Импресум

ЗА ИЗДАВАЧА:
проф. др Александар Јерков office@unilib.bg.ac.rs
Универзитетска библиотека „Светозар Марковић“
Филолошки факултет, Универзитет у Београду

ИЗДАВАЧ:
Филолошки факултет, Универзитет у Београду
Универзитетска библиотека „Светозар Марковић“
Заједница библиотека универзитета у Србији

ГЛАВНИ И ОДГОВОРНИ УРЕДНИК:
проф. др Цветана Крстев cvetana@matf.bg.ac.rs
Филолошки факултет, Катедра за библиотекарство и информатику

ОПЕРАТИВНИ УРЕДНИК:
др Александра Тртовац aleksandra@unilib.bg.ac.rs
Универзитетска библиотека „Светозар Марковић“

УРЕДНИК ЕЛЕКТРОНСКОГ ИЗДАЊА:
Јелена Андоновски andonovski@unilib.bg.ac.rs
Универзитетска библиотека „Светозар Марковић“

УРЕЂИВАЧКИ ОДБОР:
проф. др Александра Вранеш, проф. др Александар Јерков, проф. др Биљана Дојчиновић, *Филолошки факултет, Универзитет у Београду*; проф. др Елизабет Бур, *Институт за романистику, Универзитет у Лајпцигу*; проф. др Владан Девецић, *Факултет организационих наука, Универзитет у Београду*; проф. др Милена Добрева, *Факултет за медијске и когнитивне науке, Универзитет на Малти*; др Томаж Ерјавец, *Одсек технологије знања, Институт „Јожеф Стефан“, Љубљана*; проф. др Светла Коева, *Институт за бугарски језик, Бугарска академија наука*; проф. др Денис Морел, проф. др Агата Савари, *Универзитет Франсоа Рабле из Тура*; проф. др Иван Обрадовић, *Рударско-геолошки факултет, Универзитет у Београду*; проф. др Гордана Павловић Лажетић, проф. др Душко Витас, *Математички факултет, Универзитет у Београду*; проф. др Катерина Здравкова, *Факултет за информатичке науке и компјутерско инжињерство, Универзитет „Св.Кирил и Методије“ у Скопљу*

ISSN 1450-9687 (штампано издање) Београд, год. 19, бр. 1, септембар 2019.
ISSN 2217-9461 (е-издање)

ИЗРАДА ВЕБ ПОРТАЛА ЧАСОПИСА:
Јелена Андоновски
Универзитетска библиотека „Светозар Марковић“

ЛЕКТОР ЗА ЕНГЛЕСКИ ЈЕЗИК:
Тања Ивановић
Катедра за информатику и математику, Универзитет Пасау

ЛЕКТОР ЗА СРПСКИ ЈЕЗИК:
Свјетлана Ђелић
Универзитетска библиотека „Светозар Марковић“

ДИЗАЈН И ПРИПРЕМА ЗА ШТАМПУ:
Бранислава Шандрих
и **Инфо***тека* тим

РЕДАКТОР БИБЛИОГРАФИЈЕ И УДК КЛАСИФИКАЦИЈА:
Наташа Дакић
Универзитетска библиотека „Светозар Марковић“

РЕДАКТОР ДОИ БРОЈЕВА:
др Милош Утвић
Филолошки факултет, Универзитет у Београду

РЕДАКЦИЈА ЧАСОПИСА:
Часопис Инфотека
Булевар краља Александра 71, 11000 Београд

+381 11 3370–211
infotheca@unilib.rs

ШТАМПА:
Мамиго плус
Београд

Часопис излази два пута годишње

Садржај

Креирање и анализа корпуса текстова југословенских рок песама од 1967-2003.[1]

УДК 811.163.41'322

Људмила Петковић
ljudmila.petkovic@gmail.com
Универзитет у Београду
Београд, Србија

САЖЕТАК: У раду се са теоријског и практичног аспекта анализира процес образовања и обраде корпуса текстова југословенских рок песама од 1967-2003. За преузимање грађе и XML анотирање коришћене су библиотеке lyricsmaster и yattag у језику Python. Корпус је прошао фазу препроцесирања, а XSL трансформацијом генерисани су основни статистички подаци. У апликацијама Слово Мајстор и LeXimir спроведена је рестаурација дијакритика (а у другој апликацији и фреквенцијска анализа). Проналажење друштвено-политичких тема вршено је у софтверу Unitex, док су преовлађујуће теме визуализоване у TreeCloud апликацији.

КЉУЧНЕ РЕЧИ: корпусна лингвистика, југословенски рокенрол, гребање веба, обрада природних језика, копање по тексту.

РАД ПРИМЉЕН: 15. април 2019.
РАД ПРИХВАЋЕН: 19. јуни 2019.

1. Увод

1.1 Феномен југословенског рокенрола

Музички аналитичари, социолози и антрополози се једногласно слажу у оцени да је југословенски рокенрол (познатији и као *Yu рок*)

[1] Рад је проистекао из мастер тезе „Креирање и анализа корпуса текстова југословенских рок песама у периоду 1945-2003“ која је одбрањена на Универзитету у Београду дана 18.3.2019. Израдом тезе је руководила проф. др Ранка Станковић која је допринела формулацији теме, уз напомену да је 1945. година у овом научном раду замењена 1967. годином.

оставио дубоког трага на просторима бивше Југославије. У питању је музички стил који је зачет 1961. године, са појавом група Урагани, Бијеле Стријеле и Силуете, а током исте деценије основани су и састави Црвени Коњаници (1962), Златни Дечаци (1962) и Корни Група (1968) (Janjatović, 1998). Yu рок се развијао упоредо са процватом британске *beat* сцене (Раковић, 2011), коју су прославили бендови The Beatles и The Rolling Stones (Cooper and Cooper, 1993). Крајем педесетих и почетком шездесетих година, рокенрол се у Југославији поистовећивао са рокенрол и твист плесом, а не са музичким жанром (Раковић, 2011).

Од 1963. године, рокенрол у СФРЈ стиче статус жанра, који се од тада изводи првенствено помоћу електричних гитара, електричних бас гитара и бубњева (Раковић, 2011). Класични рок и рокабили су представљали веома популарне музичке форме међу Југословенима, што потврђују и обраде песама Елвиса Преслија, Бадија Холија, Чака Берија и др. (Арсенијевић et al., 2016). Седамдесете године су донеле уплив хипи покрета, уз додатно жанровско раслојавање и појаву хард-рока, прогресивног рока, арт-рока и других поджанрова. Крај седамдесетих и почетак осамдесетих година обележили су панк и нови талас, по узору на англоамеричке изворне варијанте (Арсенијевић et al., 2016). Заправо, југо-рок је пронашао своје место у друштву које је са одушевљењем прихватило производе западне културе, како посредством домаћих часописа (нпр. *Rock* и *Džuboks*) и страних радио емисија емитованих на Радију Луксембург, војним и пиратским станицама, тако и у виду филмова из земаља Запада. Такође су се пратили и западњачки модни трендови који су диктирали, између осталог, ношење дуге косе код младића или мини-сукања код девојака (Раковић, 2018).

Ипак, оно по чему се Yu рок посебно истицао били су слободарски текстови, посебно карактеристични за панк и новоталасну музику. Наиме, садржај текстова је неретко био политички ангажован, ироничан или, пак, вулгаран, па је самим тим и третиран као неподобан за тадашње југословенске друштвене прилике (Гајић, 2018). Доказ да није реч о изолованој појави представља и систематизација случајева (ауто)цензуре у Yu року.[2] Са друге стране, бројне песме тога доба опевале су отаџбину, док су у некима биле присутне и партизанске или бригадирске конотације (Божиловић, 2016). Божиловић додаје да сржЊ Yu рока чини бунт и борба рок извођача за демократију, која је преточена у неконвенционалне, једноставне, непосредне и

[2] Видети веб страну Balkanrock.

импровизоване музичке форме и текстове. Међутим, Yu рок, попут свог западњачког еквивалента, изнедрио је и уметнике који су личне животне ставове сублимирали у текстове песама са лирским, филозофским и интроспективним тоном (нпр. групе Азра, Идоли или Екатарина Велика).

1.2 Сродна истраживања

Према сазнањима аутора, студије југословенског рокенрола су се до сада заснивале превасходно на теоретским разматрањима, без примене рачунарских технологија (неки од таквих радова наведени су у одељку 1.1). Zörnig et al. (2016) спровели су једино истраживање које се бавило квантитативном анализом текстова песама из ере југословенског рокенрола. Део корпуса у наведеном раду чине текстови песама састава Рибља Чорба и његовог фронтмена, Боре Ђорђевића. У истом раду изложен је поступак рачунања фреквенција речи и лексичког варијабилитета коришћењем првенствено метрика *relative repeat rate* и *h-point*, што је омогућило класификацију корпусних текстова помоћу мултиваријационе анализе.

Са друге стране, аутоматско преузимање текстова песама са веба, њихова електронска обрада и анализа добијају све више пажње, о чему сведочи и велики број пројеката и радова. Постоји велики број корпуса текстова страних песама, нпр. The Million Song Dataset[3] (Bertin-Mahieux et al., 2011), који се могу користити у истраживањима рачунарске анализе текста. До сада је вршено генерисање листа најучесталијих речи у анотираним корпусима текстова поп (Kreyer and Mukherjee, 2007) и рок песама (Falk, 2013), а спроведено је и моделовање тема у либрету пекиншке опере (Zhang et al., 2017) и корпусу текстова нумера прикупљених са сајта SongMeanings (Lukic, s.d.).

Када је реч о текстовима песама из домена рока или сродних жанрова, рачунарска анализа њиховог садржаја представља перспективну и изузетно развијену научну праксу. Неки занимљиви резултати које су ове технике произвеле јесу екстракција структуре текстова песама групе The Beatles (Mahedero et al., 2005) или смештање текстова песама Пола Макартнија и Џона Ленона на график *емоционалног сата* (енгл. emotion clock, у раду (Whissell, 1996)). Истраживање које су спровели Petrie et al. (2008) је још један пример

[3] The Million Song Dataset (на вебу).

интересовања за рачунарску обраду текстова Битлса. Ово истраживање бави се променом доминантног расположења у текстовима њихових песама, а коришћењем стилометријске анализе открива стилистичке сличности и разлике између Ленона, Макартнија и Џорџа Харисона, као текстописаца.

Falk (2013) доноси резултате дијахронијске анализе лингвистичких особености у корпусу текстова страних рок песама од 1950. закључно са 1999. годином, на основу најфреквентнијих речи коришћених у свакој деценији. Haslam (2017) прати еволуцију тематских мотива у песмама кантаутора Леонарда Коена, што је визуализовано у виду *облака речи* (енгл. word cloud). Са становишта корпусне лингвистике, Taina et al. (2014) истражују језичке чиниоце по којима се разликују текстови песама из различитих хеви-метал поџанрова. Алати који се користе у психометријским истраживањима ради рачунања повезаности текста (Coh-Metrix)[4] и откривања емоционалних, когнитивних и структуралних особина присутних у тексту (Linguistic Inquiry and Word Count)[5] примењени су у покушају упоредне анализе текстова у рок, фолк, кантри, панк и гранџ песмама (Lightman et al., 2007). Циљ тог истраживања јесте откривање разлика у стилу писања извођача који су извршили самоубиство, и оних који нису били суицидни. Тематски удаљенији, али методолошки близак овом истраживању је рад који се бави техникама аутоматског проналажења полилексемских јединица у текстовима древних религиозних хинду поема помоћу *локалних граматика* (енгл. local grammars) у софтверу Unitex[6] (Stein, 2012).

По свему судећи, анализа текстова југословенских рок песама са становишта рачунарске лингвистике чини се као недовољно развијена област истраживања. Из тог разлога, овај рад настоји да сагледа Yu рок песме кроз призму интердисциплинарности, како би се постојећа тумачења жанра поткрепила резултатима произведеним помоћу рачунарских алата за обраду корпуса.

1.3 Теоријско-методолошки оквир рада

Једно од тежишта рада је на примени техника корпусне лингвистике. Корпусна лингвистика представља развијену научну методологију, док је многи стручњаци сматрају и дисциплином, теоријом, парадигмом и

[4] Coh-Metrix (на вебу).
[5] Linguistic Inquiry and Word Count (скр. LIWC) (на вебу).
[6] Unitex/GramLab (на вебу).

алатом (Taylor, 2008). Овај методолошки приступ бави се руковањем структурираним, машински читљивим и наменски бираним текстовима, који представљају основу за анализу различитих аспеката језика и његове употребе. У општем контексту корпусне лингвистике, „наменски бирани текстови" подразумевају колекцију одређених текстуалних јединица које су узорковане ради остваривања репрезентативности корпуса. Учесталост коришћења одређене речи или фразе, њихово проналажење у контексту кроз генерисање *конкорданце* (енгл. concordance), или екстракција метаподатака (нпр. имена аутора текста, датума објављивања, језика на коме је текст написан и сл.) из анотираног корпуса само су неки од могућих задатака корпусне лингвистике (McEnery and Hardie, 2012).

Корпусна грађа текстова Yu рок песама прикупљена је применом технике *гребања веба* (енгл. web scraping).[7] Алгоритми за аутоматско прикупљање, препроцесирање и анотирање корпуса у складу са XML синтаксом имплементирани су у програмском језику Python, коришћењем библиотека `lyricsmaster`, `xml.sax.utils` и `yattag`. XSLT трансформација XML документа у XHTML формат омогућава преглед статистичких података о корпусу.[8] Аутоматска рестаурација дијакритика вршена је у апликацијама Слово Мајстор[9] и LeXimir.[10] Методе корпусне лингвистике и обраде природних језика примењене су ради фреквенцијске анализе токена и колокација коришћењем LeXimir-a, а у софтверу Unitex конструисан је *коначни аутомат* (енгл. Finite State Automation) за екстракцију друштвено-политичких и културолошких тема. Употребљен је и алат TreeCloud (Gambette and Véronis, 2009) за визуализацију преовлађујућих тема у корпусу, у оквиру технике *копања по тексту* (Кешељ and Шипка, 2008) или *ископавања из текста*[11] (енгл. text mining), која се заснива на откривању текстуалних образаца у неструктурираним подацима.

[7] Наведени енглески термин је стандардизован у страној литератури, за разлику од чланака и радова објављених на овим просторима, у којима се срећу бројни покушаји јединственог терминолошког одређивања (нпр. *налажење података на вебу, стругање података* или *гребање веба*, да наведемо само неке). Поменути термини потичу из веб чланака „Nalaženje podataka na internetu" и „SEO optimizacija kroz rečnik najbitnijih termina."

[8] Кодови су доступни на GitHub репозиторијуму аутора (на вебу).

[9] Слово Мајстор (на вебу).

[10] LeXimir (на вебу).

[11] Термин преузет из презентације проф. др Цветане Крстев (на вебу).

Рад се састоји од пет одељака. У уводном делу размотрене су основне одлике Yu рока, релевантни радови на тему рачунарске обраде текстова домаћих и страних песама, као и методе истраживања коришћене у овом раду. Одељак 2 посвећен је опису поступка аутоматског прикупљања грађе за корпус Yu рок песама. У одељку 3 представљен је поступак полуаутоматског[12] препроцесирања и анотирања корпуса. Конкретни резултати примене корпуса изложени су у одељку 4, док одељак 5 доноси закључне коментаре и смернице за будући рад.

2. Прикупљање грађе методом *гребања веба*

Уже подручје истраживања представљеног у овом раду јесте рачунарска израда и анализа корпуса текстова југословенских рок песама, насталих у доба двеју југословенских држава – Социјалистичке Федеративне Републике Југославије (1945-1992) и Савезне Републике Југославије (1992-2003). Што се тиче критеријума прикупљања грађе, корпус чине текстови нумера на некадашњем српскохрватском језику, који је стандардизован Бечким књижевним договором 1850. године, да би се са распадом Југославије 1991. разложио на српски, хрватски и босански језик (Hentschel, 2003). У складу са тим, претраживали смо текстове песама на поменутим трима језицима, док су текстови песама на другим језицима који су били у употреби у Југославији – македонском, словеначком и језицима мањина – изостављени.

Када је реч о методи гребања веба, она се односи на употребу одређене рачунарске алатке за директно преузимање дигитализованог садржаја. Суштина ове праксе може се описати као аутоматизација свеобухватног и релативно брзог екстраховања и складиштења података са веба. Гребање веба намеће се као далеко ефикасније решење у односу на методе ручног копирања и чувања сваке информационе јединице понаособ. Осим тога, наведена техника може представљати и једино могуће решење за екстракцију података у случају када су опције „copy/paste“ онемогућене у веб читачу.

[12] Под полуаутоматским поступком подразумева се комбинација техника аутоматског и ручног препроцесирања и анотирања корпуса.

2.1 Опис извора података: сајт LyricWiki

LyricWiki[13] представља комерцијални музички веб сајт који складишти текстове песама домаћих и страних извођача. Сајт је претражив према имену извођача, албума, песама, жанрова и продуцентских кућа. Доступне су и разне листе, попут листа најпопуларнијих албума из одређене године према мишљењу уредника угледних музичких часописа, и листа албума са музиком из филмова, као и обиље других текстуалних садржаја у области музике. Како је наведено у опису овог сајта, LyricWiki је јавно доступан и има дозволу за објављивање поузданих текстова песама.

База података сајта складишти више од 2.054.289 текстова (податак од 15.06.2019).[14] Будући да LyricWiki садржи релативно велики број песама на српском и некадашњем српскохрватском језику, у почетној фази истраживања (док је сајт још увек био отворен за уређивање) преузимани су текстови у извођењу одабраних музичара из бивше Југославије.[15] Разлог више за одабир овог сајта за гребање текстуалних података јесте чињеница да се са њега могу преузимати текстови преко наменског API-ја, тј. модула `LyricWiki` библиотеке `lyricsmaster` у језику Python, о којој ће бити више речи у наредном одељку.

2.2 Функционалност библиотеке lyricsmaster у језику Python

Аутоматско преузимање текстова песама са интернета представља ефикасну и релативно често коришћену методу која претходи електронској анализи корпуса. Један од репрезентативних примера такве праксе јесте употреба програмске библиотеке `lyricsmaster`, која је доступна на веб страни PyPI репозиторијума[16] програмских пакета за рад у језику Python. Помоћу наведеног API-ја могу се прикупити текстови песама са неких од најпопуларних музичких сајтова (у даљем тексту: *провајдера*), као што су AZLyrics, Genius и др. На истој страни демонстрирано је коришћење наведеног алата да би се директно преузели и сачували текстови песама америчког репера Тупака Шакура, који су доступни на сајту LyricWiki. Ради анонимизације IP адресе корисника, постоји опција гребања сајта преко Tor Proxy Server-a.[17]

[13] LyricWiki (на вебу).
[14] Statistics (на вебу).
[15] Списак извођача почиње од веб стране Language/Serbian.
[16] lyricsmaster 2.8.1 (на вебу).
[17] Tor Project (на вебу).

Имплементација поменуте библиотеке резултовала је аутоматским екстраховањем текстуалног садржаја са различитих страница сајта LyricWiki, при чему је свака од тих страна посвећена неком од популарних југословенских извођача чији текстови песама су преузимани. Наиме, дефинисаном функцијом за гребање веба и коришћењем методе `get_lyrics()` класе `lyricsmaster.providers.LyricWiki(tor_controller=None)`[18] проналажене су жељене странице, са којих су прикупљани једино текстови песама, а не и други садржаји на истим странама (нпр. подаци у заглављу или на бочној траци веб стране). Конструисан је алгоритам који је са тих страна преузимао садржаје текстова према именима извођача, и то на следећи начин:

1. Најпре је за провајдера текстова одабран сајт LyricWiki;
2. Затим је променљивом `izvodjaci` дефинисана листа од тридесет извођача чији се текстови песама прикупљају. Имена извођача референцирамо тачно онако како су наведена на сајту (нпр. уместо `‘Yu-Grupa‘`, `‘yu grupa‘`, и сл., једино је дозвољен унос ниске `‘YU Grupa‘`). Ово се односи и на случајеве у којима два извођача деле исти назив (нпр. српска и финска група Negative). Уредници LyricWiki-ja су направили разлику између наведених група тако што су српском саставу додали одговарајућу ознаку „RS” за земљу порекла, како би се знало да је реч о бенду из Републике Србије, па је у код за гребање веба унет модификован назив `Negative (RS)`;[19]
3. Креирана је функција `korpus()` чији аргументи су елементи наведене листе извођача, и за сваког од њих је покушано преузимање текстова помоћу методе `get_lyrics()`;
4. Потом су формирани објекти `discography`, `album` и `song`, док су увођењем параметара `title` и `lyrics` добијени наслови албума (`album.title`), песама (`song.title`) и сами текстови песама (`song.lyrics`);
5. Након тога, преузети корпусни материјал чуван је на локалној меморији рачунара методом `save()`, која је примењена на објекте `discography`, `album` и `song`.[20] Подразумевана апсолутна путања јесте `/{user}/Documents/LyricsMaster/`, а сами текстови су сачувани у директоријуму `/izvodjac/album/` у формату `pesma.txt`.

[18] `get_lyrics` је главна метода наведене класе (из документације на вебу).
[19] Финска група није добила никакву ознаку, већ само назив „Negative“.
[20] `save()` је метода трију класа: `lyricsmaster.models.Discography`, `lyricsmaster.models.Album` и `lyricsmaster.models.Song`.

На LyricWiki странама са текстовима песама одређених извођача (нпр. YU групе) већина наслова била је форматирана у виду хипервеза које упућују на стране са доступним текстовима. Ипак, за неколицину наслова уопште нису били креирани текстуални садржаји (нпр. за песму „Čovek i Marsovac“), упркос формалном постојању наслова, што је изазивало обустављање рада алгоритма. Ради руковања изузецима, увели смо исказ `try` за детекцију грешке на оном месту где је уочен проблем. Исказу смо придодали одредбу `except` и исказ `continue` који отклањају грешке и омогућавају програму да настави са даљим радом. Наведене грешке сумиране су у табели 1:

Песма/Извођач	**Грешка**	**Узрок**	**Решење**
„Čovek i Marsovac“	`AttributeError`	Неактиван линк	`try-except-continue`
`‘Yu-Grupa‘`	`TypeError`	Погрешан назив	`‘YU Grupa‘`

Табела 1: Проблематични случајеви у току прикупљања текстова.

2.3 Ограничења lyricsmaster-a

За одређене саставе и соло извођаче (Рибљу Чорбу, Екатарину Велику, Азру, Филм, Џибонија, Ђорђа Балашевића и др.) није било могуће прикупити текстове помоћу `lyricsmaster`-а, било због онемогућеног приступа подацима или услед непостојања текстова песама за датог извођача на сајту LyricWiki. Такође, у фази прикупљања текстова дошло се до закључка да би корпус требало да садржи текстове песама женских извођача, али за одређене женске извођаче чији се музички стил може окарактерисати као „рок“ (Калиопи, Слађана Милошевић и сл.) није било могуће аутоматски преузети текстове са истог сајта. Са друге стране, неки извођачи се уопште нису налазили у бази података, а самим тим ни њихови текстови (као што је то случај са Мајом Оџаклијевском).

Како бисмо надоместили овај недостатак, применили смо алтернативне критеријуме за накнадни одабир извођача тако што смо проширили жанровски опсег који ће бити покривен нашом грађом. Прецизније, укључили смо у корпус и поп извођаче (нпр. Нину Бадрић и

Зану) у чијим музичким аранжманима се запажају утицаји рок музике. Поред тога, додали смо Мадам Пиано[21] и Горана Карана.[22] Корпус представљен у овом раду састоји се од песама извођача наведених у табели 2:

Бајага	Електрични Оргазам	Неверне Бебе
Бајага и Инструктори	Забрањено Пушење	Негатив
Беби Дол	Зана	Нина Бадрић
Бијело Дугме	Идоли	Октобар 1864
Ван Гог	Индекси	Партибрејкерс
Галија	Јосипа Лисац	Прљаво Казалиште
Горан Бреговић	YU Група	Рани Мраз
Горан Каран	Кербер	Смак
Дивље Јагоде	Мадам Пиано	Хари Мата Хари
Дино Мерлин	Мирзино Јато	Хаустор

Табела 2: Списак 30 извођача у корпусу.

Најстарији албум у корпусу јесте *Наше доба* (1967) групе Индекси; са друге стране, најскорије објављени албуми заступљени у истој грађи јесу албуми *Од неба до неба* групе Дивље Јагоде, односно *Collection* Нине Бадрић (оба издата 2003).

2.4 Формирање стабла директоријума за смештање грађе

Извршавањем кода за гребање веба формирана је хијерархијски организована структура података, односно стабло директоријума са својим кореном и гранама. Тако добијену дрволику организацију корпуса можемо представити у виду извода са командне линије

[21] Стручњак за Yu рок, Петар Јањатовић, уврстио је Мадам Пиано у своју енциклопедију југословенског рока (Janjatović, 1998), због значаја који је ова представница џез и етно-звука имала за развој југословенске музичке сцене.

