Универзитет у Београду

Филолошки факултет

Јелена С. Андоновски

МРЕЖА ОТВОРЕНИХ ПОДАТАКА И ЈЕЗИЧКИ РЕСУРСИ У ПРОЦЕСУ ИЗГРАДЊЕ СРПСКО-НЕМАЧКОГ ЛИТЕРАРНОГ КОРПУСА

докторска дисертација

Београд, 2019.

University of Belgrade

Faculty of Philology

Jelena S. Andonovski

LINKED OPEN DATA AND LANGUAGE RESOURCES IN CREATING SERBIAN-GERMAN LITERARY CORPUS

doctoral dissertation

Belgrade, 2019

УНИВЕРСИТЕТ В БЕЛГРАДЕ

ФИЛОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ

Елена С. Андоновски

ОТКРЫТАЯ СЕТЬ ПЕРЕДАЧИДАННЫХИ ЯЗЫКОВЫЕ РЕСУРСЫ В ПРОЦЕССЕ ПОСТРОЕНИЯ СЕРБСКО-НЕМЕЦКОГО ЛИТЕРАТУРНОГО КОРПУСА

Докторская диссертация

Белград, 2019.

Ментор:

др Цветана Крстев, редовни професор, Универзитет у Београду, Филолошки факултет

Чланови комисије:

др Милош Утвић, доцент, Универзитет у Београду, Филолошки факултет

др Јелена Костић-Томовић, редовни професор, Универзитет у Београду, Филолошки факултет

др Гордана Павловић-Лажетић, редовни професор, Универзитет у Београду, Математички факултет

др Ранка Станковић, ванредни професор, Универзитет у Београду, Рударско-геолошки факултет

Датум одбране: _____

Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса

Сажетак

Паралелни корпуси представљају врсту вишејезичних корпуса који су последњих деценија постали изузетно значајни у области обраде природних језика (Natural Language Processing - NLP) и један од важнијих ресурса за истраживаче у различитим областима лингвистике и сродним језичким дисциплинама. Са израдом ове докторске дисертације започет је рад на паралелном српско-немачком корпусу књижевних текстова, СрпНемКор. У току рада на дисертацији обрађено је четрнаест романа написаних у другој половини 20. и првој половини 21. века на српском и немачком језику. За садржај корпуса одабрано је седам романа оригинално написаних на српском и њихови еквиваленти на немачком језику и седам романа оригинално написаних на немачком (четири припадају аустријској књижевности, три припадају немачкој књижевности) и њихови еквиваленти на српском језику. У дисертацији је детаљно анализиран поступак прикупљања и одабира материјала за корпус, затим обрада текстова применом расположивих језичких алата и ресурса за оба језика, као и паралализација коришћењем одговарајућег софтвера.

Произведени паралелни корпус књижевних текстова, СрпНемКор, смештен је у дигиталну библиотеку Библиша која омогућава двојезичну претрагу комплетног текста паралелних колекција уз могућност морфолошког и семантичког проширење упита позивањем различитих лексичких и терминолошких ресурса. У дисертацији је анализирана могућност семантичког проширења упита заснована на синонимима позивањем терминолошке базе Терми. Терми је терминолошка вишејезична база која подржава развој терминолошких речника из различитих домена (математика, рачунарство, рударство, библиотекарство, рачунарска лингвистика и многи други) а до сада је омогућавала проширење упита само на српском и енглеском језику. На основу екстракције лексичких јединица из паралелне колекције СрпНемКор, база Терми је допуњена новим лексичким јединицама на српском, њиховим еквивалентима на немачком језику као и синонимима. Добијена листа преводних парова је искоришћена и

за генерисање двојезичног речника општег типа као скупа повезаних података при чему су тестиране и неке од технологија семантичког веба.

У дисертацији је анализирана и анотација именованих ентитета (имена људи, геополитичких имена, организација и сл.) у обе стране корпуса, на српском и немачком језику, уз помоћ расположивих алата за та два језика, као и могућности даље употребе добијених резултата.

Кључне речи: паралелни корпуси, анотација корпуса, обрада природних језика, дигиталне библиотеке, Библиша, лексички ресурси, терминолошки ресурси, именовани ентитети, семантички веб, отворени повезани подаци

Научна област: Филолошке науке, библиотечка информатика

Ужа научна област: Рачунарска лингвистика, корпусна лингвистика, обрада природних језика

УДК број: 811.163.41’322.2:811.112’322.2(043.3)

81’322:004.822(043.3)

025.5:004(043.3)

Linked Open Data and Language Resources in creating Serbian-German Parallel Literary Corpus

Abstract

Aligned multilingual corpora have become essential resources in multilingual Natural Language Processing (NLP) in the last decades, as well as one of the major resources for researchers in various areas of linguistics and related language disciplines. This doctoral dissertation presents a new aligned Serbian-German literary corpus, SrpNemKor, for which fourteen novels written in the second half of the 20th and the first part of the 21st century were selected. From selected novels, seven are originally written in Serbian and have their equivalents in the German language, and other seven are originally written in German (four belong to Austrian literature, three belong to German literature) and have their equivalents in the Serbian language. In the dissertation in detail is analyzed the process of the collection and selection of novels for the corpus, then texts processing using available language tools and resources for both languages, as well as parallelization using appropriate software.

The new aligned corpus, SrpNemKor, is stored in digital library Bibliša. Bibliša enables bilingual full-text search of aligned collections with the possibility of morphological and semantic query expansion by invoking appropriate lexical and terminological resources. In this research, it was specifically analyzed the possibility of semantic extension of the search queries based on synonyms by invoking the terminological database Termi. Termi is a multilingual database launched as a support for the development of terminological dictionaries in various domains (mathematics, computer science, mining, library science, computational linguistics, etc.). Until now Termi supported only the processing and representation of terms in Serbian and English. Based on the extraction of lexical units from the SrpNemKor, Termi is enriched with new lexical units in Serbian and their German equivalents with their synonyms which enables Bibliša to expand queries in German as well. The obtained list of German-Serbian translated pairs was also used to generate a general bilingual dictionary as a set of linked open data.

In the dissertation, it was analyzed the annotation of named entities (person names, organizations, locations, etc.) and are tasted the available tools for named entity recognition

both for Serbian and German in both parts of the corpus. The obtained results were analyzed for further researches in the different linguistics and informatics fields.

Keywords: parallel corpora, corpus annotation, natural language processing, digital libraries, Bibliša, lexical resources, terminological resources, named entities, semantic web, linked open data

Scientific field: Philological sciences, Library informatics

Scientific subfield: Computational linguistics, Corpus linguistics, Natural Language Processing

UDC Number: 811.163.41’322.2:811.112’322.2(043.3)

81’322:004.822(043.3)

025.5:004(043.3)

Садржај

1 Увод ... 1

2 Рачунарски језички корпуси ... 5

2.1 Корпусна лингвистика ... 5

2.2 Језички корпуси – појам, дефиниција, значај, врсте ... 8

2.2.1 Појам, дефиниција, значај ... 8

2.2.2 Врсте корпуса ... 13

2.3 Паралелни корпуси ... 16

2.3.1 Улога паралелних корпуса у истраживањима из области језика ... 23

2.3.2 Машинско превођење ... 28

2.4 Неки примери паралелних корпуса у свету ... 29

2.4.1 Корпуси некњижевних текстова ... 30

Acquis Communautaire ... 30

EuroParl ... 33

SETimes ... 34

OpenSubtitles ... 35

BabelNet ... 37

2.4.2 Корпуси књижевних текстова ... 38

Платонова Република ... 38

Орвелова 1984 (Multext-East): паралелни преводи романа *1984* Џорџа Орвела ... 40

2.5 Развој корпусне лингвистике у Србији ... 42

2.5.1 Пределектронски корпуси и услови за стварање електронских корпуса ... 42

2.5.2 Корпус савременог српског језика - СрпКор ... 47

2.5.3 Развој паралелних корпуса у Србији ... 50

Српско-француски корпус - СрпФранКор ... 51

Српско-енглески корпус - СрпЕнгКор ... 55

Библиша ... 58

Бошњачко-хрватско-српски паралелни корпус ... 66

Вишејезични Верн: паралелни преводи романа *Пут око света за 80 дана* Жила Верна ... 67

Орвелова *1984* за српски језик ... 69

3 Израда паралелних корпуса у Србији ... 71
3.1 Ауторска права, права приступа и коришћења садржаја корпуса ... 71
3.2 Прикупљање и дигитализација текстова ... 74
3.3 Доступни алати и језички ресурси за обраду текстова ... 76
3.3.1 IMS OCWB ... 76
3.3.2 Unitex ... 76
3.3.3 E-речници ... 82
3.3.4 Ворднет ... 88
3.4 Анотација корпуса ... 89
3.4.1 Структурна анотација ... 89
3.4.2 Морфосинтаксичка анотација ... 91
3.5 Паралелизација ... 92
3.6 Претрага паралелних корпуса у Корпусу савременог српског језика ... 94
4 Стварање предуслова за ефикасно проналажење информација у паралелним корпусима ... 97
4.1 Метаподаци – појам, дефиниције, врсте, значај ... 98
4.1.1 Стандарди за израду метаподатака и формати приказа ... 102
Даблинско језгро ... 104
Стандард за кодирање и пренос метаподатака ... 107
Схема за опис метаподатака објеката ... 110
Иницијатива за кодирање текста и TEI заглавље ... 111
4.1.2 Израда метаподатака у Србији и значај библиотека ... 116
4.2 Припрема текста ... 122
4.2.1 Оптичко препознавање карактера ... 122
4.2.2 Препознавања именованих ентитета ... 124
5 Семантички веб и мрежа отворених повезаних података ... 126
5.1 Семантички веб ... 127
5.1.1 О семантичком вебу ... 127
5.1.2 Структура семантичког веба ... 130
5.2 Иницијатива „Отворени повезани подаци” ... 138
5.2.1 DBpedia ... 139
5.2.2 Википодаци ... 144

5.2.3 Поступак повезивања података 150
5.2.4 Поступак објављивања повезаних података на веб 153
5.2.5 Системи за организацију знања 154
5.3 Библиотеке и семантички веб 156
5.4 Иницијативе и пројекти засновани на принципима семантичког веба 164
5.4.1 Еуропеана 164
5.4.2 Дигитална народна библиотека Америке 166
5.4.3 Оквир за библиографски опис 168
6 Српско-немачки паралелни корпус - СрпНемКор 172
6.1 Садржај корпуса СрпНемКор 173
6.2 Фазе у креирању корпуса СрпНемКор 178
6.2.1 Дигитализација текстова 178
6.2.2 Обрада немачких текстова 179
6.2.3 Обрада српских текстова коришћењем Unitex-а 185
6.2.4 Структурна анотација 187
6.2.5 Паралелизација 189
6.3 СрпНемКор у Библиши 205
6.3.1 Структура метаподатака 206
6.3.2 Допуна лексичких ресурса за двојезично претраживање 211
6.3.3 Претрага колекције СрпНемКор и анализа добијених резултата 216
6.4 Означавање именованих ентитета 223
6.5 СрпНемКор и отворени повезани подаци 233
6.5.1 Коришћени ресурси 234
6.5.2 Поступак повезивања 238
6.5.3 Двојезични речник општег типа као отворени повезани подаци 239
7 Постигнути резултати и будући рад 251
7.1 Постигнути резултати 251
7.2 План за будући рад 254
8 Библиографија 256
8.1 Библиографске референце публиковане на ћирилици 256
8.2 Библиографске референце публиковане на латиници 259
9 Додаци 290

9.1 Списак слика 290
9.2 Списак табела 292
9.3 Списак скраћеница 293
9.3.1 Скраћенице на ћирилици 293
9.3.2 Скраћенице на латиници 293
Прилози 296
Прилог 1 - Пример записа у формату COMARC/B / према стандарду ISO2709 296
Прилог 2 - Пример записа у формату MARC21 / према стандарду ISO2709 296
Прилог 3 - Пример записа у формату Даблинско језгро / XML синтакса 297
Прилог 4 - Пример записа у формату METS / XML синтакса 297
Прилог 5 - Пример записа у формату MODS / XML синтакса 299
Прилог 6 - Пример записа у формату MARC21 / XML синтакса 300
Прилог 7 - Пример записа у формату COMARC / XML синтакса 302
Прилог 8 – Пример записа у формату TEI заглавље / XML синтакса 305
Прилог 9 – Упоредни приказ метаподатака 307
Прилог 10а - Пример записа за Томаса Бернхарда у бази GND / корисничко окружење 309
Прилог 10б - Пример записа за Томаса Бернхарда у бази GND / формат RDF/Turtle 311
Прилог 11 - Пример записа за Томаса Бернхарда у бази VIAF / HTML 314
Прилог 12а - Пример записа за Томаса Бернхарда у бази LCNAF / корисничко окружење 315
Прилог 12б - Пример записа за Томаса Бернхарда у бази LCNAF / формат RDF/XML 317
Прилог 13 - Пример записа за Томаса Бернхарда у Википодацима / корисничко окружење 322
Прилог 14 - Запис у Европеани у EDM моделу / корисничко окружење 327
Прилог 15 - Пример записа за дело „Moje награде“ Томаса Бернхарда у моделу BIBFRAME / корисничко окружење 329
Биографија аутора 330
Error! Reference source not found. **.....** Error! Bookmark not defined.

1 Увод

Предмет истраживања ове докторске дисертације односи се на начине израде паралелног књижевног корпуса и експлоатацију његовог садржаја са циљем да се омогући допуна лексичких ресурса у систему двојезичне претраге паралелних колекција, али и постигне боља видљивост садржаја на вебу увезивањем припремљеног скупа података са релевантним ресурсима на вебу. У самој дисертацији објашњени су поступци израде и успостављања једног паралелног корпуса, методе припреме, обраде и упаривања (паралелизације) текстова на српском и немачком језику, процеси претраживања и проналажења информација у оквиру самог корпуса, као и поступци повезивања са другим изворима на вебу и објављивање скупа података из припремљеног корпуса према принципима и смерницама семантичког веба. Да би се објаснили поступци повезивања са другим изворима на вебу и објављивање скупа података као отворених повезаних података (Linked Open Data), предмет докторске дисертације обухвата и описивање и испитивање принципа и значаја семантичког веба и технологија које се на њему базирају, концепта „отворени повезани подаци", као и система за организацију знања (контролисани речници, тезауруси, фасетне класификације, онтологије) и њихове улоге и значаја за структурирање података. Посебан акценат стављен је на обраду и повезивање метаподатака, а затим на проналажење и екстракцију информација из добијеног корпуса коришћењем развијених језичких алата и технологија за пречишћавање текста. Примена технологија семантичког веба је демонстрирана кроз повезивање развијених ресурса са релевантним системима и другим ресурсима на вебу.

Полазећи од дефиниције појма *корпусна лингвистика,* друго поглавље ове дисертације представља језичке корпусе као значајне ресурсе у различитим областима истраживања и проучавања језика са посебним освртом на паралелне корпусе као једну од врста језичких корпуса која је последњих деценија постала изузетно значајна у областима као што су двојезична или вишејезична лексикографија, учење страних језика, превођење и машинско превођење, истраживање терминологије, лингвистичка

истраживања, упоредна изучавања два или више језика и многе друге. У другом поглављу су представљени и неки примери добре праксе паралелних корпуса у свету, као и сви паралелни корпуси које је развила Група за језичке технологије Универзитета у Београду, а који су део Корпуса савременог српског језика.

У трећем поглављу дисертације детаљно је објашњен поступак израде једног паралелног корпуса. Процес паралелизације подразумева да текстови у електронском облику буду коректно припремљени и упарени, а резултати представљени у одговарајућем стандардном формату. Сам поступак припреме текстова и креирање паралелног корпуса пролази кроз неколико фаза као што су прикупљање и дигитализација текстова, примена алата и језичких ресурса за обраду текстова пре паралелизације, анотација корпусног материјала, поступак паралелизације уз помоћ одговарајућег софтвера. Поред објашњења поступака, у поглављу су представљени алати, језички ресурси и софтвер који се у Србији користе за припрему и израду паралелних корпуса, као и начини претраге паралелних корпуса који су до сада развијени у Србији. Како се у корпусу који се анализира у дисертацији ради о делима савремених писаца односно писаца који су писали у другој половини 20. века, посебно питање које се у овом поглављу разматра јесте питање ауторских права приступа и коришћења корпусног материјала.

Четврто поглавље дисертације бави се описом и означавањем докумената са становишта њихових формалних и садржинских особина како бисмо могли да их претражујемо и проналазимо у базама података. Највећи део поглавља посвећен је формалном опису дигиталних докумената и објеката, односно додељивању метаподатака. На почетку је укратко описан појам *метаподатак*, као и врсте, улога и значај метаподатака у дигиталним репозиторијумима, различитим базама података, електронским каталозима и другим сличним системима. Затим су представљени неки од међународних стандарда који се широко користе за израду метаподатака, а који се примењују и у Србији. Поред формалног описа, дигитални документи се данас и садржински индексирају применом различитих технологија за пречишћавање текста како би на крају пуни текст дигиталне колекције био претражив. У поглављу су представљене неке од технологија оптичког препознавања карактера, технологија за аутоматско

индексирање садржаја докумената и препознавање именованих ентитета, које су примењене и на корпусу који је предмет ове дисертације. На крају поглавља представљени су стандарди за израду метаподатака и приказане неке од технологија за пречишћавање текста које се користе у Србији.

Посебан део истраживања посвећен је семантичком повезивању библиографских и структурираних метаподатака из добијеног корпуса са релевантним изворима на вебу и постављању скупа података припремљеног на основу корпусног материјала као „отворених повезаних података". У петом поглављу најпре је објашњен термин *семантички веб* и његове карактеристике, стандард који је развијен за описивање значења података и веза између података на нивоу њиховог значења (Resource Description Framework - RDF), концепт *отворени повезани подаци* (Linked Open Data - LOD), а представљени су и ресурси који су део семантичког веба и омогућавају његову реализацију. Након тога анализиране су неке иницијативе и пројекти који се заснивају на принципима семантичког веба, њихови модели података, начини претраге и приступа информацијама, као и предности овакве структуре за различите заједнице, посебно библиотечке.

У шестом поглављу дисертације представљен је нови српско-немачки паралелни корпус. У поглављу је прво представљена идеја о настанку самог корпуса, а затим и његова величина и садржај што подразумева попис свих дела од којих је корпус иницијално састављен. За одабир дела постављени су одређени критеријуми описани у самом поглављу, а за свако дело појединачно је објашњено због чега је одабрано односно који је критеријум примењен. Поред пописа дела која су постала саставни део корпуса на почетку његовог настајања, наведено је и како се до тих материјала дошло односно у којим библиотекама у земљи и ван ње је пронађен материјал неопходан за рад и истраживање. Након тога анализиран је поступак креирања самог корпуса који је подразумевао припрему електронских верзија текстова у одговарајућем формату, примену различитих језичких алата за њихову обраду, структурну и морфолошку анотацију, као и поступак паралелизације. Овако припремљен корпус смештен је у дигиталну библиотеку Библиша која омогућава двојезично претраживање садржаја

колекције позивањем разноврсних лексичких ресурса. У поглављу су анализирани лексички ресурси који се користе за претрагу српско-немачког паралелног корпуса и допуна одређених ресурса на основу добијених резултата.

Поред тога, посебна пажња посвећена је примени технологија и стандарда семантичког веба на припремљени корпус текстова. Један део односи се на повезивање ентитета односно библиографских метаподатака из корпуса који се анализира са релевантним ресурсима на вебу, док се други део односи се на припрему скупа података према принципима семантичког веба и укључивање у облак „отворених повезаних података". За ове потребе на основу паралелне колекције генерисан је узорак двојезичног речника општег типа.

У самом закључку су наведени резултати постигнути радом на овом корпусу и њихов значај, која су решења понуђена за повезивање података и ресурса на мрежи, као и колико ова комплексна организација података доприноси бољој видљивости садржаја једног паралелног корпуса на вебу и добијању релевантнијих резултата претраге. У дисертацији је посебан осврт дат на библиотекарство као струку у којој постоји доста искуства у раду са уређеним системима за опис различитих садржаја као и са разноврсним ресурсима (електронски каталози, нормативне датотеке, дигитални репозиторијуми, тезауруси и друго) који могу да се употребе за повезивање података на нивоу њиховог значења и омогуће међусобно умрежавање.

