ELRA.

Ranka Stanković, Cvetana Krstev, and Duško Vitas. 2024. SrpELTeC: A Serbian Literary Corpus for Distant Reading. *Primerjalna književnost*, 47(2):45–63.

Ranka Stanković, Cvetana Krstev, Duško Vitas, Nikola Vulović, and Olivera Kitanović. 2016. Keyword-Based Search on Bilingual Digital Libraries. In *Semantic Keyword-Based Search on Structured Data Sources - Second COST Action IC1302 International KEYSTONE Conference, IKC 2016, Cluj-Napoca, Romania, September 8-9, 2016. Revised 285 Selected Papers*, pages 112–123. Springer.

Ranka Stanković, Cvetana Krstev, Branislava Šandrih Todorović, and Mihailo Škorić. 2021. Annotation of the Serbian ELTeC Collection. *Infotheca - Journal for Digital Humanities*, 21(2):43–59.

Ranka Stanković, Ivan Obradović, and Aleksandra Trtovac. 2012b. An Approach to Development of Bilingual Lexical Resources. In *Fifth Balkan Conference in Informatics BCI 2012, Novi Sad, Serbia, September 16-20, 2012*, pages 101–104. Novi Sad: Faculty of Sciences, Department of Mathematics and Informatics.

Aleksandra Trtovac, 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.

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

Јелена Андоновски. 2021. *Паралелни српско-немачки корпус књижевних текстова: израда, проналажење информација и семантички веб.* Београд: Универзитетска библиотека „Светозар Марковић“. ISBN: 978-86-7301-167-7.

Александра Тртовац. 2017. *Проналажење информација у дигиталним библиотекама.* Београд: Универзитетска библиотека „Светозар Марковић“. ISBN: 978-86-7301-103-5.

Nouns and Adjective–Noun Collocations in the Serbian Subcorpus of the Infoteka Parallel Corpus: A Textometric Analysis

Jelena Andonovski, Aleksandra Trtovac

Abstract

This paper presents textometric analyzes Serbian part of the domain-specific parallel corpus of the scientific journal Infotheca, Infoteka-sr. The research focuses on comparing two periods: the phase oriented toward library and information science (up to 2014) – subcorpus BI and the phase focused on digital humanities (from 2016 onward) – subcorpus DH.Using the TXM tool, a frequency analysis of nouns and adjective–noun collocations was conducted, along with a specificity analysis of terms. The results indicate a shift in the journal's thematic focus, from traditional library and information science concepts toward concepts related to digital humanities, language resources, and technologies. The Infoteka-srcorpus is confirmed as a relevant digital resource for further research in the fields of library and information scienceand digital humanities.

Keywords: Infotheca, textometry, library and information science, digital humanities, corpus, language resources, frequency analysis.

Works are licensed under the terms and conditions of the Creative Commons Attribution 4.0 International License.

https://creativecommons.org/licenses/by/4.0/