[22] Горан Каран такође није репрезентативан пример рок извођача, али не одступа превише од задатог оквира. Иако је Каран стекао популарност извођењем поп песама „са далматинским призвуком,“ он је започео каријеру као рок певач (податак преузет из његове биографије на вебу).

оперативног система, након покретања кода у програмском језику Bash:[23]

```
alias tree="find . -print | sed -e 's;[^/]*/;|____;g;s;____|; |;g'"
```

На слици 1 дат је парцијални приказ наведене структуре који је преузет са апликације Terminal на оперативном систему macOS, након што је претходно извршена навигација до директоријума `LyricsMaster`, где је смештен корпус текстова:

Слика 1: Стабло директоријума „LyricsMaster“ са командне линије.

Као што се може закључити на основу слике 1, корени директоријум носи подразумевани назив „LyricsMaster”, унутар којег су смештени директоријуми са називима извођачâ (нпр. Мирзино Јато или YU Група). Затим се за сваког извођача наводе називи албума (*Шећер и мед, YU Група, Рим 1994*, итд.), у којима су садржани текстови песама у .txt формату, уједно и крајњи чворови у структури дрвета (нпр. „Апсолутно твој”, „Црни лептир” итд.). Међу наведеним подацима формално се јавља и датотека „.DS_Store”, која није ни од каквог значаја за даљу обраду и анализу корпуса, и која ће стога у каснијим фазама бити елиминисана. Након процедуре гребања сајта, прикупљени садржаји текстова у .txt формату, који су ускладиштени у засебним директоријумима, у наредној

[23] По узору на Bash код (на вебу).

етапи припреме корпуса обједињени су у јединствену XML датотеку и анотирани.

3. Препроцесирање корпуса

3.1 DTD спецификација

Пре аутоматског генерисања корпуса у XML формату, састављена је *декларација типа документа* (скр. DTD), која спецификује логичку структуру XML документа, на коју се он реферише, и у односу на коју се проверава његова валидност. Овај корак је такође важан јер представља предуслов за даљу обраду XML документа, попут генерисања основних статистичких података из њега у прегледном XHTML формату коришћењем XSLT процесора, о чему ће бити речи у одељку 4.1. У валидном XML документу корпуса, примера ради, DTD декларише да корени елемент `<exYuPesme>` може имати више аутора, али да је његова вредност атрибута фиксирана празна ниска.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="tabele-css-classes.xsl"?>
<!DOCTYPE exYuPesme [
<!ELEMENT exYuPesme (autor)+>
<!ATTLIST exYuPesme
xmlns CDATA #FIXED ''> ... ]
```

Слика 1: Декларација типа документа – одломак.

3.2 Проналажење свих текстова коришћењем модула os

Први корак у процесу генерисања XML документа представља примена модула `os` са методом `listdir()`, из стандардне програмске библиотеке језика Python. Када се метода позове из кореног директоријума „LyricsMaster,“ она враћа листу директоријума, тј. имена извођача. Да би се прикупили називи поддиректоријума (наслови албума датог извођача) и датотека (називи песама на тим албумима), било је потребно форматирати ниске са апсолутним путањама помоћу `format()` функције. Крајњи делови путања, тј. све оно што се налази иза последње косе црте – *основни назив* (енгл. base name), представља вредности тражених категорија. У том поступку коришћене су две функције `os.path` модула: помоћу функције `join()`, елементи листе директоријума и датотека спојени су са крајевима текућих путања

директоријума, како би се приступило тим елементима. Друга функција, `isfile()`, комбинована је са исказима контроле тока `if` и `if not` како би се исправно утврдило да ли се генерише листа директоријума или датотека. Илустрације ради, у наставку су дате путање на основу којих можемо екстраховати име групе Смак, наслов њиховог албума *Црна дама* и песму са тог албума – „Даире“:

```
../../LyricsMaster/Smak
../../LyricsMaster/Smak/Crna-Dama
../../LyricsMaster/Smak/Crna-Dama/Daire.txt
```

Затим су учитани сами текстови песама помоћу функција `open()` и `read()`, при чему је задржана стиховна структура текстова која садржи нове редове коришћењем функције `split('\n')[:-1]`. Суштина целокупне описане процедуре јесте могућност правилног распоређивања података приликом анотирања датотеке у XML језику.

3.3 Елиминација сувишног садржаја

У корпусу је запажен известан број текстова песама написаних искључиво на страним језицима. Будући да је тема овог рада обрада корпуса текстова на српском, односно српскохрватском језику, извршено је ручно уклањање наведених нумера или, пак, читавих албума из корпусних директоријума. У овом одељку наводимо неке карактеристичне примере пречишћавања корпуса. Конкретно, избачене су песме на грчком, македонском, ромском, енглеском, португалском и пољском језику. Специфичан случај представља присуство вишејезичности у текстовима песама Горана Бреговића као соло извођача, који је и аутор песама «Κέρνα μας», са албума *Alkohol: šljivovica & champagne*; „7/8 & 11/8“, „Ederlezi“, „TV screen“, „Ausência“ (албум *Ederlezi*) и „To nie ptak“ (*Kayah & Bregović*). На албуму са музиком из филма *Arizona Dream* све песме су на енглеском језику, тако да је и он уклоњен.

Из корпуса су такође искључени текстови песама који су грешком приписани неком аутору. Примера ради, међу текстовима нумера на LyricWiki страни Нине Бадрић налази се и песма „Ubila si del mene“, која припада словеначком *boy* бенду Game Over. Крагујевачкој групи Смак такође су погрешно додељене две песме (“Muistoja” и “Myrsky”) истоимене групе из Финске. Поред тога, на албуму *Zašto ne volim sneg* групе Смак појављује се и неколико инструменталних нумера. Нису у

обзир узете ни песме за које LyricWiki још увек нема садржај (уместо којег се јављају три тачке), као у случају нумере „Ne mogu da kapiram“, бенда Партибрејкерс. Такође су се јавили и дупликати једног текста под различитим насловима. Примера ради, за групу Забрањено Пушење задржан је текст за песму „Počasna salva“ са тим насловом, док су дупликати са лажним насловима „Manijak“, „Vuk“ и „Ujka Sem“ брисани.

Нарочиту пажњу је требало обратити на одабир песама са концертних или компилацијских албума. Наиме, првобитно је замишљено да се такви директоријуми одмах одстрањују, јер је било очекивано да ће се у њима налазити нумере које већ постоје у корпусу. Ипак, уочено је присуство одређеног броја необјављених нумера на концертним албумима (нпр. „Na vrhovima prstiju“ са албума *Neka svemir čuje nemir* групе Бајага и Инструктори). Са албума највећих хитова Нине Бадрић (*Collection*) задржан је мали број јединствених песама. Још једну врсту дупликата представљале су и неке обраде песама, као за песму „Tako ti je mala moja kad ljubi Bosanac“ Бијелог Дугмета из 1975. године, коју је обрадила група Забрањено Пушење 1998. године.

3.4 Анотирање и препроцесирање корпуса у XML формату

`Yattag`[24] представља програмску библиотеку за рад у језику Python којом се аутоматски могу додавати HTML и XML етикете приликом структурирања докумената. Овај API аутоматски додаје отворене и затворене изломљене заграде, и сваки почетни таг праћен је затвореним тагом. На овај начин, анотатор може брже и лакше обележавати текстове, јер програм смањује могућност јављања синтакстичких грешака. `Yattag` кроз модул `indent` такође подржава и аутоматско увлачење према општој хијерархијској структури XML документа, при чему се величина увученог простора може модификовати. Будући да је поступак анотације требало спровести на великом корпусу, идеја је била искористити функционалности наведене библиотеке ради ефикаснијег обележавања у складу са дефинисаним правилима означавања садржаја текстова. Та правила подразумевала су укључивање XML етикета елемената и атрибута за описивање одговарајућих делова садржаја корпуса по следећем поступку:

– Корени елемент дефинисан је етикетом `<exYuPesme>`;

[24] Yattag (на вебу).

– Аутори, односно извођачи, дефинисани су елементом `<autor>`, чији су атрибути `ime`, `brojAlbuma`,[25] `pol`, `zanr`, `rodnoMesto` (текстописци се не бележе, јер се не наводе ни на сајту);
– Албуми су дефинисани елементом `<album>`, који има атрибуте `naziv`, `godina` и `izdavac`;
– Песме су дефинисане елементом `<pesma>`, који има атрибут `naslovPesme`, док је за стихове резервисана етикета елемента `<li>`.

Пошто метод `tag()` креира XML етикете, њему су као аргументи прослеђени само жељени називи етикета за означавање елемената и њихових атрибута. Нпр. албум као елемент има атрибуте за назив, годину објављивања и продуцентску кућу, што је дефинисано на следећи начин: `with tag('album', naziv=album, godina="", izdavac="")`:. Са друге стране, метод `text()` генерише текстуални садржај који није назив етикете. У циљу скраћивања кода, коришћена је `Doc` инстанца класе `yattag.Doc` и здружена метода `tagtext()`, која тој инстанци додаје садржај који наведена метода производи (називе етикета, атрибута и обичан текст). Након дефинисања свих неопходних параметара, примењен је `getvalue()` метод, како би се целокупан садржај претворио у велику ниску карактера. Овим путем текстови су смештени у нову XML датотеку.

Након пробног генерисања XML корпуса, у њему је уочен специјални карактер амперсенд (`&`), који је, по правилу, замењен карактером излазне секвенце (`&amp;`). То је учињено да се амперсенд не би третирао као почетак референце ентитета и како би се у потпуности испоштовали принципи правилног XML структурирања. Због тога је у код за анотирање корпуса уметнута функција `escape()` модула `xml.sax.saxutils`. Иако се у корпусу често јавља и симбол за апостроф (`'`), он у процесу аутоматске замене карактера није замењен својим XML еквивалентом `&apos;`, што није ометало генерисање исправно формираног XML документа. Ради веће прегледности, аутоматски су уклоњене и цртице у називима песама и албума које су остале приликом гребања веба (нпр. промена од „Da-Sam-Pekar“ у „Da Sam Pekar“).

Што се тиче ручног обележавања корпуса, елементу `<autor>` додате су вредности атрибута `zanr`: поп, рок и *world music*.[26] Осим жанра, додате су и вредности атрибута `rodnoMesto` за родно место извођача и `izdavac` за продуцентску кућу која је издала албум. Пример

[25] Подр. број албума обухваћених текстовима песама укључених у корпус.
[26] Према класификацији жанрова доступних на веб страни Discogs.

полуаутоматски обележеног текста поменуте песме са албума *Концерт код Хајдучке чесме* групе Бијело Дугме, може се видети на слици 2.

```
<autor ime="Bijelo Dugme" brojAlbuma="13" pol="Grupa" zanr="Rok" rodnoMesto="Sarajevo">
    <album naziv="Koncert kod hajdučke česme" godina="1977" izdavac="Jugoton">
        <pesma naslovPesme="Da Sam Pekar">
            <li>Da sam pekar, mala moja</li>
            <li>Ne znam bi l'□ me htjela</li>
            <li>Kad bi noću bila sama</li>
            <li>Zemičke bi jela</li>
```

Слика 2: Одломак анотираног корпуса у софтверу oXygen XML Editor.

3.5 Аутоматска рестаурација дијакритика – софтвер LeXimir

Нормализација корпуса и његова припрема за рачунарску анализу не представља нимало тривијалан задатак, а, по правилу, доприноси генерисању информативнијих резултата у поређењу са рачунарском анализом непрепроцесираног текста. Прикупљени текстови песама у Yu корпусу били су прилично неуједначени по питању употребе писма: већина текстова била је записана на латиници, а није мали број оних у којима специјална латинична слова (č, ć, ž, đ, š) нису садржала потребне дијакритичке знаке. Са друге стране, мањи број песама био је првобитно забележен на ћирилици. Почетна замисао била је да се целокупан корпус преслови на ћирилицу; од те идеје се одустало јер се у текстовима на српском језику појављују и одломци на страним језицима (нпр. “la musique c’est fantastiqu / prepare la revolution / et la femme est tres jolie / tre jolie comme un bonbon“).[27] Стога је одлучено да у поступку транслитерације текстови написани деградираном латиницом и ћирилицом буду аутоматски претворени у стандардну српску латиницу са дијакритичким знацима. Преслovљени текст корпуса у .txt формату коришћен је у поступку екстракције лексичких јединица из одељка 4.2.

Да би се одабрао најпогоднији алат, обављена је рестаурација дијакритика у апликацијама Слово Мајстор и LeXimir. Други алат користи електронске морфолошке речнике (са валидним облицима речи), локалне граматике и Корпус савременог српског језика (Krstev

[27] Према изворном тексту, са штампарским и правописним грешкама.

et al., 2018). Након тестирања обеју метода, дошло се до закључка да Слово Мајстор солидно решава проблем претварања деградиране у стандардизовану латиницу, уз напомену да је неким речима погрешно додао дијакритичке знаке, што је нарушило семантику речи. Тако је од глаголског облика „isecka“ погрешно створена реч „isečka“. Добра страна ове апликације јесте могућност накнадне исправке погрешно транслитерираних речи.

Са друге стране, LeXimir, који користи функционалности Unitex софтвера за обраду корпуса, исту реч је оставио у исправном облику. На слици 3 може се видети да су потенцијални кандидати за одабир адекватног облика реч „isečka“, што је флективни облик именице „исечак“ у генитиву једнине, и „isecka“, односно 3. л. јд. аориста глагола „исецкати“. Алгоритам за исправљање на основу левог и десног текстуалног контекста потенцијалне речи за исправљање врши *дезамбигуацију* (разрешавање вишезначности). Одабрана је друга опција, где заменица „га“ може да стоји уз глагол „исецка“, а не и уз флективни облик именице „исечка“, јер би друга конструкција била граматички неисправна. Ипак, ручна евалуација је неопходна и након примене описаног алата, јер у неким речима (као нпр. у речи „oci“ у контексту „Trljam oci sanjive“) нису додати дијакритички знаци тамо где је требало да постоје.

Potencijalni kandidati	Levi kontekst	Ispr.	Desni kontekst
*7(uže(72)_uze(61))	********	uze	********
*7(Isečka(5)_Isecka(1))	s I ujeo kuvara Kuvar uze, uze nož	Isecka	ga na dva, na tri komada Došli, do
*7(oči(818)_oci(24))	žurno Kafu pijem, nestajem Trljam	oci	sanjive Da mi ne bi zaspale Vrata

Слика 3: Рестаурација дијакритика у алату LeXimir.

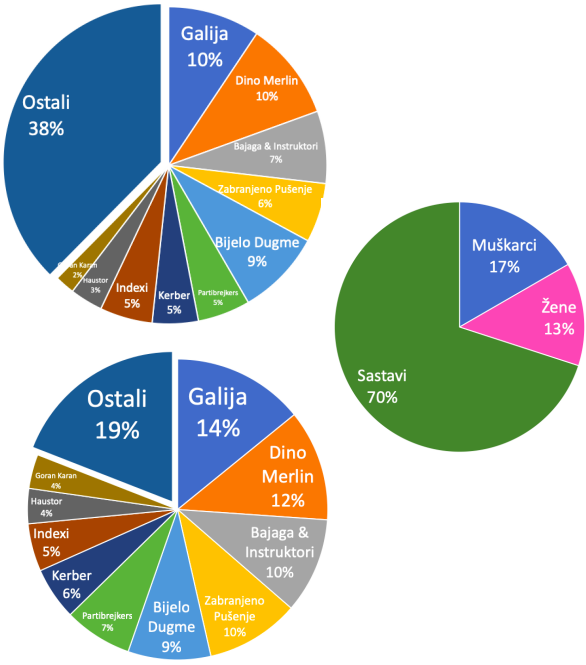
4. Рачунарска анализа корпуса

4.1 Статистички подаци о корпусу

DTD спецификација омогућила је постављање упитâ кроз навигацију над датим документом помоћу XSLT језика за трансформацију XML документа у друге формате. Нпр., следећи XSLT израз, за сваки елемент `<autor>` у документу, бира вредност атрибута `ime`, чиме се генерише листа свих извођача у корпусу:

```
<xsl:for-each select="//autor">
        <xsl:value-of select="@ime" />
</xsl:for-each>
```

Подаци из XML документа су трансформисани у XHTML табелу са слике 4, где се може видети укупан број аутора, песама и албума у корпусу. Кружни графикони са исте илустрације описују (у смеру казаљке на сату, одозго надоле): заступљеност извођача у корпусу према броју албума, број мушких, женских и групних извођача, као и удео по броју песама.



Statistika

Ukupno: 30	*Ukupno: 953*	*Ukupno: 147*
Autori	**Pesme**	**Albumi**
Galija	135	14
Dino Merlin	114	15
Bajaga Instruktori	98	11
Zabranjeno Pusenje	96	9
Bijelo Dugme	85	13
Partibrejkers	70	8
Kerber	54	7
Indexi	49	8
Haustor	36	5
Goran Karan	35	3
Prljavo Kazaliste	25	6
Negative	22	2
Hari Mata Hari	22	7
Smak	21	6
Nina Badric	20	5
Idoli	15	1
YU Grupa	14	7
Rani Mraz	13	2
Divlje Jagode	6	4
Madame Piano	5	1
Goran Bregovic	4	2
Van Gogh	2	2
Josipa Lisac	2	2
Bajaga	2	1
Zana	2	2
Oktobar 1864	2	1
Mirzino Jato	1	1
Neverne Bebe	1	1
Elektricni Orgazam	1	1
Bebi Dol	1	1

Слика 4: Статистички подаци о Yu корпусу.

У LeXimir-у су такође произведени основни статистички подаци о Yu корпусу. Након рестаурације дијакритика, корпус садржи 248.807

токена, 16964 јединствене лексичке јединице, 116.972 просте форме (16.909 различитих) и 268 бројева (10 различитих). LeXimir нуди подршку и за приказ флективних облика токена, лематизованих облика, врста речи, фреквенција речи и колокација. Резултати фреквенцијске анализе могу се филтрирати према врсти речи. На основу извезене .xlsx датотеке можемо утврдити нпр. које су најфреквентније именице или колокације чији је главни члан именица. Слика 5 у наставку представља парцијални приказ резултата који сведочи о присуству најучесталијих именичких токена, међу којима се јављају речи *dan, noć, ljubav, srce* итд. Слика 6 доноси преглед одређених колокација у вези са ратом (*svetski rat, vojnu muziku, ratne filmove*).

Oblik	Lema	POS	Freq.
dan	dan	N	299
Al	Al	N	231
noć	noć	N	228
do	do	N	224
ljubav	ljubav	N	201
srce	srce	N	196
život	život	N	195
put	put	N	184
meni	mena	N	164
meni	meni	N	164
kraj	kraj	N	155
noći	noć	N	147
biti	bit	N	132
bila	bilo	N	124
grad	grad	N	120

Слика 5: Токени.

svetski rat	svetski rat	N	3
novi svet	Novi svet	N	3
noćne ptice	noćna ptica	N	3
tam-tam	tam-tam	N	3
prošlog vremena	prošlo vreme	N	2
vojnu muziku	vojna muzika	N	2
železnička stanica	železnička stanica	N	2
crno grožđe	crno grožđe	N	2
sunčev zrak	sunčev zrak	N	2
malog medveda	Mali Medved	N	2
zlatne medalje	zlatna medalja	N	2
morske obale	morska obala	N	2
svetla budućnost	svetla budućnost	N	2
ratne filmove	ratni film	N	2

Слика 6: Колокације.

4.2 Екстракција тема из корпуса

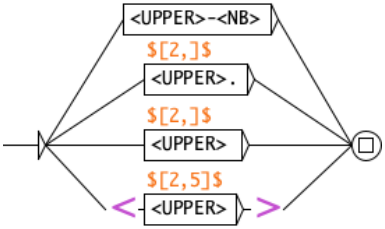
Познато је да су у текстовима песама југо-рока биле присутне друштвено-политичке и културолошке теме, те је у софтверу Unitex конструисан граф коначног аутомата за препознавање индикатора наведених тема. Помоћу регуларних израза и лексичких маски трагало се за лексичким обрасцима у виду скраћеница[28] састављених од:

– мајускуле, цртице и низа бројева (нпр. *B-52*);
– секвенце мајускула + тачка која се понавља најмање два пута (*K.P.*);

[28] Видети веб страну Скраћенице и објашњења

– секвенце мајускула + размак која се понавља најмање два пута (*ES EF ER JOT*);
– секвенце која садржи од две до пет мајускула (*CZ, CIA*).

Међу њима су се издвојили називи државних органа (нпр. *AVNOJ*), спортских клубова (*PFC*), назива телевизијских канала (*BBC*) итд. Граф са слике 7, осим скраћених назива представљених у резултату конкорданце на слици 8, такође препознаје и токене: *CZ, ES EF ER JOT, FK, TV, JRT, K.P., KGB, KK, MUP, O.K., PC, PFC, SUP* и *TAS*.

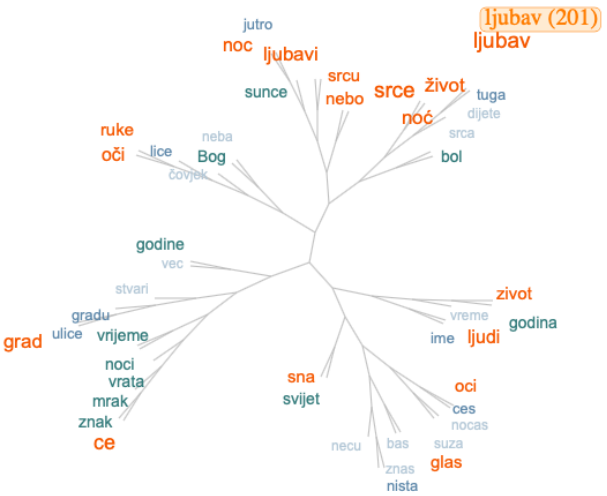


Слика 7: Граф за препознавање скраћеница.

Од istorijskog AVNOJ-a Do izbjegličkog
Od istorijskog AVNOJ-a Do izbjegličkog
Od istorijskog AVNOJ-a Do izbjegličkog
Od istorijskog AVNOJ-a Do izbjegličkog
nebu je pisalo B-52 I veliki buketi od
nebu je pisalo B-52 I veliki buketi od
šao Haše!{S} I BBC na mome radiju, Osl
šao Haše!{S} I BBC na mome radiju, Osl
niskim terenom BMW-a kešom plaćenog, o
KGB zaspala je CIA odmara se naša mili
, penjao se na CK i pevao pesnu protiv
iti stan Preko CNN-a gledao sam Mufu S

Слика 8: Одломак из конкорданце.

Да би се утврдило да ли су друштвено ангажоване теме статистички значајне у корпусу, генерисан је визуални приказ преовлађујућих тема у виду *облака дрвета* (енгл. tree cloud) коришћењем алата TreeCloud, уз напомену да је анализа вршена на нелематизованим облицима. Овај програм комбинује графичко представљање података у виду облака речи са дрволиком структуром, тако да се, осим фреквенције речи, пружа увид и у начин груписања токена, према њиховом растојању у овој структури (видети слику 9). У апликацију је уграђена и листа *стоп речи* (енгл. stop words) за српски језик, којој су, за потребе визуализације датог корпуса, придодате још неке речи. Креиран је визуални приказ у коме се издвајају кластери са значењем осећања, у виду речи *tuga, ljubav/-i, bol, srce/-a*. Од урбаних тема, присутне су речи *grad/-u, ulice*, делови тела су означени речима *ruke, lice, oči, srce/-a/-u*. Мотив времена је такође фреквентан у корпусу (*život, vr(ij)eme, godina/-e*). Према овом приказу, друштвено ангажоване теме нису у довољној мери заступљене у корпусу.



Слика 9: Облак дрвета за цео корпус.

5. Закључна разматрања

У раду је представљен пројекат електронског формирања и обраде текстова југословенских рок песама од 1967. до 2003. У језику Python извршено је аутоматско прикупљање садржаја помоћу библиотеке `lyricsmaster` са сајта LyricWiki, а препроцесирани корпус је и аутоматски анотиран по правилима XML синтаксе коришћењем `yattag` алата, уз ручно додавање вредности неких атрибута. Спроведена је и аутоматска рестаурација дијакритика. Приказана је и XSL трансформација корпуса у XHTML формат, уз проналажење друштвено-политичких и културолошких тема у софтверу Unitex и визуализацију преовлађујућих тема у програму TreeCloud.

Даљи рад укључивао би евалуацију корпуса ради исправљања правописних и штампарских грешака (нпр. *Dunav* уместо *dunav*, *Avale* уместо *Aavale* и сл.). Такође, у плану је примена електронских морфолошких речника именованих ентитета, у циљу екстракције властитих имена музичара, политичара, спортиста и осталих личности које су извршиле значајан утицај на југословенско друштво.

Литература

Bertin-Mahieux, Thierry, Daniel PW Ellis, Brian Whitman and Paul Lamere. “The million song dataset”. In *Proceedings of the 12th International Conference on Music Information Retrieval*, 2011, 591–596. Датум приступа: 19. октобар 2019, http://ismir2011.ismir.net/papers/OS6-1.pdf.

Cooper, Laura E. and B. L. Cooper. “The pendulum of cultural imperialism: Popular music interchanges between the United States and Britain, 1943-1967”. *Journal of Popular Culture* Vol. 27 no. 3(1993): 61. Датум приступа: 19. октобар 2019, https://sci-hub.tw/10.1111/j.0022-3840.1993.00061.x#.

Falk, Johanna. “We will rock you : A diachronic corpus-based analysis of linguistic features in rock lyrics”. 2013. Датум приступа: 19. октобар 2019, http://www.diva-portal.org/smash/get/diva2:605003/FULLTEXT02.pdf.

Gambette, Philippe and Jean Véronis. “Visualising a Text with a Tree Cloud”. *IFCS’09: International Federation of Classification Societies Conference* (2009): 561–569, 19. октобар 2019, https://hal-lirmm.ccsd.cnrs.fr/lirmm-00373643/file/2009GambetteVeronis.pdf

Haslam, Thomas J. “Mapping the Great Divide in the Lyrics of Leonard Cohen”. *Rupkatha Journal on Interdisciplinary Studies in Humanities* Vol. 9, no. 1(2017): 1–10. Датум приступа: 19. октобар 2019, http://rupkatha.com/V9/n1/v9n1s01.pdf

Hentschel, Elke. “The expression of gender in Serbian”. In *Gender across languages: The linguistic representation of women and men*, Vol. 3, 287–309. John Benjamins Publishing Company, 2003. Датум приступа: 19. октобар 2019, https://epdf.tips/gender-across-languages-volume-iii-the-linguistic-representation-of-women-and-me.html

Janjatović, Petar. *Ilustrovana YU rock enciklopedija: 1960-1997*. Geopoetika, 1998. Датум приступа: 19. октобар 2019, https://monoskop.org/images/c/ca/Janjatovic_Petar_Ilustrovana_YU_Rock_Enciklopedija_1960-1997.pdf

Kreyer, Rolf and Joybrato Mukherjee. “The style of pop song lyrics: A corpus-linguistic pilot study”. *Anglia-Zeitschrift für englische Philologie* Vol. 125, no. 1(2007): 31–58. Датум приступа: 19. октобар 2019, https://doi.org/10.1515/ANGL.2007.31

Krstev, Cvetana, Ranka Stanković and Duško Vitas. “Knowledge and Rule-Based Diacritic Restoration in Serbian”. *Proceedings*

of the Third International Conference Computational Linguistics in Bulgaria, 41–51. 2018. Датум приступа: 19. октобар 2019, https://www.researchgate.net/publication/328416358_Knowledge_and_Rule-Based_Diacritic_Restoration_in_Serbian

Lightman, Erin J., Philip M. McCarthy, David F. Dufty and Danielle S. McNamara. "Using computational text analysis tools to compare the lyrics of suicidal and non-suicidal songwriters". In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 29, 2007. Датум приступа: 19. октобар 2019, https://cloudfront.escholarship.org/dist/prd/content/qt0dh4553j/qt0dh4553j.pdf.

Lukic, Alen. "A Comparison of Topic Modeling Approaches for a Comprehensive Corpus of Song Lyrics", Tech report, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, s.d. Датум приступа: 19. октобар 2019, http://alenlukic.com/assets/docs/lyric_topic_modeling.pdf

Mahedero, Jose P.G., Álvaro Martínez, Pedro Cano, Markus Koppenberger and Fabien Gouyon. "Natural language processing of lyrics". In *Proceedings of the 13th annual ACM international conference on Multimedia*, 475–478. 2005. Датум приступа: 19. октобар 2019, https://www.researchgate.net/profile/Pedro_Cano5/publication/221573745_Natural_language_processing_of_lyrics/links/00b7d52826f623edfb000000.pdf

McEnery, Tony and Andrew Hardie. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, 2012. Датум приступа: 19. октобар 2019, http://gen.lib.rus.ec/book/index.php?md5=33c1c5b6d73ea816dfb2a034f73bb176

Petrie, Keith J., James W. Pennebaker and Borge Sivertsen, "Things We Said Today: A Linguistic Analysis of the Beatles", *Psychology of Aesthetics, Creativity, and the Arts* Vol. 2, no. 4(2008): 197. Датум приступа: 19. октобар 2019, https://www.uvm.edu/pdodds/files/papers/others/2008/petrie2008a.pdf.

Stein, Daniel. "Multi-Word Expressions in the Spanish Bhagavad Gita, Extracted with Local Grammars Based on Semantic Classes". In LREC '2012 Workshop: LRE-Rel, Language Resources and Evaluation for Religious Texts, 88–93. 2012. Датум приступа: 19. октобар 2019, https://www.academia.edu/26035335/Linguistic_and_Semantic_Annotation_in_Religious_Memento_mori_Literature

Taina, Jesse et al.. "Keywords in heavy metal lyrics: A Data-Driven Corpus Study into the Lyrics of Five Heavy Metal Subgenres", 2014. Датум

приступа: 19. октобар 2019, https://helda.helsinki.fi/bitstream/handle/10138/136524/keywords.pdf?sequence=1.

Taylor, Charlotte. “What is corpus linguistics? What the data says”. *ICAME journal* Vol. 32 (2008): 179–200. Датум приступа: 19. октобар 2019, http://sro.sussex.ac.uk/id/eprint/53389/1/what_is_corpus_linguistics.pdf

Whissell, Cynthia. “Traditional and Emotional Stylometric Analysis of the Songs of Beatles Paul McCartney and John Lennon”. *Computers and the Humanities* Vol. 30, no. 3 (1996): 257–265. Датум приступа: 19. октобар 2019, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.461.7171&rep=rep1&type=pdf.

Zhang, Shuo, Rafael Caro Repetto and Xavier Serra. “Understanding the expressive functions of jingju metrical patterns through lyrics text mining”. In *18th International Society for Music Information Retrieval Conference*, 397–403. 2017. Датум приступа: 19. октобар 2019, https://repositori.upf.edu/bitstream/handle/10230/32652/Zhang_ISMIR2017_unde.pdf?sequence=1&isAllowed=y

Zörnig, Peter, Emmerich Kelih and Ladislav Fuks. “Classification of Serbian texts based on lexical characteristics and multivariate statistical analysis”. *Glottotheory* Vol. 7, no. 1(2016): 41–66. Датум приступа: 19. октобар 2019, http://homepage.univie.ac.at/emmerich.kelih/wp-content/uploads/2016_Zoernig_Kelih_Fuks_www.pdf.

Арсенијевић, Александра, Милена Обрадовић and Михаило Шкорић. “Израда мултимедијалног документа „YU рок сцена””. *Инфотека – часопис за дигиталну хуманистику* Vol. 16, no. 1-2a(2016):113-129, датум приступа: 19. октобар 2019, https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2016.16.1_2.6_sr

Божиловић, Никола. “Култура сећања и југословенски рокенрол”. *Култура* no. 152(2016): 257–280. Датум приступа: 19. октобар 2019, https://scindeks-clanci.ceon.rs/data/pdf/0023-5164/2016/0023-51641652257B.pdf

Гајић, Златомир. “Рок поезија Бранимира Штулића и њени медијски одјеци”. PhD thesis, Универзитет у Новом Саду, Филозофски факултет, 2018. Датум приступа: 19. октобар 2019, http://nardus.mpn.gov.rs/bitstream/handle/123456789/9926/Disertacija17602.pdf?sequence=1&isAllowed=y

Кешељ, Владо and Данко Шипка. “Приступ изградњи стемера и лематизатора за језике с богатом флексијом и оскудним ресурсима заснован на обухватању суфикса”. *ИНФОтека – часопис за*

библиотекарство и информатику Vol. 9, no. 1-2(2008): 21–31. Датум приступа: 19. октобар 2019, http://infoteka.bg.ac.rs/pdf/Srp/2008/04%20Vlado-Danko_Stemeri.pdf

Раковић, Александар. "Бит мода, рокенрол и генерацијски сукоб у Југославији 1965-1967". *Етноантрополошки проблеми* Vol. 6, no. 3(2011): 745–762. Датум приступа: 19. октобар 2019, https://www.eap-iea.org/index.php/eap/article/view/604/594

Раковић, Александар. "Рокенрол у Социјалистичкој Југославији: од забаве градске омладине до националне културе". In *Сан о граду : зборник радова*, Андрићев институт, 427–439. 2018. Датум приступа: 19. октобар 2019, http://doi.fil.bg.ac.rs/pdf/eb_book/2018/ai_san_o_gradu/ai_san_o_gradu-2018-ch18.pdf

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 TXM 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 (Pincemin, 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, TXM (Heiden et al., 2010; Heiden, 2010) је програм отвореног кода[1], веома коришћен у истраживањима различитих области друштвено-хуманистичких наука (историја, књижевност, географија, лингвистика, социологија, политичке науке). Графичко корисничко окружење TXM-а заснива се на коришћењу CQP[2] (енг. *Corpus Query Processor*) претраживача и R[3] статистичког пакета. TXM 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 XML[4] (енг. *eXtensible Markup Language*) докумената, који могу бити кодирани у складу са TEI[5] (енг. *Text Encoding Initiative*) упутствима (XML ili 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-TXM 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 TXM-а изграђује облик текста у HTML формату, на основу ког је у сваком кораку анализе могућ повратак тексту.

Иако TXM 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 TXM-а, резултати примене алата обраде природних језика виђени су као анотације додате XML-TEI репрезентацији текста.

Осим изградње XML-TXM формата, врши се конверзија и генерисање CWB[6] (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).

TXM 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 извезе у табеларном или графичком облику ради даљег истраживања и уређивања помоћу неког другог алата.

Циљ овога рада је представљање текуће верзије srpELTeC корпуса[7] и илустрација могућности текстометријског приступа и примене TXM алата за анализу и визуелни приказ резултата. Спроведена 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 сканирање 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 речницима (Krstev, 2008). За ручну корекцију преосталих грешака и означавање текстова основним структурним јединицама ангажован је велики број волонтера[9]. У овој фази вршена је и припрема метаподатака који ће се употребити за касniју 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 TXM 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 патријахалном 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	Хаџи Ђера	1904	65.554
Јанковић, Милица	Калуђер из Русије*	1919	8.279
Комарчић, Лазар	Драгоцена 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 srpELTeC корпус због дужине

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

Текстови srpELTeC корпуса сачувани су у XML формату, у складу са TEI смерницама, који је веома користан за каснију анализу. Заглавље TEI 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 TEI 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 TXM 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 TXM окружењу коришћени за поделу корпуса на партиције, креирање поткорпуса, као и за претраживање текстова.

Ради анализе srpELTeC корпуса, колекција текстова у XML формату увезена је у TXM окружење помоћу XML-TEI Zero + CSV модула. Етикетирање текстова srpELTeC 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 TXM 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 TXM 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 TXM 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:

$$Prob_{specif}(card\{A \in V | A \in p\} = f) = \frac{C_F^f \times C_{T-F}^{t-f}}{C_T^t}, \text{ где је}$$
$$C_n^k = \frac{n!}{k!(n-k)!}$$
$$n! = 1 \times 2 \times 3 \times ... \times (n-1) \times n$$

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

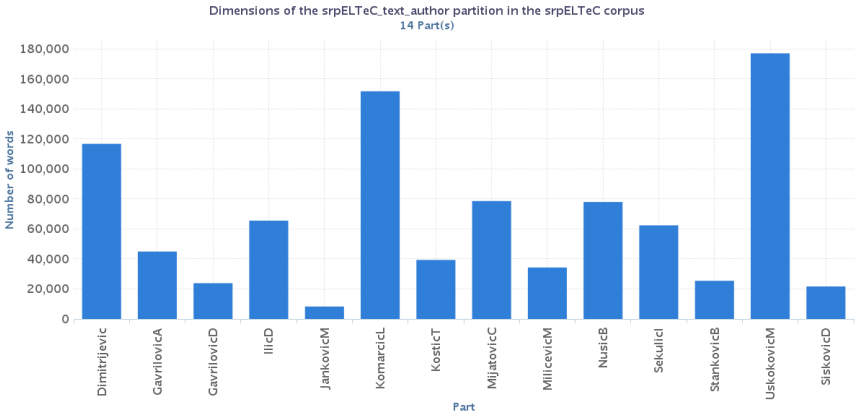
омогућава боље разумевање података и поједностављује њихову интерпретацију.

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).



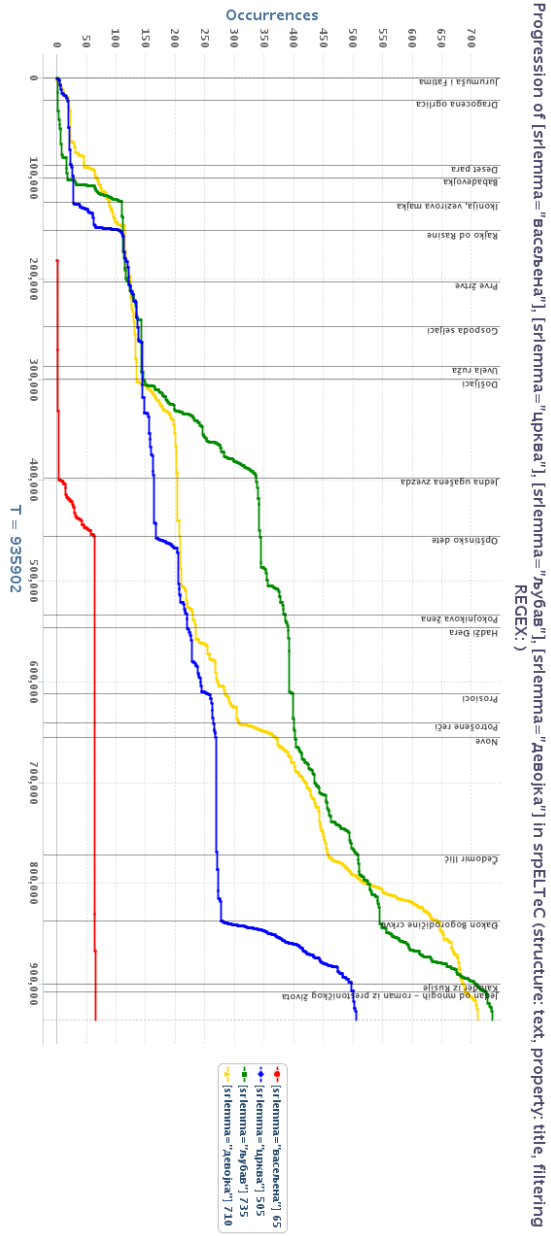
Слика 1. Димензије партиција srpELTeC корпуса креираних на основу ауторстава

.

На основу података о фреквентности TXM 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 *Ђакон 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 *Ђакон Богородичине цркве*.

3.2 Индекс специфичности

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

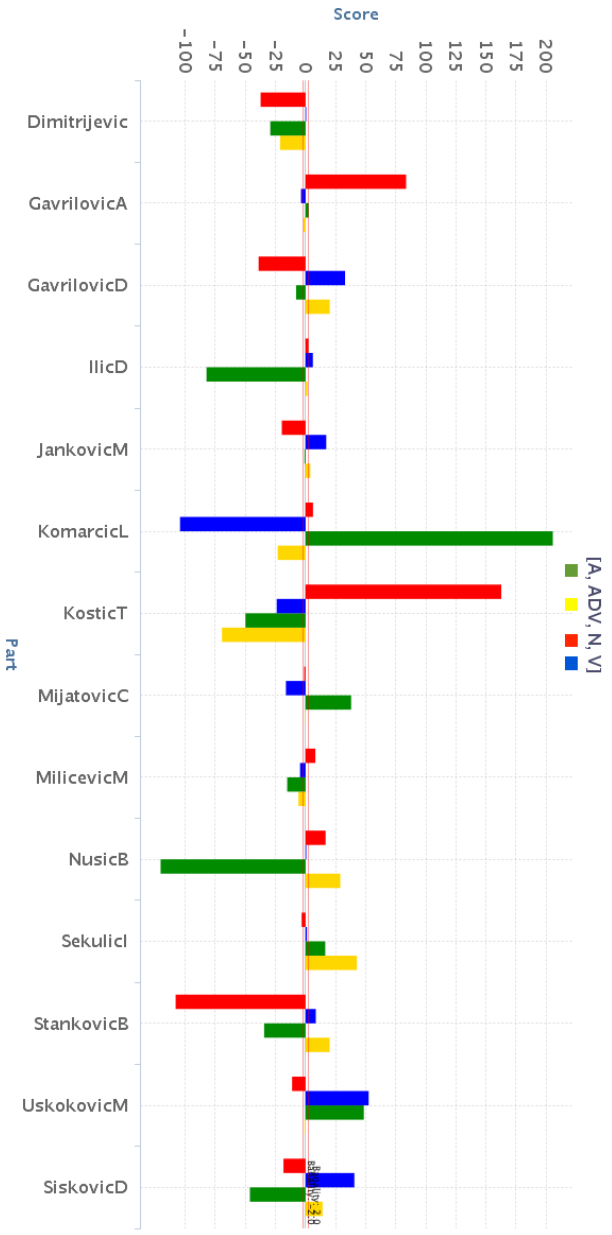


Слика 2. Графички приказ употребе појединих лема у текстовима srpELTeC корпуса

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 srpELTeC 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

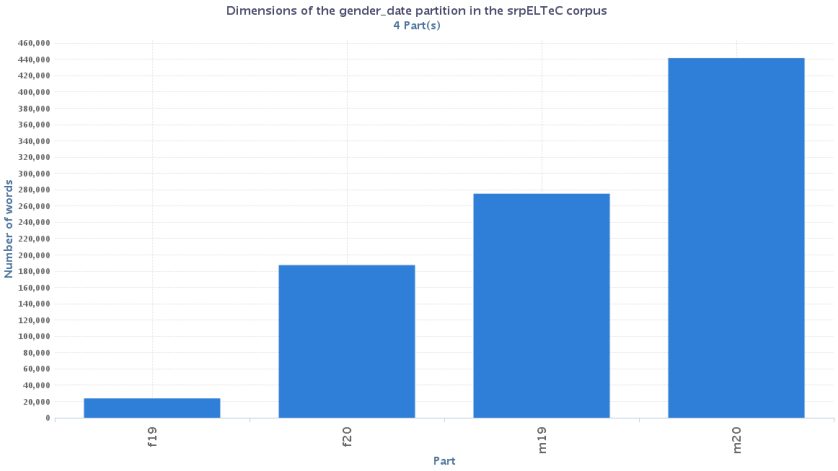
Аутор	f_N	S_N	f_V	S_V	f_A	S_A	f_{ADV}	S_{ADV}
Uskokovic M	35331	-11.2	35446	52.5	13502	48.4	9577	-0.8
Komarcic L	31869	6.2	25443	-104.4	13193	205.6	7481	-23.1
Dimitrijevic J	22330	-37.4	21938	0.5	7066	-29.3	5687	-21.2
Nusic B	16926	16.6	14675	0.6	3812	-120.5	4952	28.9
Mijatovic C	15985	-1.2	13862	-16.3	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 корусу (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).

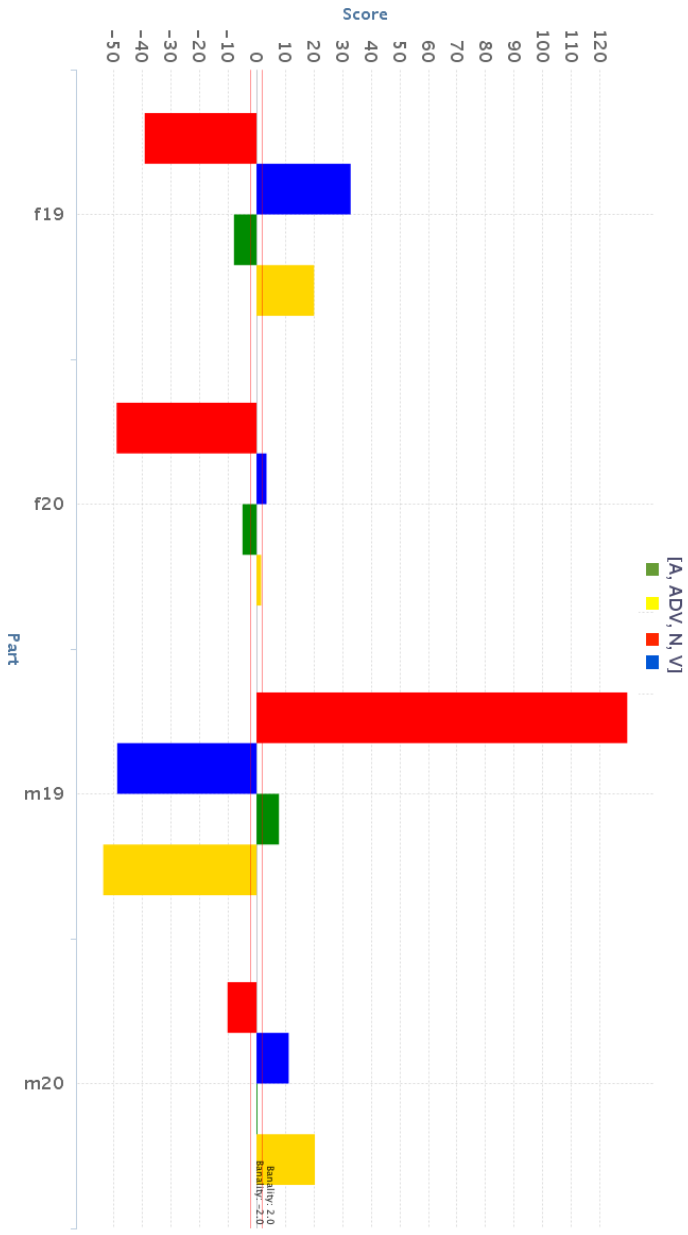
На основу података о учесталости употребе врста речи, а према χ^2 расподели, спроведена је факторска анализа међузависности, представљена у дводимензионалном графичком облику (слика 6). На добијеној факторској мапи је уочљиво да су глаголи и прилози чешће коришћени у партицијама *f19* и *m20*, па се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



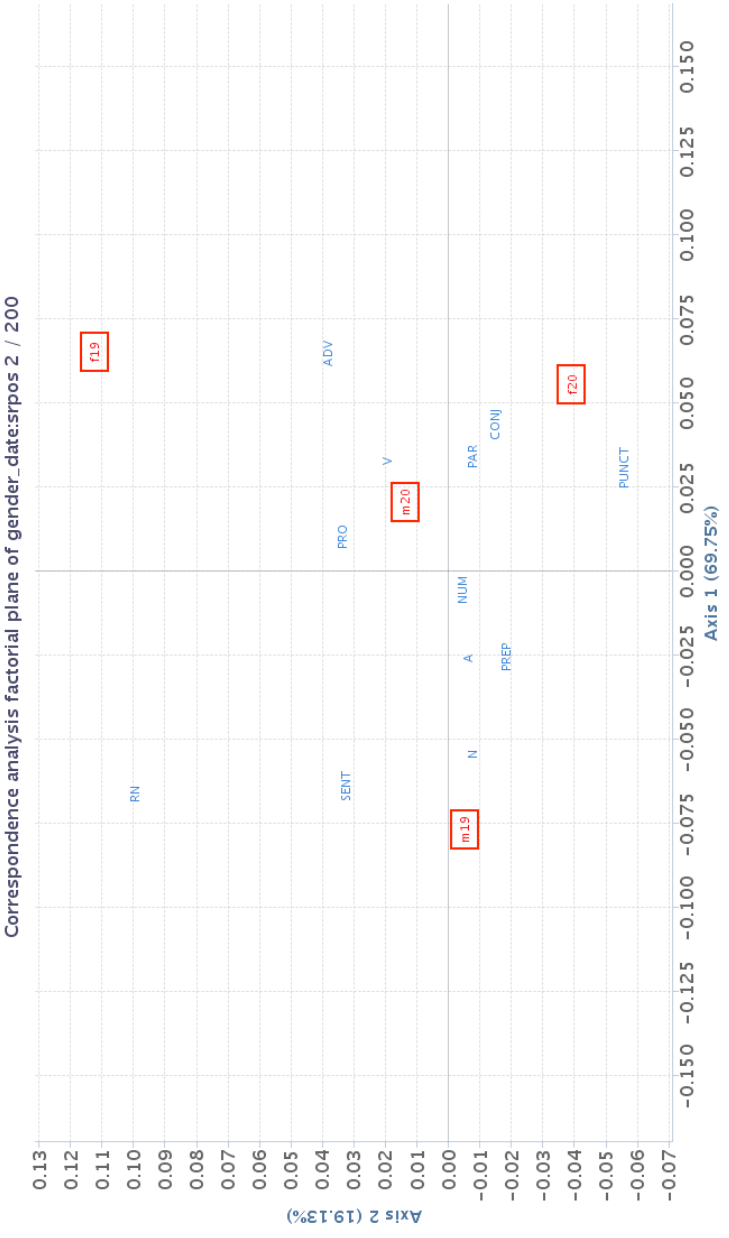
Слика 4. Димензије партиција srpELTeC корпуса креираних на основу информација о полу аутора и веку у коме је дело објављено

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 srpELTeC 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

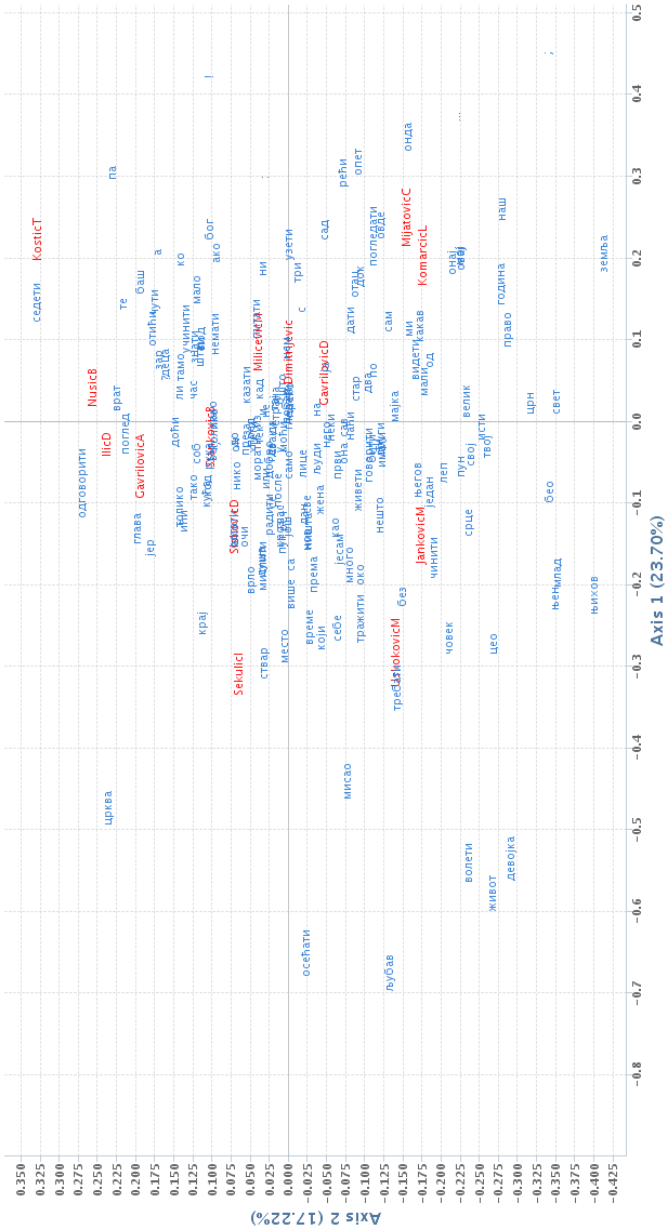
Табела 5. Фреквенција употребе и *индекс специфичности* именица (N), глагола (V), придева (A) и прилога (ADV) по партицијама креираним на основу пола аутора и временског раздобља

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.

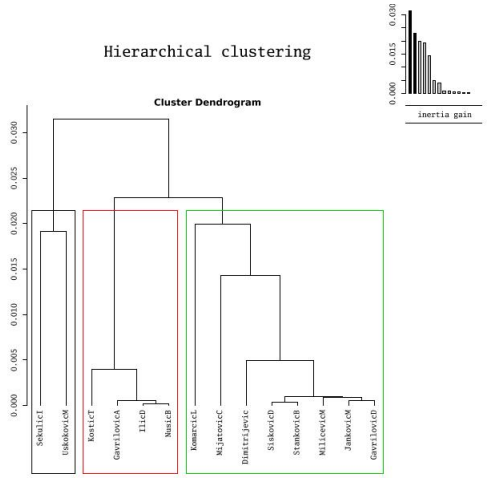
У последњем кораку урађена је и кластер анализа матрице добијене претходно спроведеном факторском анализом међузависности. Дијаграм у форми стабла (слика 8) приказује хијерархијско груписање тј. уређеност релација које постоје између текстова аутора и лема коришћених у њима. Оваква класификација текстова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



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 кванитативном 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”. PhD. thesis, Univerzitet u Beogradu, Filološki fakultet: Beograd, 2014

Utvić, Miloš. “Annotating the corpus of contemporary Serbian”. *INFOtheca* Vol. 12, no. 2 (2011): 39–51

Класификација докумената из медицинског домена екстраховањем таксономских концепата из MeSH онтологије

УДК 004.82:025.43MESH

Михаило Шкорић
mihailo.skoric@rgf.bg.ac.rs
Универзитет у Београду
Београд, Србија

Мауро Драгони
dragoni@fbk.eu
Фондација „Бруно Кеслер“
Тренто, Италија

САЖЕТАК: Рад је настао као одговор на задатак класификације медицинских докумената, постављен током летње школе *Keyword Search in Big Linked Data*, одржане у оквиру COST акције Keystone 2017. године на Технолошком универзитету у Бечу. У њему се приказују резултати специфичног приступа класификацији заснованог на креирању минималних сурогата текста. Као основа класификације узета је MeSH онтологија, заснована на тезаурусу *Medical Subject Headings*. У текстовима, претходно класификованим помоћу таксономије ове онтологије, најпре се проналазе појмови од важности, а потом се замењују таксономским референцама. Тако екстраховане референце користе за класификацију унутар MeSH таксономије помоћу простог алгоритма, а резултати се евалуирају у односу на ручно класификоване документе.

КЉУЧНЕ РЕЧИ: класификација докумената, MeSH, онтологије, екстракција информација

РАД ПРИМЉЕН: 21. април 2019.
РАД ПРИХВАЋЕН: 30. август 2019.

1. Увод

1.1 О задатку

Овај рад описује решење проблема који је задао предавач на једнодневном хакатону летње школе Keystone 3rd training school:

Keyword Search in Big Linked Data,[1] одржаној у Бечу 21–25. августа 2017. у оквиру *COST* акције IC1302 - *Keyword Search in Big Linked Data*. Проблем се састоји у класификовању 10.000 докумената из дигиталне колекције Националне медицинске библиотеке Сједињених Америчких Држава (слика 1), које је обезбедио предавач, док је на учесницима било да те документе класификују у класе предефинисане у MeSH (*Medical Subject Headings*) онтологији (Dragoni, 2017).

> *... Goserelin in the adjuvant treatment of breast cancer An update of the Zoladex Early Breast Cancer Research Association (ZEBRA) trial was presented by Professor R Blamey (Nottingham City Hospital, UK). Goserelin was found to be better ... Results were presented by the Austrian Breast and Colorectal Cancer Study Group comparing ...*

Слика 1. Одломак једног од докумената који треба класификовати

Класификација у овом раду направљена је, сходно правилима, на основу таксономије медицинских појмова из верзије MeSH онтологије из 2016. године.[2] Онтологија се може претраживати на вебу[3], где се могу наћи предефинисани упити међу којима су они за класе и предикате, или где се новим SPARQL упитима могу добити други жељени подаци.

Документација, RDF тројке и преузимање примера су такође доступни је на вебу,[4] док за прегледање предиката постоји и шира опција[5] која нуди табеларни приказ предиката, њихов опис и XML етикету која је посебно важна ако се користи локална копија MeSH онтологије у виду XML серијализације. Онтологија се састоји од 56.309 медицинских концепата, описаних и систематски сврстаних у хијерархијско дрво (слика 2).

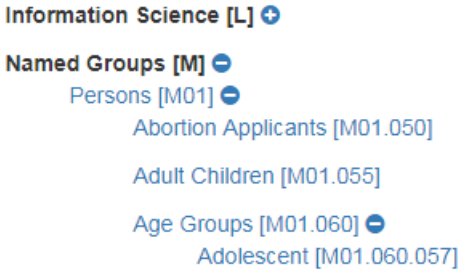
[1] Keyword Search in Big Linked Data (на вебу)

[2] Класификација докумената из медицинског домена заснована на онтологијама је предмет истраживања бројних тимова, при чему се користе различити приступи, али као онтологија се најчешће користи управо MeSH.

[3] Приступна тачка за претраживање је на вебу

[4] Документација, RDF тројке и преузимање примера на вебу

[5] Прегледање предиката на вебу



Слика 2. Приказ исечка хијерархијског дрвета.

Концепти из онтологије су хијерархијски поређани, а сваки од њих има додељен и идентификатор који се састоји од блокова цифара раздвојених тачкама, а који описују надређене концепте у опадајућем редоследу, од највишег до најнижег у хијерархији. Класе за класификацију се, конкретно у овом задатку, односе на други ниво хијерархије дрвета – има их укупно 1.718 – а препознају се тако што се њихови идентификатори састоје два блока, на пример [M01.055] *Adult children* (слика 2), где први блок *M01* означава да му је надређен чвор [M01] *Persons*, а други блок *055* је јединствен међу сабратним чворовима.

1.2 О класификацији текста

Проблем класификације докумената, у најопштијем смислу, јавља се у две варијанте: класификација у непознатом, и класификација у познатом, ограниченом домену класа. За обе варијанте проблем се своди на одређивање сличности два документа, при чему је уобичајено коришћење такозваног Дајсовог[6] индекса, односно коефицијента сличности.

$$sim(D_i, D_j) = \frac{2|S_i \bigcap S_j|}{|S_i| + |S_J|} \tag{1}$$

Он нам каже да ако је S_i скуп термина додељених документу D_i, а S_j скуп термина додељених документу D_j, онда се овај индекс може дефинисати као двоструки број заједничких термина према укупном броју термина у оба документа (ако је S скуп, онда је $|S|$ број елемената

[6] Lee Raymond Dice – амерички биолог (1887–1977)

скупа). Ако документа немају заједничких термина у том случају је $sim(D_i, D_j) = 0$ и одражава минималну сличност двају докумената, а уколико два документа имају додељене потпуно исте скупове термина онда је $sim(D_i, D_j) = 1$ и одражава максималну сличност. Када је домен класа унапред познат и ограничен, што је наш случај, проблем се своди на проналажење одговарајуће класе са највећом sim(документ, класа докумената) вредношћу, за сваки документ који је предмет класификације.[7]

Проблем који се јавља приликом израчунавања коефицијента сличности између текстова је висока цена, која се мора платити било процесорском моћи или временом извршења. Управо због тога први корак у класификацији најчешће представља креирање сурогата текста, где се он преводи у векторски простор који чине вектори речи са фреквенцијама појављивања. Речи у индексу се некада добијају стемовањем, некада лематизацијом, некада заменом синонима, или хипонима хиперонимима, како би се сурогати текста додатно смањили и како би се убрзало време извршења израчунавања. За квалитетну класификацију у овом поступку, неопходно је да сурогати буду исправно креирани и да верно представљају текст.

У радовима (Trieschnigg et al., 2009) и (Elberrichi et al., 2012) тестирано је више метода класификације медицинских докумената заснованих на MeSH онтологији или тезаурусу. Направљени класификатори користили су:

- Само MeSH тезаурус (*'Thesaurus-oriented' classifiers*);
- Тренинг скуп за изградњу експлицитних модела за сваки MeSH концепт (*'Concept-oriented' classifiers*);
- Ручно креиране анотације докумената слично као код обичних класификатора текста, да се одреди одговарајући концепт (*'K-Nearest Neighbor' classifier*);
- Хибридне и ручно профињене системе који комбинује више приступа (*'Hybrid' classifiers*).

У оба рада је закључено да *K-Nearest Neighbor* класификатор (KNN) даје најбоље резултате, али да је, упркос својим предностима, много спорији од класификатора заснованих на тезаурусу, као и да се са растом скупа докумената за тестирање његове перформансе додатно смањују, што није било пожељно при решавању добијеног задатка.

[7] Проналажење информација (на вебу)

У овом раду експериментише се једноставним приступом класификацији да би се оценио значај благовременог управљања великом количином података, као и употребна вредност семантике похрањене у MeSH онтологији. Циљ је био направити класификатор који би био брз и једноставан, како би се решио проблем велике количине текста коју треба класификовати. Примењује се драстична сумаризација докумената и самих класа – класе (концепти другог нивоа онтологије) сведене су на један једини термин – њихов назив. Са друге стране, документи су сведени само на појављивања термина (назива концепата из MeSH онтологије) који се уз једноставно мапирање (похрањено у њиховим идентификаторима у самој онтологији) поистовећују са термином који означава неку класу, њима надређени објекат. Ово умногоме олакшава и убрзава процес израчунавања сличности, јер свака класа сада има само један термин. На овај начин документ ће помоћу Дајсовог индекса увек бити класификован у класу чији (једини) термин се највише пута у њему јавља, чиме се израчунавање избегава и своди на проналажење најфреквентнијег термина у сурогату текста.

2. Поставка експеримента

Циљ експеримента је провера могућности и успешности класификације медицинских докумената на основу таксономије из MeSH онтологије и као и на основу система заснованог на правилима – појављивању термина везаних за концепте из MeSH онтологије у документима који треба да се класификују. Ток експеримента се може поделити на пет сукцесивних етапа.

1. **Екстракција таксономије из MeSH онтологије коришћењем SPARQL упита**. Ова етапа је неопходна како би се прикупила листа идентификатора и утврдила таксономска позиција термина употребљених у документима, као и позиција њихових чворова у односу на позиције чворова могућих класа.
2. **Конверзија докумената у векторе идентификатора концепата који се у њима јављају**.Ова етапа омогућава да документима доделимо атрибуте који су непосредно и нераскидиво повезани са класама у које те документе треба сврстати.
3. **Уклањање шумова**. Ова етапа треба да омогући и обезбеди што боље резултате при класификацији докумената.

4. **Класификација докумената у класе на основу њихових вектора идентификатора и једноставних скупова правила**.
5. **Евалуација успешности класификација докумената за сваки од коришћених скупова правила**. Ова етапа нам омогућава да сагледамо и упоредимо различита правила класификације, као и да установимо да ли се неки скупови правила могу сматрати успешним, који су то скупови и колико су успешни.

У наредним поглављима, етапе експеримента биће описане детаљније, како би се стекао бољи увид у коришћене методе и добијене резултате.

2.1 Екстракција таксономије концепата из MeSH онтологије

Екстракција матрице назива концепата и њихових идентификатора позиција у класификационом дрвету обавља се помоћу SPARQL упита. Како у овој онтологији постоје тројке који се састоје од концепта, предиката и објекта тог предиката, а који одражава позицију у таксономији, овај део задатка своди се на екстракцију субјекта и објекта за сваку од ових тројки.

Најпре је потребно одредити име предиката у онтологији који одражава идентификатор позиције у таксономији облика **[A-Z][0-9][0-9]([.][0-9][0-9][0-9])***.[8] Коришћен је једноставан SPARQL упит, који за било који концепт из онтологије излистава све предикате и објекте свих тројки из MeSH 2016 онтологије чији је тај концепт део (слика 3). Упит смо илустровали концептом (*mesh2016:D049916*).