2 Рачунарски језички корпуси

2.1 Корпусна лингвистика

Корпусна лингвистика се у најширем смислу може дефинисати као научна област која се бави истраживањима језика заснованим на корпусу. У фокусу истраживања су процедуре и методе за проучавање језика, помоћу којих се суштински мења приступ истраживача проучавању језика. Осим тога, може да обликује и редефинише низ теорија о језику, а омогућава и да се постојеће теорије језика примене у развоју корпуса и разноврсних језичких алата и ресурса (McEnery and Hardie 2012, 1). Она подразумева активности као што су припрема и изградња језичких рачунарских корпуса, развој алата за анализу корпусног садржаја, коришћење корпуса за опис речника и граматике једног језика, учење и подучавање језика, као и у аутоматској обради података за обраду и разумевање природних језика. Корпусна лингвистика се не може упоредити са другим областима лингвистике јер се не односи на област проучавања самог језика већ на алате и методологију које се примењују у лингвистичким истраживањима. Она лако комбинује друге области лингвистике (може се проучавати фонетика, синтакса, социолингвистика и други аспекти) коришћењем рачунарских језичких корпуса и развијених алата за анализу корпусног садржаја (Leech 1992, 105-106).

Сам термин *корпусна лингвистика* први пут се појавио почетком осамдесетих година 20. века. Међутим, истраживања језика заснована на корпусу имају дужу историју. Већ у првој половини 20. века лингвисти су користили обичне кутије за складиштење текстуалних колекција на папирним листићима. Иако таква колекција писаног или транскрибованог текста није довољно репрезентативна за данашње поимање корпуса, „примењена емпиријска методологија, заснована на прикупљеним подацима, била је корпусно базирана” (McEnery, Xiao and Tono 2006, 3). Овакви такозвани пределектронски корпуси су углавном креирани за потребе различитих истраживачких пројеката са специфичним циљем. Због времена потребног за прикупљање, обраду и анализу материјала ови корпуси су углавном били мали, а коришћени су пре свега за потребе

истраживања из конкретних језичких области као што је, на пример, фонетика, док их је само мали број истраживача користио и за шира истраживања у области језика.

Крајем педесетих и почетком шездесетих година 20. века радови америчког лингвисте Ноама Чомског (Noam Chomsky) померили су фокус лингвистичких истраживања са апстрактног описа језика ка теоретским приказима (McEnery and Wilson 2001, 6). Посебну прекретницу у лингвистици представља књига *Синтаксичке структуре* објављена 1957. године у којој је Чомски представио нове идеје које су за кратак временски период, поред потискивања дескриптивне лингвистике, пажњу лингвиста преусмериле са емпиризма на рационализам. Такође, Чомски је увео појмове *језичка способност* (linguistic competence) и *говорна делатност* (linguistic performance). Језичку способност је дефинисао као способност човека, односно његово интерно знање, да разуме и произведе језик, док је говорну делатност дефинисао као спољашњу манифестацију тог знања (Lyons 1996, 11-12). Постављајући своју хијерархију (Хијерархија Чомског) у оквиру које класификује формалне језике према њиховој генеративној моћи (Jäger and Roger 2012), Чомски примећује да структура реченице може бити рекурзивна те да је број реченица једног језика потенцијално бесконачан. Ово је довело до критике дотадашње претпоставке дескриптивних лингвиста да све реченице једног језика могу да се прикупе и преброје односно да је њихов број коначан. Оно што Чомски прихвата као коначно јесу синтаксичка правила којима се на једини начин може објаснити граматика једног језика.

Сагледавајући бесконачне могућности генерисања природног језика са једне стране, а посматрајући језички корпус као коначан скуп његове говорне делатности са друге стране, Чомски је обеснажио корпус као извор садржаја за лингвистичка истраживања. Он и његови следбеници прокламовали су да је „корпус сиромашан извор одређеног делокруга истраживања" (Dobrić 2009, 360), да „корпус не може никада бити користан алат за лингвисте чији је прави задатак да истраже језичку способност пре него говорну делатност... Корпус је по својој природи колекција испољених језичких исказа; то је употреба језика и као такав мора бити сиромашан водич обликовања језичке способности" (McEnery and Wilson 2011, 6). Његова критика је у том тренутку била велики

изазов за лингвисте који су желели да користе језичке корпусе у својим истраживањима те је довела до великих дебата и дискусија о корисности корпуса за лингвистичка истраживања.

Развој и усавршавање рачунара и рачунарских технологија допринео је усавршавању и развоју електронских корпуса. Предност рачунара у процесу стварања корпуса огледа се у брзини обраде и складиштењу садржаја, једноставнијем и систематичнијем процесу претраживања, избору и екстракцији информација, сортирању и форматирању података, као и могућности допуњавања текстова метаподацима корисним за анализу корпуса. Први електронски корпус који настаје почетком шездесетих година 20. века у САД је *Стандардни корпус савременог америчког енглеског језика Универзитета Браун* (The Brown University Standard Corpus of Present-Day American English) познатији и као *Браунов корпус* (Francis and Kučera 1964). Браунов корпус су креирали Винтроп Нелсон Франсис (Winthrop Nelson Francis) и Хенри Кучера (Henry Kučera) на Универзитету Браун (Brown University) из Провиденса, САД. Идеја је била да се направи корпус од милион речи савременог америчког енглеског језика који би се користио на рачунару и који би на неки начин био модел за припрему, обраду и представљање одабраног текстуалног садржаја било ког језика.

На конференцији лингвиста одржаној 1963. године на Универзитету Браун донета је одлука о укупном броју узорака текстова који ће бити део корпуса, њиховој величини и расподели међу разним жанровима писаног прозног текста. Тако је у корпус ушло 500 узорака текста, сваки приближне величине око 2000 речи, подељених у 15 жанрова: информативни (новине, магазини, научни радови, есеји итд.) и књижевни текстови (Kučera 2002, 307). Сваки узорак текста почиње од почетка реченице, али не нужно од прве реченице пасуса или неке веће структурне јединице, а завршава се на крају прве реченице текста после 2000. речи (Francis and Kučera 1964). Године 1964. креатори корпуса објавили су приручник у коме је описана структура корпуса, процес обраде текстова, објашњен је поступак разрешења ауторских права и дате су основне информације о његовим техничким карактеристикама. Прерађена издања приручника објављена су 1971. и 1979. године, а издање из 1979. је уједно и допуњено (Francis and

Kučera 1964). Постоји шест верзија Брауновог корпуса: основна верзија корпуса, „огољена" верзија – без интерпункцијских знакова, етикетирана верзија, Берген – верзија 1 (Bergen Form I) и Берген – верзија 2 (Bergen Form II) и MARC верзија.

Када је реч о новијој историји корпусне лингвистике, почетком августа 1991. године на међународном симпозијуму британских, холандских, шведских и норвешких лингвиста у Стокхолму (Svartvik 1991) формирана је заједница корпусних лингвиста која је током деведесетих година прошлог века успела да учврсти свој рад и објави многобројне публикације, а од 1996. године започела је и са издавањем свог часописа *The International Journal of Corpus Linguistics*.

Иако се корпусна лингвистика као дисциплина развија већ деценијама, корпусни лингвисти још увек немају прецизну дефиницију шта би тачно она представљала. На питање *Шта је корпусна лингвистика?* постоје различити одговори од оних који је дефинишу као алат, метод, методологију, методолошки приступ, дисциплину, теорију, теоријски приступ, парадигму (теоријску или методолошку), до оних који је тумаче као комбинацију свега наведеног (Taylor 2008, 180). Данас се корпусна лингвистика пре сматра за методологију или скуп методологија него за посебну теоријску дисциплину (McEnery, Xiao and Tono 2006, 3-12), али је и даље отворено питање да ли се под тим подразумева и нешто више.

2.2 Језички корпуси – појам, дефиниција, значај, врсте

2.2.1 Појам, дефиниција, значај

Корпуси представљају основу корпусне лингвистике. Реч *корпус* потиче од латинске речи *corpus* која значи *тело*, а метафорички је преузета да опише колекције језичких и комуникационих података. Према најједноставнијој дефиницији корпус представља велику колекцију текстова (McEnery and Wilson 1993, 1), док се у лингвистици под корпусом, у најширем смислу, подразумева емпиријски материјал намењен истраживању језика (Витас и Поповић 2003, 221). „Корпус је колекција језичких текстова или њихових делова у електронском формату одабраних према неком критеријуму, који представљају језик или језички варијетет као извор података за језичка истраживања" (Sinclair 2005).

Иако постоји велики број дефиниција корпуса, готово у свима се корпус сагледава као „колекција аутентичних машински читљивих текстова који представљају репрезентативни узорак неког језика или језичког варијетета и могу бити анотирани различитим облицима лингвистичких информација" (McEnery, Xiao and Tono 2006, 5). Иако се прави разлика између пределектронских и електронских корпуса термин *корпус* се данас углавном односи на електронске корпусе и савремена истраживања језика не могу да се замисле без њих.

Технички процес креирања корпуса знатно је олакшан захваљујући савременим рачунарским технологијама. Процес обраде, анализе и складиштења прикупљеног материјала захтева мање времена него раније, прецизан је и доследан, те и величина корпуса не представља велики проблем. Највећи проблем приликом процеса креирања корпуса јесте да се утврди узорак текста који ће бити обрађен. У том поступку постављају се следећа питања: коју врсту садржаја одабрати, у којој мери би поједини садржаји требало да буду заступљени (величина корпуса), која би била дужина појединачних текстова (да ли уносити текстове у целини или њихове делове одређене дужине), који временски период настанка текстова обухватити (поготово када су у питању савремени преводи текстова насталих пре утврђеног временског периода).

Сва претходно наведена питања и питање репрезентативности корпуса међу првима је разматрао амерички лингвиста Даглас Бајбер (Douglas Biber) који је дефинисао да „свака категорија одабраног узорка мора да садржи знатан и разноврстан број различитих текстова тако да одабрани узорак не нарушава претходно утврђен критеријум категоризације садржаја" (Biber 1993, 243-244). Према Дејвиду Лију (David Lee) категоризација текстова за корпус се првенствено заснива на критеријумима као што су где и када је настао текст, ко га је написао и за кога и о чему је текст него на лингвистичким карактеристикама самог материјала. Која врста текстова ће бити укључена у корпус као и које величине треба да буде одабрана врста текстова у великој мери одређује да ли ће корпусни узорак бити довољно репрезентативан за циљну групу истраживача језика (Aston 2001, 74). Са друге стране, сврха корпуса је и да омогући

екстраховање жељених информација које могу да се искористе у новим областима рачунарске лингвистике и обради природних језика.

Иако је сам процес обраде прикупљеног материјала олакшан захваљујући рачунарским технологијама и у тој процедури може да дође до одређених проблема. Рачунари могу да обрађују само податке који су већ у електронском облику, такозване машински читљиве податке. Таквих података још увек нема у довољној мери односно велики број текстова који се бирају за корпус нису у електронском облику те их је неопходно конвертовати у жељени формат. Такође, рачунару недостаје добар део уобичајеног људског предзнања о језику (на пример, где су границе речи) што захтева додатно развијање и тестирање програма. Сви ови проблеми на одређен начин одређују време потребно за стварање једног језичког корпуса али и његову величину и квалитет.

Језички корпуси представљају један од важнијих ресурса за истраживања у области лингвистике и сродним језичким дисциплинама. Они се користе у свим лингвистичким дисциплинама као помоћно средство за анализу књижевно-уметничких дела или текстова који припадају неком другом посебном функционалном стилу (новински, административни, разговорни, научни и научно-популарни), а посебно место заузима статистичка анализа језика. Неке од области лингвистике у оквиру којих се истраживања базирају на корпусима су (Dobrić 2009, 362):

1. Лексикографија: корпусно засновани приступ омогућава испитивање лингвистичких и нелингвистичких корелација одређених речи што може да помогне у изградњи речника.
2. Социолингвистика: корпусно засновани приступ омогућава истраживање дијалеката и регистара који претходно нису разматрани.
3. Анализа дискурса: корпусно засновани приступ омогућава да се са довољно великим узорком језика утврди не само његова структура, већ и карактеристике језика.
4. Морфологија: резултати истраживања корпусно заснованог приступа могу добро да представе фреквенцију, дистрибуцију и функцију различитих варијанти речи.

5. Фонологија: корпусно засновани приступ омогућава увид у различита средства фонетске дистрибуције.
6. Семантика: ниједан други постојећи приступ осим овог не може да омогући тако потпун увид у значење речи.
7. Синтакса: истраживање структуре језика на овај начин може да да емпиријски доказ о структури реченица, односно, самом језику.
8. Компаративна и контрастивна лингвистика: корпусно засновани приступ показује сличности и разлике у употреби између језика, за шта су посебно значајни паралелни корпуси.
9. Методика наставе: корпуси могу помоћи при дизајнирању материјала и активности за учење језика.
10. Когнитивна лингвистика: аутентична употреба језика смештена у корпусе даје увид у начин на који ментални процеси утичу на комуникацију и на језик у целини.

У зависности од нивоа анотације, корпуси омогућавају корисницима да уоче различите примере употребе језика, да приликом истраживања, позивањем скупа речи једног језика и њихових граматичких облика, сагледају фреквентност појављивања одређене речи или фразе постављене кроз упите за претрагу, у којим се све облицима и варијантама постављена реч или фраза појављује, као и у каквој је семантичкој корелацији са другим речима и фразама и њиховим облицима у датом корпусу. Такође, са новим технологијама повезивања садржаја на вебу могуће је сагледати фреквентност појављивања речи или фразе у разним базама података, дигиталним репозиторијумима, онтологијама, контролисаним речницима и сличним изворима.

Један од најутицајнијих примера националних корпуса јесте Британски национални корпус (British National Corpus - BNC)[1]. BNC је општи, синхрони корпус британског енглеског језика од 100 милиона речи настао у периоду од 1991. до 1994. године, а јавности је први пут представљен 1995. године. Пројекат је покренут са циљем да се направи добар пример за сличне подухвате изградње националних корпуса других језика

[1] „British National Corpus“, приступљено 26.3.2019, http://corpus.byu.edu/bnc/

те су га подржале академске институције, приватни издавачи, државне институције Велике Британије, али и британска влада која је финансирала велики део његове изградње (Aston and Burnard 1998, 28-32).

BNC је састављен од писаних и говорних текстова који су даље разврстани у мање категорије и подкатегорије (Burnard 1995). Највећи део корпуса, скоро 90%, чине писани текстови који су категоризовани према три критеријума („домен", „време" и „медијум"), док мањи део корпуса представљају говорни текстови који су категоризовани према два критеријума („демографски" и „контекстно вођени") (British National Corpus 2016). Део корпуса са писаним текстовима садржи делове из регионалних и националних новина, специјализованих часописа и периодичних издања за све узрасте, академских књига и популарне књижевности, објављена и необјављена писма, школске и универзитетске писане радове, као и многе друге текстове, а део корпуса са говорним текстовима садржи транскрипцију неформалних разговора које су снимали волонтери, као и формалних разговора са пословних састанака, састанака владе, стручних предавања, током радио-емисија и друго (Aston and Burnard 1998). Говорни текстови преузети су из Лонгмановог говорног корпуса (Longman Spoken Corpus - LSC), једног од три корпуса насталих крајем осамдесетих година 20. века у организацији издавачке куће Лонгман и Универзитета у Ланкастеру (Lancaster University).

За кодирање BNC одабране су Смернице Иницијативе за кодирање текста (Guidelines of the Text Encoding Initiative - TEI)[2] као би се истовремено означила логичка структура текстова (поглавља, пасуси, листе, итд.) али и резултати добијени приликом аутоматске анотације граматичких категорија уз помоћ система CLAWS[3] (British National Corpus 2016). Такође, потпуна класификација, контекстуалне и библиографске информације означене су према препорукама TEI. Корпус се може претраживати преко кључних речи, фраза или преко колокација.

Након завршетка изградње корпуса нису додавани нови текстови али су зато настале још две верзије овог корпуса, BNC World 2001. и BNC XML Edition 2007. године, као и два поткорпуса BNC Sampler (BNC Sampler 2008) (колекција писаних и говорних текстова,

[2] TEI: Guidelines, приступљено 26.3.2019, http://www.tei-c.org/Guidelines/
[3] CLAWS part-of-speech tagger for English, приступљено 26.3.2019, http://ucrel.lancs.ac.uk/claws/

свака од по милион речи) и BNC Baby (Reference Guide to BNC Baby 2008) (Burnard 2008) (четири поткорпуса текстова различитог жанра, сваки представљен са по милион речи) (Burnard 2016).

2.2.2 Врсте корпуса

Приликом креирања корпуса група истраживача и техничких стручњака који на њему раде одређују његов садржај, методологију рада, језик на ком ће бити заступљени текстови, временски период који ће бити обухваћен што све одређује врсту корпуса. Класификација корпуса може се извршити на основу више параметара од којих су најважнији носач, обим (величина), домен, намена, период, извор, начин анотације и број укључених језика (Витас и Поповић 2003, 223).

У зависности од носача, корпуси могу бити пределектронски и електронски о чему је већ било речи у одељку 2.2.1. Пределектронски корпуси су углавном били у папирној форми. Са појавом рачунара почињу да се развијају електронски корпуси који су постали један од главних ресурса у различитим областима истраживања језика. Рачунари су процес развоја, одржавања и креирања корпуса доста олакшали и поједноставили али и унапредили начин њиховог коришћења.

Према домену и намени корпуси могу да се поделе на опште и специјализоване корпусе. Општи корпуси могу да се користе за различите врсте лингвистичких истраживања као што су лексикографска, граматичка, семантичка, прагматичка, социолингвистичка. Специјализовани корпуси, са друге стране, настају ради неког специфичног лингвистичког истраживања (дијалекатски корпуси, регионални корпуси, нестандардни корпуси, корпуси језика као нематерњег, и друго) или као помоћно средство у рачунарској лингвистици и обради природног језика (корпуси за тренирање алата за аутоматску морфосинтаксичку анализу текста, аутоматско препознавање и генерисање говора, и друго) (Utvić 2013, 25).

Следећи параметар на основу кога се корпуси могу разликовати јесте њихов обим односно величина. Према овом параметру корпуси се могу поделити на статичке и динамичке. Код статичких корпуса претходно утврђена величина остаје фиксирана и по

завршетку креирања они се не допуњују додатним текстовима. За разлику од њих динамички корпуси се непрекидно допуњују новим текстовима. Посебну врсту динамичких корпуса представљају опортунистички корпуси који се константно допуњују свим текстовима до којих креатори корпуса могу да дођу, али тако да се задовољи основна намена корпуса.

На основу временског периода корпуси се могу поделити на синхроне и дијахроне. Синхрони корпуси садрже текстове који се односе на специфичан временски период који се посматра као целина, док дијахрони корпуси садрже текстове који су настали у неком дужем временском периоду што омогућава истраживачима да прате развој једног језика и уочавају његове промене кроз време. Дијахрони корпуси се често називају и историјским корпусима.

Приликом креирања корпуса може се обрађивати материјал у писаном, говорном или електронском формату (блогови, форуми, друштвене мреже, и слично) те се корпуси могу поделити и на основу извора односно медијума текстова. Од свега наведеног корпуси говорних текстова су најмање заступљени због времена и новца који је потребан да се говорни текст транскрибује и пренесе у машински читљив облик. Посебна врста корпуса су мултимодални корпуси који представљају дигитализоване колекције језичког и комуникационог материјала забележеног на више медијума (Allwood 2008, 207-208).