```
PREFIX mesh2016: <http://id.nlm.nih.gov/mesh/2016/>
SELECT DISTINCT ?predikat ?objekat
FROM <http://id.nlm.nih.gov/mesh/2016>
WHERE  mesh2016:D049916 ?predikat ?objekat
ORDER BY ?class
```

Слика 3. SPARQL упит за добављање таксономијие свих концепата из MeSH 2016 онтологије.

[8] Овај регуларни израз описује конструкцију која се састоји од обавезног дела (велико слово, цифра, цифра) и опционог дела (тачка, цифра, цифра, цифра), који може да се понавља.

На основу упита добијен је излаз из кога се, поред осталог, закључује да је тражени предикат *meshv:treeNumber*, јер садржи блокове цифара који описују хијерархију (слика 4). Овај податак се користи у наредном упиту који има циљ да излиста све називе концепата (термине) и њихове *meshv:treeNumber* вредности.

rdf:type; meshv:TopicalDescriptor
rdfs:label; Polyplacophora
meshv:identifier; D049916
meshv:dateEstablished; 2006-01-06
meshv:historyNote; 2006
meshv:publicMeSHNote; 2006
meshv:treeNumber; mesh2016:B01.050.500.644.600

Слика 4. Неки од Излаза SPARQL упита, међу којима су и жељени предикат и објекат

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX meshv: <http://id.nlm.nih.gov/mesh/vocab#>
PREFIX mesh2016: <http://id.nlm.nih.gov/mesh/2016/>
SELECT DISTINCT ?naziv ?treeNumber
FROM <http://id.nlm.nih.gov/mesh/2016>
WHERE ?koncept rdfs:label ?naziv .
?koncept meshv:treeNumber ?treeNumber
ORDER BY DESC(STRLEN(?naziv))
```

Слика 5. SPARQL upit за добављање листе назива и позиција свих концепата из онтологије MeSH 2016.

Други SPARQL упит прибавља називе концепата на основу предиката *rdfs:label*, а затим тражи и *mesh:treeNumber* тог истог концепта. Термини су на излазу поређани по величини назива од најдужег до најкраћег, како би се тим редом претраживали у документима да би се избегло да дужа поклапања не буду препозната због краћих (слика 5).

Резултат овог упита је CSV датотека чији сваки ред садржи назив (*rdfs:label*) и таксономски идентификатор (*treeNumber*) сваког од концепата (слика 6). Ова датотека биће коришћена у наредном кораку, у којем се у документима проналазе називи концепата, тј. термини, након чега се замењују идентификаторима чворова у таксономском дрвету. Треба истаћи да се могу наћи и једночлани термини као и вишечлани (нпр. *Bacteria* и *Gram-Negative Bacteria*), али имајући у виду редослед примене замена (по дужини ниске, од најдуже до најкраће) неће доћи до погрешне замене и препознавања само дела термина.

```
Ganglia;A08.340
Neurons;A08.675
Malleus;A09.246.397.247.524
Cochlea;A09.246.631.246
Eyelids;A09.371.337
Choroid;A09.371.894.223
Tissues;A10
Chorion;A10.615.284.473
Muscles;A10.690
```

Слика 6. Примери линија CSV документа који садржи називе и идентификаторе чворова концепата.

2.2 Конверзија докумената у векторе концепата

Ова етапа састоји се од два корака. Најпре се у свим документима проналазе и замењују претходно излистани термини, а затим се уклања преостали текст како би се документи трансформисали у листу вектора идентификатора. Никаква нормализација ни документа ни термина није рађена, што за енглески језик, који нема богат флективни систем може да буде прихватљиво, али би за флекитвно богат језик, као што је српски, била неопходна претходна лематизација или неки други вид нормализације оба ресурса.

Проналажење концепата из онтологије у тексту. Како је оно што желимо да постигнемо да у документима нешто (термине

који означавају концепте) пронађемо и потом нечим (одговарајућим таксономским идентификаторима) заменимо, при чему те две ствари имамо излистане заједно у претходно генерисаној датотеци, било је могуће листу из датотеке директно трансформисати у C# функцију која би то радила.

Ово је постигнуто заменом карактера **;** у листи са ниском каракатера **", "** затим конкатенацијом (или дописивањем) ниске карактера **doc = doc.Replace("** на почетак сваког новог реда и ниске карактера **");** на крај сваког реда, где се doc односи на варијаблу која представља комплетан садржај документа (слика 7).[9]

```
doc = doc.Replace("Ganglia", "A08.340");
doc = doc.Replace("Neurons", "A08.675");
doc = doc.Replace("Malleus", "A09.246.397.247.524");
doc = doc.Replace("Cochlea", "A09.246.631.246");
doc = doc.Replace("Eyelids", "A09.371.337");
doc = doc.Replace("Choroid", "A09.371.894.223");
doc = doc.Replace("Tissues", "A10");
doc = doc.Replace("Chorion", "A10.615.284.473");
doc = doc.Replace("Muscles", "A10.690");
```

Слика 7. Одломак C# скрипте за проналажење и замену која се заснива на претходно наведеним концептима и идентификаторима (слика 6)

За замену у свим документима које треба класификовати припремљен је други C# код који учитава један по један документ за класификацију и на њима примењује скрипту генерисану у претходном кораку како би концепти били пронађени према именима и замењени идентификатором чвора из онтологије. Ова етапа је најдужа и временски најзахтевнија јер се за наш експеримент подразумева примењивање 56.309 replace израза над 10.000 докумената што даје укупно 563.090.000 трансформација. Даље истраживање може ићи у правцу коришћења коначних аутомата и трансдуктора за решење овог

[9] Накнадним размишљањем закључили смо да би се сличним приступом могао генерисати и другачији вид замене заснован на петљама или регуларним изразима, што би убрзало замене и умањило ефекте вишеструког парсирања истог документа.

проблема, које је комплексније за имплементацију, али брже врши обраде овог типа.

Трансформација докумената у векторе идентификатора (креирање сурогата) Након што су документи успешно обележени идентификаторима концепата који се у њима појављују, било је неопходно документе очистити од осталог, неупареног текста. Како се ово не би радило појединачно за сваки документ, они су спојени у један, величине ∼230MB, са новим редовима као границом између докумената. Информације од важности – називи докумената (**[0-9]+[.]txt**), идентификатори у њима (**[A-Z][0-9][0-9]([.][0-9][0-9][0-9])***) и ознаке за нови ред (**[\r\n]+**) - проналазе се помоћу регуларног израза (**[A-Z][0-9][0-9]([.][0-9][0-9][0-9])***)|(**[0-9]+[.]txt**)|(**[\r\n]+**), док се све остало уклања. Овим се величина датотеке смањила преко 450 пута (нова величина: ∼0.5MB).

По завршетку трансформације добија се датотека у којој сваки нови ред представља нови документ: ред започиње називом документа (без екстензије) иза кога следи тачка зарез, а затим сви идентификатори концепата који су у њему пронађени одвојени зарезима (слика 8).

```
2875592;M01.060.116
2875593;D13.444.308,D13.444.308
2875594;C04.557.465.625.650.510,D13.444.735,D13.444.735
2875595;A01.236,A01.236;A01.236
2875596;D13.444.735
2875598;
2875599;D02.455.612
```

Слика 8. Одломак датотеке која садржи називе докумената и све идентификаторе у њима.

Илустроваћемо на једном примеру трансформацију једног од полазних докумената по корацима. У делу документа са Слике 1 препознати су неки термини (слика 9), који су потом замењени идентификаторима (слика 10). Уочава се да је у примеру *Colorectal* дошло до препознавања дела речи и тако је ниска **Color** погрешно

замењена са **G01.590.540.199**[10] јер се термини *colorectal cancer* и *colorectal* не налазе као појмови у коришћеној верзији онтологије. Слика 11 приказује коначан сурогат тескта са Слике 1.

> ... **Goserelin** in the adjuvant treatment of breast cancer An update of the Zoladex Early **Breast** Cancer **Research Association** (ZEBRA) trial was presented by Professor R Blamey (Nottingham City Hospital, UK). **Goserelin** was found to be better ... Results were presented by the **Austrian Breast** and **Color**ectal Cancer Study Group comparing ...

Слика 9. Одломак оригиналног документа са оболеженим терминима који се налазе у MeSH онтологији.

> ... **D06.472.699.327.740.320.340** in the adjuvant treatment of breast cancer An update of the Zoladex Early **A01.236** Cancer **H01.770.644 F02.463.425.069** (ZEBRA) trial was presented by Professor R Blamey (Nottingham City Hospital, UK). **D06.472.699.327.740.320.340** was found to be better... Results were presented by the **Z01.542.088 A01.236** and **G01.590.540.199**ectal Cancer Study Group comparing ...

Слика 10. Одломак документа у коме су термини из MeSH онтологије замењени својим таксономским идентификаторима.

> ...D06.472.699.327.740.320.340;A01.236;H01.770.644;F02.463.425.069;D06.472.699.327.740.320.340;A01.236;G01.590.540.199;...

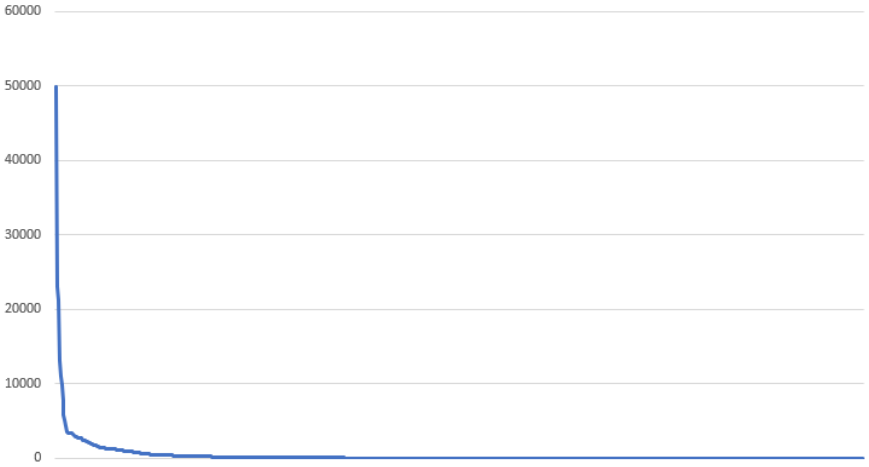
Слика 11. Коначан сурогат одломка оригиналног документа.