Следећи начин поделе корпуса јесте на неетикетиране и етикетиране (анотиране) корпусе. Неетикетирани корпуси садрже чист текст без додатних садржаја и анотација (етикета) док етикетирани (анотирани) корпуси поред чистог текста садрже и различита структурна и морфосинтаксичка обележја. Приликом креирања корпуса одређени његови делови (текстови, логичке целине у оквиру текста, токени[4]) могу бити анотирани, односно могу им се придружити додатне информације. Анотација или означавање делова корпуса подразумева следеће (Утвић 2011, 40):

1. придруживање одговарајућих библиографских референци, података о креирању и ажурирању електронске верзије текста, као и статистичких података о тексту;

[4] Токен представља појединачно појављивање неке речи чија фреквентност спада у поље квантитативне анализе.

2. структурну анотацију – обележавање логичке структуре текста (поглавља, наслова, пасуса, реченица) (структурна анотација је детаљније објашњена у поглављу 3 одељак 3.4.1 и поглављу 6 одељак 6.2.4);
3. додељивање свакој корпусној речи информације о врсти речи (именица, придев, глагол, и друго), леми (номинатив једнине именице, инфинитив глагола, и друго), вредностима флективних категорија (род, број, падеж, глаголски облик, глаголски вид, и друго), (творбеној) основи, префиксима, инфиксима и суфиксима, начину изговора (акценат), границама слогова;
4. семантичку анотацију – додељивање ознака одговарајућег значења сваком токену;
5. синтаксичку анотацију – додељивање информација о функцији у реченици (субјекат, предикат, објекат, глаголска одредба);
6. анотацију кореференције која се може реализовати на нивоу дискурса којом се у тексту означавају релације кореференције (анафорички и катафорички односи) између корпусних речи[5];
7. прагматичку анотацију – обележавање говорног чина у тексту прагматичким односно стилистичким информацијама.

Детаљно анотирани корпуси пружају велике могућности за квалитативна и квантитативна истраживања различитих аспеката језика, од нивоа фонологије до нивоа синтаксе и анализе дискурса (Костић 2003, 260).

На основу тога да ли су текстови у оквиру истог корпуса на једном или више језика корпуси се могу разврстати на једнојезичне и вишејезичне. Једнојезични корпуси садрже узорак текста на једном језику намењен одређеним областима истраживања датог језика. Овакви корпуси могу да садрже једну врсту текстова или текстуалне узорке различитог типа (књижевни текстови, новински текстови, и слично) на једном језику који су на одређени начин обрађени и припремљени за даљу употребу.

[5] У корпусној лингвистици је уобичајено да се под *корпусном речју* подразумева низ карактера (корпусног) текста између два узастопна *сепаратора*, при чему се скуп сепаратора дефинише као скуп неалфанумеричких карактера (Утвић 2014).

Вишејезични корпуси представљају посебну врсту корпуса у чијем саставу су текстови написани на различитим језицима, односно садрже један или више оригиналних текстова и њихове преводе на један или више језика. У литератури најчешће су изједначени са двојезичним корпусима који су, у ствари, само једна њихова врста. Када се говори о вишејезичним корпусима они се могу разврстати на паралелне и упоредне корпусе. Према утврђеној дефиницији паралелни корпуси су састављени од текстова на изворном језику и њихових превода на један или више циљних језика, док су упоредни корпуси углавном састављени од текстова на различитим језицима који су одабрани према одређеном критеријуму (одређени временски период, одређена врста текста, одређени домен итд.) односно текстова који припадају истом жанру (McEnery and Xiao 2008, 19-20).

2.3 Паралелни корпуси

Паралелни корпуси представљају врсту вишејезичних корпуса који су последњих деценија постали изузетно значајни за различите области истраживања језика као што су двојезична или вишејезична лексикографија, учење страних језика, процеси превођења и машинског превођења, истраживање терминологије, лингвистичка истраживања, упоредна изучавања два или више језика, и тако даље. Под паралелним корпусима подразумевају се корпуси који садрже један или више оригиналних текстова и њихове преводе на један или више језика (Töny 2016, 11). Поред вишејезичних, односно двојезичних, паралелних корпуса постоје и корпуси са паралелним текстовима на једном језику који подразумевају да корпусни садржај чине различита издања истог текста на одабраном језику било да су у питању различита издања текста који је оригинално настао на том језику или су то различита издања превода неког текста на том језику. Значај и вредност паралелних корпуса расте са следећим карактеристикама (Erjavec et al. 2005, 529-531):

1. *величина*: велики корпуси не дају само статистички поузданије резултате већ омогућавају да се открију и језички феномени које је немогуће видети у корпусима мањег обима;

2. *број језика*: корисност корпуса расте са растом броја језика јер су текстови на тим језицима међусобно упарени, а велике паралелне колекције могу тако да садрже и упарене садржаје написане на такозваним „малим" језицима до којих је иначе тешко доћи;
3. *лингвистичка анотација*: омогућава да се из великог садржаја корпуса извуку различите граматичке категорије и њихове варијанте и сагледају њихови парови у другим језицима;
4. *семантичка анотација*: корисна је за класификацију докумената (или њихових делова, на пример, речи) у оквиру неког хијерархијског концепта што се даље може употребити за повезивање података и њихово коришћење (на пример, у оквиру парадигме семантичког веба).

Вишејезични паралелни корпуси најчешће су састављени од скупа двојезичних паралелизованих корпуса те ће у наставку бити описани двојезични паралелни корпуси, њихова структура и фазе формирања. Основни елемент двојезичног паралелног корпуса је битекст или паралелизовани текст. Под битекстом се подразумева постојање две семантички еквивалентне верзије истог текста на два обично различита језика (Laporte, Vitas and Krstev 2006, 111), односно битекст представља текст и његов превод, односно преводе, представљене на такав начин да је између елемената њиховог логичког исказа успостављена експлицитна веза, на пример, на нивоу пасуса или реченица (Vitas 2010, 273). Везе се успостављају између различитих структурних елемената текста на изворном и њихових еквивалената на циљном језику који се још називају и варијанте јединица превођења (Translation Unit Variant - TUV). Варијанте јединица превођења могу бити цео документ, поглавље, пасус, реченица или реч. Две варијанте јединица превођења чине једну јединицу превођења (Translation Unit - TU) те се тако двојезични паралелни корпус може дефинисати и као скуп јединица превођења. Процес успостављања веза између одговарајућих варијанти јединица превођења, односно формирање сета јединица превођења назива се паралелизација односно упаривање текстова. Потпуно паралелизовање текстова остварено је када су јединице текста једног језика у потпуности поравнате са јединицама текста другог језика (Ристовић 2012, 56).

Процес паралелизације подразумева да су текстови у електронском облику, коректно припремљени и упарени, а резултати представљени у одговарајућем стандардном формату. За илустрацију овог процеса навешћемо поступак припреме текстова и креирање паралелног корпуса који је разрадила Група за језичке технологије Универзитета у Београду, а њега чине следеће фазе:

1. припрема и сегментација текста на јединице упаривања (сегменте),
2. упаривање сегмената,
3. визуелизација паралелизованих текстова, контрола и корекција упаривања,
4. генерисање паралелизованог текста у формату TMX (Translation Memory eXchange – Преводилачка меморија за размену), више о овом формату касније у овом одељку,
5. разлагање текста у формату TMX на појединачне текстове у формату XML,
6. вертикализација појединачних текстова и креирање корпуса.

Све претходно поменуте фазе у поступку креирања једног паралелног корпуса подразумевају примену разноврсних рачунарских технологија и апликација, софтвера, као и бројних језичких ресурса за обраду и анализу одабраног текстуалног узорка. Како би се одабрани текстови могли обрађивати на овакав начин они морају бити у електронском облику. Са једне стране, постоји могућност да су текстови већ у електронском облику, доступни преко веба, односно да су настали као е-текстови (born-digital). Они могу бити бесплатно доступни те је довољно преузети их, прилагодити формат и даље обрадити или могу бити део неке базе података за коју је потребна ауторизација или можда чак одређена накнада за преузимање комплетног текста. Са друге стране, ако су текстови доступни само у папирном облику они се конвертују у електронски облик како би се даље могли обрађивати уз помоћ рачунарских технологија. То може подразумевати њихово прекуцавање или сканирање. Прекуцавање представља боље решење када су текстови у лошем стању и када процес сканирања не би дао жељене резултате или када сканер није доступан. Ипак, сканирање се чешће примењује. Сам процес сканирања као производ даје „слике” документа у одређеном сликовном формату, које се могу листати, прегледати, читати, али напредна употреба таквих „слика”, у смислу претраге целовитог

текста и повезивања садржаја са другим изворима на вебу, није могућа. Због тога се користи метода оптичког препознавања карактера (Optical Character Recognition - OCR) (оптичко препознавање карактера је детаљније објашњено у поглављу 4 одељак 4.2.1) која подразумева конверзију, односно превођење, слика текста (руком писаног, куцаног на машини или штампаног) у електронски текст. За оптичко препознавање карактера користи се данас разноврстан софтвер који омогућава и производњу рашчитаног текста у различитим форматима прикладним за даљу обраду.

Да би текстови могли бити аутоматски упарени коришћењем одговарајућег програма, они морају бити структурно анотирани одговарајућим обележјима проширивог језика за обележавање (eXtensible Markup Language - XML), а у складу са Смерницама Иницијативе за кодирање текста (Text Encoding Initiative - TEI). XML је најраспрострањенији формат за размену података између различитих рачунара, оперативних система и апликативних програма (Крстев и Витас 2008, 231). Настао је 1996. године са циљем да се олакша аутоматска обрада докумената и података. Настао је као замена за Стандардни генерализовани језик за обележавање (Standard Generalized Markup Language - SGML), мета-језик, који је омогућавао дефинисање конкретних специфичних језика за обележавање у складу са одређеним правилима (A brief SGML tutorial 2016) (ISO 8879:1986). Основна идеја је била да се створи језик, довољно једноставан, за структурирање информација тако да људи могу да их читају, а разноврсне апликације да их аутоматски обрађују (Extensible Markup Language 2016). Структура и детаљне карактеристике XML-а објашњене су у поглављу 3 одељак 3.4.1, док је структура XML датотека корпуса који се анализира детаљније објашњена у поглављу 6 одељак 6.2.4.

Иницијатива за кодирање текста (Text Encoding Initiative - TEI) је међународни пројекат који су 1987. године започеле три асоцијације: Асоцијација за коришћење рачунара у друштвеним наукама (Association for Computer and the Humanities – ACH)[6], Асоцијација за употребу рачунарских метода у филолошким и лингвистичким истраживањима (Association of Literary and Linguistic Computing – ALLC)[7] и Асоцијација за

6 „Association for Computer and the Humanities“, pristupljeno 26.3.2019, http://www.ach.org/

[7] Данас носи назив Европска асоцијација за дигиталну хуманистику (European Association for Digital Humanities – EADH). Доступно на: http://eadh.org/

рачунарску лингвистику (Association for Computational Linguistics – ACL)[8]. Циљ пројекта је био развијање, одржавање и објављивање хардверски и софтверски независних правила за обележавање електронских текстова (Text Encoding Initiative 2016). Као резултат рада на пројекту, 1994. године објављене су прве Смернице за кодирање и размену електронских текстова (Guidelines for Electronic Text Encoding and Interchange). У основи развијене TEI шеме налазио се тада нови стандард за означавање текста, SGML, помоћу кога су утврђене етикете и правила којима се могу описати структура и друга својства докумената. Дефинисана TEI правила су модуларно конципирана тако да корисници могу да одаберу делове који су им потребни за кодирање одређеног документа.

Због комплексности шеме временом је развијена њена једноставнија верзија, такозвана „TEI лајт" верзија, која се данас широко примењује у библиотекама. Са развојем XML језика, кодна шема SGML је замењена кодном шемом XML, а временом су се и Смернице модификовале, те је од новембра 2007. године у употреби P5 верзија.[9] Поред детаља о томе како треба кодирати текст, ове смернице дефинишу и део уграђеног заглавља који садржи метаподатке о делу, такозвано TEI заглавље. Структура TEI заглавља заједно са поступком израде метаподатака детаљније је објашњена у поглављу 4 одељак Иницијатива за кодирање текста и TEI заглавље.

Структурна анотација материјала одабраног за корпус подразумева означавање логичке структуре текста (поглавља, пасуса, реченица, речи) XML ознакама у складу са одређеним правилима. Поступак структурне анотације може бити урађен и аутоматски ако постоје одговарајући ресурси. Међутим, често је потребна ручна корекција аутоматске анотације те се у ту сврху могу користити и различити едитори који омогућавају уређивање текстова у формату XML, проверу њихове добре формираности и валидацију (детаљније објашњено у поглављу 3 одељак 3.4.1) у односу на одговарајућу дефиницију типа документа (Document Type Definition - DTD). Поступак паралелизације текстова који је развила Група за језичке технологије Универзитета у Београду, а који је описан у поглављу 3 одељак 3.5, предвиђа анотацију текстова за паралелни корпус на нивоу одељака, пасуса и сегмената (у пракси најчешће реченица). Као варијанте јединица превођења користе се

[8] Доступно на: http://www.aclweb.org/

[9] P5 верзија Смерница доступна је на: http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

сегменти, али се упаривање одговарајућих делова текста врши на сва три нивоа ради веће прецизности што је потребно и за рад програма за паралелизацију. Структурна анотација текстова за паралелни корпус који је предмет дисертације објашњена је у поглављу 6 одељак 6.2.4.

Овако припремљени текстови паралелизују се применом различитих програмских алата који аутоматски производе упарене сегменте. Приликом паралелизације најчешће се тежи постизању 1-1 упаривању сегмената, односно успостављању директне везе сегмената текста на изворном са еквивалентним сегментима текста на циљном језику. Међутим, због различите структуре изворног и циљног текста у пракси може доћи до погрешно упарених варијанти јединица превођења што захтева, најчешће ручну, корекцију добијених битекста. Погрешно упарене варијанте јединица превођења могу се добити из више разлога (Vitas and Krstev 2006, 150-151):

1. Због разлике између оригиналног и преведеног текста у погледу броја реченица или пасуса.
2. Због изостављања неких делова у преведеном тексту услед пропуста преводиоца приликом превођења или издавача приликом припреме текста за објављивање.
3. Због различитости у означавању пасуса. Врло често оригинални и преведени текст не садрже исти број пасуса. Иако је пасус логичка јединица структуре једног документа преводиоци врло често један пасус поделе на два или више мањих пасуса.
4. Због разлике у сегментацији реченица. Ово је проблем који се најчешће јавља углавном због правописних ознака које утичу на процес аутоматске сегментације. Програм који се користи за аутоматску сегментацију неће увек препознати када је крај реченице ако је она завршена неким знаком који системом није предвиђен као ознака за крај реченице. Тада долази до разлике у броју реченица и могу се десити следећи случајеви:
    a. сегмент изворног језика је једна реченица, а одговарајући сегмент циљног језика су две или више реченица и обрнуто;

b. сегмент изворног језика је једна реченица, а одговарајући сегмент циљног језика је део реченице и обрнуто;
c. сегмент изворног језика је једна реченица, а одговарајући сегмент циљног језика не постоји и обрнуто;
d. сегмент изворног и еквивалентан сегмент циљног језика се састоје од две или више реченица које нису у истом редоследу.

У оваквим случајевима најчешће се приступа ручној контроли и корекцији погрешно упарених сегмената како би се постигло 1-1 упаривање, уколико је могуће.

Следећи корак у процесу паралелизације је генерисање формата TMX паралелизованог текста. Формат TMX развила је Асоцијација за локализацију индустријских стандарда (Localisation Industry Standards Association - LISA). TMX је ISO стандард (ISO 24616:2018) за складиштење такозваних преводилачких меморија (Translation Memories) и њихову размену између различитих софтверских преводилачких алата, као и између различитих фирми које се баве одржавањем преводилачких меморија (TMX 2005). Преводилачке меморије представљају збирке одредница у којима је текст изворног језика повезан са еквивалентним преводом текста циљног језика односно произведени TMX документ састављен је од добијених јединица превођења. У оквиру стандарда TMX користе се и следећи ISO стандарди: ISO 639 за језик, ISO 3166 за земље и регионе и ISO 8601 за датуме и време.

Овако генерисани формат се у следећем кораку разлаже на појединачне текстове у формату XML, на XML документа за сваки појединачни језик који садржи информације о јединицама превођења. Добијене појединачне XML датотеке користе се као улазни текстови за паралелизовани корпус. За сваки од језика паралелних текстова креира се посебан корпус између којих се успостављају посебне везе које омогућавају паралелну претрагу. Пре креирања паралелног корпуса обавља се вертикализација XML документа која подразумева да се у једном реду налази само један токен (реч, број, знак интерпункције, XML-етикета, XML-коментар и слично) (Obradović, Stanković i Utvić 2008, 563-565). Поступак генерисања формата TMX паралелизованог текста, разлагање датотеке

на појединачне текстове у формату XML и вертикализација текстова за корпус који је предмет ове дисертације детаљније су објашњени у поглављу 6 одељак 6.2.5.

Након вертикализације, као последњи корак, креира се корпус применом програма за успостављање одговарајућих веза између добијених појединачних корпуса, а на основу постојећих јединица превођења. За паралелне корпусе као што су СрпФранКор и СрпЕнгКор (детаљније у поглављу 2.5.3) се користио програмски пакет IMS OCWB (Institut für Maschinelle Sprachverarbeitung Open Corpus Workbench)[10] (детаљније објашњено у поглављу 3 одељак 3.3.1) који омогућава креирање корпуса са структурном и морфолошком анотацијом, индексирање текстова и њихову паралелну претрагу коришћењем регуларних израза. Корпус који је предмет дисертације је смештен у дигиталну библиотеку Библиша што је детаљније објашњено у поглављу 6 одељак 6.3.

Претрага корпуса може бити на више начина. Један од начина је коришћење регуларних израза, а као резултат претраге задатог упита добијају се све ниске које одговарају постављеном регуларном изразу. Регуларни изрази и поступак претраге постојећих корпуса детаљније су објашњени у поглављу 3 одељак 3.6. Претрага корпуса може бити и преко метаподатака и двојезичних упита за претрагу комплетног текста у виду кључних речи. О метаподацима ће детаљније бити речи у поглављу 4 одељак 4.1, док су структура метаподатака корпуса који се анализира у дисертацији и претрага преко метаподатака објашњени у поглављу 6 одељци 6.3.1 и 6.3.3. Двојезична претрага преко упита у виду кључних речи над корпусом који је предмет дисертације детаљније је представљена у поглављу 6 одељак 6.3.3.

2.3.1 Улога паралелних корпуса у истраживањима из области језика

Језички корпуси представљају један од важнијих ресурса за истраживања у области лингвистике и сродним језичким дисциплинама било да је реч о морфологији, синтакси, семантици, дискурсу, прагматици, лексикологији, лексикографији, као помоћно средство за анализу књижевно-уметничких дела или текстова који припадају неком другом посебном функционалном стилу (новински, административни, разговорни, научни и

[10] Institut für Machinelle Sprachverarbeitung Open Corpus Workbench, http://cwb.sourceforge.net/

научно-популарни), а посебно место заузима статистичка анализа језика. У зависности од нивоа анотације, корпуси омогућавају истраживачима да уоче различите примере употребе језика, да позивањем сета речи једног језика и њихових граматичких облика приликом истраживања сагледају фреквентност појављивања одређене речи или фразе постављене кроз упите за претрагу, у којим се све облицима и варијантама постављена реч или фраза појављује, као и у каквој је семантичкој корелацији са другим речима и фразама и њиховим облицима у датом корпусу. Такође, са новим технологијама повезивања садржаја на вебу могуће је сагледати фреквентност појављивања речи или фразе у разним базама података, репозиторијумима, онтологијама, контролисаним речницима и сличним изворима.

Последњих пар деценија развијен је велики број вишејезичних паралелних корпуса на различитим језицима који су своју примену нашли у свим поменутим лингвистичким дисциплинама. Њихова предност је што садрже преводе истог текста на два или више језика што даље омогућава истраживачима да сагледају све предности корпуса, претходно наведене, упоредо на два или више језика. Они су постали центар истраживања у области обраде природних језика како у теоријским областима лингвистике као што су контрастивна лингвистика и лексикографија, тако и у практичним језичким областима као што су превођење, терминолошка екстракција или производња преводилачких меморија (Laporte et al. 2006, 110).

Са једне стране, језичким професионалцима као што су лексикографи и писци, коришћење паралелних корпуса омогућава да постигну боље резултате у раду, док се са друге стране паралелни корпуси могу користити у настави страних језика. Употреба корпуса и алата за анализу корпуса у наставним процесима омогућава ученицима и студентима да постану бољи познаваоци језика користећи развијене рачунарске алате за анализу текста, засноване на савременим технологијама, који су постали правило, а не изузетак у свакодневном раду (Bernardini et al. 2003, 4-5).

Паралелни корпуси су посебно значајан извор у вишејезичној лексикографији јер омогућавају корисницима да сагледају упарене еквиваленте одређених сегмената текста као и шири контекст који не постоји у класичним двојезичним речницима (Simeon 2002,

212). Двојезична или вишејезична лексикографија представља област примењене лингвистике чији је примарни циљ превођење уз помоћ вишејезичних речника. Вишејезични речници, вишејезичне лексичке базе података и вишејезични лексикони нека су од средства којима истраживачи располажу у оквиру вишејезичне лексикографије. Двојезични и вишејезични речници и лексикони нису само ресурси неопходни за превођење текстова већ представљају лингвистичка средства која много говоре о језику и значењу уопште.

Данас, једнојезични и двојезични речници, како општи тако и терминолошки, као и енциклопедије доступни су не само на папиру него и у електронском облику (Mosavi Miangah 2006, 43). Они су добар извор за разрешавање двосмисленог значења речи што је неопходно за развој алата за обраду природних језика. Волфганг Тојберт (Wolfgang Teubert) нагласио је позитивни утицај коришћења електронских корпуса на динамику и квалитет лексикографских истраживања и активности. Истраживања у области двојезичне и вишејезичне лексикографије достиже нови ниво квалитета експлоатацијом корпуса (Teubert 1996, 242). Веза између вишејезичне лексикографије и рачунарске технологије омогућава стварање разноврсних електронских вишејезичних речника неопходних у процесу превођења, али и у процесима стварања апликација за обраду природних језика.

Основна примена вишејезичне лексикографије јесте у области превођења. Постоје две врсте поступка превођења. Једна је превођење страног текста на матерњи језик преводиоца, док други вид превођења представља превођење текста са матерњег језика преводиоца на неки од страних језика. У оба случаја преводилац мора да пронађе еквивалентан превод преносећи исто значење из оригиналног текста у духу жељеног језика. Када су у питању службена документа, научни радови, приручници, правна документа и слично не води се рачуна толико о духу самог језика колико о техничким концептима са утврђеном терминологијом у различитим језицима. Технички термини и њихови еквивалентни преводи у другом језику смештени су у вишејезичне терминолошке базе података.

Речници представљају основно средство у процеса превођења. Вишејезична лексикографија у оквиру корпусне лингвистике подразумева испитивање лингвистичких и

нелингвистичких односа између одређених речи које помажу у изградњи речника. Идеја је да се у корпусу циљног језика пронађе потребна информација, односно да се пронађе адекватан превод за одговарајућу реч који затим треба уклопи у контекст. Тада преводилац може да користи синтаксичку конструкцију коју је пронашао у жељеном резултату претраге. Коришћењем упарених сегмената из паралелних корпуса двојезична и вишејезична лексикографија достиже нови ниво квалитета. Међутим, постојање самог корпуса није довољно. Задатак вишејезичне лексикографије је да дизајнира алате који ће бити примењени на корпусу, да добро организује добијене резултате и адекватно их представи преводиоцу.

Данашњи преводиоци имају могућност да истражују и користе информације садржане у паралелним корпусима користећи разноврсне алате за анализирање садржаја корпуса (Mosavi Miangah 2006, 46). Кроз међусобно упарене структурне јединице текста у једном паралелном корпусу преводиоци имају могућност да сагледају које све варијанте превода одређеног сегмента текста (речи, фразе, реченице) постоје у другом језику. Федерико Занетин (Federico Zanettin) је још деведесетих година 20. века нагласио значај вишејезичних корпуса у области превођења истичући потребу за повећањем броја паралелних вишејезичних корпуса, као и потребу за сталном едукацијом корисника у коришћењу корпусних ресурса посебно у образовне сврхе (Zanettin 2002, 10).

Паралелни вишејезични корпуси последњих деценија нашли су велику примену у настави страних језика. Пионир у употреби паралелних корпуса у настави учења страних језика јесте Тим Џонс (Tim Johns) који је осамдесетих година 20. века представио упарене конкорданце на часу страног језика. Паскалева и Михов су у оквиру Лабораторије за лингвистичко моделовање у Софији вршили испитивања употребе вишејезичних корпуса у процесу наставе страног језика, фокусирајући се углавном на развој преводилачких способности учесника у истраживању, истичући тако предности употребе рачунарских алата за креирање и уређивање садржаја корпуса из угла креатора, као и њихову употребу из угла корисника (Paskaleva and Mihov 1997, 43). На Универзитету Северне Аризоне Ренди Репен (Randi Reppen) спровела је експериментално истраживање примене корпуса у настави. Резултати експеримента представљени су у књизи “Using Corpora in the

Language Classroom" (Reppen 2014) у којој је ауторка истакла „неопходну постепену преквалификацију корпуса у примени кроз клизни прелаз од корпуса као генератора текстуалних колекција, преко корпуса као активатора разноврсних језичких активности, ка корпусу као дисеминатору нових практичних методичких инстанци експлоатације у директном наставном процесу" (Ристовић 2016, 16).

Јединствен пример из праксе употребе паралелних корпуса у настави страних језика код нас представља нови наставни метод чији је аутор је др Зоран Ристовић, професор енглеског језика у ОШ „Мито Игумановић" у Косјерићу. Идеја Зорана Ристовића била је да испита иновативни приступ у настави, учењу и усвајању страног језика међу децом школског узраста кроз експлоатацију једног савременог језичког алата као што је паралелни корпус, као и да анализира резултате које доноси овакав метод рада. Према Ристовићу „контекстуално богат паралелизовани улазни језички садржај омогућава остваривање нелинеарног учења, а корпусно засновани језички задаци пружају окружење за епизодичне инстанце свесног и несвесног усвајања морфологије, синтаксе и граматике циљног страног језика" (Ристовић 2016, 5).

Резултати експеримента представљени су у докторској дисертацији *Кумулативни ефекти експлоатације вишејезичних корпуса у настави страних језика* (Ристовић 2016). За експеримент припремљен је паралелни српско-енглески корпус књижевних текстова Пола Остера (Paul Auster) и Ернеста Хемингвеја (Ernest Hemingway). Књижевни текстови Пола Остера представљају основу овог корпуса са 220.000 речи, док део корпуса са текстовима Ернеста Хемингвеја садржи 30.000 речи. „Корпусно подржана настава, учење и усвајање циљног страног језика манифестовала се, кроз укључивање корпусних пакета у непосредни наставни процес, иницирање реконструктивних, конфронтативних и продуктивних активности у оквиру корпусних дуала, и праћење и у извесној мери кроз репликативне циклусе циљане експлоатације кориговање и усмеравање кумулативних тенденција и ефеката током експлоатације у учењу и усвајању циљног страног језика, као хибридни модел наставе и учења страног језика" (Ристовић 2016, 357).

2.3.2 Машинско превођење

Паралелни корпуси имају важну улогу било као основа за изградњу нових платформи, алата и софтвера за аутоматско превођење, било као додатни ресурси којима се унапређују и надограђују већ постојећи системи и алати за машинско превођење. Задатак машинског превођења јесте да се на основу успостављеног алгоритма одабере реченица на циљном језику која је највероватнији превод дате реченице на изворном језику. Озбиљна истраживања из области машинског превођења започета су после Другог светског рата са појавом рачунара, односно педесетих година 20. века (Slocum 1985, 1), али су тек осамдесетих година, са развојем и напретком рачунарских технологија и система за обраду података, ови системи почели да се користе у комерцијалне сврхе (Cribb 2000, 562). Први систем за машинско превођење развијен је 1954. године у сарадњи IBM-а и Универзитета Џорџтаун (Georgetown University) као демонстративни пример ове идеје. Временом машинско превођење имплементирано је у различите области, а један од првих је систем Mateo које је у Канади коришћен од 1977. године за превођење временских прогноза са енглеског на француски језик (Hutchins 1997, 115).

Последњих деценија доста је рађено на развоју система за машинско превођење који користе упарене јединице паралелних корпуса за изградњу модела превођења. Системи за машинско превођење засновани на вероватноћи преводилачких модела тестирани су, најопштије говорећи, коришћењем паралелних корпуса који су поравнати на нивоу реченица (Callison-Burch et al. 2004b). У већини статистичких приступа система за машинско превођење основне јединице превођења су фразе, док су основне компоненте преводилачких система модели који процењују преводилачке могућности упарених фраза изворног и циљног језика (Kalchbrenner and Blunsom 2013, 1700). Један од високо развијених система за машинско превођење заснован на оваквом приступу јесте Moses[11]. Moses је систем за машинско превођење који омогућава аутоматско тестирање модела превођења за било који одабрани језички пар (Koehn et al. 2007, 177). Када се једном утврди модел превођења алгоритам за претрагу на основу њега брзо проналази најпогоднији превод између понуђених избора за сваки задати пример. Систем је

[11] Moses, http://www.statmt.org/moses/

оригинално развијен на Универзитет у Единбургу и данас га користи доста академских институција као основну инфраструктуру у истраживањима из области машинског превођења (Koehn and Schroeder 2007, 224). Moses користи фразно заснован приступ (Koehn et al. 2003) који подразумева да се улазна реченица дели на текстуалне одломке, који су често на нивоу фразе, које се затим упарују са одломцима циљног језика. Састоји се од компонената за обраду података, тестирања језичких модела и модела превођења, али и од алата за подешавање ових модела коришћењем минималне стопе грешака приликом тестирања и евалуације добијених превода (Koehn et al. 2007, 178).

Поред традиционално фразно заснованих система за машинско превођење (Callison-Burch et al. 2004a) у последњих пар година ради се на неуронском машинском превођењу (Sutskever et al. 2014) (Cho et al. 2014), новом приступу у овој области. У односу на фразно засноване системе за машинско превођење неуронско машинско превођење покушава да изгради и тестира јединствену, велику неуронску мрежу која чита реченице одабраног корпуса текстова и производи исправан превод (Bahdanau et al. 2014). Већина ових система заснована је на принципу кодирања и декодирања језичког пара који подразумева да се реченица изворног језика кодира у вектор фиксне дужине на основу кога декодери аутоматски производе жељени превод на циљном језику (Hermann and Blunsom 2014).

2.4 Неки примери паралелних корпуса у свету

Паралелни корпуси као колекције текстова на два или више језика најчешће садрже административне, правне или новинске текстове, те заправо представљају корпусе подјезика. Паралелни корпуси књижевних текстова су били знатно већи али се последњих година интензивно ради и на њиховом креирању. Књижевни текстови су најчешће прецизно преведени и пружају добар увид у језик и граматичке феномене који се не могу тако детаљно сагледати у другим врстама текстова. Такође, поравнати корпуси књижевних текстова у односу на корпусе подјезика нуде више могућности за истраживања из различитих области језика: учење језика, лексикографија, упоредне студије језика. Вишејезични паралелни корпуси књижевних текстова најчешће су много

мањи у односу на корпусе правних или новинских текстова јер њихова припрема захтева више времена и рада, али је и њихов број последњих година у порасту.

2.4.1 Корпуси некњижевних текстова

Acquis Communautaire

Acquis Communautaire (Acquis) је широко прихваћен француски назив за скуп правних докумената и обавеза који све државе чланице обавезује и повезује унутар Европске уније, једном речју правна регулатива Европске уније (ЕУ). Све државе чланице ЕУ морају да поштују ову правну регулативу, а све земље кандидати за чланство морају да је прихвате како би постале део ЕУ. Ова колекција се стално развија и обухвата документа која су писана од 1958. године до данас: садржаје, начела и политичке циљеве уговора ЕУ; законодавство ЕУ; усвојене стандарде који се односе на законодавство ЕУ и међународне споразуме; декларације и резолуције; међународне споразуме; акта и заједничке циљеве (Miller 2011).

Колекција Acquis послужила је за формирање вишејезичног паралелног корпуса под називом *JRC Collection of the Acquis Communautaire* (JRC-Acquis) који садржи документе из ове колекције на 22 званична језика Уније. Као и сва друга званична документа Европске комисије и Европског парламента и документа у колекцији Acquis су класификована коришћењем тезауруса EuroVoc,[12] вишејезичног, хијерархијски организованог портала различитих речника за ручно класификовање докумената (Steinberger et al. 2006).

[12] EuroVoc је вишејезични, мултидисциплинарни тезаурус који садржи близу 7.000 хијерархијски организованих предметних домена који се користе за класификацију и проналажење званичних докумената Европске уније и Европског парламента. Тезаурус садржи термине на 23 језика Европске уније (бугарски, грчки, дански, енглески, естонски, италијански, летонски, литвански, мађарски, малтешки, немачки, пољски, португалски, румунски, словачки, словеначки, фински, француски, холандски, хрватски, чешки, шведски и шпански), уз додатак македонског (mk), албанског (shqip-sq) и српског (sr). EuroVoc користи онтолошки заснован програм за управљање тезаурусом, као и технологије семантичког веба које су у складу са препорукама веб конзорцијума (W3C) и са најновијим трендовима у развоју стандарда за израду тезауруса. Тезаурус је организован у класе термина који су хијерархијски повезани везама „надређени термин“ - „подређени термин“, као и „повезани термин“ којим се повезују термини који дати термин додатно описују али нису са њим у хијерархијском односу. Доступно на: http://eurovoc.europa.eu/.

Корпус JRC-Acquis садржи преко 4 милиона поравнатих докумената, око 20.000 докумената по језику. Процес компилирања корпуса састојао се из неколико корака (Erjavec et al. 2005, 532):

1. *Преузимање текстова*: сви текстови који су припремани за корпус преузети су са званичне веб стране ЕУ[13]. Приликом формирања корпуса бирана су документа која имају јединствени идентификатор, CELEX ID број, и за која постоји превод бар на десет званичних језика од којих три морају да буду језици држава које су постале чланице ЕУ до 2004. године. CELEX ID број је посебно структуриран број фиксне дужине којим су кодиране одређене информације, као што су тип документа, година издавања и слично.
2. *Идентификација језика докумената*: након одабира документа идентификован је језик текста. За идентификацију језика текста коришћен је одговарајући софтвер који је сам одбацивао документа која нису на очекиваном језику. За мали број докумената текстови на одређеном језику су били, у ствари, непреведени енглески текстови те такви текстови нису уврштени у корпус.
3. *Структурирање текстова*: скоро сваки текст у колекцији Acquis има структуру која се састоји од наслова, тела текста, потписа и анекса који су означени одговарајућим XML етикетама у складу са препорукама TEI конзорцијума. Потпис документа је посебно интересантан део сваког текста јер садржи податке о личним именима, именима места, датумима, референцама на друга документа те је врло погодан за препознавање и обраду именованих ентитета.
4. *Лингвистичка анотација текстова*: текстови су анотирани морфосинтаксичким ознакама, а речима су додељене одговарајуће леме и морфосинтаксички описи.
5. *Паралелизација*: поступак паралелизације урађен је на нивоу пасуса и реченица, а за паралелизацију су коришћени алати Vanilla (Danielsson and Ridings 1997) и HunAlign (Varga et al. 2005).

[13] European Commission: International Cooperation and Development, http://ec.europa.eu/development/body/legislation/policy-papers2_en.htm

Као што је већ објашњено у одељку 2.3 текстови за паралелизацију морају бити у одговарајућем формату, структурно анотирани. Прва фаза у припреми корпуса подразумевала је преузимање текстова који су из затим формата HTML конвертовани у формат XML у складу са смерницама TEI конзорцијума које су биле важеће у тренутку формирања корпуса (P4) при чему су сви текстови морали да буду кодирани у Unicode-у, трансформациони формат UTF-8. Процес паралелизације подразумевао је упаривање одговарајућих сегмената текстова уз помоћ наведених алата који омогућавају упаривање на више начина: 1-1, 1-2 (подела једног сегмента на два), 2-1 (комбинација два сегмента са једним), 0-1 (додавање реченице) и 1-0 (брисање реченице). Након завршеног процеса паралелизације постигнуто је 1-1 упаривање сегмената од скоро 90%.

Поред процеса паралелизације текстова урађена је и анотација корпуса која је подразумевала токенизацију (идентификација речи и интерпункцијских знакова у тексту), означавање врстама речи (додељивање морфосинтаксичких карактеристика) и лематизацију (додељивање канонског облика речи - леме) (Tufiş et al. 2009, 40). За процес анотације коришћени су ажурирани лингвистички ресурси развијени у оквиру пројекта MULTEXT-East (Multilingual Text Tools and Corpora for Eastern and Central European Languages)[14] који су имплементирани у посебан програм намењен означавању вишејезичног садржаја. Прво је урађена токенизација, затим су речи добијене токенизацијом означене одговарајућим морфосинтаксичким описом и на крају су тим речима додељене леме. Програм може да произведе излаз у неколико формата, укључујући формат табеле и одговарајући XML формат (Erjavec et al. 2005, 533).

С формалне тачке, формирани корпус се састоји из два дела: докумената и поравнања. Сва документа су груписана према језику, док сви текстови једног језика чине појединачан TEI корпус. Сваки од ових појединачних корпуса састоји се из TEI заглавља, метаподатака о језичком корпусу и самих докумената. Када је реч о документима, сваки који је део корпуса садржи своје TEI заглавље које поред осталих метаподатака садржи и релевантну URL адресу пуне верзије текста и кôд из EuroVoc тезауруса, као и сам текст документа (Steinberger et al. 2006, 5). Посебан део корпуса чине датотеке поравнатих

[14] MULTEXT-East Home Page, pristupljeno 26.3.2019, http://nl.ijs.si/ME/

текстова које не садрже текстове као целине већ су састављене од показивача ка упареним варијантама јединица превођења, односно, од добијених јединица превођења. Овај корпус послужио је за израду система за машинско превођење за 462 језичка пара (Koehn, Birch and Steinberger 2009, 65-66).

EuroParl

EuroParl[15] је паралелни корпус који садржи документа Европског парламента од 1996. године на 21 званичном језику Европске уније. Рад на корпусу започет је са циљем спровођења истраживања у области статистичког машинског превођења односно генерисања реченица из поравнатих текстова за стварање система за статистичко машинско превођење. Међутим, од тренутка када је 2001. године иницијално представљен, корпус је почео да се користи и у другим областима обраде природних језика: отклањање вишезначности, екстракција информација и слично. Последња верзија корпуса, V7, објављена је 15.05.2012. године. EuroParl садржи 20 појединачних паралелних корпуса, а поравнање је урађено са енглеском верзијом докумената („European Parliament" 2016).

Израда корпуса EuroParl урађена је у пет корака (Koehn 2005, 79-81):

1. *Прикупљање такозваних сирових текстова.* Сви текстови преузети су са веб сајта Европског парламента у HTML формату. URL адреса сваког документа садржи релевантну информацију за идентификацију као што су језик документа, дан и број дискусије, као и број исказа.
2. *Упаривање докумената.* Први корак у овој фази подразумевао је сортирање текстова према тематици, а затим и упаривање докумената на појединачним језицима са њиховом енглеском верзијом. У овој фази урађено је само упаривање целих докумената без процеса сегментације.
3. *Сегментација текстова на реченице*.
4. *Припрема корпуса за системе статистичког машинског превођења (нормализација, токенизација)*.

[15] European Parliament Proceedings Parallel Corpus 1996-2011, http://www.statmt.org/europarl/

5. *Упаривање сегмената (реченица).* Процес упаривања сегмената урађен је уз помоћ одговарајућег алгоритма који је описан у (Gale and Church 1993).