[10] Оваква грешка могла би се избећи претходном токенизацијом текста.

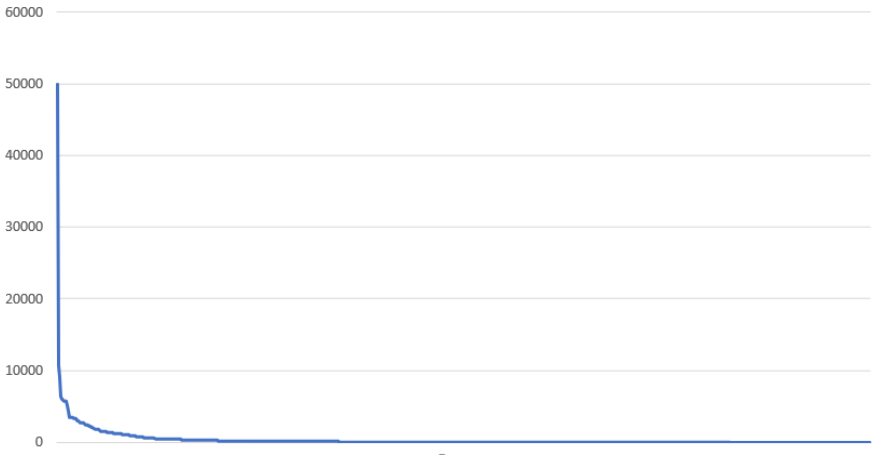
2.3 Уклањање стоп речи и шумова

Пре преласка на класификацију било је потребно је детектовати могуће шумове у виду погрешних обележја или појмова који се јављају у превеликом број докумената те нису дискриминаторни, као и концепата који се често јављају, а тешко да могу успешно класификовати документ. За сваку класу избројан је укупан број понављања, што је пружило увид у неравномерну расподелу (Слика 12).

Први идентификатори који су додати у листу стоп речи и одстрањени јесу они који се односе на географске локације побројане у онтологији под класом Z01 — географске локације као и хомоними попут термина *back*, који се појављује у документима 11.440 пута, очигледно не увек да би означио део људског тела (леђа). На Слици 13 се, међутим, види да је неравномерна расподела фреквенција класа остала и након овог корака.



Слика 12. Расподела фреквенција пре уклањања шумова.



Слика 13. Фреквенција класа у документима после уклањања стоп речи.

2.4 Класификација докумената према идентификаторима

Над документима који су предмет класификације примењена су два поступка обраде, па су тако настала два тест скупа. Контролна метода је била замена идентификатора њиховом класом, општијом хијерархијском ознаком. У експерименталној методи је узета у обзир и дужина идентификатора, па су они замењивани одређеним бројем понављања надређене класе у зависности од њихове дужине, како би се тестирала претпоставка да је коришћење уже специјализованих термина од веће важности за одређивање класе докумената. Дакле, уместо да идентификатори буду скраћени на прва два блока цифара, у обзир је узета њихова дужина, тј. дубина сваког од концепата у дрвету. Идентификатори су замењени одређеним бројем идентификатора класа на основу њихове дужине, на пример: идентификатор **D04.345.295.750.650.700** замењен је помоћу одговарајућег регуларног израза са **D04.345,D04.345** што је еквивалентно појављивању двају концепата који припадају класи **D04.345** у том документу. Начин мапирања дужине концепата у одговарајући број понављања дат је у табели 1. Ту се види да се термини поистовећују са највише четири понављања (у случају да имају преко осам блокова) термина који означава неку класу. Након

примене ових корака, у сурогатима докумената сада се појављују само идентификатори класа, које треба једноставно пребројати.

број додатних блокова (преко 2)	0	1	2	3	4	5	6	7	8	9	10
број резултованих понављања	1	1	2	2	2	3	3	3	4	4	4

Табела 1. Мапирање броја слогова вишка и броја резултујућих идентификатора класа

Након што су тест скупови успешно направљени, за класификацију докумената припремљен је једноставан програм који као улаз захтева датотеку са нискама које означавају класе (прва два блока цифара) који су препознати. Програм једноставно пребројава класе које се јављају у сурогату документа и враћа ону која се најчешће јавља, а уколико постоји више класа које се у документу јављају са истом фреквенцијом, враћа се класа која се прва појављује, што је логично јер је у сурогатима задржан редослед појављивања термина. Такође, неопходно је да сваком документу буде додељен неки идентификатор, па се у случају да неки документ нема додељене идентификаторе, њему додељује један од најопштијих – **H02.403** – који означава медицину.

Када је низ идентификатора у сваком документу сведен на једну класу, она је узета за резултат класификације и прослеђена на евалуацију.

3. Евалуација и упоређивање резултата

Експерименти су извршени над текстовима из TREC Clinical Decision Support 2016 скупа (Подршка клиничком одлучивању).[11] Циљ је био да се документи класификују на основу термина који означавају концепте из онтологије MESH, а који се у њима јављају. Анотирање коришћено у евалуацији ручно је спровео експертски тим. Примењене су уобичајене метрике средње просечне прецизности, одзива[12] и F-

[11] TREC Clinical Decision Support 2016 (на вебу)

[12] Као погрешно негативно класификовани документи узети су релевантни документи који нису укључени у рангирање.

Тест скуп	Узимање дужине идентиф. у обзир	Прецизност @10	Средња просечна прецизност	Одзив	F-мера
1	да	0.58	0.0060	0.0696	0.0108
2	не	0.46	0.0057	0.0648	0.0103

Табела 2. Резултати евалуације класификације.

мере, као и метрика прецизност@10, која одражава успешност враћања релевантних докумената у првих 10 резултата за неки упит.

Табела 2 показује да је узимања дужине идентификатора у обзир донело мало побољшање резултата (5% побољшана прецизност и F-мера, 7% побољшан одзив, 12% прецизност@10).

Узимајући у обзир добијене вредности, одмах се може приметити да је прецизност необично ниска у односу на одзив. Примењивањем детаљније анализе података, примећује се јако велики број лажних погодака (false positive), па са тим опада и прецизност дате стратегије. Овај резултат не изненађује него је у складу са тренутним стандардима у пољу класификације медицинских докумената (Calì et al., 2017). Главни проблем у концептно оријентисаном проналажењу информација у биомедицинској сфери је велики број погрешно негативно класификованих докумената, што води изузетно ниском одзиву. Мала прецизност је тако у овом раду прихватљива јер је надомештена вишим одзивом и многи релевантни документи су враћени на највишим позицијама, са вредностима за прецизност@10 чак до 58%. Ипак и овде постоји простор за напредак.

4. Закључак

У овом раду је представљен приступ класификацији докумената, који се заснива на креирању минималних сурогата текста. У медицинским текстовима најпре се проналазе појмови од значаја који се потом замењују таксономским референцама. Тако екстраховане референце, користе за класификацију текстова помоћу простог алгоритма коришћењем класа из MeSH онтологије, а резултати се потом евалуирају у односу анотацију експертског тима.

Прелиминарни резултати показују погодност понуђеног приступа решењу овог комплексног задатка. Будући рад фокусираће се на смањење лажних погодака како би се унапредиле перформансе система.

Класификација заснована на онтологијама не зависи од домена у ком се примењује, али свакако зависи од расположивих ресурса, конкретно онтологије или таксономије која се користи за класификацију (Rakesh et al., 2001), тако да једном успостављен систем може наићи и на ширу примену. Када је у питању класификација (медицинских) докумената за српски језик, неопходно је најпре припремити ресурсе, при чему би свакако од користи била Међународна класификација болести – Шифарник болести МКБ 10. У оквиру истраживања на Рударско-геолошком факултету креирана је таксономија са овом класификацијом (Kolonja et al., 2016), где је одређен број термина повезан са енглеским и латинским еквивалентима, што омогућава проширење претраге назива концепата и њиховог проналажења у документима. Ипак, треба имати у виду богату морфологију српског језика која би захтевала допуну приступа коришћењем лексичких ресурса специфичних за област медицине за нормализацију текста пре класификације или индексирања, што би припомогло препознавању већег броја таксономских појмова у документима (Stanković et al., 2015).

5. Захвалност

Овај рад је настао у оквиру летње школе *Keyword Search in Big Linked Data* (Претраживање кључним речима велике количине података), организоване у оквиру COST акције Keystone од 21. до 25. августа 2017. године на Технолошком универзитету у Бечу.

Литература

Rakesh et al., Agrawal. “Multilevel Taxonomy Based on Features Derived from Training Documents Classification using Fisher Values as Discrimination Values”, U.S. Patent No. 6,233,575, 2001

Calì, Andrea, Dorian Gorgan и Martin Ugarte. *Semantic Keyword-Based Search on Structured Data Sources: COST Action IC1302 Second International KEYSTONE Conference, IKC 2016, Cluj-Napoca, Romania, September 8–9, 2016, Revised Selected Papers*, Vol. 10151, 2017

Dragoni, Mauro. "3rd KEYSTONE Summer School", 2017, URL http://ifs.tuwien.ac.at/keystone.school/slides/Dragoni_SemanticSearch.pptx

Elberrichi, Zakaria, Malika Taibi и Amel Belaggoun. "Multilingual Medical Documents Classification Based on MesH Domain Ontology". *International Journal of Computer Science Issues* Vol. 9 (2012)

Kolonja, Ljiljana, Ranka Stanković, Ivan Obradović, Olivera Kitanović и Aleksandar Cvjetić. "Development of terminological resources for expert knowledge: a case study in mining". *Knowledge Management Research & Practice* Vol. 14, no. 4 (2016): 445–456

Stanković, Ranka, Cvetana Krstev, Ivan Obradović и Olivera Kitanović. "Indexing of Textual Databases Based on Lexical Resources: A Case Study for Serbian". У *Semantic Keyword-based Search on Structured Data Sources*, Cardoso, Jorge, Francesco Guerra, Geert-Jan Houben, Alexandre Miguel Pinto и Yannis Velegrakis, 167–181. Cham: Springer International Publishing, 2015

Trieschnigg, Dolf, Piotr Pezik, Viv Lee, Franciska de Jong, Wessel Kraaij et al.. "MeSH Up: Effective MeSH Text Classification for Improved Document Retrieval". *Bioinformatics (Oxford, England)* Vol. 25 (2009): 1412–8

Организовање дигиталног репозиторијума у Институту за српски језик САНУ

УДК 061.6: 811.163.41]:004.738.5

Владимир Живановић
vladimirludwig@yahoo.com
Институт
за српски језик САНУ
Београд, Србија

САЖЕТАК: Овим радом желели бисмо да представимо случај прикључивања Института за српски језик САНУ (ИСЈ САНУ) свету дигиталних репозиторијума. Године 2018. Министарство просвете, науке и технолошког развоја Републике Србије усвојило је Платформу којом обавезује све јавно финансиране институције да своје научне радове депонују у неки репозиторијум. Намера нам је да представимо основне фазе успешно спроведене Платформе Министарства у ИСЈ САНУ. У периоду од једне године у ИСЈ САНУ припремљено је преко 100.000 страница дигиталног текста које чине 2.500 записа пуног текста. Овим подухватом реализованим у оквиру концепта отворене науке омогућено је да српска језичка хуманистика постане видљива у свету.

КЉУЧНЕ РЕЧИ: дигитални репозиторијум, институционални репозиторијум, ДАИС, отворена наука, самоархивирање, дигитална хуманистика.

РАД ПРИМЉЕН: 10. мај 2019.
РАД ПРИХВАЋЕН: 18. јуни 2019.

1. Увод у процес дигитализације

Године 2018. Министарство просвете, науке и технолошког развоја усваја Платформу за отворену науку (Платформа, 2018). Ова Платформа, која се заснива на принципима отворене науке и

смерницама Европске комисије у тој области, уводи обавезу за појединце и институције да део своје научне продукције који је јавно финансиран учини доступним широј јавности депоновањем у неки дигитални репозиторијум. Како би се поштовала „транспарентност научне комуникације и методологије“ и обезбедила доступност научне продукције, било је потребно приступити развоју одговарајуће дигиталне инфраструктуре, која у Србији није постојала.

Отворени приступ у овом документу дефинише се као „право сваког корисника интернета да без финансијских издатака чита, преузима, чува, штампа и користи дигитални садржај публикација“. Поред ових права, корисници ће уз поштовање ауторских права, тј. лиценце која је придружена депонованом раду и уз правилно навођење извора, суделовати у ширењу идеја отворене науке, користећи неки од „институционалних/тематских/националних репозиторијума“. Тип репозиторијума за депоновање научне продукције није строго одређен па је тако остављено институцијама да нађу адекватно решење у складу са својим могућностима. На основу ове Платформе, обавеза је да свака од обухваћених институција усвоји „локалну“ политику, тј. одговарајући документ о спровођењу Платформе којим ће регулисати њено спровођење на институционалном нивоу.

У исто време, током припреме ове Платформе, и Српска академија наука и уметности (САНУ), у сарадњи са Рачунарским центром Универзитета у Београду (РЦУБ) ради на развијању сопственог репозиторијума који би био намењен Академији и припадајућим институтима. Репозиторијум је назван Дигитални архив издања САНУ и Института – ДАИС.[1] Према дефиницији Џозефа Бранина, репозиторијуми представљају „моделе система и услуга који су осмишљени да прикупе, организују, ускладиште, деле и чувају дигиталне информације и знања институције“ (Branin, 2004-2005, 237). Сврха дигиталног репозиторијума ДАИС је да омогући Академији, као и Институтима којима је Академија оснивач, да у њему трајно похрањују своју научну продукцију те тако јавно представљају резултате свога научног рада.

Развој комуникационих процеса омогућава већу доступност научних радова у пуном тексту. Општа доступност је и даље идеал према коме се тежи, али научне институције подстакнуте бројним иницијативама из домена отворене науке, као и праксом великих сервиса технолошких

[1] ДАИС (на вебу)

компанија (Google books и др.) почеле су да заснивају своје дигиталне репозиторијуме. Један од резултата успешно вођеног репозиторијума је повећање видљивости научних радова, посебно из области српске хуманистике. Због комерцијалног потенцијала техничких и природних наука, њихова видљивост у свету је много већа у односу на хуманистички домен. Оваква несразмера између природних и техничких наука није оправдана, с обзиром да оно што је локално представља једну културу на глобалном нивоу.

За разлику од класичне библиотеке, формирање дигиталног репозиторијума захтева техничку припрему. ДАИС је настао прилагођавањем софтвера отвореног кода DSpace.[2] Прилагођавање овог софтвера подразумевало је његову локализацију, адаптацију корисничког интерфејса, усклађивање са смерницама конзорцијума OpenAIRE за дигиталне репозиторијуме, интеграцију са сервисима ORCID и Altmetric.com и развој додатних екстерних апликација које омогућавају нормирање имена, као и преузимање записа и масован унос и корекције метаподатака.[3]

Разлог за стварање оваквих софтвера почива на развоју библиотечког пословања. Бодвајн и Бранскофска, са МИТ-а, тврде да универзитетске заједнице зависе од својих библиотека које им омогућавају стални приступ истраживању и научном раду, и то их ставља у позицију да траже решења за проблем складиштења и преузимања интелектуалног рада на дуже стазе (Baudoin and Branschofsky, 2004, 32). Имплементирање и развој софтвера, као и његово прилагођавање специфичним потребама српске научне продукције извршио је РЦУБ. У рачунарском центру се стално ради на одржавању репозиторијума и његовом континуираном усавршавању.[4]

Потребно је уочити разлику између дигиталне колекције и дигиталног репозиторијума. Док колекција остаје локална, какве су нпр. колекције дигиталних објеката које на своје веб-сајтове

[2] Dspace софтвер отвореног кода је развио Масачусетски институт за технологију (МИТ) заједно с фирмом Hewlett-Packard 2002. године. Његов квалитет види се у лаком и отвореном приступу различитим типовима дигиталних објеката: тексту, сликама, аудио- или видео-материјалу.

[3] За више о DSpace софтверу као ослонцу за развој дигиталних репозиторију видети (Rajović and Ševkušić, 2018)

[4] Прилагођавање софтвера отвореног кода локалним потребама је највећи изазов у почетним фазама организовања репозиторијума и захтева квалитетну техничку подршку. На коду софтвера се континуирано ради.

постављају институције или појединци, она не задовољава стандарде интероперабилности (преко DOI броја, других перзистентних идентификатора, метаподатака и атрибута, структуре и стандарда за описивање документа, протокола за размену метаподатака и др. (Van de Sompel and Nelson, 2015)). Дигитални репозиторијум пак представља дигиталну колекцију која је организована тако да се његов садржај може трајно архивирати, чувати и дисеминирати кроз оријентисане циљеве и протоколе компатибилне с другим базама података. Његова сврха је „представљање интелектуалног учинка“ и промовисање доступности (Winter and Bowen-Chang, 2010, 320).

Искуство ИСЈ САНУ у организовању и вођењу дигиталног репозиторијума може послужити другим институцијама које се укључују у организовање дигиталних колекција с пуним текстом у отвореном приступу.

2. Структура

Како би институционални репозиторијум заживео, његово попуњавање је прва највећа препрека. Зато смо у ИСЈ САНУ приступили двема фазама попуњавања. Правилно организован репозиторијум повећава видљивост научне продукције одређене установе, те тиме омогућава и бољу цитираност истраживача (Piwowar and Haustein, 2018).

Прва фаза подразумева прикупљање, дигитализовање и постављање дигиталног садржаја часописа и монографија ИСЈ САНУ. Историја периодичних публикација Института веома је дуга. Два од четири научна часописа које Институт објављује излазе дуже од сто година.[5] Објављени томови ових часописа броје неколико стотина свезака. Поред серијских публикација у првој фази у институционални репозиторијум уносили смо научне радове истакнутих научника Института. Ту су убројане колеге које су у пензији, а неки од њих нису више с нама. Уношењем ових дигитализованих дела успешно се испуњава неколико критеријума за функционисање репозиторијума: успоставља се добра пракса будућег рада, учествује се у формирању навика будућих

[5] Први број часописа *Српски дијакетолошки зборник* изашао је 1904. год., *Јужнословенског филолога* 1913., часописа *Наш језик* 1933. године, и *Лингвистичких актуелности* 2000.

корисника, представља се научна продукција у најбољем светлу.[6] Тиме се покрива и дијахронија научне продукције Института.

Друга фаза, која представља синхронијски ниво, подразумева објављивање текуће научне продукције, најкасније 18 месеци од датума објављивања научног чланка или монографије, и дефинисана је Правилником који је усвојен на нивоу Института. Тиме се испуњавају услови које прописује Платформа Министарства.

Институције које приступају формирању репозиторијума морале би да дефинишу своје циљеве те да јасно идентификују скуп особина добро обученог особља које ће учествовати у процесу одржавања репозиторијума (Winter and Bowen-Chang, 2010, 323). Обука кадрова за рад у репозиторијуму може да представља изазов. Недостатак иницијативе и ентузијазма код прихватања другачијег приступа научним публикацијама, те неразвијена свест о његовом значају, може да се одрази у првим фазама организовања репозиторијума. Веома је важно да руководећи кадар, као и сарадници научне установе, увиде важност успостављања репозиторијума.

Сваки научни сарадник у Институту дужан је да уноси своје научне радове кроз поступак самоархивирања. За истраживаче је организована обука, у виду радионица које садрже и практични део (самосталан унос метаподатака и депоновање објеката), а припремљено је и детаљно упутство у виду PowerPoint презентације које могу да користе у самосталном раду. Након што истраживач депонује рад и опише га свим припадајућим метаподацима, библиотекар као администратор контролише тачност уноса и одобрава запис, и једини има могућност да накнадно мења метаподатке или уклони дигитални објекат.[7]

Скуп метаподатака подразумева уношење основних података о документу, текст сажетка и навођење кључних речи, те одређивање степена доступности документа и типа лиценце. Такође се обавезно уносе и подаци о пројекту у оквиру ког је рад финансиран, па се, поред тога, подаци о типу и верзији документа, доступности садржаја и лиценци уносе у складу са актуелним смерницама за дигиталне репозиторијуме конзорцијума OpenAIRE. Домен који задобија све већи

[6] За разлику од техничких наука у лингвистици велики део старих научних радова не губе на значају зато што су за истраживања и даље релевантни.

[7] ДАИС репозиторијум претпоставља више нивоа приступа и дозвола за управљање садржајем тако да администратор има могућност да ограничи овлашћења корисницима када самоархивирају, и тиме обезбеди контролу квалитета, тачност података и интегритет базе.

значај је надгледање поштовања ауторских права и ступање у контакт са часописима и издавачима ради регулисања ауторских права.

Такође, библиотекар користи додатне сервисе ДАИС-а, као што су сервис за масовно едитовање метаподатака, нормативну базу, и сервис за преузимање постојећих дигиталних садржаја из других репозиторијума и сл. Тим за развој репозиторијума РЦУБ-а је самостално развио апликацију Ellena2 у којој су интегрисани нормативна датотека, сервис за масовно кориговање метаподатака и сервис за масовни увоз метаподатака. У раној фази развоја, постојале су две одвојене апликације: Ellena (нормативна датотека и масовно кориговање метаподатака) и MultiLoad (преузимање записа из других DSpace репозиторијума и масовни увоз метаподатака у XML или RIS формату). Почетком 2019. године, ове две апликације су интегрисане у апликацију Ellena2. Разлог зашто се приступило изради самосталне апликације јесу недостаци платформе DSpace, која сама по себи не садржи модул са нормативном датотеком. Поред тога, иницијални тестови су показали да систем за преузимање метаподатака који је већ био уграђен у DSpace не даје задовољавајуће резултате.

Руковођење садржајем омогућава сарадницима да репозиторијум користе у ширем распону него што их Платформа отворене науке обавезује јер ДАИС подржава унос и обраду више типова докумената и формата. Поред научног чланка, поглавља у монографији, монографије и сл., сарадници су у могућности да депонују и препринт верзију научног рада, докторске или магистарске тезе, извештаје и све друге садржаје који се називају „сивом“ литературом.[8] У DSpace-у индексира се сав читљив текст из депонованих докумената, па се самим тим може и претраживати. Рангирање индексираних докумената врши се преко формалних параметара, тј. по релевантности, наслову и датуму. DSpace подржава похрањивање практично свих формата. У DSpace-у не постоји систем за прегледање докумената на онај начин на који он функционише у Omeki и другим платформама које су првенствено намењене приказивању културне баштине и дигитализоване грађе. DSpace је намењен презентацији научних садржаја, а похрањене датотеке се могу снимити на рачунар корисника и оданде отворити помоћу одговарајуће апликације. PDF, JPG и слични формати отварају се у новом прозору интернет претраживача. DSpace се може надоградити тако да у оквиру

[8] За више о научној „сивој“ литератури видети (Ferreras-Fernández and Merlo-Vega, 2015) и (Ćirković, 2018)

саме платформе приказује садржаје (нпр. листање књига), али нама, а ни корисницима, за сада то није потребно.

3. Пракса

Пред крај 2017. године у ИСЈ САНУ приступило се раду на попуњавању ДАИС-а. С обзиром на релевантност часописа ИСЈ САНУ у целом славистичком свету, један део годишњака већ је био скениран у оквиру сервиса Google Books, али није био јавно доступан.[9] Више америчких универзитета је у оквиру својих колекција у сарадњи са Гуглом дигитализовало наше часописе. Те дигиталне примерке смо прикупили и они су чинили отприлике 50% од укупног броја свезака часописа. Остатак је требало дигитализовати. Приступило се послу који је подразумевао организовање, скенирање, редиговање дигиталних копија, рад на оптичком препознавању карактера садржаја и његово постављање у базу. На тај начин је за годину дана припремљено преко 50.000 страница текста. Недостатак инфраструктуре у одређеним сегментима посла надомештила је сарадња са сервисом Google Books: захваљујући томе што смо скениране публикације учинили доступним и на овом сервису, успели смо да скенирани материјал, који је био веома обиман у меморијском смислу, смањимо на оптималне величине,[10] као и да обезбедимо оптичко препознавање текста.[11]

У првој фази досадашњу продукцију Института учинили смо јавно доступном депоновањем целих годишта периодичних публикација (као

[9] Једна од основних мисија сервиса Google Books је организација информација на светском нивоу. Дигитализацијом светске библиотечке баштине компанија Google постаје универзална библиотека светског знања. За више о партнерском програму са сервисом Google Books видети (Шевкушић, 2013)

[10] Како би учинили наша документа читљивим и смањили величину фајлова (јер су неки имали и по 600 Mb) а то све оптерећујуе како репозиторијум и његове ресурсе, тако и крајњег корисника који жели да добије документ оптималне величине за кратко време. На пример, једно годиште часописа запремало је у просеку 200 и 300 Mb, али је након Гугловог оптичког препознавања карактера документ смањен на отприлике 30 Mb.

[11] До сада није урађена евалуација успешности оптичког препознавања текста. Када буду биле доступне и приступачне услуге домаће иницијативе која се бави оптичким препознавањем текста, посебно ћириличног, радо ћемо поново приступити OCR-у, те у репозиторијум депоновати документа са прецизније препознатим текстом.

једног документа), док је за касније остало да се та годишта поделе на чланке, који би се потом појединачно депоновали и тако учинили видљивијим. Поред часописа и монографска продукција је унета у ДАИС и обухвата преко 60 свезака. Овим је део прве фазе био готов и Институт је у ДАИС-у имао представљену ретроспективу своје издавачке делатности (250 томова часописа и монографија).

Други сегмент у првој фази успостављања ДАИС-а представљао је сакупљање и припремање те објављивање чланака истакнутих научника ИСЈ САНУ. У првих годину дана је у репозиторијум унесено више од 1.200 научних чланака.[12] Мали број ових чланака је већ постојао у дигиталној форми те је цео посао дигитализације требало да прође процес физичке припреме и рада на скенирању, сређивању скенираних докумената и креирању метаподатака. Према Фостеру и Гибонсу институционални репозиторијум „без садржине, је исто што и низ празних полица“ (Gibbons and Foster, 2005, пар. 4).[13] Зато је за формирање корисничких навика и успостављање доброг примера од суштинске важности оснаживање почетне динамике једног репозиторијума.

Креирање и организовање дигиталног садржаја у ДАИС-у усклађено је са захтевима Европске комисије у вези са отвореним приступом публикацијама, и омогућава даље ширење научних информација преко европских и међународних портала као што су OpenAire, BASE, CORE, Google Scholar. Једно од поља за унос метаподатака бележи име пројекта на коме је рад настао. То поље, преко линка, повезује све радове једног пројекта у ДАИС-у, али и у бази OpenAire, која на једном месту прикупља податке о резултатима свих пројеката Европске комисије, као и податке о резултатима националних пројекат из десетак земаља, међу којима је и Србија (OpenAire, 2019). Такође, коришћењем идентификатора ORCID (Open Researcher and Contributor ID) на једном месту се окупљају и повезују све публикације аутора, и када су објављене под различитим варијантама именима.

[12] Међу њима су чланци академика Ирене Грицкат, академика Митра Пешикана, др Ивана Поповића, др Берислава Николића, научног саветника др Егона Фекета, проф. др Димитрија Е. Стефановића, академика Александра Ломе, научног саветника др Јасне Влајић-Поповић, научног саветника др Стане Ристић, научног саветника др Срета Танасића и др.

[13] Као пример добре праксе узели бисмо репозиторијум Масачусетског технолошког института (МИТ) који до 2019. године у свом репозиторијуму поседује преко 106.000 дигиталних записа (DSpace, 2019).

При одређивању профила дигиталне колекције ИСЈ САНУ се одлучио за обезбеђивање пуног текста у отвореном приступу за сва своја издања, као и за објављене радове запослених уз ограничења која одређују ауторска права других издавача. Одлучено је да се на издања Института примењују CC лиценце (Creative Commons). Овај једноставан и стандардизован начин одређивања ауторских права има више нивоа. Институт се одлучио за модул CC BY-NC-ND (Ауторство – Некомерцијално – Без прерада), што подразумева да је обавезно навођење имена аутора и да аутор даје дозволу за преузимање дела и његову даљу дистрибуцију. Ова најрестриктивнија од слободних лиценци не дозвољава даљу прераду дела нити комерцијалну употребу.

Софтвер DSpace у својој архитектури омогућава стварање колекција у дигиталном репозиторијуму. Ове колекције категоришу садржај, те га тематски, формално или на други начин разврставају. ИСЈ САНУ у ДАИС-у тренутно има седам колекција. Поред четири колекције које разврставају периодику Института, две су намењене монографским серијама, док је једна општа. Општа колекција сабира зборнике Института, друге материјале и радове истраживача објављене изван Института. Проблем класификације увек намеће разна решења. Грађа је могла бити и другачије разврстана, али оптимални приступ води рачуна и о лаком сналажењу крајњих корисника.

4. Закључак

На пољу дигиталне хуманистике у ИСЈ САНУ за релативно кратко време дошло је до великих промена. У пракси, институционални репозиторијум је заживео и служи за више намена. Главна намена је промоција концепта отворене науке и представљање научне продукције установе, чиме се испуњавају услови које је Министарство прописало Платформом. Оно што је јавно финансирано требало би да буде и јавно доступно. Други правац којим се институционални репозиторијум креће јесте и представљање ретроспективне научне продукције. Објављено је готово све што је ИСЈ САНУ публиковао. Тиме се стари научни радови који су и даље релевантни чине лако доступним. Библиотека радом репозиторијума партиципира у научном истраживању. Према Бодвајн и Бранскофској, заснивањем дигиталног репозиторијума мења се начин на који размишљамо о животном циклусу научног истраживања (Baudoin and Branschofsky, 2004, 43).

Захваљујући успешној сарадњи библиотекара и руководства Института, као и научних сарадника, дигитални репозиторијум је заснован, попуњен научним радовима и спреман да прати будућу научну продукцију. Кроз успешну сарадњу свих фактора у овом сложеном процесу гради се и промовише кредибилитет успостављеног репозиторијума. Важно је напоменути да су сарадници Института показали спремност да се укључе у цели подухват што се одражава у резултатима. Сада је њихова продукција скоро у целости у слободном приступу. Проценат садржаја у отвореном приступу у колекцијама ИСЈ САНУ је 100%, што је велики успех, имајући у виду да је отворени приступ у хуманистичким наукама у свету мање заступљен и спорије се развија него у другим дисциплинама. Према извештајима начињеним за потребе Европске комисије, највеће ограничење код отвореног приступа је баш у оквиру друштвених наука и области хуманистике (Archambault and Roberge, 2014, 20).

Литература

"DSpace". Duraspace. Приступљено 12.03.2019, https://duraspace.org/dspace/about/

"OpenAire". European Union/European Commission. Приступљено 08.03.2019, https://www.openaire.eu

"Платформа за отворену науку". Министарство просвете, науке и технолошког развоја. Приступљено 21.02.2018, http://www.mpn.gov.rs/wp-content/uploads/2018/07/Platforma-za-otvorenu-nauku.pdf

Archambault, É., D. Amyot, P. Deschamps, A. Nicol, F. Provencher, L. Rebout, and G. Roberge. "Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Levels – 1996–2013". *Science-Metrix* (2014). Приступљено 20.03.2019, http://science-metrix.com/en/publications/reports/proportion-of-open-access-papers-published-in-peer-reviewed-journals-at-the

Baudoin, Patsy and Margret Branschofsky. "Implementing an Institutional Repository: The DSpace Experience at MIT". *Science & Technology Libraries* Vol. 24, no. 1/2 (2004): 31–45.

Branin, Joseph. *Encyclopedia of Library and Information Science*. Institutional Repositories. New York, N.Y.: Marcel Dekker, 2004-2005.

Ferreras-Fernández, Tránsito, Francisco J. García-Peñalvo and José A. Merlo-Vega. “Open Access Repositories as Channel of Publication Scientific Grey Literature”. У *TEEM ’15 Proceedings of the* 3^{rd} *International Conference on Technological Ecosystems for Enhancing Multiculturality*, 419–426. New York, NY, USA: ACM, 2015.

Gibbons, Susan and Nancy Fried Foster. “Understanding Faculty to Improve Content Recruitment for Institutional Repositories”. *D-Lib Magazine* Vol. 11, no. 1 (2005). Приступљено 25.03.2019, http://www.dlib.org/dlib/january05/foster/01foster.html

Piwowar, Heather, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West and Stefanie Haustein. “The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles”. *PeerJ* Vol. 6 (2018). Приступљено 01.04.2019, https://peerj.com/articles/4375/;6:e4375, DOI:10.7717/peerj.4375

Rajović, Vasilije, Biljana Kosanović and Milica Ševkušić. “DSpace – institutional repositories – dissemination of research results: A local case study”. У *Primena slobodnog softvera i otvorenog hardvera PSSOH 2018*. Belgrade, Serbia: University of Belgrade – School of Electrical Engineering and Academic Mind, 2018.

Van de Sompel, Herbert and Michael L. Nelson. “Reminiscing About 15 Years of Interoperability Efforts”. *D-Lib Magazine* Vol. 21, no. 11/12 (2015). Приступљено 15.03.2019, http://www.dlib.org/dlib/november15/vandesompel/11vandesompel.html

Winter, Marsha and Portia Bowen-Chang. “Dealing with DSpace: The Experience at the University of the West Indies, St. Augustine”. *New Library World* Vol. 111, no. 7/8 (2010): 320–332.

Ćirković, S. “Grey Literature – The Chameleon of Information Resources”. *Infotheca - Journal For Digital Humanities*, Vol. 18, no. 1 (2018): 75–83. doi:10.18485/infotheca.2018.18.1.5

Шевкушић, Милица, “Партнерски програм Гугл књиге као платформа отвореног приступа у научним библиотекама”. У *Отворен приступ знању у библиотекама, организатори конференције Библиотекарско друштво Србије [и] Народна библиотека Србије*, 187–207. Београд: Библиотекарско друштво Србије, 2013.

Дигитална библиотека „Велики рат“ – развој и резултати

УДК 027.54(497.111):004.738.5

Биљана Калезић
biljana.kalezic@nb.rs

Народна библиотека Србије
Београд, Србија

САЖЕТАК: Током рада на пројекту Europeana collections 1914–1918, стручни тим Народне библиотеке Србије је донео одлуку да се формира посебна, тематска дигитална библиотека, која ће садржати грађу насталу у периоду Првог светског рата, везану за Србију и српски народ. У раду је дат преглед тока и фаза рада на развоју портала и дигиталне библиотеке *Велики рат*. Захваљујући сервису Google Analytics, изведени су закључци и дати основни статистички подаци о коришћењу и корисницима ове дигиталне библиотеке.

КЉУЧНЕ РЕЧИ: дигитална библиотека, Велики рат, Први светски рат, Народна библиотека Србије, развој, корисници, статистика.

РАД ПРИМЉЕН: 15. април 2019.
РАД ПРИХВАЋЕН: 29. мај 2019.

1. Увод

Поводом обележавања стогодишњице од почетка Првог светског рата, Међународна фондација *Europeana*[1] је 2012. године покренула неколико пројеката: *Europeana collections 1914–1918*[2], *Europeana 1914–1918*[3] и *EFG1914*.[4] Сви ови пројекти су имали заједнички циљ – да се дигитализују и јавно доступним учине публикације и објекти настали у периоду 1914–1918. године, као и грађа из каснијег периода која је тематски везана за Први светски рат (Калезић, 2014).

[1] Europeana (приступљено 27.03.2019.)
[2] Europeana collections 1914–1918 (приступљено 27.03.2019.)
[3] Europeana 1914–1918 (приступљено 27.03.2019.)
[4] European Film Gateway 1914 (приступљено 27.03.2019.)

Идеја пројекта *Europeana collections 1914–1918* је да се из девет националних библиотека земаља које су на различитим странама учествовале у овом историјском конфликту, дигитализује грађа из њихових фондова и постане расположива онлајн. Библиотечка грађа из овог периода, до тог момента доступна за коришћење само у читаоницама поменутих библиотека, неретко је у веома лошем физичком стању с обзиром на старост и квалитет израде, па је њихова дигитализација значајно допринела физичком очувању и заштити овог осетљивог материјала. На овај начин је дигитализовано и за будућност сачувано преко 400.000 публикација којима је омогућен слободан приступ на сајту пројекта. Међу дванаест равноправних партнера у овом пројекту је и Народна библиотека Србије. Један од резултата њеног учешћа је портал *Велики рат*,[5] чији највећи део чини дигитална библиотека. Према смерницама пројекта, у ову дигиталну библиотеку су ушли сви материјали који су настали у периоду Првог светског рата, а везани су за Краљевину Србију и српски народ. Рад на развоју портала је почео у новембру 2012. године, а прва верзија је инсталирана 5. јануара 2013. године. Јавна промоција је одржана на Дан Народне бибилиотеке Србије, 28. фебруара 2013. године.

2. Дигитална библиотека *Велики рат* – селекција грађе

Врсте грађе које су укључене у дигиталну библиотеку *Велики рат* дефинисане су пројектом *Europeana collections 1914–1918*: књиге, периодичне публикације, ратни дневници, музикалије, дечја литература, фотографије, постери, пропагандни материјали и све остало што се и иначе може наћи у фондовима националних библиотека и њихових дигиталних двојника. Међутим, селекција грађе која ће бити дигитализована и укључена у дигиталну библиотеку *Велики рат* је у великој мери била условљена специфичним положајем Краљевине Србије и српског народа у Првом светском рату.

Иначе неразвијена издавачка продукција у Србији с почетка 20. века, претрпела је додатни ударац због свеопште оскудице, када се Београд, у којем је штампано око 80% тадашњих издања, нашао на линији фронта више од годину дана. Међутим, окупација Србије 1915. године, повлачење војске и државног апарата преко Албаније у

[5] Велики рат (приступљено 10.04.2019.)

Грчку и касније расипање избеглица широм Западне Европе и света довели су до отварања нових центара издавачке продукције српског народа. Основане су српске штампарије на Крфу и у Солуну, а посебно је активна штампарија српских војних инвалида у Бизерти у Тунису. У универзитетским центрима земаља Западне Европе (Француској, Великој Британији, Италији, Швајцарској) појављују се научна издања и школски листови на српском или домицилним језицима, док пројугословенска емиграција у Јужној и Северној Америци издаје велики број публикација посвећених рату и идејама уређења југословенског простора након рата. У исто време, у окупираној Србији аустроугарске и бугарске власти издају своје новине, објаве становништву итд. У Аустроугарској монархији издаван је велики број пропагандних брошура које су за циљ имале да учврсте лојалност словенског, а посебно српског становништва Монархији (Калезић, 2014).

Имајући све ове историјске околности у виду, за формирање и селекцију грађе за ову дигиталну библиотеку, било је неопходно формирати мултидисциплинарни тим, састављен од библиотекара, историчара, информатичара, професора и предавача историје. Циљеви који су поставили пред себе су да грађа буде што обухватнија, али и да се презентује на такав начин да претраживање и коришћење грађе, било да су у питању почетници или напреднији корисници дигиталних библиотека, буде једноставно. Свеобухватност грађе не би могла да се постигне да није било сарадње са установама и институцијама из земље и света које су дигитализацијом грађе из својих колекција значајно обогатиле ову дигиталну збирку. Неке од институција са којима је Народна библиотека Србије остварила сарадњу су: Архив Србије, Народна и универзитетска библиотека Републике Српске, Војни музеј, Народна скупштина Републике Србије и бројне друге (Калезић, 2014). Поред поменуте грађе, у поседу Народне библиотеке Србије, у Одељењу посебних фондова, чува се велики број дневника, писама, фотографија и остале грађе која пружа увид у свакодневни живот окупираног становништва, као и војника на фронту, у болницама и заробљеништву. Укључивањем и ове грађе у дигиталну библиотеку *Велики рат*, испуњен је циљ који је пред себе поставио стручни тим пројекта – дати корисницима могућност увида и дубљег разумевања многобројних друштвених процеса у датом периоду, а који су значајно изменили слику Србије, Европе, па и света. У складу са тим, а узимајући у обзир потребе будућих корисника, одлучено је да се у дигиталну

библиотеку унесе и грађа важна за проучавање овог периода, а која је објављена и након Првог светског рата, до 1941. године.

3. Формирање колекције

Након обављеног посла на селекцији грађе, отпочео је рад на њеној дигитализацији. На тим пословима у почетку су радили запослени из Одељења за дигитализацију Народне библиотеке Србије. Како су обим посла и притисак да се дигитализација заврши у одређеном року расли, тако је постало јасно да је неопходна помоћ са стране. Недуго затим је потписан уговор са *Медијским архивом Ебарт*, који је као подизвођач наставио започету дигитализацију. Договорено је да резолуција скениране грађе буде 300 dpi. Уз грађу из фондова Народне библиотеке Србије, овде су скенирани и материјали партнерских установа које нису биле у могућности да саме изврше дигитализацију. Поједине установе које су биле у могућности да тоураде по смерницама које су добиле од Народне библиотеке Србије, доставиле су готове дигиталне копије. Рад на дигитализацији је трајао око 2 године. Данас, дигитална колекција *Великог рата* броји 9.698 докумената свих категорија грађе, на 96.473 странице.

За унос и приказ дигитализоване грађе изабрана је платформа отвореног приступа Omeka v. 1.5.3.[6] Примењена шема за метаподатке је Dublin Core Extended.[7] Уз сваку унету јединицу грађе стоји одговарајући опис, који је преко COBISS ID броја повезан са библиографским описом у локалној бази Народне библиотеке Србије. Метаподаци су у систем делом преузети аутоматски, а делом су ручно уношени. Предност оваквог начина рада је у томе што су аутоматски преузети описи и подаци већ прошли библиографску контролу, чиме је смањена могућност нетачних и непроверених информација. Грађа која је на основу сарадње са другим институцијама унета у дигиталну библиотеку је такође описана у складу са важећим стандардима (Калезић и Михаиловић, 2014). Посао на уношењу грађе која се састоји од великог броја везаних дигиталних објеката, као што су књиге или периодичне публикације, олакшан је инсталацијом Dropbox апликације. Наиме, техничке могућности инсталиране верзије Omeke нису дозвољавале истовремени унос више дигитализованих страница, већ се морала

[6] Omeka v. 1.5.3 (приступљено 23.05.2019.)
[7] Dublin Core Extended (приступљено 23.05.2019.)

уносити свака засебно. Dropbox је омогућио да се у Omeku унесе читава фасцикла са свим дигитализованим страницама једног документа.

Шема метаподатака је посебно развијена за периодичне публикације. Основни опис за појединачни наслов је преузет из система COBISS, али за опис појединачних бројева периодике требало је унети додатне податке. Уз наслов појединачних бројева унет је и датум, а попуњавањем додатних поља у шеми метаподатака омогућена је детаљнија идентификација и повезивање са основним описом за читав наслов, па су попуњена и следећа поља: *Description* – у које су унете година и број, *Date* – где је унет датум и поље *Is part of* са линком који води ка опису целокупног наслова, тј. колекције.

С обзиром на то да апликација за календар није могла да буде инсталирана, а са циљем да се каснија претрага додатно растерети, искоришћене су Omeka *Tag* ознаке за месец и годину издања, које имају следећу структуру: римским бројем је уношена ознака за месец, а арапским за годину. Уз помоћ ових ознака постало је могуће да се уз опис појединачног наслова формира део *Доступна грађа*, о чему ће касније бити речи.

Метаподаци за картографску грађу су додатно обогаћени (Glišović and Gardašević, 2015). У жељи да се превазиђе језичка баријера и корисницима пруже додатне информације о дигитализованим мапама, донета је одлука да се у пољу под називом *Spatial Coverage* унесе URI.[8] У овом случају, као URI је искоришћен линк ка једној од географских база података под називом GeoNames.[9] Кликом на линк добијају се географски подаци о локацији, нпр. подаци о становништву, региону, географским координатама, различитим верзијама имена и сл. Због различитих ограничења, у претрази картографске грађе није могуће искористити све верзије имена географских локација из базе GeoNames, али за већину топонима је у пољу *Alternative Title* унет облик на енглеском језику и на тај начин олакшана претрага корисницима који се не користе српским језиком (Glišović and Gardašević, 2015).

У време када је писана документација за пројекат, крајем 2012. године, техничке могућности програма Omeka нису дозвољавале претрагу преко комплетног текста докумената. Свесни ове мањкавости, стручни тим се потрудио да на поменуте начине унеколико надокнади овај недостатак.

[8] URI – Uniform Resorce Identifier
[9] GeoNames (приступљено 10.04.2019.)

4. Претраживање и приказ резултата

Сва грађа у дигиталној библиотеци *Велики рат* је разврстана у колекције ради лакшег сналажења и прегледнијег приказивања података. Приликом формирања колекција примењени су и формални и садржински принцип. Формални је пратио основну поделу грађе према њиховој врсти, а садржински се показао као нужан. Тако је грађа подељена на следеће колекције: књиге, периодика, рукописна грађа, плакати, слике, картографска грађа и разно. Међутим, чланови стручног тима су, на основу својих искустава у раду са другим дигиталним библиотекама, али и кроз истраживања историјске грађе, сматрали да би било корисно направити потколекције, јер би колекције биле теже претраживе у случају да је формални принцип расподеле грађе остао и једини. Тако, на пример, у колекцији *Књиге* имамо поделу на неколико потколекција, од којих су неке: *Књижевност, Правни прописи, Велики рат Србије, Библиотека Напред*. Ово је урађено са намером да се на једном месту групишу сродне публикације и корисницима омогући лакше сналажење и упоредна анализа појединих наслова.

На насловном екрану се налази поље за основну претрагу по кључној речи. Такође, корисници могу самостално да се крећу кроз појединачне колекције. Уз појединачне наслове периодичних публикација, на десној страни екрана се налази део *Доступна грађа*, где се преко наведених година и месеци директно може приступити грађи из одговарајућег периода.

Међутим, захваљујући могућностима напредне претраге, корисници могу значајно сузити критеријуме за претрагу, практично по свим елементима описа једног дигитализованог објекта, као што су, на пример: година издања, аутор, врста грађе итд. Даље сужавање претраге може да се изврши захваљујући филтерима за колекције или поменутих *Tag* ознака. За претраживање помоћу *Tag* ознаке треба водити рачуна о формату упита који мора да се поклапа са структуром ових ознака, па тако, приликом претраживања, ознака месеца мора бити унета римским бројем, година арапским, а одвајају се зарезом нпр. II, 1915.

На постављени упит, добија се страница са погоцима који одговарају постављеним параметрима за претрагу са малом сликом и насловом са годином издања. Кликом на било који од ова два, отвара се страница са детаљним описом тражене публикације и умањеним приказом прве стране дигитализоване грађе. Кликом на слику отвара се дигитални

објекат и, у зависности од врсте грађе, омогућено је даље листање преко стрелица, али и прескакање на жељену страну. Грађу је могуће зумирати, а десни клик омогућава да се страница сачува као слика на рачунару корисника. Општим условима за коришћење грађе[10] су дефинисани начини њене употребе.

5. Додатни садржаји

Од самих почетака рада на овом пројекту, постојао је план да *Велики рат* буде коришћен као помоћно средство у школама. На идеју да се уз дигиталну библиотеку формира и одељак *За наставу*, дошло се захваљујући појединим члановима стручног тима који су имали вишегодишње искуство у просвети. Кроз овај сегмент наставницима и ученицима је омогућено да на једном месту пронађу и као наставно средство користе материјале везане за Први светски рат. Наставници ту могу да пронађу педагошке савете, образовне стандарде за крај основношколског образовања из области историје, препоручену грађу по темама и саме припреме за часове са линковима ка грађи на порталима *Велики рат* и *Europeana*. Иако је тематски уско везана за Први светски рат, ова дигитална библиотека може да, поред очигледне примене у настави историје, има своју примену и у настави српског језика и књижевности, верској настави, ликовном образовању, али и другим предметима (Ковачевић и др., 2016).

Да би се овај сегмент портала *Велики рат* додатно промовисао осмишљен је семинар „Дигиталне библиотеке посвећене Првом светском рату као наставна средства“. Семинар је акредитован од стране Завода за унапређивање образовања и васпитања и одржан је у пет наврата током 2014. и 2015. године. Полазници семинара су били махом наставници и професори историје из основних и средњих школа. Током семинара су упознати са колекцијама и могућностима претраживања најзначајнијих дигиталних библиотека посвећених Првом светском рату, са посебним освртом на *Велики рат*, као и начинима да се ови садржаји приближе ученицима и код њих подстакну истраживачки дух.

Виртуелне изложбе су још један сегмент којим се грађа из дигиталне библиотеке презентује корисницима и пружа могућност упознавања са догађајима, темама и личностима које су на свој начин обележиле овај

[10] Општи услови за коришћење грађе (приступљено 27.03.2019.)

период. Актуелна изложба на порталу је *Црњански у рату*, која је уједно и једина за сада.

На почетној страни се налази и рубрика *На данашњи дан*, у којој се аутоматски и насумично приказује грађа означена одговарајућим датумом.

Међу додатним садржајима се издваја интерактивна мапа преко које корисници могу да, помоћу геовизуелизације на мапи света, пронађу место издања одређене грађе за време и непосредно по окончању рата. Интерактивна мапа је израђена на Google платформи. Одабиром региона или града, долази се до директног линка ка публикацији, а употребом филтера за годину издања, врсту грађе и језик публикације, претраживање преко мапе је могуће додатно ограничити (Калезић и Михаиловић, 2014).