EuroParl је направљен за тестирање система за статистичко машинско превођење где се везе између изворног и циљног језика испитују на основу поравнатих реченица та два језика (van Halteren 2008, 937). На корпусу је примењена методологија фразно заснованог модела у статистичком машинском превођењу која подразумева извлачење упарених језичких фраза из датог паралелног корпуса на основу чега се креира табела вероватноће превода ових фраза. Постоји више метода за реализацију ове методологије, а неколико њих почиње аутоматским добијањем поравнатих речи на основу чега следи прикупљање фраза у виду језичких парова (Koehn 2005, 82). EuroParl појединачни паралелни корпуси могу се бесплатно преузети са званичне веб стране корпуса.

SETimes

SETimes (South-East European Times)[16] је била веб страна која је објављивала вести из Југоисточне Европе (са подручја Албаније, Босне и Херцеговине, Бугарске, Грчке, Македоније, Румуније, Србије, Турске, Хрватске и Црне Горе) односно са Балканског полуострва на десет језика. На основу материјала са сајта креиран је вишејезични паралелни корпус SETimes (Tyers and Alperen 2010). Корпус је настао по угледу на већ поменуте корпусе, JRC-Acquis и EuroParl, а за потребе тестирања система за машинско превођење и различитих истраживања из области обраде природних језика у оквиру вишејезичних истраживања. Корпус, поред материјала на енглеском језику, садржи и материјале на језицима такозване балканске језичке заједнице (Balkan Sprachbund/Balkan language area) који су сродни по лексичким и граматичким карактеристикама али пре свега на основу географске близине.

Материјал за корпус прикупљен је са SETimes сајта коришћењем XENU веб-паука. Прво су прикупљени материјали на енглеском језику на основу којих су прикупљени и материјали на другим језицима. Преузети материјали кодирани су у UTF-8 трансформационом формату, а сегментација на реченице извршена је коришћењем алата

[16] SETimes је био доступан на http://www.setimes.com/

SentParBreaker. Процес паралелизације урађен је коришћењем алата HunAlign, а садржај је поравнат на нивоу реченица са енглеским језиком, али и између заступљених језика. Овако добијен корпус искоришћен је за тестирање фразно заснованог система за статистичко машинско превођење Moses (детаљније објашњено у одељку 2.3.2 овог поглавља). У тренутку писања дисертације званична веб страна SETimes није била у функцији, али се паралелни корпуси вести могу бесплатно преузети[17].

OpenSubtitles

OpenSubtitles[18] представља базу података која садржи близу 4 милиона превода филмова и серија на преко 60 светских језика. Како преводи, са лингвистичке тачке гледишта, покривају широк и интересантан опус жанрова, од колоквијалног језика или сленга до наративног дискурса, ова база података послужила је за израду паралелног корпуса *OpenSubtitles*[19]. Процес израде овог корпуса прошао је кроз фазе претходне обраде и паралелизације које су у појединим сегментима биле специфичне због природе материјала који се анализирао и обрађивао.

Пре поступка претходне обраде све датотеке превода[20] конвертоване су у датотеку са проширењем .srt, а за сваки превод прикупљене су следеће информације:

1. јединствени идентификатор,
2. листа датотека (у случају када се филмови налазе на више од једног CD-а),
3. кôд за језик филма и превода и формат превода,
4. генеричке информације о филму као што су наслов, година производње, идентификатор на сајту IMDb[21],
5. разноврсни атрибути као што су датум ажурирања датотеке, број преузимања и корисничка процена.

[17] Паралелни корпуси су доступни на http://nlp.ffzg.hr/resources/corpora/setimes/
[18] „OpenSubtitles“, pristupljeno 26.3.2019, http://www.opensubtitles.org/en/search
[19] Корпус је доступан на: http://opus.lingfil.uu.se/OpenSubtitles2016.php
[20] Датотека која садржи податке о преводу: редне бројеве сегмента превода, почетак и крај сегмената превода, текст превода. Датотеку превода овог формата подржавају DivX и DVD видео формати.
[21] Internet Movie Database, филмска интернет база података која садржи информације о познатим светским филмским и телевизијским личностима, филмовима, телевизијским емисијама, серијама, рекламама и видео-играма. Доступно на: http://www.imdb.com/

Поступак претходне обраде превода одабраног филмског и серијског материјала подразумевао је неколико корака, а као излаз добијене су XML датотеке (по једна за сваки превод) које се састоје од структурираних листа реченица над којима је извршена токенизација. Први корак у овој фази било је кодирање текста. Како база OpenSubtitles не прописује кôд за текстове, било је неопходно конвертовати их из локалног кодирања у Unicode који је одабран као најпогоднија кодна шема. У следећем кораку извршена је сегментација и токенизација. Датотеке превода структуриране су у блокове који представљају кратке сегменте текстова, са временом почетка и краја приказивања, а који представљају редове од 40 до 50 карактера (највише два реда на екрану) у периоду приказивања од 1 до 6 секунди (Aziz, Sousa and Specia 2012, 103). За процес токенизације и сегментације на реченице написан је алгоритам који прво препознаје редове текстова над којима се врши токенизација, а затим и крај реченице формирајући блокове превода (Lison and Tiedemann 2016, 924-926). Након токенизације и сегментације урађена је исправка грешака (текстуалних и грешке у акцентима у одређеним језицима) које су препознате у поступку OCR-а, а затим и генерисање метаподатака о сваком преводу:

1. генерички атрибути као што су година производње, оригинални језик, трајање и жанр филма или серијске епизоде;
2. атрибути превода као што су језик превода, датум ажурирања, оцена превода на сајту OpenSubtitles и трајање превода;
3. вероватноћа да се одређени језик превода слаже са језиком који је коришћен као језик превода текста;
4. карактеристике процеса конверзије као што су број извучених реченица, број токена, број грешака препознатих системом OCR и датотека кодирања.

Процес паралелизације је такође прошао кроз неколико корака. Први корак подразумевао је означавање парова превода које треба упарити, а затим је примењен алгоритам временског преклапања (Tiedemann 2007) уз помоћ кога је извршено упаривање реченица на основу дужине трајања текстуалних блокова превода који се појављују на екрану. Процесом упаривања добијено је 1689 битекстова са укупно 50 милиона упарених реченица (Lison and Tiedemann 2016, 927). Квалитет добијених

битекстова тестиран је у систему Moses (детаљно објашњен у одељку 2.3.2 овог поглавља). По истом принципу урађена је и паралелизација алтернативних превода за језике за које они постоје. OpenSubtitles омогућава корисницима претраживање превода за филмове и серије који се могу преузети уз одговарајућу регистрацију која се обавља на самој веб страни базе.

BabelNet

Корпус дефиниција BabelNet[22] је вишејезична лексичка семантичка мрежа састављена од аутоматски интегрисаних лексикографских и енциклопедијских садржаја који се преко одговарајућег алгоритма аутоматски интегришу из лексикона и речника као што су Ворднет[23], Википодаци[24], Википедија[25], Викиречник[26] и OmegaWiki[27] (Navigli and Ponzetto 2012, 218). Систем омогућава и аутоматско повезивање садржаја између поменутих ресурса стварајући BabelNet синсетове и релације између њих. BabelNet садржи дефиниције на преко 250 светских језика чинећи тако велики и свеобухватан језички ресурс погодан за отклањање семантичке двосмислености (Camacho-Collados et al. 2016, 1701), али и тестирање система за машинско превођење. Може се користити као енциклопедијски речник, семантичка мрежа или база знања.

Процес отклањања семантичке двосмислености дефиниција урађен је аутоматски из паралелизованих дефиниција. У овом процесу коришћени су следећи алати: Babelfy[28] за одређивање значења дефиниција (и генерално било ког текста) после обављене

[22] BabelNet, приступљено 26.3.2019, http://babelnet.org/

[23] Ворднет је машински читљива лексичка база података створена као допуна за конвенционалне речнике. Не садржи неке карактеристике конвенционалних речника као што су дефиниције свих врста речи, изговор, акценат, етимологију, белешке о употреби, док са друге стране, садржи својства које се не могу наћи у конвенционалним лексикографским делима. База је организована преко чворова, синсетова (скупова речи које у неком контексту имају исто значење) и релација између њих. База је подељена у делове који су организовани као хијерархијска мрежа чворова која се успоставља на основу постојања релације подређености и надређености између појмова које ти чворови представљају. Поред релације подређен-надређен успостављене су и релације део-целина, члан-целина и многе друге. Доступно на: https://wordnet.princeton.edu/

[24] Wikidata, https://www.wikidata.org/wiki/Wikidata:Main_Page

[25] Wikipedia, https://www.wikipedia.org/

[26] Wiktionary, https://www.wiktionary.org/

[27] OmegaWiki, http://www.omegawiki.org/

[28] Babelfy, http://babelfy.org/

претходне обраде и анотације, затим напредан вишејезични систем за разрешавање значења речи и повезивање ентитета и алат NASARI[29] за репрезентацију семантичких вектора. Цео процес разрешавања значења прошао је кроз два корака. У првом кораку све дефиниције прикупљене су заједно, груписане према сродности и уз помоћ одговарајућег вишејезичног система одређено је њихово значење. У следећем кораку урађено је разрешавање вишезначности (дезамбигуација) добијених резултата према семантичкој сличности (Camacho-Collados et al. 2016, 1706). BabelNet се може бесплатно претраживати, а могуће је и бесплатно преузети сетови података.

2.4.2 Корпуси књижевних текстова

Платонова Република

Платонова Република је вишејезични паралелни корпус настао у оквиру пројекта TELRI (Trans-European Language Resources Infrastructure).[30] Резултат рада на пројекту је CD-ROM који садржи вишејезичне стандардизоване језичке ресурсе за велики број језика, углавном из земаља које нису у том тренутну биле део Европске уније, а за које су овакви ресурси још увек били ретки: паралелне корпусе, лексичке ресурсе и алате за обраду природних језика (Erjavec, Lawson and Romary 1998, 981).

Добијени CD-ROM се састоји из два тома. Садржај првог тома настао је у потпуности у оквиру акције TELRI у процесу припреме заједничког корпуса, док други том садржи резултате европског пројекта Multext-East који ће бити представљен у следећем одељку. Оба тома су слична у погледу врсте ресурса које нуде и система кодирања који користе, а разликују се у погледу организационе структуре коју садрже.

Први том добијеног CD-ROM-а представља паралелни корпус дела *Република* грчког филозофа Платона које је паралелизовано на 21 европски језик. Из практичних разлога за припрему паралелног корпуса одабран је један текст за који су постојали преводи на већини језика учесника пројекта. Осим Естоније и Албаније где превод овог

[29] Доступно на: http://lcl.uniroma1.it/nasari/
[30] Trans-European Language Resources Infrastructure, http://telri.nytud.hu/

дела није постојао, свака од земаља партнера понудила је одговарајући превод, а на неким језицима, као што су енглески и чешки, постојало је и више различитих превода.

Приликом прикупљања доступних превода дела *Република* на пројектним језицима било је различитих ситуација. Свака доступна електронска верзија превода која је постојала искоришћена је, док су остале сканиране и урађено је оптичко препознавање карактера. У случајевима када је квалитет штампаног текста био лош текст је ручно прекуцаван. Када је реч о енглеском језику постојало је више доступних верзија текста, али са друге стране, на пример, није могла да се пронађе потпуна верзија текста на летонском језику. Пронађено је више делова из различитих превода на летонском који су посебно обрађивани и на крају спојени у један текст. У неким случајевима једини доступни примерци налазили су се у библиотекама или у приватном власништву. Како су поједини примерци из библиотека били оштећени, процес сканирања се није могао применити те су текстови прекуцавани. Такође, како нико од учесника у пројекту није познавао антички грчки, корпус је остао без оригиналне верзије дела односно изворног текста па сви текстови који су заступљени у корпусу представљају текстове на циљним језицима (Erjavec, Lawson and Romary 1998, 982) који су поравнати са енглеском верзијом.

Друго питање које је разматрано у оквиру пројекта јесте питање ауторских права. За све текстове који су још увек били под заштитом ауторских права била је потребна дозвола за њихову употребу за потребе академског истраживања. Дозволе за употребу текстова добијене су на основу писма које је саставила радна група у оквиру пројекта за пројектне партнере. У случају када текст није више био под заштитом ауторских права текстови су слободно коришћени без захтевања потребне дозволе.

Искуство приликом израде овог корпуса показало је да процес паралелизације није једноставан задатак, посебно ако је у питању текст са комплексном језичком структуром за који је тешко пронаћи оригиналну верзију (Vitas and Krstev 2006, 148). Разлике у преводима за поједине језике у односу на одабрани енглески превод правиле су велике проблеме током процеса паралелизације који због тога није дао убедљиве крајње резултате.

Како је реч о класичном филозофском тексту насталом пре више од два миленијума многи преводи пратили су специфичну структуру оригиналног текста коју је било тешко обрадити. За кодирање текстова коришћен је језик SGML, а у складу са тада важећим смерницама TEI P3. Текстови су анотирани до нивоа реченице, процес паралелизације урађен је аутоматски системом Vanilla, а добијени резултати су ручно проверени и поправљени. Тако произведени текстови су првенствено коришћени за тестирање софтвера за паралелизацију, док су добијени паралелни текстови коришћени за испитивање специфичних карактеристика одређеног језичког пара, као и за упоређивање превода одређених речи или фраза. У оквиру пројекта урађена је и паралелизација са српским преводом овог дела (Vitas et al. 1998).

Орвелова 1984 (Multext-East): паралелни преводи романа *1984* Џорџа Орвела

Орвелова 1984 је вишејезични паралелни корпус развијен у оквиру добро познатог вишејезичног лингвистичког ресурса Multext-East. Multext-East је развијен као део пројекта Multext - Језички ресурси и евалуација (Language Resources and Evaluation - LRE) који је трајао од 1995. до 1997. године, а финансиран је у оквиру Copernicus програма (Dimitrova et al. 1998, 315). Како је за западно-европске језике велики број стандардизованих језичких ресурса и алата већ постојао или је био у развоју, Multext-East пројекат је покренут да би се развили стандардизовани језички ресурси и алати за екстракцију информација из језичких корпуса CEE језика (Central and Eastern European Languages) који би били бесплатно доступни за потребе лингвистичких истраживања (Erjavec et al. 2003, 25). Конкретни циљеви пројекта били су: тестирање и прилагођавање језичких стандарда, изградња анотираног вишејезичног корпуса, развој морфолошких и лексичких ресурса и прилагођавање Multext корпусних алата (Erjavec and Ide 1998, 971). Постављени циљеви били су и резултати прве верзије пројекта који су остали језгро и у његовим наредним верзијама.

Добијени анотирани вишејезични корпус састојао се од текстуалног и говорног корпуса. Текстуални део чинили су један паралелни и два упоредна корпуса који су заједно садржали у просеку око 100.000 речи по језику. Паралелни корпус чинили су

паралелизовани текстови романа Џорџа Орвела (George Orwell) *1984* на седам језика, енглески и шест језика партнера на пројекту (Dimitrova et al. 1998, 315). Овај роман је одабран због доступности оригиналне верзије и других верзија у дигиталном облику у Оксфордској текстуалној архиви (Oxford Text Archive)[31] захваљујући Европској корпусној иницијативи (European Corpus Initiative). Корпус се састојао од одвојених датотека поравнатих текстова оригиналне верзије романа на енглеском са верзијама на осталим језицима што је у последњој верзији Multext-East-а промењено. Текстови су поравнати до нивоа реченице и анотирани лингвистичким информацијама.

Упоредни вишејезични корпус састојао се од два подскупа, „књижевност" и „вести", за сваки од шест језика (Erjavec and Ide 1998, 972), а поређење је урађено у погледу броја и величине текстова (Dimitrova et al. 1998, 317). Део „књижевности" садржао је упоредне текстове једног романа или упоредне делове из више романа, док је део „вести" садржао упоредне чланке из дневних новина. Говорни корпус чинио је мали део Multext-East-а, а чинило га је четрдесет кратких одломака од по пет тематски повезаних реченица који су преведени са енглеског.

Од морфолошких и лексичких ресурса развијени су морфосинтаксички речници који за све језике, осим за естонски и мађарски, садрже пуну флективну парадигму за подскуп лема које се појављују у корпусу. Сваки лексички улаз састоји се из три дела: форма речи – флективна форма речи онако како се појављује у тексту, лема – основна форма речи и морфосинтаксички опис (Morphosyntactic Description - MSD) (Erjavec 2004, 2546). Ови речници искоришћени су за анотацију развијених вишејезичних корпуса.

За кодирање корпусних текстова усвојен је Стандард за кодирање корпуса (Corpus Encoding Standard – CES), апликација језика SGML (Erjavec 2004, 2544), који је заснован на TEI препорукама за електронско кодирање и размену текстова. За MSD, коришћен у речницима и корпусу, развијене су морфосинтаксичке спецификације које дефинишу валидну граматику, синтаксу и семантику морфосинтаксичког описа (MSD), а такође, утврђују шта је валидан MSD за сваки од језика пројектних партнера и шта значи (Erjavec et al. 2003, 27). Multext-East морфосинтаксичке спецификације развијене су према већ

[31] Oxford Text Archive, http://ota.ox.ac.uk/

постојећим спецификацијама за шест западноевропских језика EUMULTEXT формулисаним у оквиру пројекта и у складу са препорукама Експертске саветодавне групе о стандардима језичког инжењерства (Expert Advisory Group on Language Engineering Standards – EAGLES)[32]. Спецификације се састоје из три дела: уводна материја, заједничка спецификација и спецификације за сваки од језика пројектних партнера појединачно.

Након завршетка пројекта Multext-East велики број других пројеката помогао је да се постојећи језички ресурси ажурирају и да се додају нови. Последња, четврта, верзија V4 реализована је маја 2010. године и она садржи ресурсе за седамнаест језика поред енглеског: бугарски, дијалекат, естонски, енглески, литвански, мађарски, македонски, персијски, пољски, резијански, румунски, руски, словачки, словеначки, српски, украјински, хрватски и чешки. У последњој верзији ресурси су прекодирани у формат XML чиме је олакшана валидација и обрада података, а цео процес кодирања ажуриран је у складу са тренутно важећим издањем TEI P5 препорука. Када је реч о паралелном корпусу *1984* урађена је паралелизација између свих језика аутоматским индуковањем из појединачних корпуса са паралелизованим енглеским језиком (Erjavec 2010, 1537). Морфосинтаксичка спецификација и анотирани Орвел за српски језик додати су Multext-East ресурсима у његовој трећој верзији која је изашла 2004. године (Krstev et al. 2004, 432) о чему ће бити речи у одељку Орвелова 1984 за српски језик.

2.5 Развој корпусне лингвистике у Србији

2.5.1 Пределектронски корпуси и услови за стварање електронских корпуса

Интензиван развој корпусне лингвистике од осамдесетих година 20. века и препознавање њеног значаја у области рачунарске лингвистике утицали су на то да се и у Србији започне са изградњом и развојем корпуса. Међутим, и пре дефинисања термина *корпусна лингвистика* у Србији су развијани језички ресурси за рачунарску обраду српског језика. Тако је у периоду од 1957. до 1962. године у Институту за експерименталну фонетику и патологију говора у Београду развијен пределектронски дијахрони корпус,

[32]Expert Advisory Group on Language Engineering Standards, http://www.ilc.cnr.it/EAGLES/home.html

Корпус српског језика, који обухвата текстове из периода од 12. до 20. века. Пројектом је руководио Ђорђе Костић, а на пројекту је радило око 400 сарадника, међу којима је било око 80 лингвиста. Корпус је састављен од 11 милиона ручно анотираних речи што га чини једним од првих језичких корпуса у свету. Свакој корпусној речи придружена је лема и информације о морфолошким категоријама (род, број, падеж, лице, глаголско време и друго) (Kostić 2014, 35). Рад на пројекту је обустављен 1962. године, а ресурси су остали недигитализовани све до средине деведесетих година 20. века.