6. Резултати

После нешто више од шест година постојања и рада дигиталне библиотеке *Велики рат*, захваљујући алату *Google Analytics*[11] могуће је извршити анализу статистичких података о корисницима и коришћењу и остварeним резултатима. *Google Analytics* омогућава праћење броја корисника, њихову географску локацију, дужину трајања појединачне посете, број страница које су корисници прегледали приликом једне посете и разне друге параметре.

Тако је од промоције Великог рата до 31.12.2018. године укупан остварени број посета износио 102.550. Укупан број појединачних корисника је 58.588, а од тог броја је 17,7% сталних корисника. Током овог периода прегледано је укупно 2.081.557 страница, тј. нешто више од 20 страница по посети. Просечно трајање једне посете је око 6 минута. У табели 1 је дат преглед ових података по годинама, почев од 28.2.2013. до краја 2018. године.

У прва три месеца текуће године 2.238 корисника (од тога 17,9% сталних) су остварили 3.332 појединачне посете и прегледали 58.763 стране, уз просечну дужину трајања посете од 5.16 минута.

По географској локацији корисника, као што је и очекивано, највећи број их је из Србије и земаља региона, али међу првих 10 земаља по

[11] Google Analytics је бесплатан сервис компаније Google који служи за статистичка праћења посета веб-сајту (приступљено 10.04.2019.)

Година	**Број корисника**	**Број остварених посета**	**Број посећених страница (по посети)**	**Просечно трајање посете (у минутима)**
2013	7.169	14.571	295.229	6,01
2014	11.797	24.465	657.023	8,09
2015	10.792	18.222	348.851	5,37
2016	10.338	15.605	274.453	5,07
2017	7.635	12.921	222.684	5,04
2018	10.854	16.766	283.317	4,46

Табела 1. Подаци о коришћењу дигиталне библиотеке *Велики рат*

броју корисника, по подацима Google Analytics, налазе се и земље које су у Првом светском рату биле главне учеснице, на обе стране.

Остали демографски подаци које је могуће анализирати су пол и старост корисника. Од укупног броја корисника 54,15% их је мушког пола, а 45,85% женског. Што се старости тиче, највећи број, њих 33,5%, припада групи која има између 25 и 34 године, а следе их корисници старости између 18 и 24 године – 27,5%. Удео корисника старости од 35 до 44 године је 15,5%, од 45 до 54 је 12,5%, а оних од 55 до 64 и оних преко 65 године старости је по 5,5%.

Google Analytics прати и податке о уређајима коришћеним за приступ *Великом рату*. Највећи број посета је обављен са стоних рачунара, тј. око 75% свих посета, са мобилног телефона око 22% и тек око 3% посетилаца користи таблет.

Подаци о коришћењу грађе, у периоду 28.3.2013. – 31.3.2019, по Google Analytics, показују да је највише посета имала монографија *Ратни албум 1914–1918*, Андре Поповића, објављена 1926. године са 10.087 појединачних посета. Следе је прва и четрнаеста књига колекције *„Велики рат Србије за ослобођење и уједињење Срба, Хрвата и Словенаца“* са 8.651, односно 4.779 посета. Ова колекција, издавана од 1924. до 1937. године, у којој су објављена документа Врховне команде Краљевине Србије, представља непроцењив извор за сваког ко се бави проучавањем периода Првог светског рата, па стога не изненађује посећеност појединачних књига.

Слика 1. Географска локација корисника (28.3.2013. – 31.3.2019.)

Међу посећенијим страницама је и албум са сличицама – *Светски рат у сликама : 1914–1918 : Млечна чоколада Шонда* из тридесетих година прошлог века који бележи 4.165 посета. Од периодичних публикација најпосећеније је пето годиште часописа Жена из 1918. – 3.661 посета. Разгледница из Француске *Serbie : la revanche ou la mort* има 3.071 посету, а разгледница са описом „Поднаредник Момчило Гавриђ од 10 година, којему су родитеље убили аустријски војници, излази на рапорт са својим другом пред официра“ – 2.987.

Монографска публикација *Трагични дани Београда,* аутора Јована Миодраговића из 1915. године, је најпосећенија публикација објављена током ратног периода са 2.953 посете.

7. Закључак

Анализирани подаци о корисницима и коришћењу портала *Велики рат* показују да је ова дигитална библиотека, са просечним бројем месечних посета од око 900, веома вредан и значајан извор за проучавање периода Првог светског рата. Старосна структура корисника показује да све више њих даје предност лако доступним и претраживим дигиталним изворима информација у односу на оне традиционалне. Све ово указује на потребу да стручни тим настави и даље одговорно да ради на развоју и промоцији портала *Велики рат*. Иако је основни пројекат завршен, рад на унапређивању и обогаћивању се наставља, како континуираним додавањем скениране грађе, тако и одржавањем предавања на којима,

Земља	**Број корисника**	**Број остварених посета**	**Број посећених страница (по посети)**	**Просечно трајање посете (у минутима)**
Србија	41.524	79.551	21,36	6,12
БиХ	3.017	4.075	21,08	6,32
САД	1.552	1.954	8,16	2,20
Црна Гора	1.103	1.752	21,43	6,33
Хрватска	1.059	2.135	24,78	10,03
Немачка	903	1.179	13,62	3,39
Бугарска	888	1.158	19,64	4,40
Француска	724	866	9,68	2,58
Велика Британија	707	819	8,75	2,35
Македонија	704	1068	18,94	5,34

Табела 2. Преглед броја корисника по географској локацији и подаци о коришћењу дигиталне библиотеке *Велики рат* (28.2.2013 – 31.3.2019.) – првих 10 места

у сусрету са корисницима, постојећим и новим, тим долази до важних сазнања и идеја за будућност портала.

Литература

Glišović, Jelena and Stanislava Gardašević. "Cartographic Collection of National Library of Serbia throughout History until the Digital Present". *e-Perimetron* Vol. 10, no. 2 (2015): 73–86. Преузето 01.04.2019, http://www.e-perimetron.org/Vol_10_2/Glisovic_Gardasevic.pdf

Калезић, Немања. "Дигитална библиотека 'Велики рат'". Српске студије". Vol. 5 (2014): 389–391.

Калезић, Немања и Наташа Михаиловић. "Народна библиотека Србије, Интерактивни портал Велики рат". Нови Сад: Заједница матичних библиотека Србије, Градска библиотека, 2014, 147–150.

Ковачевић, Огњен, Наташа Михаиловић и Катарина Стаменић-Станојевић. *Приручник за примену дигитализованих историјских извора у настави*. Београд: Народна библиотека Србије, 2016.

Мултимедијални документ „Књижевност на филму“

УДК 004.55:378.147]:02(497.11)

Катарина Меглић
meglic.katarina@gmail.com
Филолошки факултет
Универзитет у Београду
Београд, Србија

САЖЕТАК: У оквиру предмета Мултимедијални документи, студенти четврте године на Катедри за библиотекарство и информатику Филолошког факултета Универзитета у Београду, школске 2017/2018 године обрађивали су тему „Књижевност на филму“. Студенти су имали прилику да знања стечена током школовања практично примене у оквиру овог пројекта. Прикупљен, класификован, обрађен материјал и дизајн обједињени су у овом мултимедијалном документу чији је акценат на екранизовању књижевних дела узимајући у обзир домаћу кинематографију.

КЉУЧНЕ РЕЧИ: мултимедијални документ, библиотекарство и информатика, књижевност, филм, екранизација књижевних дела.

РАД ПРИМЉЕН: 28. август 2019.
РАД ПРИХВАЋЕН: 10. септембар 2019.

„Ви сигурно знате ону причу о две козе које су управо јеле ролне филма прављеног према једном бестселеру, и једна коза каже другој: Ја више волим књигу.“

Алфред Хичкок

1. Увод

Књижевност *је реч која изазива помисао на* **књиге**. „Књижевност је врста уметности чије је изражајно средство реч. За исказивање мисли, осећања, карактера, доживљаја и догађаја књижевност се служи

језиком. Најпростије речено, књижевност означава сва писана дела која имају естетичку вредност.“ (Simon and Delyse, 2014, 2)

Филм *јесте документарно или играно дело снимљено на траци* (Pars pro toto, 2001, 2). Филм је врста уметности која се служи сопственим језиком, граматиком и стилом. Са усклађеним редоследом фотографских призора у кретању који при пројекцији стварају илузију реалности, филм поред сопствених изражајних средстава користи и средства других уметности попут књижевности, сликарства, позоришта и музике (Што је филм, 2019).

У Србији је први филм приказан јуна 1896. године, у кафани „Код златног крста“, док је први српски филм „Живот и дело Бесмртног Вожда Карађорђа“ снимљен далеке 1911. године.[1]

Иако се према пореклу и трајању не могу поредити, **књижевност и филм** преплићу се узајамно. Речи и слике чине величанствен спој књижевности и кинематографије. Данас стојимо опчињени пред остварењима која прожимају књижевност и филмску уметност и указују на њихов комплексан однос.

Предмет Мултимедијални документи на Катедри за библиотекарство и информатику, уведен је школске 2009/2010. године залагањем професора, ментора др Цветане Крстев. У оквиру овог предмета студенти су били у могућности да стечена знања из области библиотекарства, музеологије и архивистике сучељавају са знањима из области информатике, програмирања, база података, дигиталног текста итд. Комбиновањем свих аспеката настао је, девети по реду мултимедијални документ, под називом „Књижевност на филму“.

2. Израда мултимедијалног документа

Појам *мултимедијалан, -лна, -лно* Клајн и Шипка дефинишу као појам који комбинује текст, графику, филм, видео и музику тако да делују као целина (Клајн и Шипка, 2006, 799). Према Студији Хоиц-Божић, човек је у стању да запамти око 20% података ако их је само чуо, 40% ако их је и видео и чуо, те 75% ако их је видео, чуо и активно користио (Шта је мултимедија и хипермедија, 2019, 6). Имајући у виду да живимо у времену мултимедија, морали смо добро да размислимо како да на интерактиван, едукативан и забаван начин укрстимо стечена

[1] Више можете прочитати у делу „Кинематографија у Србији 1896-1941“ Стевана Јовичића, доступном на вебу (приступљено: 29.6.2019)

знања и садржај представимо другима. Дакле, како да интерактивно представимо информације.

2.1 Проналажење грађе

Први сегмент израде овог мултимедијалног документа односио се на то да студенти пронађу и попишу што више књижевних дела и њихових домаћих филмских остварења. Током проналажења и пописивања, асистент Бранислава Шандрих направила је табелу у оквиру Гугл диска како би сви студенти могли да приступе табели и попуњавају је у складу са текућим и предстојећим задацима (слика 1).

B	C	D	E	F	G
Naziv knjige	**Naziv domaćeg filma**	**1. zadatak (20. X)**	**2. zadatak (3. XI)**	**3. zadatak (1. XII 2017)**	**ID**
Daleko je Sunce	Daleko je Sunce	Katarina Meglić	Aleksandra Bakić	Aleksandra Bakić	daleko_je_sunce
Kiklop	Kiklop	Kristina Milojević	Ana Đoković	Aleksandra Bakić	kiklop
Maratonci trče počasni krug	Maratonci trče počasni krug	Katarina Meglić	~~Nastasija Jovanović~~	Aleksandra Bakić	maratonci_trce_pocasni_krug
Nikoletina Bursać	Doživljaji Nikoletine Bursaća	Kristina Milojević	Kristina Milojević	Aleksandra Bojković	dozivljaji_nikoletine_bursaca
Posljednji Stipančići	Posljednji Stipančići	Aleksandra Bojković	~~Ana Marija Nedeljku~~	Aleksandra Bojković	posledngji_stipancici
Prokleta Avlija	Prokleta Avlija	Tatjana Vučković	Milica Nešić	Aleksandra Bojković	prokleta_avlija
Dervis i smrt	Dervis i smrt	Mladen Bajić	Anđela Bibić	Ana Đoković	dervis_i_smrt
Glasam za ljubav	Glasam za ljubav	Aleksandra Bojković	Aleksandra Bojković	Ana Đoković	glasam_za_ljubav
Ranjeni orao	Ranjeni orao	Katarina Meglić	Milica Nešić	Ana Đoković	ranjeni_orao
Marš na Drinu	Marš na Drinu	Bojana Bratić	~~Nastasija Jovanović~~	Ana Đorđević	
Ne okreci se sine	Ne okreci se sine	Mladen Bajić	Gavrilo Milešević	Ana Đorđević	ne_okreci_se_sine
Nož	Nož	Kristina Milojević	Kristina Milojević	Ana Đorđević	noz

Слика 1. Табела за унос пронађене грађе и задужења

2.2 Класификација података

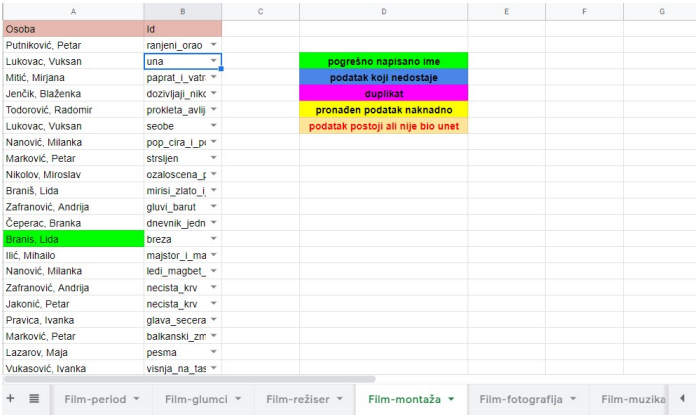
Након пронађене 152 књиге и њима одговарајућих филмова други сегмент био је упоређивање наслова књига са називима филмова, навођење аутора књиге и режисера филма, навођење године првог издања књиге и године када је филм снимљен, проналажење и постављање трајног линка записа за књигу у Узајамном каталогу Републике Србије COBIB.SR.

Ови подаци морали су бити унети у табелу како би студенти отпочели израду индекса.

2.3 Индекси

Израда индекса била је веома битан део рада на пројекту јер су се каснији задаци заснивали управо на њима. Први корак био је

упоређивање и усаглашавање наслова књиге и назива филма. Други корак био је формирање назива индекса који се уписивао у табелу како би студенти који су имали додатна задужења могли лакше да уносе податке у табелу. Индекси су се распоређивали по рубрикама филм-период, филм-глумци, филм-режисер, филм-монтажа, филм-музика, филм-фотографија, књига-писац, књига-година првог издања, књига-дечја (слика 2).



Слика 2. Табела са унетим индексима

Након провере индекса требало их је унети у базу „интерне“ Википедије како би се на крају повезали са индексима на сајту.

2.4 Прикупљање података и метаподатака

Следећи корак био је прикупљање основних података и метаподатака. Што се тиче књига, основни подаци обухватали су:

– Назив публикације,
– Име и презиме аутора,
– Везу ка запису у COBIB.SR,
– Годину(е) издања,
– Редни број издања,
– Да ли постоје преводи и на које стране језике,

– Критике,
– Да ли је књига награђивана и
– Рубрика „Остало“ која је обухватала следеће – да ли је из књиге преузет цитат или позната реплика (реченица коју треба забележити), жанр књиге и слично.

Најпре су истакнута дела Стевана Сремца, Бранислава Нушића, Иве Андрића, Меше Селимовића – сва дела са којима су се студенти сретали у досадашњем школовању попут: *Дервиш и смрт, Нечиста крв, Проклета авлија, Орлови рано лете, Поп Ћира и поп Спира* итд.

Мање позната или непозната књижевна дела овој генерацији студената била су *Последњи Стипанчићи, Мирис, злато и тамјан, Дукат и дивљи пелин и Ближњи, док су више интересовања показали за књиге Мајстор и Маргарита, Лајање на звезде, Шешир професора Косте Вујића и Општинско дете.*

Студенти су одређене публикације морали да прочитају због пројекта како би прикупили валидне податке зарад што квалитетније обраде података. То су: *Бановић Страхиња, Дружба Пере Кржице и Добар ветар, Плава птицо.*

Када су у питању филмови, основни подаци односили су се на:

– Назив филма,
– Име и презиме режисера,
– Везу ка IMDB-у,[2]
– Година када је филм снимљен,
– Продуцентску кућу,
– Списак главних глумаца,
– Музику,
– Монтажу,
– Фотографију,
– Сценарио,
– Везу ка промотивном видео запису на сајту YouTube (уколико промотивни запис постоји),
– Везу ка чланку на сајту Википедија (уколико чланак постоји),
– Везу ка постеру,
– И рубрика „Остало“ у којој се налазе подаци – да ли је филм учествовао на неком фестивалу (Пулски фестивал, Нишки фестивал и сл.) и да ли је филм награђиван.

[2] IMDB (на вебу)

Због пројекта студенти су морали да погледају филмове унете у табелу, а настале током 50-их, 60-их и 70-их година прошлог века како би употпунили базу неопходним подацима. Највише недоумица било је у вези са филмовима сниманим у периоду од 1951. до 1960. године због недостатка података, али и њихове тачности.

Најпривлачнији део израде овог дела задатка студентима је био да провере да ли се наслов публикације подудара са називом филма. Нпр. филм *Свети Георгије убива аждаху* у режији Срђана Драгојевића рађен је по књизи Душана Ковачевића *Балкански змајеви*. Поред овога имамо: *Вецвудов прибор за чај* Милорада Павића екранизован под називом *Византијско плаво*, Николај Љесков и његова Леди Магбет Мценског округа приказана као *Сибирска Леди Магбет*, док је књига Излет у небо Гроздане Олујић филмована као *Чудна девојка* у режији Јована Живановића. Такође *Добри дух Загреба* продуциран је под насловом *Ритам злочина*, Папрат и ватра Антонија Исаковића као *Три*, док књига *Мемоарски записи Ивана Шибла* постаје филм *Кад чујеш звона*.

2.5 Додатна задужења студената

Када су подаци прикупљени, обједињени и обрађени, студентима су у трећем сегменту рада расподељена додатна задужења која су подразумевала проверу података и оформљених индекса.

Податке и индексе о аутору књиге проверавала је Јелена Аћимовић, наслов је контролисала Тијана Вучићевић, док је за годину издања књиге била задужена Настасја Јовановић.

Рубрику глумци прегледале су: Јована Андоновска, Тијана Савић и Сава Тепавчевић контролисале су рубрику глумци.

Наслов филма прегледала је Ана-Марија Недељку. Проверу режисера радила је Ана Ђорђевић, музику Милица Васковић, фотографију Марија Шарчевић и монтажу Катарина Меглић. Период у којем је филм настао прегледала је Кристина Милојевић.

Пошто су подаци проверени, и утврђено је да су прецизни и тачни, професорка Крстев је студенте поделила у групе.

2.6 Програмирање и дизајн веб-стране

У оквиру четвртог сегмента рада, тим за кодирање веб-стране чинили су Александра Бојковић, Бојана Братић, Никола Миловановић и Гаврило Милешевић. Задатак је био да на основу прикупљених и

обрађених података израде веб-сајт. Усагласили су се да за израду сајта користе HTML (HyperText Markup Language) и CSS (Cascading Style Sheets), без напреднијег програмирања. На овакав потез утицао је изглед ранијих пројеката. Претходни радови попут „Жил Верн: Пут око света за осамдесет дана“ (Перић et al., 2014), „ YU рок сцена“ (Обрадовић et al., 2016), „Пролазим улицом твојом...“ (Суботић et al., 2018) богати су анимацијама, сликама, текстовима и визуелним аспектом уопште, стога је донета одлука за класичан изглед веб-странице, са индексима повезаним путем хипервеза. Сматрали су да за обиље информација и велики број индекса треба применити далеко захтевније програмирање које превазилази њихове вештине и могућности. На заједничком састанку донета је одлука да студенти програмирају без помоћи колега са других факултета, али и без помоћи других институција.

У групи за Дизајн били су Анђела Бибић, Ана Ђоковић, Настасја Јовановић и Јелена Аћимовић. Задатак групе био је да осмисле изглед сајта. Групи за кодирање веб-стране доставиле су фотографије и одредиле боју за позадину сајта. За позадину је одабрана црна боја. Ненаметљива и елегантна, истиче текст. Фотографије које су коришћене у сврху овог пројекта су у формату .jpg, а обрађивале су се у програму LightShot,[3] углавном ради подешавања величине.

2.7 Новина у оквиру Мултимедијалних докумената – Квиз

Новина у односу на досадашње пројекте у оквиру предмета Мултимедијални документи јесте одељак „Квиз“.[4]

За рубрику „Квиз“ у групи су биле Катарина Меглић, Кристина Милојевић, Ана Кнежевић, Милица Нешић и Александра Бакић. Креативност у овој групи није мањкала. Након осмишљених питања, која су представљала лакши део задатка на ред је дошао избор програма за израду квиза. Поред Online Quiz Creator, Survey Monkey, ProProfs Quiz Maker, Hot Potatoes избор је сужен на FyreBox и iSpring QuizMaker програме.

Изузев FyreBox-а и iSpring QuizMaker-а, остали програми нису се учинили довољно добрим јер поред компликованог регистровања на сајт имали смо потешкоћа због ограниченог броја питања, заштите ауторских

[3] LightShot (на вебу)
[4] Квиз (приступљено: 1.8.2019)

права, суженог избора комбинација питања, али и доплате за разне додатке у оквиру програма.

Програм FyreBox[5] има образац по којем се израђује квиз. Предодређена је позадина, величина и врста слова. Да бисте осмислили квиз морате се регистровати на сајт. Могу се написати инструкције за кориснике и ограничити време играња, али за уметање аудио-визуелног садржаја захтева се доплата.

С обзиром на сва ограничења на која смо наишли, одлучено је да се користи програм iSpring QuizMaker 9,2 који се студентима учинио као интерактиван и интересантан, уз то и лако савладив с обзиром да је сличан PowerPoint-у.

Након лаке и брзе регистрације, комплетно упутство за инсталирање програма и израду квиза добија се на имејл.

Слике, видео или аудио записе је лако уметнути, а може се и подесити жељена позадина. Корисник сам бира да ли ће на крају квиза бити приказани резултати, односно да ли је и колико поена освојио процентуално.

Последња варијанта квиза „Игранка без престанка“ посебно се истиче због тога што представља варијацију свих питања, одговара се све док се дају тачни одговори, а након погрешног одговора тзв. „ланац питања“ се прекида и квиз се враћа на почетак.

Осмишљено је неколико комбинација квиза које су постављене на веб, корисник може приступити било којој варијанти и проверити своје знање (слика 3).

2.8 Повезивање садржаја и „интерна“ Википедија

„Обрада преосталих тема“ последња је, али не и мање битна група, у којој су били Јована Бобета, Невена Симоновић, Ана Ђорђевић, Андријана Михајловић и Јована Марковић. Ова група проверавала је, последњи пут, индексе – поређала их абецедно и у сарадњи са колегама из групе за кодирање веба повезала комплетан садржај.

Повезивање комплетног садржаја одвијало се на следећи начин: Студенти су користили шему као што је приказано на слици (слика 4). Након што су неопходни подаци и индекси унети, све је сачувано у формату .html како би се на крају све објединило у једном фолдеру.

Студенти са Катедре за библиотекарство и информатику су у оквиру предмета Дигитални текст похађали у Универзитетској библиотеци

[5] FyreBox (на вебу)

Слика 3. Једно од питања у квизу

„Светозар Марковић “ праксу која се односила на уређивање странице Википедије.

С обзиром да су већ имали искуства на овом пољу, идеја је била да направе нове чланке или уреде постојеће на Википедији, а потом их повежу са страницом коју израђују. Из техничких разлога није дошло до реализације те идеје, па смо уз помоћ асистента Браниславе Шандрих оформили „интерну“ Википедију.

„Интерна“ Википедија јесте платформа која је веома слична оригиналној Википедији и састоји се од прикупљених података.

На почетку се може унети наслов књиге или филма како би се добиле тражене информације. Након што се уђе у жељени одељак, при врху стоји име књиге и повлаком одвојен назив филма како би се разрешиле недоумице уколико се наслови књиге и филма разликују. Затим следе фотографије, информације како о књизи тако и о филму, и на крају занимљивости, анегдоте и цитати везани за књигу и филм. Док су се студенти у групама бавили својим задацима, истовремено су прикупљене податке и проверене индексе објединили и поставили на „интерну“ Википедију. Повезане су све странице и крајњи корак био је коначни изглед веб-стране.

Слика 4. Приказ шеме на основу коју се уносе индекси

Направљене су рубрике „Почетна страна“ са основним подацима и музиком у позадини.

Младен Бајић, задужен за одабир композиције пројекта, предложио је неколико песама. Студенти су једногласно одабрали песму „Флојд“. Сматрали су да та песма одговара задатој теми и представља одраз духа генерације – динамичност и жељу за успехом.