Кроз сарадњу Института за експерименталну фонетику и патологију говора и Лабораторије за експерименталну психологију при Филозофском факултету Универзитета у Београду у периоду од 1996. до 2003. године обновљен је рад на *Корпусу српског језика*. Пројекат је обновио академик Александар Костић, професор Филозофског факултета Универзитета у Београду, син Ђорђа Костића. Постојећи ресурси, који обухватају текстове до 1957. године, су дигитализовани како би могли да се укључе у *Корпус српског језика,* а корпус је затим допуњен и текстовима савременог српског језика. Постоји званичан сајт корпуса, али сам корпус није јавно доступан преко веба. Александар Костић је објавио папирна издања листи учестаности неколико средњевековних текстова уз које су приложене и електронске верзије тих текстова (са могућношћу претраге текстова по морфолошким категоријама) и листе учестаности (уз могућност претраге по азбучном реду, фреквенцији и друго) (Утвић 2013, 54).

Материјал корпуса обухвата српски језик у распону од осам векова и подељен је на пет временских узорака (Костић 2003, 262):

1. Први временски узорак садржи текстове настале у периоду од 12. од 18. века и састоји се од 500.000 речи. Обухвата дела светог Саве, Доментијана, Теодосија, архиепископа Данила II, Григорија Цамблака, патријарха Пајсија, Стефана Првовенчаног и *Старе српске повеље и писма* у издању Љубомира Стојановића.
2. Други временски узорак обухвата српску књижевност 18. века и састоји се од милион речи. Обухвата дела Доситеја Обрадовића, Милована Видаковића, Јоакима Вујића, Герасима Зелића и Јована Стерије Поповића.

3. Трећи временски узорак обухвата сабрана дела Вука Стефановића Караџића и састоји се од 1.600.000 речи. Обухвата српске народне песме, српске народне приповетке, превод *Новог завета*, Вуков *Рјечник*, историјско-етнографске списе, језичко-полемичке списе и Вукову преписку.
4. Четврти временски узорак обухвата текстове из друге половине 19. века, после усвајања Вукове реформе. Обухваћена су дела Бранка Радичевића, Петра Петровића Његоша, Јована Јовановића Змаја, Ђуре Јакшића, Марка Миљанова и једно дело Лазе Костића.
5. Пети временски узорак обухвата текстове на савременом српском језику и састоји се од 7 милиона речи. Подељен је на пет подузорка: поетски текстови (215 књига), литерарна проза (126 књига), текстови дневне штампе, научни (136 књига) и политички текстови.

Крајем седамдесетих година 20. века у Математичком институту Српске академије наука и уметности[33] започиње са радом стални Семинар за математичку и рачунарску лингвистику са којим је започео и развој рачунарске лингвистичке школе у Србији чији је оснивач и руководилац др Душко Витас, професор Математичког факултета Универзитета у Београду. У оквиру ове школе настаје и Група за језичке технологије која окупља истраживаче са Универзитета у Београду[34] која је своју делатност усмерила на развој језичких ресурса и алата за аутоматску обраду српског језика, корпуса и електронских речника. Већ на самом почетку рада Група је успоставила контакте са некима од врхунских стручњака из ове области као што су Волфганг Тојберт (Wolfgang Teubert) из Института за немачки језик у Манхајму (Institut für Deutsche Sprache)[35] и Морис Грос (Maurice Gross), руководилац Лабораторије за аутоматску документацију и лингвистику (Laboratoire d'Automatique Documentaire et Linguistique - LADL)[36].

После 2001. године Група за језичке технологије је почела интензивно да учествује у домаћим и међународним пројектима. У својим приступима обједињује лингвистичка

[33] Математички институт Српске академије наука и уметности, http://www.mi.sanu.ac.rs/
[34] Група за језичке технологије, Математички факултет, Катедра за рачунарство и информатику, http://www.racunarstvo.matf.bg.ac.rs/?content=nauka_jezicke_tehnologije
[35] Institut für Deutsche Sprache, http://www1.ids-mannheim.de/
[36] Laboratoire d'Automatique Documentaire et Linguistique, http://ladl.univ-mlv.fr/

знања и статистичке анализе како би што боље и веродостојније одговорила на специфичности језика и домена примене. Међу расположивим ресурсима посебно треба издвојити електронски морфолошки речник српског језика (преко 200.000 улаза, преко 7 милиона одредница односно близу 3 милиона различитих облика речи) (детаљније у поглављу 3 одељак 3.3.3), Корпус савременог српског језика (СрпКор) (око 122 милиона речи) (детаљније одељак 2.5.2) и српску семантичку мрежу Ворднет (више од 22.000 синсетова) (детаљније одељак 3.3.4). Поред тога, Група ради и на развоју паралелних корпуса о којима ће бити више речи у наредном одељку, затим на креирању прве банке синтаксних стабала, препознавању именованих ентитета и екстракцији терминологије из различитих домена, анализи ставова и мњења, тематској класификацији, претраживању информација и машинском превођењу. Новембра 2013. године Група је прославила 35 година постојања пригодном конференцијом (Pavlović-Lažetić et al. 2014). Током овог периода организовани су многобројни семинари, одржане су многе важне конференције и реализовани пројекти од међународног и националног значаја. Група активно сарађује са истраживачима с многих факултета Универзитета у Београду (Група за језичке технологије 2016).

Поред Групе за језичке технологије важно је напоменути да на Филолошком факултету Универзитета у Београду постоје неки курсеви из области рачунарске лингвистике. На Катедри за општу лингвистику у оквиру курсева на основним студијама постоји стручно-апликативни предмет Корпусна лингвистика у оквиру којег се студенти оспособљавају за практичан рад са постојећим корпусима и упознају са најважнијим питањима која се односе на конструкцију корпуса. На Катедри за библиотекарство и информатику у оквиру курсева на основним студијама постоји наставни предмет, као изборни, Језичке технологије[37] који студентима омогућава да се упознају са врстама дигиталних језичких ресурса и текућим међународним стандардима за њихову реализацију и да се обуче за њихово активно коришћење и израду. Фокус курса јесте коришћење постојећих дигиталних ресурса за српски језик и изградња нових. Поред основних студија, на мастер академским студијама на истој Катедри постоји изборни

[37] Програм предмета доступан је на: http://poincare.matf.bg.ac.rs/~cvetana/Nastava/1819/JT12-Prog-1819.pdf

предмет Напредне језичке технологије[38] који студентима омогућава да се упознају са основним методама у обради природних језика закључно са синтаксичким парсирањем, са посебним осврtом на применљивост и употребљивост ових метода на српски језик. На докторским академским студијама у оквиру модула Језик постоје предмети Лексичко препознавање у обради природних језика[39], у оквиру којег се студенти упознају са алатима за обраду природних језика као што су електронски морфолошки речници и графови, и Математичка лингвистика.

На Математичком факултету Универзитета у Београду на мастер академским студијама у оквиру студијског програма Информатика постоји предмет Алгоритми текста[40] који омогућава студентима да се упуте у област која представља основу за разумевање значајних савремених истраживачких подручја, као што су обрада природних језика или биоинформатика. Током курса студенти се упознају са теоријом алгоритама текста, аспектима њихове имплементације и примене. На истом факултету на мастер академским студијама у оквиру студијског програма Математика постоји предмет Теорија језика и аутомата[41] који омогућава студентима да стекну основна знања из теорије формалних језика и аутомата. На докторским академским студијама у оквиру студијског програма Информатика постоји предмет Алгоритми текста – напредни концепти[42] у оквиру којег студенти могу да продубе знања потребна за разумевање значајних савремених истраживачких подручја. У оквиру истог студијског програма постоји и предмет Обрада природних језика[43] у оквиру којег се студенти упознају са основним методама у обради природних језика, заснованих на моделирању знања о језику и на статистичким моделима природних језика. Курс уводи студенте у основне проблеме анализе природних језика и методе њиховог решавања са илустрацијом српског језика. На мастер академским студијама при Универзитету у Београду постоји предмет Рачунарство у

[38] Програм предмета доступан је на: http://poincare.matf.bg.ac.rs/~cvetana/Nastava/1516/NJT-Prog-1516.pdf
[39] Програм предмета доступан је на: http://poincare.matf.bg.ac.rs/~cvetana/Nastava/1516/nastava1516-new.html
[40] Програм предмета доступан је на: http://www.math.rs/files/R315_-_Algoritmi_teksta.pdf
[41] Програм предмета доступан је на: http://www.math.rs/files/R311_-_Teorija_jezika_i_automata.pdf
[42] Програм предмета доступан је на: http://www.math.rs/files/R415_-_Algoritmi_teksta_-_napredni_koncepti.pdf
[43] Програм предмета доступан је на: http://www.math.rs/files/R465_-_Obrada_prirodnih_jezika.pdf

друштвеним наукама[44], док на докторским академским студијама при Универзитету постоји предмет Интелигентни системи[45].

Године 2014. у Београду је на иницијативу чланова Групе за језичке технологије основано Друштво за језичке технологије - JePTex[46], а чланови Друштва су поред старих чланова Групе за језичке технологије и нови млади истраживачи из области рачунарске и корпусне лингвистике. Друштво је основано са циљем реализације различитих програма, пројеката, скупова, семинара, конференција и слично у домену примене језичких технологија. Сваког месеца Друштво организује Семинар и предавања страних и домаћих предавача на тему примене језичких технологија у различитим доменима. У марту 2019. године Друштво је организовало скуп „Serbian Unitex Day"[47] на коме су домаћи предавачи представили примену система Unitex (детаљније представљено у поглављу 3, одељак 3.3.2) у Србији.

2.5.2 Корпус савременог српског језика - СрпКор

Идеја о формирању корпуса савременог српског, односно српскохрватског, језика потиче из 1978. године када је одржана прва југословенска конференција о рачунарској обради лингвистичких података. Током периода који обухвата више од три деценије рада креатори Корпуса савременог српског језика (СрпКор) су учешћем у различитим пројектима у периоду од 1981. до 2013. године успели да преусмере средства мањег или већег обима на изградњу СрпКор-а која су коришћена у различитим фазама рада: набавка рачунарске опреме, ангажовање додатних људских ресурса и тако даље (Утвић 2014, 243). На Математичком факултету Универзитета у Београду је као први корак конструисан систем АУРОРА (Vitas 1982) који је генерисао конкорданце и различите врсте индекса за задати текст, а чије су перформансе биле упоредиве са водећим системом тога доба, системом COCOA (Computations in Commutative Algebra)[48] (Korpus savremenog srpskog jezika 2016).

[44] Програм предмета доступан је на: https://www.bg.ac.rs/sr/studije/studije-uni/racunarstvo.php
[45] Програм предмета доступан је на: https://bg.ac.rs/sr/studije/studije-uni/inteligentni-sistemi.php
[46]Друштво за језичке технологије, http://jerteh.rs/index.php/
[47] „Serbian Unitex Day", http://jerteh.rs/?page_id=1793
[48] Computations in Commutative Algebra, http://cocoa.dima.unige.it/

За развој корпуса савременог српског језика било је потребно време да се разреше нека важна питања специфична за сам српски језик: подједнака употреба два писма (ћирилице и латинице), употреба различитих кодних шема за електронске текстове (ISO 646 IRV, ISO 8859-2 и 8859-5, WindowsCP 1250 и 1251, Unicode и друге), употреба два изговора (екавски и ијекавски), проблеми дефинисања обима српског језика у широј заједници који је некада био познат као српскохрватски (Krstev and Vitas 2005, 19).

У оквиру пројекта *Математичка и рачунарска лингвистика*, који је реализован од 1981. до 1985. године, постигнути су и први резултати у погледу изградње корпуса:

1. формирана је прва колекција тестова у дигиталном облику која се састојала првенствено од литерарних текстова, уџбеника и стручне литературе;
2. направљени су први експерименти у морфолошком генерисању српскохрватског језика;
3. урађена су прва истраживања на подручју корпусне лингвистике (анализе језика уџбеника, језика закона, итд);
4. успостављени су контакти са водећим европским истраживачима са подручја корпусне лингвистике, посебно са Волфгангом Тојбертом (Wolfgang Teubert) из Института за немачки језик у Манхајму и групом професора Петера Сгала са Карловог универзитета у Прагу, као и истраживачима из Загреба (СРЦЕ, Филозофски факултет) и Љубљане (Институт "Јожеф Стефан").

Занимљиво је да су у овом периоду, користећи систем АУРОРА, састављени и обрађени и први паралелни корпуси (српско-словеначки подјезик упутстава за лекове (Krstev i dr. 1988), српско-хрватско-словеначки на узорку савезних закона (Krstev i Vitas 1994)). Такође, већ 1989. је био припремљен српски превод стандарда о SGML-у. Укључивањем у европске пројекте Група је проширила круг европских лабораторија за лингвистичка истраживања са којима је сарађивала те је у том периоду и успостављен контакт са лабораторијом LADL професора Мориса Гроса (Maurice Gross). СрпКор је коначно постављен на веб кроз научно-истраживачки пројекат *Интеракција текста и речника* који је реализован од 2002. до 2005. године.[49]

[49] Корпус савременог српског језика доступан је на: http://www.korpus.matf.bg.ac.rs

Прва верзија СрпКор под називом *Неетикетирани корпус савременог српског језика* (НЕТК) са 22,2 милиона речи јавности је постала доступна 2003. године. У (Krstev and Vitas 2005) су детаљно описани параметри и структура почетне верзије СрпКор-а. Друга верзија СрпКор-а са 113 милиона корпусних речи је постављена је 2011. године, а према подацима из 2013. године текућа верзија СрпКор-а садржи 122 милиона корпусних речи, а припремљено је укупно 5.058 текстова насталих од 1910. до 2005. године од чега је 4.890 укључено у корпус.[50] Текстови су подељени по функционалним стиловима, по статусу текста у односу на језик на коме је текст оригинално настао и по временском периоду у коме је текст настао односно у коме је објављен (Утвић 2014, 246). СрпКор је анотиран библиографским информацијама о корпусним текстовима и морфосинтаксичким информацијама (врста речи и лема). Када је реч о структурној анотацији корпуса сам СрпКор није структурно анотиран у смислу постојања XML-етикета иако су поједини корпусни текстови структурно анотирани односно садрже информацију о логичкој структури текста (Utvić 2013, 259). Међутим, постоје информације о крајевима сегмената (реченица), пошто поједини знаци интерпункције имају ознаку за крај реченице (SENT) као вредност „леме".

Веб-сумећа СрпКор-а нуди једноставну и напредну претрагу. Једноставна претрага омогућава основне опције претраживања уношењем једноставног регуларног израза. Напредна претрага може се подесити према следећим параметрима: разликују се велика и мала слова, подразумевани атрибут корпусне позиције (корпусна реч, лема), аутор(и), функционални стилови (административни, литература, научни, новински, остало, сви стилови) и домаћи аутор или превод (оригинал, превод, све). Пре покретања претраге може се подесити и како ће изгледати резултати претраге и то према следећим параметрима: сортирање (по резултату и десном контексту, без интерпункције; по резултату и десном контексту, са интерпункцијом), леви и десни контекст резултата, приказ свих резултата (по 100 на страни) и приказ случајно одабраних резултата (сваки n-ти). Напредна претрага користи могућности упитног језика CQP (CQP Query Language) (IMS CWB/CQP 2016) (детаљније објашњено у поглављу 3 одељак 3.3.1) заснованог на

[50] Приказани подаци о Корпусу савременог српског језика датирају из 2013. године и објављени су у (Utvić 2013) и (Утвић 2014)

регуларним изразима и поред карактеристика једноставне претраге омогућава и претрагу додатних позиционих атрибута, прецизније речено претрагу по морфосинтаксичкој анотацији. Као резултат претраге СрпКор-а добијају се конкорданце које корисник може да прегледа страну по страну и приступи свакој страни у резултатима претраге, а може и да изабере које конкорданце жели да задржи у приказу, као и да сачува издвојене конкорданце. Конкорданце се аутоматски генеришу уз помоћ програмског алата конкорданцер. Сам назив у ширем смислу у корпусној лингвистици означава систем за креирање и анализу корпуса, док у ужем смислу представља само једну од компоненти тог система.

2.5.3 Развој паралелних корпуса у Србији

Поред једнојезичног Корпуса савременог српског језика Група за језичке технологије деценијама ради на развоју једнојезичних и вишејезичних паралелних корпуса српског језика чији значај за потребе истраживача у области језика дуго није био препознат (Krstev and Vitas 2005, 15). Без обзира на потешкоће на које је наилазила, Група је постигла значајне резултате у изградњи и коришћењу језичких корпуса те је до сада развијено неколико различитих врста једнојезичних, двојезичних и вишејезичних паралелних корпуса. Корпуси се не могу преузимати у целости, али се сваком од њих може приступити након дозвољене ауторизације односно добијањем корисничког имена и шифре за приступ.

Изградња паралелних корпуса на Математичком факултету Универзитета у Београду започета је учешћем у пројекту TELRI и производњом CD-а „East meets West – A Compendium of Multilingual Resources" што је детаљније описано у претходном одељку овог поглавља. Након учешћа у овим пројектима и користећи стечена искуства Група је наставила да ради на развоју паралелних корпуса који обухватају српски језик тако да данас постоје два већа корпуса, француско-српски (СрпФранКор) и енглеско-српски (СрпЕнгКор), дигитална библиотека Библиша која садржи више паралелних двојезичних колекција, вишејезична колекција Вишејезични Верн, као и српско-српски/хрватски паралелни корпус. Ови паралелни корпуси примарно су развијени за потребе

лингвистичких и лексикографских истраживања и у њихов садржај је укључен и знатан број књижевних текстова.

Српско-француски корпус - СрпФранКор

Након учешћа у изради вишејезичних паралелних корпуса *Република* и *1984* Математички факултет Универзитета у Београду наставио је са израдом паралелних корпуса који обухватају српски језик. Први двојезични корпус са којим је започет рад је српско-француски паралелни корпус, СрпФранКор. Корпус у коме је француски постао први, а српски други језик, састављен је углавном од књижевних и новинских текстова, али и неких савремених текстова из области филозофије, социологије, етнологије и науке.

Највећи део СрпФранКор-а састоји се од класичних дела француске књижевности насталих од краја 18. века па до данас, док су сви преводи на српски рађени после 1926. године. Корпус садржи текстове који су бирани према својој књижевној вредности, али и према доступности у електронском формату. Велики део француских текстова већ је постојао у електронском формату и тако су и преузети, док су остали сканирани и поправљани. Када је реч о српским текстовима на некима је примењена техника оптичког препознавања карактера, неки су прекуцани, док су неки преузети од преводилаца (Vitas and Krstev 2006, 149).

Овај део корпуса иницијално је обухватио следеће романе: Волтеров (Voltair) *Кандид* (*Candide*), *Пут око света за осамдесет дана* (*Le tour du monde en quatre-vingt jours*) Жила Верна (Jules Verne), *Бувар и Пекише* (*Bouvard et Pécuchet*) Густава Флобера (Gustave Flaubert) и *Жена и њена играчка* (*La Femme et son pantin*) Пјера Лиуса (Pierre Louÿs), као и ажуриране верзије дела *Републике* и *1984* (Vitas and Krstev 2004, 249). Временом су додата дела Онореа де Балзака (*Honoré de Balzac*), Алберта Камија (*Albert Camus*) и Амина Малуфа (*Amin Maalouf*). Поред дела француских писаца корпус је допуњен и делима српских писаца који су преведени на француски: *Све звери што су са тобом* (*L'arche de Boba*) Бобе Благојевић (прво дело српског писца које је укључено у корпус), затим су ту и дела Иве Андрића, Данила Киша, Растка Петровића, Боре Станковића и многих других.

Један део корпуса чине и новински текстови преузети из француског часописа „Le Monde Diplomatique”[51] који излази једном месечно и његове преводе на српскохрватски језик. Прикупљање текстова започето је у марту 2001. године када је започето и објављивање српског превода. Француски текстови редовно су прикупљани са веб стране самог часописа, док су преводи чланака директно преузимани од издавача или самог преводиоца (Vitas and Krstev 2004, 249).