Поред одељка „Почетна страна“ налази се одељак „Књиге“ са подрубрикама „Писци“ и „Период“, потом следи рубрика „Филмови“ која се састоји од подрубрика „Глумци“, „Режисери“, „Фотографија“, „Монтажа“, „Музика“ и „Период“. Одељак „О-Пројекту“ са кратким информацијама о пројекту, као и рубрика „Квиз“ где се може разонодити и решавати квиз.

3. Закључак

Разлика овог мултимедијалног документа у односу на досадашње јесте у индексу који је кључан за претрагу садржаја. Као будући библиотекари, студенти су се определили за абецедни поредак индекса да би се посетиоци сајта лакше сналазили.

Рад на пројекту „Књижевност на филму“ студентима је представљао својеврсни изазов, али и непроцењиво искуство. Изазов јер је требало

употребити до сад стечено знање на студијама, а искуство јер као крајњи резултат стоји успешно остварен пројекат. Дилема књига или филм – читати или гледати, јавља се и код оних који више читају књиге и код оних који више гледају филмове.

Управо та дилема била је главни мотив студентима због тога што су кроз овај пројекат објединили познавање књижевности, кинематографије, библиотекарства и других разноврсних области које се тичу ове теме. Појединачан допринос свакога од нас обогатио је тимски рад генерације.

Иако је тема била комплексна, студенти су пажљивим одабиром информација и метаподатака, залагањем на изради мултимедијалног документа, приближили тему будућим генерацијама.

Изјаве захвалности

Није било лако младе људе усмерити, привући им пажњу и мотивисати их, и зато велику захвалност дугујемо професорки др Цветани Крстев и асистенту Бранислави Шандрих.

Својим саветима и помоћу допринеле су да превазиђемо препреке које је пројекат носио и научимо како да применимо стечено знање.

Уз неизмерну подршку и идеје, а заједничким залагањем и уједињавањем знања као резултат настао је пројекат „Књижевност на филму“ – који је доступан јавности на сајту професорке Крстев у одељку „Радови студената“.[6]

Литература

“Шта је мултимедија и хипермедија”, преузето 8.9.2019, http://poincare.matf.bg.ac.rs/~vladaf/Courses/PmfB1%20M%20MNR/Predavanja/mnr02-1-Multimedija%20i%20hipermedija.pdf

“Што је филм”, преузето 8.9.2019, http://www.kinovalli.net/sto-je-film

Pars pro toto : део за целину. Поглавље Филм. Београд: Центар за аудиовизуелне медије – Медија фокус, 2001, 1–12

Simon, Ryan и Ryan Delyse. *What is Literature*. Foundation: Fundamentals of Literature and Drama, 2014, преузето 2.12.2018, https://resource.acu.edu.au/siryan/academy/Instructions%20page2.htm

[6] Сајт професорке Крстев (приступљено: 09.04.2019)

Јовичић, Стеван. “Кинематографија у Србији 1896-1941”, 2010, преузето 29.6.2019, http://www.suedslavistik-online.de/02/jovicic.pdf

Клајн, Иван и Милан Шипка. *Велики речник страних речи и израза*. Нови Сад: Прометеј, 2006

Обрадовић, Милена, Александра Арсенијевић и Михајло Шкорић. “Израда мултимедијалног документа ‘YU рок сцена‘”. *Инфотека* Vol. 16 (2016), no. 1: 113–129, преузето 8.9.2019, http://infoteka.bg.ac.rs/pdf/Srp/2016/infoteka-2016-16-1_2-6.pdf

Перић, Катарина, Кристина Гогић и Ана Николић. “Израда мултимедијалног документа „Пут око света за 80 дана”“. *Инфотека* Vol. 15 (2014), no. 2: 56–62, преузето 8.9.2019, http://infoteka.bg.ac.rs/pdf/Srp/2014-2/Srp2014-2INFOTHECA_XV_2_april_56-62.pdf

Суботић, Слађана, Филип Станковић, Сања Сланкаменац, Растислав Марковић, Анастасија Мандић et al.. “Израда мултимедијалног документа „Пролазим улицом твојом“”. *Инфотека* Vol. 18 (2018), no. 1: 88–97, преузето 8.9.2019, http://infoteka.bg.ac.rs/pdf/Srp/2018/infoteka-2018-18-1-6.pdf

Публика у фокусу – демократизација дигитализације у библиотекама

Јелена Андоновски
andonovski@unilib.rs

Никола Крсмановић
krsmanovic@ubsm.rs

Универзитетска библиотека „Светозар Марковић“ Београд, Србија

РАД ПРИМЉЕН: 5. јуни 2019.
РАД ПРИХВАЋЕН: 8. јули 2019.

Дигитализацијом је, у најопштијем смислу, до сада омогућено трајно чување само мањег дела материјала културног наслеђа, како у свету, тако и у Србији. Још мањи проценат овако сачуваних материјала је обрађен и стога ефективно доступан и пријемчив корисницима. До сада су институције контролисале процесе дигитализације у оба сегмента и тиме дефинисале токове креирања информација на основу употребе дигиталних материјала културног наслеђа. Технолошке могућности данас чине да процесе дигитализације могу равноправно контролисати и појединци, чиме се отварају бројна питања везана за дигитализацију књиге и културног наслеђа. Универзитетска библиотека „Светозар Марковић“ у Београду придружени је партнер пројектног тима H2020 READ[1] (Пројекат за препознавање и обогаћивање архивских докумената) који се реализује од 2016. до јуна 2019. године. Циљ пројекта је развијање платформе за аутоматско препознавање, транскрипцију и претраживање историјских докумената (Транскрибус) која је намењена истраживачима хуманистичких наука, архивистима, библиотекарима, као и стручњацима из области информационих технологија. Основа и главни предуслов за коришћење услуга у оквиру Транскрибуса јесте дигитализација документа. У циљу поједностављивања и убрзавања процеса дигитализације истраживачи у

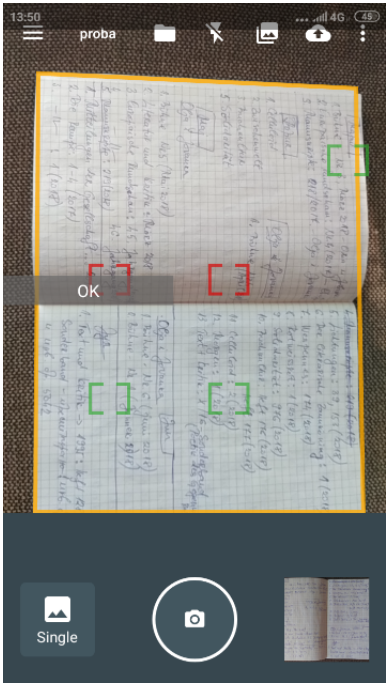
[1] Recognition and Enrichment of Archival Documents, на вебу

лабораторији Computer Vision Lab на Техничком универзитету у Бечу[2] развијају мобилну апликацију DocScan и преносни уређај ScanTent. Апликација DocScan је отвореног кода и доступна је на платформи GitHub, као и на GooglePlay продавници, садржи приказ уживо, детектује страницу документа у реалном времену, врши процену квалитета снимка (прилагођавање фокуса) и омогућава режим рада у серији, односно аутоматско фотографисање приликом окретања страница. На овај начин је учињен значајан помак у демократизацији дигитализације будући да путем ове апликације мобилни андроид телефон постаје уређај за скенирање. Са друге стране, преносни уређај ScanTent функционише као постоље за мобилни телефон када је у режиму рада скенирања.

Делови програма H2020 READ реализовани су у Универзитетској библиотеци „Светозар Марковић“ у Београду у оквиру пројеката које је финансирало Министарство културе и информисања Републике Србије: „Нови хоризонт дигитализације“ за 2016. годину, „Рашчитана стара српска ћирилица: оживљена руком писана прошлост“ за 2017. годину, „Рашчитаност старе српске ћирилице: историја и традиција надохват руке“ за 2018. Поред тога, у сарадњи са Градском библиотеком у Новом Саду Универзитетска библиотека реализовала је пројекат „Демократизација дигитализације у библиотекама“, подржан кроз јавни конкурс Фондације „Публика у фокусу“ средствима опредељеним из буџета АП Војводине за 2017. годину.[3] Пројекат је имао за циљ обуку библиотекара, али и укључивање што већег броја корисника у процес дигитализације библиотечке грађе како би библиотечка и архивска грађа постале доступније. У току октобра, новембра и децембра 2018. године, као и у току марта, априла и маја 2019. године одржано је преко 30 радионица посвећених упознавању полазника са основним концептима и идејама о демократизацији дигитализације, као и технологијама које ово омогућавају. Радионице су одржане у Универзитетској библиотеци у Београду и Градској библиотеци у Новом Саду, укључујући и њене огранке. Радионице је одржао тим Универзитетске библиотеке за развој и унапређење нових технологија у области дигитализације. У децембру 2018. године (20. и 21. децембра) одржане су две централне националне радионице у Универзитетској библиотеци „Светозар Марковић“ у Београду и Историјском архиву у Новом Саду на којима су полазници

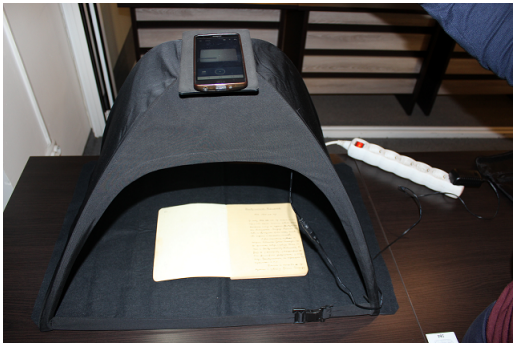
[2] TU Wien, Faculty of Informatics, Institute of Visual Computing & Human-Centered Technology, Computer Vision Lab на вебу

[3] О реализацији пројекта „Демократизација у дигитализацији“ више на вебу и на вебу



Слика 1. Мобилна апликација DocScan

.



Слика 2. Преносни уређај ScanTent

.

имали прилике да раде са сопственом опремом и чују предавања др Гинтера Милбергера (Günter Mühlberger), руководиоца паневропског тима за развој нових технологија у дигитализацији у оквиру пројекта READ.

Слика 3. Национална радионица у Београду

.

Слика 4. Национална радионица у Новом Саду

.

Даљи рад и усавршавање за коришћење поменутих технологија наставља се у Универзитетској библиотеци у Београду у оквиру

пројекта „Раскош потпуно рашчитане српске ћирилице: живо рукописно наслеђе од Бранковића до Обреновића“ које финансира Министарство за културу и информисање за 2019. годину и у оквиру акредитованог програма за стручно усавршавање библиотекара „Демократизација дигитализације у библиотекама“ аутора Наташе Дакић, Александре Тртовац и Јелене Андоновски.[4] На Европском нивоу, након завршетка пројекта H2020 READ биће основана паневропска Колаборатива како би се омогућила дугорочна одрживост резултата пројекта.

[4] Више о акредитованом курсу „Демократизација дигитализације у библиотекама“ на вебу

Прелазак на каталогизацију са нормативном контролом у систему COBISS.SR

Светлана Пуцаревић
svetlana.pucarevic
@unilib.bg.ac.rs

Сњежана Фурунџић
snjezana.furundzic
@unilib.bg.ac.rs

Универзитетске библиотеке „Светозар Марковић“ Београд, Србија

РАД ПРИМЉЕН: 6. мај 2019.
РАД ПРИХВАЋЕН: 17. мај 2019.

После система COBISS.SI (Словенија) и COBISS.BG (Бугарска) и у систему COBISS.SR (Србија) уведена је каталогизација са нормативном контролом. У нашем систему тако се успоставља нормативна контрола личних имена, а предвиђена је у неким наредним корацима и нормативна контрола колективних тела и предметних одредница. Нормативна датотека личних имена у домаћем систему зове се CONOR.SR. Преко базе CONOR.SR врши се повезивање библиографских и нормативних записа. База садржи нормативне записе с подацима о личним именима аутора који су јединствено одређени за употребу у целом ситему каталогизације и то је јединствена база коју користе и надограђују све библиотеке које учествују у систему. Поцес преласка на каталогизацију са нормативном контролом подразумево је припремне кораке који су предузеле Библиотека Матице српске, Народна библиотека Србије и Универзитетска библиотека „Светозар Марковић“, као библиотеке оснивачи и пуноправне чланице узајамног система COBISS.SR. Иницијална база CONOR.SR формирана је 2013. године и у периоду 2013. до 2018. године вршена је редакција записа у бази. Потпуни прелазак на нов начин рада подразумевао је следеће кораке:

Једнодневни курс за каталогизаторе

Током 2018. године обављена је обука предавача из три поменуте библиотеке и образовање каталогизатора из свих библиотека у систему

COBISS.SR. У Универзитетској библиотеци „Светозар Марковић” предавачи су одржали 12 курсева за око 250 каталогизатора из универзитетских и факултетских библиотека у Београду, Нишу и Крагујевцу. Рад у нормативној бази је хијерархијски и основну привилегију имају сви запослени који раде на каталогизацији било које врсте грађе.

Курс за редакторе нормативних записа

У Народној библиотеци Србије организован је курс за ажурирање података у нормативним записима намењен тзв. „средњим“ редакторима, који су добили више привилегије. Највише привилегије добили су тзв. „топ“ редактори.

Ажурирање постојеће базе CONOR.SR

Година 2018. и почетак 2019. били су кључни период у погледу редакције записа и брисања дупликата у постојећој бази CONOR. Посебна пажња усмерена је на утврђивање и брисање дупликата који касније неће моћи да се бришу јер је програм тако подешен. Брисање записа након успостављања CONOR-а значило би да одређен библиографски запис који је повезан за конкретан нормативни запис том приликом остаје без одреднице (нормативне приступне тачке) и у испису у OPAC-у не би био коректан.

Важно је било да се редигује што више записа за имена личности које су значајне за нашу културу, као и записи за ауторе за чија имена је повезан највећи број библиографских записа.

Универзитетска библиотека „Светозар Марковић“ била је задужена за редакцију нормативних записа домаћих истраживача са чијим именима је повезан највећи број библиографских записа у систему COBISS. Разрешавањем и брисањем великог броја дуплираних записа у CONOR-у база је срећена што је чини једноставнијом за употребу и повезивање нормативних и библиографских записа.

Тестна база за повезивање библиографских и нормативних записа успостављена је почетком априла 2019, када су извршене провере и вежбе поступка рада пре самог преласка на праву базу.

У периоду од 13. до 14. априла 2019. ИЗУМ (Институт Информацијских Знаности, Марибор) је извршио инсталацију програмске опреме за нормативну контролу и повезивање CONOR-а

са COBIB-ом и локалним базама података. Од 15. априла 2019. база CONOR је активна и у нашем систему COBISS.SR.

Сви обучени каталогизатори за каталогизацију са нормативном контролом прешли су на нови начин рада. Приликом креирања записа нема више ручног уноса одреднице у блок 7XX, већ се запис за аутора:

- повлачи из базе CONOR ако одговарајући запис већ постоји
- креира и похрањује у базу CONOR уколико запис за аутора не постоји.

Због различитих потреба библиотека у систему COBISS и различитог вођења одредница, база CONOR.SR садржи парове одредница, односно нормативних приступних тачака. Прва нормативна приступна тачка се увек наводи ћирилицом, а њен пар латиницом у паралелним пољима 200. На примерима то изгледа овако:

Пример 1. – домаћи аутор

200 1 <7>cb - ćirilica - srpska <a>Црњански <b>Милош <f>1893-1977
200 1 <7>ba – latinica <a>Crnjanski < b>Miloš <f>1893-1977

Пример 2. – страни аутор чије се име изворно пише латиницом

200 1 <7>cb - ćirilica – srpska <a> Ирби <b> Аделина Паулина < f> 1831-1911
200 1 <7>ba – latinica <a> Irby <b> Adelina Paulina <f> 1831-1911

Пример 3. – страни аутор чије се име изворно пише ћирилицом

200 1 <7>ca - ćirilica - nije specifikovana <a>Достоевский <b>Фёдор Михайлович< f>1821-1881
200 1 <7>ba - latinica <a>Dostoevskij <b>Fedor Mihajlovič <f>1821-1881

За ове ауторе, увек је неопходан унос и упутнице, односно варијантне приступне тачке у пољу 400.

400 1 <7> cb – ćirilica - srpska <a> Достојевски <b>Фјодор Михајилович <f>1821-1881

Тренутан број записа у нормативној бази CONOR.SR око 52.000 записа.[1]

Потпуну синхронизацију система и повезивање библиографских записа који су креирани пре 15. априла 2019. са базом CONOR, извршиће Институт информацијских знаности (ИЗУМ) до краја јуна 2019.

Рад са нормативним записима предуслов је за повезивање података из локалних и узајамног каталога са системом Отворених повезаних података (Linked Open Data). Верујемо да ћемо бити успешни у новом процесу рада и да ће нормативна контрола унапредити и квалитет библиографских записа и њихову унифицираност.

[1] Стање на дан 25.4.2019.

УПУТСТВО ЗА АУТОРЕ

Сви радови у часопису *Инфотека* објављују се на енглеском и на српском језику у истом издању. Аутори треба да доставе своје радове на једном од ова два језика; тек након обавештења о прихватању рада, очекује се да аутор пошаље превод (за српске ауторе; за све друге ауторе часопис ће обезбедити превод са енглеског језика на српски језик). Осим у штампаном издању, сви радови се обајављују и онлајн у отвореном приступу.

КАТЕГОРИЗАЦИЈА РАДОВА

За документе прихваћене за објављивање и који подлежу рецензији примењује се следећа категоризација у часопису:

1. Научни чланци:
   - Оригинални научни рад (са претходно необјављиваним резултатима сопствених истраживања научном методом);
   - прегледни рад (који садржи оригиналан, детаљан и критички приказ истраживачког проблема или подручја у коме се допринос аутора показује и аутоцитатима);
   - претходно саопштење (оригинални научни рад прелиминарног карактера и мањег обима);
   - научна расправа и осврти на одређену тему заснована на научној аргументацији.
2. Стручни чланци који представљају искуства корисна за унапређење професионалне праксе
3. Информативни прилози могу бити:
   - уводници и коментари;
   - прикази књига, рачунарског програма, база података, стандарда и сл.;
   - научног догађаја, јубилеја.

Радови класификовани као научни морају имати бар две позитивне рецензије. Ставови Редакције не морају се подударати са ставовима у објављеним радовима. Рад се не може прештампавати нити објавити под сличним насловом нити у измењеном облику.

ЕЛЕМЕНТИ РУКОПИСА

За научне и стручне радове, потребно је доставити следеће податке:

1. Радови не би требало да буду дужи од 15 страна А4 формата, писаних врстом слова Times New Roman величине 12pt. У случају дужих радова, потребно је ступити у контакт са уредницима часописа.
2. Имена и презимена свих аутора требало би да буду написана оним редом којим ће се појавити у објављеном раду.
3. Пуно име једног или више аутора, без назнака титула и диплома, треба да буде достављено уз e-mail адресу, као и уз пун, званичан назив институције којој припадају (у сложеним организацијама се наводи укупна хијерархија назива, одозго надоле).
4. Наводи се и датум слања рада.
5. Аутори треба да предложе категорију свог рада, премда коначну одлуку о категоризацији доноси главни уредник.
6. Информативни сажетак, НЕ ДУЖИ ОД 200 РЕЧИ, који језгровито представља суштину рада, циљ истраживања и примењене методе и пружа главне закључке, треба да буде достављен уз рад. Сажетак треба написати на оба језика који се користе за објављивање часописа. У сажетку, аутори треба да користе термине који се као стандардни често користе за индексирање и претраживање чланака.
7. Аутори треба да доставе најмање 3 и не више од 10 кључних речи, одвојених зарезима, које означавају главне појмове представљене у раду. Списак кључних речи треба доставити на оба језика који се користе за објављивање часописа.
8. Ако рад произилази из мастер рада или докторске тезе, аутори треба да доставе назив тог рада или тезе, као и датум када су рад или теза предати и назив надлежне институције.
9. Ако рад представља резултат учешћа аутора у неком пројекту или програму, аутори би требало да укажу захвалност институцији која је била задужена за финансирање у посебном одељку под називом „Изјаве захвалности" на крају рада, пре одељка под називом „Литература". У истом том одељку требало би навести имена особа које су помогле у изради рада.
10. Ако је неки рад представљен на конференцији, али није изашао у зборнику радова са те конференције, то такође треба нагласити у засебној белешци.
11. Аутори могу да користе фусноте, док употреба напомена на самом крају текста није дозвољена; ипак, треба избегавати употребу предугачких фуснота. Аутори могу да убаце додатке у своје радове.
12. Реферисани материјал треба да буде излистан у одељку под називом „Литература" на крају рада. У списку реферисаних радова аутори треба да наведу све информације које су неопходне за проналажење радова на које се реферише. У овом одељку треба да се наведу све библиографске јединице реферисане у раду – такође, овде не треба да се појави ниједна библиографска јединица која у раду није реферисана.

ПРАВИЛА УРЕЂИВАЊА ПРИХВАЋЕНИХ РАДОВА

1. Препоручује се и подржава слање радове припремљених коришћењем пакета LaTeX, коришћењем стила часописа (стил часописа и сви потрени пакети могу се преузети са сајта часописа). Аутори који нису упознати са коришћењем овог пакета треба да припреме свој рад у програму Word (формати .docx, .doc, .rtf или .txt). Аутори који користе програм Word на треба на посебан начин да га форматирају – пребацивање у LaTeX ће обавити редакција часописа.
2. Радови писани на српском језику треба да буду откуцани ЋИРИЛИЧНИМ писмом јер ће на том писму бити и штампани. Изузетак су једино делови текста за које је коришћење другог писма, на пример латиничног, погодније. Све врсте писама треба представити користећи Unicode UTF-8 кодни распоред.
3. Наслов рада не би требало да буде написан великим словима. Дужина наслова треба да буде у разумним границама – пожељно је да то буде мање од 100 карактера. За све наслове, аутори треба да доставе скраћене верзије тих наслова које ће бити коришћене у заглављима.
4. Искошена слова (Italic) је могуће користити за истицање у тексту, док се подебљана слова (Bold) или искошена подебљана (Italic bold) могу користити ако је то неопходно. Употребу подвучених слова (Underlined) треба избегавати. Молимо да не истичете целе реченице, нити целе пасусе.
5. Рад може бити подељен у одељке и пододељке, али не треба користити више од два нивоа наслова одељака. Сви одељци ће бити на одговарајући начин нумерисани. Додаци треба да дођу на крај рада, а ако их има више и они ће бити означени. Ако се користе листе, не треба користити више од два нивоа угњежености.
6. Сви пасуси треба да буду раздвојени једним празним редом (једним притиском на дугме Enter).
7. Код припреме табела и слика треба водити рачуна да се радови штампају у A5 формату, тако да треба избегавати сувише широке табеле. Све илустрације треба припремити у неком од формата који сажимањем не губе на квалитету, на пример .png, .tif или .jpg и у резолуцији од најмање 300 dpi.
8. Молимо ауторе да, ако је то могуће, додају везу на екран са кога је неки снимак екрана узет. Препоручујемо да аутори приликом чувања снимка екрана, или дела екрана, користе Zoom опцију претраживача или неког другог програма. За дијаграме који су произведени помоћу програма Excel, молимо да доставите оригинални .xls документ.
9. Све табеле, илустрације, дијаграме и фотографије треба припремити као засебне датотеке, црнобеле за штампање, а у боји за онлајн верзију. Наслови испод табела, илустрација, дијаграма или фотографија треба да остану у

тексту. Свака датотека треба да носи исто име као главни текст, уз опис врсте материјала коме је додат редни број у тексту. На пример, датотека у којој се налази четврта по реду слика у раду под називом „Пример“ треба да буде именована Пример_слика_4.

10. Молимо да додате све потребне датотеке у којима су објашњени неки посебни аспекти форматирања Вашег рада.
11. Уколико се у раду појављују URL адресе веб страница на чији се садржај текст реферише, њих би требало наводити у фуснотама, уз обавезно навођење датума када је страници приступљено.

БИБЛИОГРАФИЈА И ЦИТИРАЊЕ

1. Коришћену литературу навести на крају текста, у оквиру ненумерисаног одељка Литература. Овај одељак треба да садржи сва дела наведена у тексту, и ништа више од тога. Реферисану литературу не треба наводити у фуснотама у оквиру самог текста.
2. Сва дела треба сложити по азбучном реду имена аутора, уредника или издавача (ако аутор није наведен), а уколико се јавља више дела истог аутора њих треба поређати у хронолошком редоследу.
3. За израду референци користити Чикаго стил – (Chicago Style) (`www.chicagomanualofstyle.org`).
4. Наводити пуне, а не скраћене наслове часописа или акрониме.
5. Сви аутори, без обзира да ли предају рукопис припремљен коришћењем LaTeX-a или Word-a, припремиће све библиографске јединице из одељка литература и у BibTeX према шаблонима који су наведене уз примере на сајту Инфотеке (`http://infoteka.bg.ac.rs/index.php/sr/upu-s-v-z-u-r`).