Поступак припреме СрпФранКор-а прошао је кроз три фазе. Прва фаза обухватила је припрему текстова за процес паралелизације. Како процес захтева претходно добро форматиран XML текст тако су и текстови за овај корпус анотирани одговарајућим XML ознакама, а у складу са TEI препорукама. Процес припреме XML докумената урађен је полуаутоматски. Наслови су анотирани ручно етикетом <head>, док су пасуси анотирани аутоматски, односно, полу-аутоматски етикетом <p>. Процес сегментације на реченице (етикета <seg>) урађен је аутоматски уз помоћ коначног трансдуктора *Sentence.grf* који је имплементиран у систем Unitex (Krstev 2008, 174). За процес паралелизације коришћен је прво алат Vanilla (Danielsson and Ridings 1997) који упарује текстове који су сегментирани на најмање два нивоа, касније алат Xalign (више у 3.5). Након аутоматски упарених сегмената урађене су ручне корекције погрешно упарених варијанти јединица превођења, а добијени битекстови су произведени у формату TMX уз помоћ система WS4LR (Work Station for Language Resources), софтверским алатом који омогућава интеграцију различитих лексичких ресурса (Krstev et al. 2006b, 1693-1694), као и у формату HTML који омогућава визуелизацију. У трећој фази сви текстови су прикупљени и обрађени коришћењем система IMS/CQP (детаљније у поглављу 3 одељак 3.3.1) који омогућава анализу и претрагу целог корпуса.

Могућности претраге СрпФранКор-а илустроваћемо следећим упитима[52]: [Ll]xubav[a-z]* (Слика 1) и amour (Слика 2). Уз све добијене конкорданце као резултат претраге на српском стоји паралелна конкорданца на француском језику и обрнуто.

[51] Le Monde Diplomatique http://www.monde-diplomatique.fr/

[52] Текуће верзије корпуса као и претрага користе такозвани АУРОРА начин кодирања: š/ш – sx, ž/ж – zx, č/ч – cy, ć/ћ – cx, đ/ђ – dx, dž/џ – dy, lj/љ – lx, nj/њ – nx.

17017: Jer , ne treba da se varate , gospodine , vi koji čitate romene , a možda uzimate i učešća u igrama sa banjskim poludevojkama , naše Andaluskinje nemaju ni volje ni smisla za poročnu <ljubav> .

FR: Car ne vous y trompez pas , jeune Français , lecteur de romans et acteur peut - être d ' intrigues particulières avec les demi - virginités des villes d ' eaux , nos Andalouses n ' ont ni le goût , ni l ' intuition de l ' amour artificiel

169759: međutim oni zaslužuju našu <ljubav> .

FR: Cependant , ils méritent notre amour

85956: učinih to njima za <ljubav> .

FR: je le fis pour leur plaire

26755: Eto koliko mašta može da spreči ženu da ne pozna <ljubav> !

FR: Jusqu ' où l ' imagination des femmes peut - elle les aveugler sur l ' amour viril

157360: Kićanka crvene kape klanjala se zaljubljeno , i njegov uzdrhtali glas , njegovo dobro lice preklinjali su svirepog da se sažali na njenu <ljubav> .

FR: La mèche du bonnet rouge s ' inclinait amoureusement ; et sa voix tremblante , et sa figure bonne conxuraient le cruel de prendre en pitié sa flamme

54090: ali , je li <ljubav> bolest od koje se može izlečiti ?

FR: mais l ' amour est - il un mal dont on puisse guérir

5918: Ja sam isuviše poštovao <ljubav> da bih išao kud bilo , i gotovo nikad nisam zagrlio ženu koju nisam strasno voleo .

FR: Je respectais trop l ' amour pour fréquenter les arrière - boutiques , et je n ' ai presque jamais possédé une femme que je n ' eusse aimée passionnément

33510: - Šta znači za jednog oca <ljubav> dece koja mu smetaju ?

FR: - Que fait à un père l ' amour d ' enfants qui le gênent

22392: Ja sam dobra hrišćanka , ali Bog štiti iskrenu <ljubav> , i ja ću otići u raj pre mnogih udatih žena .

Слика 1. Резултат претраге регуларним изразом „[Ll]xubav[a-z]” у СрпФранКор-у*

224118: - - " Ah ! ah ! fructus belli ! - - ce sont des syphilides , mon bonhomme ! soignez - vous ! diable ! ne badinons pas avec l ' <amour> . "

SR: - - Aha , fructus belli ! to su sifilide , dobri moj čoveče , ponegujte se ! Đavola , ne treba se šaliti s ljubavlju

211462: " Mais si tu es tué , mon <amour> ?

SR: - - Ali ako te ubiju , ljubavi moja

19060: Si je le lui avais demandé , elle ne l ' eût sans doute pas permis , car je commençais à douter que cette nuit d ' entretiens s ' achevât jamais en nuit d ' <amour> :

SR: Da sam tražio , zacelo mi ne bi dopustila , jer sam već počeo pomišljati da se ova noć razgovora neće završiti kao noć ljubavi

181460: Les deux hommes se récrièrent ; et un dialogue s ' en suivit sur les femmes , sur l ' <amour> .

SR: Dva čoveka se usprotiviše , i usledi dijalog o ženama , o ljubavi

209582: Cependant , ils méritent notre <amour> .

SR: međutim oni zaslužuju našu ljubav

383568: - Milord , dit - elle , soyez indulgent pour mon humble irréalité et , avant d ' en dédaigner le rêve , rappelez - vous la compagne humaine qui vous oblige à recourir , fût - ce à un fantôme , pour vous racheter l ' <Amour> .

SR: - - Milorde - - reče ona - - budite blagonakloni mojoj skromnoj nerealnosti i pre no što prezete njen san , setite se čovečjeg društva koje vas je nateralo da čak i priviđenje potražite ne biste li iskupili ljubav

194743: Oh ! laisse - moi dormir et rêver sur ton sein , Doña Sol ! ma beauté ! mon <amour> !

SR: - - Oh , pusti me da usnem i sanjam na grudima tvojim Dona Sol , lepoto moja , ljubavi moja

Слика 2 Резултат претраге регуларним изразом „amour” у СрпФранКор-у

Прве анализе корпусног садржаја показале су да се у добијеном корпусу могу пронаћи решења за многе преводе који не постоје у двојезичним француско-српским речницима. Такође, корпус даје анализу стратегије превођења која даље омогућава решавање лексичког јаза или двосмислености у оригиналном тексту, као и проналажење недоследности у преводу (Vitas and Krstev 2006, 154).

Корпус садржи 31 књижевни текст од чега је 28 текстова оригинално написано на француском језику и преведено на српски (један са два превода), 2 текста оригинално написана на српском језику која су преведена на француски и један енглески роман који је преведен на француски и српски језик. Мерено у корпусним речима тренутна величина корпуса је 1.738.752 корпусне речи од чега је 953.935 у француском делу, док је 784.817 у српском делу корпуса.

Поред СрпФранКор-а значајно је поменути још један паралелни француско-српски корпус односно француско-српско-енглески корпус, ParColLab[53] (Balvet et al. 2014). ParColLab је нови паралелни корпус од скоро 6 милиона речи који садржи изворне текстове и њихове преводе на три европска језика (француски, српски и енглески) од којих је сваки подједнако и изворни и циљни језик (Miletic et al. 2015). Корпус тренутно садржи књижевне текстове (са идејом да се прошири текстовима са интернета, преводима из филмова и серија, техничком документацијом и многим другим) који су упарени на нивоу пасуса и реченица и анотирани одговарајућим XML обележјима у складу са тренутно важећом верзијом TEI Смерница за кодирање текста.[54]

Корпус ParColLab настао је као резултат вишегодишње сарадње лингвиста и информатичара, као и стручњака за рачунарску обраду природних језика из Француске и Србије. Пројекат координира Дејан Стошић, доцент и истраживач на Универзитету Тулуз Жан Жорес (Université Toulouse Jean Jaurès), члан истраживачке екипе CLLE-ERSS из Тулуза. Корпус је намењен истраживачима за проучавање различитих области лингвистике ова три језика, али и у наставне и педагошке сврхе (употреба у оквиру наставе и учења ова три језика као страних, школовању преводилаца, припреми наставног материјала). Корпус је бесплатно доступан преко интернета уз отворен кориснички налог.

[53] ParColLab, http://parcolab.univ-tlse2.fr/en/

[54] На сајту http://parcolab.univ-tlse2.fr/en/about/content/ налази се списак свих књижевних текстова који су тренутно саставни део корпуса.

Српско-енглески корпус - СрпЕнгКор

Након рада на СрпФранКор-у Група за језичке технологије започела је рад на паралелном српско-енглеском корпусу, СрпЕнгКор. Прво питање које је разматрано пре почетка рада на корпусу јесте одабир текстова. Одлука је зависила од два фактора:

1. доступност текстова на оба језика у електронском формату
2. разрешење питања ауторских права.

Први текст у корпусу био је *1984* Џорџа Орвела преузет из истоименог вишејезичног корпуса. Након разрешења постављених питања рад на корпусу је настављен паралелизацијом дела Џејн Остин (Jane Austen). Дела ове списатељице су одабрана, са једне стране, због доступности превода на доста светских језика који би у будућности могли да се додају, а са друге стране, питање разрешења ауторских права се није постављало. За почетак рада на корпусу преузети су оригинални текстови шест романа са сајта *The Republic of Pemberley*[55], док су српске текстове прекуцали студенти Катедре за библиотекарство и информатику као део практичног рада у оквиру предмета Информатичка писменост. Све грешке у прекуцаним текстовима пронађене су и поправљене применом постојећих електронских морфолошких речника српског језика (Krstev and Vitas 2011, 498).

Како би се наставило са аутоматском паралелизацијом, сви текстови су анотирани основним XML етикетама до нивоа реченице (поглавља - , наслов - <head>, пасус - <p> и реченица - <seg>). Процес сегментације урађен је аутоматски и за српске и за енглеске текстове коришћењем система Unitex и одговарајућег графа *Sentence.grf* за сегментацију на реченице за оба језика. Поглавља, наслови и пасуси су углавном ручно анотирани. За процес паралелизације коришћен је програмски пакет ACIDE са интегрисаним алатима XAlign и Concordancier. Овај програмски пакет је детаљно објашњен у поглављу 3 одељак 3.5.

Приликом креирања СрпЕнгКор-а циљ процеса паралелизације био је да се у што већој мери постигне 1-1 упаривање сегмената. Уз помоћ претходно наведених алата идентификовани су погрешно упарени сегменти, сегменти који недостају или

[55] The Republic of Pemberley, http://pemberley.com/

недоследност између упарених сегмената изворног текста и његовог превода. Контрола и корекција упарених сегмената урађена је ручно при чему је готово у потпуности постигнуто 1-1 упаривање. Све даље фазе у процесу припреме паралелног корпуса објашњене у одељку 2.3, примењене су и у изради СрпЕнгКор-а.

У току даљег рада овај корпус допуњен је и другим делима класичне енглеске књижевности као што су дела Томаса Хардија (Thomas Hardy) и Ернеста Хемингвеја (Ernest Hemingway) али и делима савремених писаца као што су Ден Браун (Dan Brown), Џ. К. Роулинг (J. K. Rowling). Поред дела на енглеском језику која су преведена код нас корпус обухвата и дела наших писаца који су преведени на енглески језик као што су дела Данила Киша, Драгана Великића, Светислава Басаре и других.

Поред књижевних дела саставни део корпуса су и новински чланци из корпуса SETimes при чему је формиран поткорпус BALKANTIMES, који садржи вести из Југоисточне Европе на десет језика: бошњачком, хрватском, енглеском, македонском, српском, турском, албанском, бугарском, грчком и румунском. Поткорпус садржи вести из економије, дипломатије, филма, туризма, спорта, науке и текуће вести. Сваки текст који је део поткорпуса, припремио је један студент са Катедре за библиотекарство и информатику или Катедре за општу лингвистику Филолошког факултета Универзитета у Београду, као део семинарског рада у оквиру предмета *Информатика 4* и *Примењена лингвистика* у периоду од школске 2003/2004. до 2009/2010. године.

У оквиру пројекта Intera (Integrated European Language data Repository Area)[56] (Gavrilidou et al. 2004) СрпЕнгКор је допуњен паралелним текстовима из области права, пословања, образовања и здравствене заштите при чему је формиран поткорпус SELFEH (Serbian-English Law Finance Education and Health). Intera је двогодишњи пројекат који је покренула Европска Унија како би се изградио интегрисан вишејезични европски језички ресурс повезивањем међународних, националних и регионалних центара података (Gavrilidou et al. 2004, 97). Такође, корпус је допуњен и неким текстовима из корпуса Acquis communautaire (детаљније у одељку Acquis Communautaire). Српски текстови припремљени у оквиру овог пројекта су аутоматски лематизирани и морфолошки

[56] Intera, http://www.mpi.nl/intera/

анотирани коришћењем лексичких ресурса за српски уз накнадну контролу и корекцију, док су енглески текстови означени коришћењем система TreeTagger-a (Schmid 1994). SELFEH садржи преко 150 паралелних текстова из поменутих области са укупно милион корпусних речи по језику и је коришћен за тестирање различитих тагера за српски језик, али и у експериментима за машинско превођење и екстракцију термина.

Као закључак се може извести да СрпЕнгКор садржи текстове оригинално написане на енглеском језику који су преведени на српски, текстове оригинално написане на српском језику који су преведени на енглески, као и поравнате енглеске и српске преводе текстова који су оригинално написани на француском језику. Мерено у корпусним речним тренутна величина СрпЕнгКор-а је 4.420.711 корпусних речи од чега је 2.330.742 у енглеском, док је 2.089.969 у српском делу корпуса. Од тога поткорпус BALKANTIMES броји 780.614 корпусних речи од чега је 388.933 у енглеском делу, док је у 381.681 у српском делу корпуса. Могућност претраге СрпЕнгКор-а илустроваћемо упитом „[Zz]dravlx[a-z]* radnika" (Слика 3) и „occupational health" (Слика 4).

29591: Sa druge strane , neke jedinice medicine rada smatraju da je teško da prošire svoje aktivnosti na životnu sredinu , pa se radije ograničavaju na probleme <zdravlja radnika> .

EN: On the other hand , some occupational health units find it difficult to expand their activities into environmental health and prefer to limit themselves to issues of workers ' health

37263: To stvara osnovu za jačanje političke volje da se stvore preduslovi za poboljšanje uslova na radu , a time i <zdravlja radnika> .

EN: This creates a basis for strengthening political will to create the prerequisites for improvement of working conditions , and thereby workers ' health

21084: Prema tome , nova svetska strategija za zdravlje na radu je veoma značajna za SZO i druge organizacije koje se bave problemima <zdravlja radnika> .

EN: Thus a new global strategy for health at work is very relevant to WHO and other organizations interested in dealing with workers ' health issues

19979: Predložena Opšta strategija medicine rada za sve predstavlja kratku analizu situacije na osnovu dostupnih pokazatelja <zdravlja radnika> , identifikuje očigledne potrebe za razvojem medicine rada i zaštite na radu , uključujući i prioritetne oblasti i na nacionalnom i na međunarodnom planu , i predlaže prioritetne akcije u okviru Programa za zdravlje radnika SZO .

EN: This proposed Global Strategy on Occupational Health for All presents a short situation analysis by using available occupational health indicators , identifies the most evident needs for the development of occupational health and safety , including the priority areas at both national and international levels , and proposes the priority actions for WHO ' s Workers ' Health Programme

151422: Savremene službe medicine rada iz svake relevantne profesije , discipline ili nauke - bilo da se radi o biomedicinskim ili ekološkim - uzimaju potrebne elemente koje integrišu u sveobuhvatni multidisciplinarni pristup usmeren na očuvanje i unapređenje <zdravlja radnika> kroz aktivnosti vezane i za radnu sredinu i za same radnike .

EN: Modern occupational health services draw from each relevant profession , discipline or science - be it biomedical or environmental - all the required elements and integrates them into a comprehensive multidisciplinary approach aimed at the protection and promotion of workers ' health through actions

26664: Međutim , mehanizmi prevencije i kontrole profesionalnih rizika na radu još su nerazvijeniji i mnoge potrebe koje se tiču <zdravlja radnika> nisu zadovoljene .

EN: Mechanisms for prevention and control of occupational hazards are , however , less developed and many of the needs of workers ' health are not met

Слика 3. Резултат претраге регуларним изразом „[Zz]dravlx[a-z] radnika" у СрпЕнгКор-у*

Слика 4. Резултат претраге регуларним изразом „occupational health" у СрпЕнгКор-у

Библиша

Библиша[57] је веб апликација коју је развила Група за језичке технологије Универзитета у Београду са циљем унапређења могућности претраживања вишејезичних дигиталних библиотека електронских часописа (Stanković et al. 2012, 1710). Дигитална библиотека Библише тренутно садржи једанаест текстуалних колекција упарених докумената: пет колекција поравнатих чланака из домаћих научних часописа који излазе на српском и енглеском језику (Инфотека[58], Подземни радови[59], Архитектура и урбанизам[60], Стоматолошки гласник Србије[61], Management); две колекције поравнатих техничких извештаја са пројеката (BEAKTEL TEMPUS, CESAR); три паралелна доменска корпуса: Intera, EIEner (паралелни корпус текстова из области енергетике) и Mining (паралелни корпус текстова из области рударства), и један корпус књижевних текстова СрпНемКор (паралелни корпус књижевних текстова на српском и немачком језику) који је

[57] Biblisha, http://jerteh.rs/biblisha/
[58] Инфотека, http://infoteka.bg.ac.rs/index.php/sr
[59] Подземни радови, http://www.rgf.rs/publikacije/PodzemniRadovi/?lang=sr
[60] Архитектура и урбанизам, http://www.iaus.ac.rs/code/navigate.aspx?Id=111
[61] Стоматолошки гласник, http://www.stomglas.org.rs/

и предмет ове докторске дисертације. Сви текстови анотирани су одговарајућим XML етикетама, поравнати до нивоа реченица (постигнуто је 1-1 упаривање), извезени у формату TMX, а потом увезени у базу. Сви поравнати текстови смештени су у базу података NoSQL MongoDB[62] која подржава динамичке упите за претрагу дигиталних колекција користећи језик за постављање упита заснован на документима са основном јединицом складиштења у JSON формату (Stanković et al. 2017).

Дигитални објекти смештени у дигиталну библиотеку Библише описани су одговарајућим метаподацима. Како је Библиша иницијално креирана као дигитална библиотека електронских часописа тако је структура метаподатака првобитно била дизајнирана да опише сам часопис, сваки број појединачно и сваки чланак појединачно на следећи начин (Stanković et al. 2012, 1711):

1. *метаподаци часописа*: наслов часописа, међународни стандардни број за серијске публикације (International Standard Serial Number - ISSN) и URL колекције;
2. *метаподаци за сваки број часописа*: годиште, број, месец и година објављивања;
3. *метаподаци за сваки чланак*:
    a. описни метаподаци: име(на) аутора, афилијација, адреса електронске поште, наслов чланка, његова категоризација и пагинација у оквиру броја
    b. метаподаци о садржају: сажетак, кључне речи и УДК[63] број.

У првој верзији Библише, метаподаци су били у формату XML и смештани заједно са документима у MarkLogic NoSQL базу података. База података MarkLogic[64] представља систем за управљање великим базама неструктурираних и полуструктурираних података

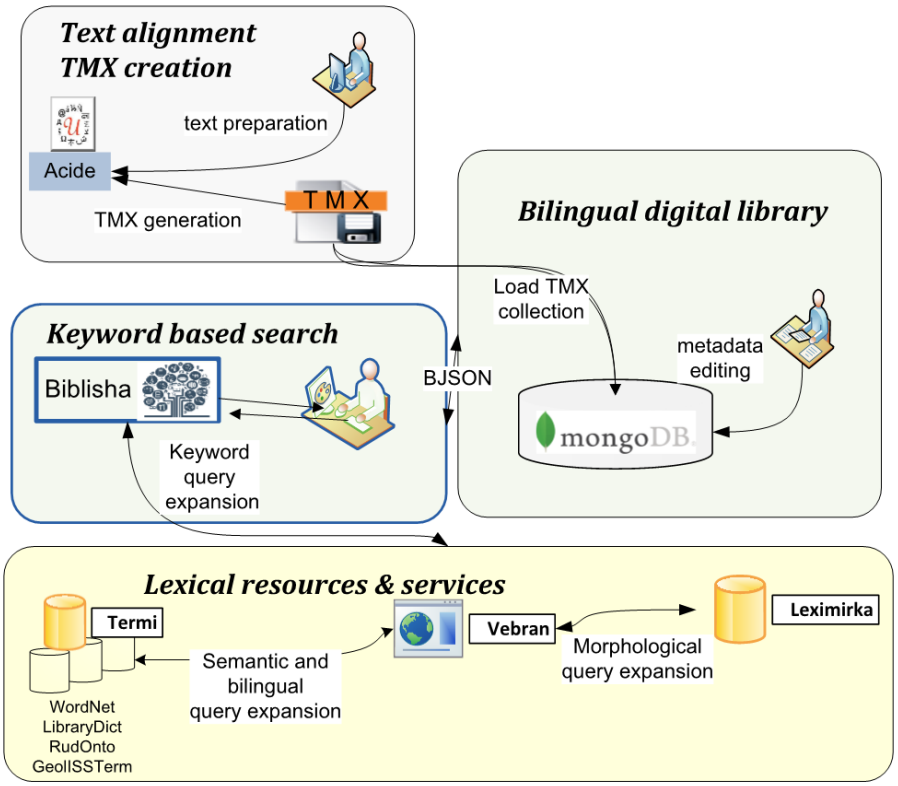
[62] MongoDB, https://www.mongodb.com/

[63] УДК (Универзална децимална класификација) је међународни класификациони систем заснован на децималном систему који омогућава да се уз помоћ цифара одреди садржај и предмет публикације која се описује и њено место у систему људског знања. Користи се за стручну класификацију свих врста грађе у библиотечком систему. Управо због цифарске ознаке садржаја грађе УДК представља добар алат за индексирање и претраживање података.

[64] MarkLogic, https://www.marklogic.com/

(Stanković et al. 2015, 1771). Са даљим развојем Библише успостављена је нова база података MongoDB (Stanković et al. 2017), а метаподаци и паралелни текстови су конвертовани у JSON формат са одговарајућом схемом која садржи сетове елемената за опис свих текстуалних колекција, њихових потколекција и објеката. Постављањем корпуса СрпНемКор у Библишу структура метаподатака прилагођена је опису и књижевних текстова што је детаљније објашњено у поглављу 6 одељак 6.3.

Структура алата Библиша је комплексна и састоји се из неколико компоненти: лексички ресурси (интегрисани за проширење корисничких упита за претрагу), текстуалне колекције (документа на два језика поравната до нивоа реченице), веб сервиси (интегрисани за приступ лексичким ресурсима) и веб сумећа (развијена за потребе корисника) (Слика 5)(Stanković et al. 2015, 1769).



Слика 5. Структура Библише

Приликом постављања упита за претрагу, Библиша позива одговарајуће лексичке ресурсе који омогућавају морфолошко и семантичко проширење постављених упита. Лексички ресурси који се позивају могу се поделити у три веће групе:

1. електронски морфолошки речници српског језика (детаљније објашњени у поглављу 3 одељак 3.3.3) за генерисање свих флективних облика кључних речи датих у упиту за претрагу на српском језику;
2. српски и енглески Ворднет за семантичко проширење постављених упита на српском и енглеском језику. Ворднет је детаљније описан у поглављу 3 одељак 3.3.4;
3. терминолошке базе података из различитих домена и онтологије:

   а. Библиотекарски терминолошки речник (Kovačević et al. 2004) је терминолошки речник који се користи у теорији и пракси библиотечко-информационих наука и сродних области. Иницијално је направљен као српско-енглески/енглеско-српски, али је временом урађена и немачка верзија речника. Електронска верзија речника тренутно садржи 40.000 одредница (приближно 14.000 на српском, 12.400 на енглеском и 14.000 на немачком језику); 900 дефиниција или анотација термина који су део библиотечких стандарда; 2.300 акронима међународних и националних организација и институција; 190 адреса релевантних веб локација.[65]

   б. GeolISSTerm је тезаурус термина из геологије на српском и енглеском језику развијен на Рударско-геолошком факултету Универзитета у Београду на основу GeolISS базе (Geologic Information System of Serbia - Геолошки информациони систем Србије). Развијен је за потребе Министарства за животну средину, рударство и просторно планирање Републике Србије у циљу израде стандардног речника геолошких термина са логички конзистентним описима, тумачењима и класификацијом геолошких јединица, геолошких структура,

[65] Електронска верзија речника доступна је на: http://rbi.nb.rs/srlat/dict.html. У Библишу је уграђена прва верзија речника на српско-енглеском односно енглеско-српском.

минерала и хидролошких ресурса, али и других геолошких карактеристика у домену примењених дисциплина (Станковић и др. 2011).[66]

в. RudOnto је комплексни терминолошки ресурс развијен на Рударско-геолошком факултету Универзитета у Београду са циљем да покрије већи део стручне терминологије из области рударског инжењерства и геологије и да постане будући референтни извор електронског формата за рударску терминологију на српском језику (Stanković et al., 2014). RudOnto тренутно садржи стручне термине упоредо на српском и енглеском и мали број њихових еквивалената на другим језицима.[67]

г. Termi је терминолошка вишејезична база односно терминолошки речник развијен у оквиру пројекта BEAKTEL. Апликација је развијена да подржи развој терминолошких речника из различитих домена (математика, рачунарство, рударство, библиотекарство, рачунарска лингвистика и многи други).[68] О структури базе Termi детаљније у поглављу 6, одељци 6.3.2 и 6.3.3.

Значајна карактеристика Библише је што омогућава вишејезичну претрагу комплетних поравнатих текстова корпуса са сумеђом која је потпуно прилагођена корисницима (user-friendly). Библиша омогућава корисницима две врсте претраге: претрагу преко метаподатака и претрагу пуног текста. Претрага преко метаподатака је једнојезична и омогућава корисницима да комбинују више поља за претрагу при чему дефинишу језик претраге из листе која се налази на почетку претраживача и бирају да ли желе да претражују све расположиве колекције или неку одређену. Након тога корисници користе образац са следећим предефинисаним пољима: речи из наслова, име аутора, кључне речи, речи из сажетка, речи из текста (Слика 6). Као резултат корисници добијају

[66] Електронска верзија тезауруса доступна је на http://geoliss.mre.gov.rs/recnik/

[67] База је свим заинтересованим корисницима доступна јавно и бесплатно без регистрације на: http://rudonto.rgf.bg.ac.rs/

[68] Termi је терминолошка вишејезична база односно терминолошки речник развијен у оквиру пројекта BEAKTEL. Апликација је развијена да подржи развој терминолошких речника из различитих домена (математика, рачунарство, рударство, библиотекарство, рачунарска лингвистика и многи други). Апликација је доступна на адреси http://termi.rgf.bg.ac.rs/. Странице за прелиставање и претрагу су јавно доступне, док је за приступ страницама за ажурирање и уређивање профила потребно имати кориснички налог са посебним привилегијама.

листу докумената који одговарају постављеном упиту и могућност прегледа комплетних метаподатака који описују документ, приступ тексту у TMX формату, приступ комплетном тексту или делу текста у формату PDF где год је то могуће и везу на „врећу речи" (BOW - bag of words) који је векторска репрезентација документа за потребе рангирања документа (Слика 7).

Слика 6. Окружење у Библиши за претрагу преко метаподатака

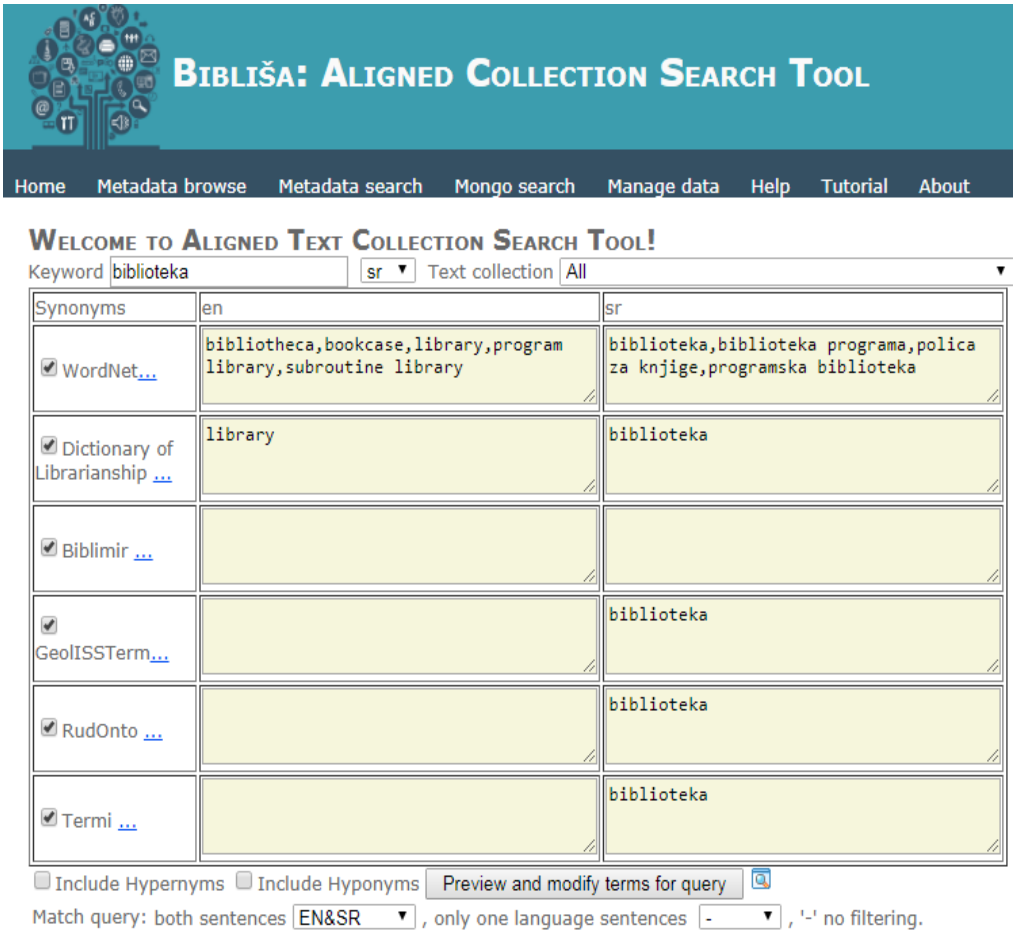
Слика 7 Резултати претраге преко кључне речи „biblioteka"

Поред једнојезичне претраге преко метаподатака Библиша омогућава корисницима двојезичну претрагу комплетног текста колекција уз могућност морфолошког и семантичког проширење упита позивањем различитих лексичких ресурса. Основу за морфолошко проширивање упита чине морфолошки електронски речници, као и системи правила за промену вишечланих речи, коначни аутомати и уграђене функције базе података MongoDB.

Када је реч о семантичком двојезичном проширењу упита, систем се ослања на српски и енглески Ворднет и поменуте терминолошке базе података и онтологије. За проширење упита је у основи задужен веб сервис VebRan заједно са веб апликацијом VebRanka[69] који покреће софтверски алат LeXimir[70] за синхронизовано коришћење разноврсних језичких ресурса (Stanković et al. 2011) (Stanković et al. 2012, 1712) (Stanković et al. 2015, 1773) (Stanković et al. 2017, 173).

Корисници дефинишу упит уношењем кључних речи у поље за претрагу и као и код претраге преко метаподатака дефинишу језик претраге из листе која се налази на почетку претраживача и бирају колекцију коју желе да претражују. Уз ове параметре комбинују се и лексички ресурси интегрисани у Библишу ради семантичког проширења упита. Као резултат добијају се листе кључних речи распоређене према одговарајућим лексичким ресурсима, а корисници могу да их користе онакве какве јесу или да их уређују брисањем или додавањем нових. На пример, ако се као упит за претрагу постави „biblioteka” на српском језику над свим понуђеним колекцијама и означе се сви расположиви лексички ресурси за семантичко проширење добијају се резултати које илуструје Слика 8. Добијени резултати претраге могу се проширити надређеним и подређеним појмовима.

[69] VebRanka је веб апликација за проширивање упита на вебу. Доступно је на: http://hlt.rgf.bg.ac.rs/vebranka/
[70] LeXimir је софтверски алат за изградњу, одржавање и руковање лексичким ресурсима. Доступно је на: http://korpus.matf.bg.ac.rs/soft/LeXimir.html



Слика 8. Окружење у Библиши за претрагу комплетног текста са резултатима за упит „biblioteka”

На основу добијених резултата претраге производе се конкорданце поравнатих сегмената у којима се јавља одговор на упит. Свака произведена конкорданца садржи информацију о изворном документу и везу ка метаподацима који додатно описују овај документ (Слика 9). Конкорданце могу приказивати одговор на упит на два начина. Први начин омогућава да сагледамо све конкорданце у којима се одговор на упит јавља у текстовима на два одабрана језика, на пример “EN&SR” . Ово значи да имамо могућност да видимо одговор на упит у једном језику и њене еквиваленте у другом језику, односно као резултат генеришу се конкорданце које садрже одговор на упит на један или на оба одабрана језика једне паралелне колекције. Други начин је да сагледамо све конкорданце у којима се одговор на упит јавља у текстовима на једном од језика

одабране паралелне колекције али не и на другом. На основу овога можемо утврдити у којим се све облицима кључна реч из упита може појавити у једном језику и да ли постоје њени еквиваленти у другом језику и у ком облику. Српско-немачки паралелни корпус, који је предмет ове дисертације, смештен је у Библишу. У поглављу 6 одељак 6.3 објашњено је који лексички ресурси се позивају приликом претраге, на који начин су одређени лексички ресурси допуњени на основу садржаја српско-немачког паралелног корпуса и како изгледају резултати претраге.

	Number of concordances (en/de/fr): 2251	Broj konkordansi (sr): 2251
Zaman Shuvaa et al., 2011, vol. XII:1, ID: 1.2011.1.3 metadata	Bangladesh National **Library** has 5 staff employed for digitization, National **Library** of the Republic of Indonesia has 12 staff, and the National **Library** of the Philippines has 26 staff employed for digitization.	U Nacionalna **biblioteci** Bangladeša petoro ljudi je angažovano na digitalizaciji, u Nacionalnoj **biblioteci** Republike Indonezije dvanaest, u Nacionalnoj **biblioteci** Filipina dvadeset i šest.
Jordan, 2011, XII:2, ID: 1.2011.2.2 metadata	There are CBS installations at the Royal **Library** of the Netherlands, the German National **Library** and the National **Library** of Australia, and at ABES (Agence bibliographique de l'enseignement superieur) in France, to name a few.	Instalacije CBS sistema postoje na primer u Kraljevskoj **biblioteci** Holandije, Nacionalnoj **biblioteci** Nemačke, Nacionalnoj **biblioteci** Australije i u ABES (Agence bibliographique de l'enseignement superieur) u Francuskoj.
Dakić, 2007, vol. VIII:1/2, ID: 1.2007.1/2.1 metadata	Until the foundation of united national bibliographic-information system COBISS.SR, neither in the University **library** "Svetozar Marković", nor in the National **library** of Serbia and **Library** of Matica srpska existed the unified way of education, except state professional examination, which librarians are due to pass after one year of working in the **library.**	U Univezitetskoj **biblioteci** "Svetozar Marković", kao i u Narodnoj **biblioteci** Srbije i **Biblioteci** Matice srpske, do osnivanja jedinstvenog bibliotečko- informacionog sistema COBISS.SR nije postojao jedinstveni vid obuke, osim stručnog bibliotekarskog ispita koji su bibliotekari dužni da polože posle godinu dana rada u struci.
Trtovac, 2010, vol. XI:2, ID: 1.2010.2.4 metadata	Special thanks to the Theatre Museum of Serbia, **Library** of the Yugoslav Film Archive, Radio Belgrade, Radio Television Serbia, Zvezdara Theatre, Bitef Theatre, Atelier 212, the Yugoslav Drama Theatre, National Theatre in Belgrade, **Library** of the Serbian National Theatre in Novi Sad, Sterija Theatre Festival in Novi Sad, the Cultural Center and **Library,** "Jovan Tomic " in Nova Varos, The Sabac Theatre.	Naročito se zahvaljujemo Muzeju pozorišne umetnosti Srbije, **Biblioteci** Jugoslovenske kinoteke, Radio Beogradu, Radio televiziji Srbije, Zvezdara teatru, Bitef teatru, Ateljeu 212, Jugoslovenskom dramskom pozorištu, Narodnom pozorištu u Beogradu, **Biblioteci** Srpskog narodnog pozorišta u Novom Sadu, Sterijinom pozorju u Novom Sadu, Domu kulture i **Biblioteci** "Jovan Tomić" u Novoj Varoši, Šabačkom pozorištu.
Adžić, 2013, vol. XIV:1, ID: 1.2013.1.6 metadata	**n100** Or, would it be good to join the Digital National **Library** of Serbia?	**n100** Ili se pridružiti Narodnoj **biblioteci,** to jest Digitalnoj Narodnoj **biblioteci** Srbije?
Vasiljević, 2011, XII:2, ID: 1.2011.2.6 metadata	In the National **Library** of the Netherlands we've seen amazing book collection and we got more familiar with Europeana project.	U holandskoj Nacionalnoj **biblioteci** smo se upoznali sa bogatom kolekcijom koju ta **biblioteka** poseduje, kao i sa projektom Europeana čije se sedište nalazi u ovoj **biblioteci.**
Dakić, 2007, vol. VIII:1/2, ID: 1.2007.1/2.1 metadata	Commission for issuing licenses, which consists of members elected from employees in NBS and UBSM, examines the records.	Proveru ovih zapisa vrši Komisija za dodelu licenci, koja se sastoji od članova izabranih iz redova stalno zaposlenih u Narodnoj **biblioteci** Srbije i Univerzitetskoj **biblioteci** "Svetozar Marković".

Слика 9. Произведене конкорданце за упит „biblioteka" са опцијом "EN&SR"

Бошњачко-хрватско-српски паралелни корпус

Поред до сада описаних корпуса који се развијају у Србији, од корпуса који се развијају у региону треба поменути и паралелни корпус бошњачко-хрватско-српског језика (hrWaC) Николе Љубешића и сарадника (Ljubešic and Klubička 2014) који је развијен у оквиру Групе за језичке технологије на Универзитету у Загребу. Корпус обухвата веб материјал који је прикупљен са три највиша интернет домена (top-level domain - TLD): бошњачког (bs), хрватског (hr) и српског (rs). За креирање корпуса коришћен је алат

SpiderLing[71], веб-трагач који представља софтвер за прикупљање текстова са веба. Релевантне URL адресе за бошњачки и српски прикупљене су коришћењем Google Search API упита, а за URL адресе хрватског језика коришћене су насловне стране веб домена које су прикупљене приликом израде прве верзије корпуса hrWaC[72] (Ljubešic and Klubička 2014, 29-30).

Процес анотације сва три корпуса подразумевао је додељивање лема, морфосинтаксичког описа и синтаксу анотацију. За процес анотације коришћени су алати који су већ тестирани на анотираном корпусу SETimes.HR[73]: за лематизацију је коришћен алат CST's Lemmatiser[74], за морфосинтаксички опис алат HunPos[75], а за синтаксну анотацију алат mate-tools[76]. Посебан изазов представљала је сличност између поменутих језика те је развијен посебан метод који запажа разлике између сличних језика, заснован на *n*-грамском језичком моделу и примењен на овом корпусу. Метода испитује фреквенцију појављивања речи над корпусом текстова преузетих са сваког домена појединачно и пореди их са језичким вокабуларом дефинисаним над сваким од интернет домена (Ljubešic and Klubička 2014, 32). На сличан начин развијен је и *n*-грамски модел на нивоу карактера и *n*-грамски модел на нивоу речи како би се испитао квалитет садржаја докумената који су бирани за овај корпус. Прецизније речено, применом овог модела извршена је евалуација садржаја како би се у корпус укључили текстови који су потенцијално квалитетан материјал за различите врсте истраживања (Vitas et al. 2016).

Вишејезични Верн: паралелни преводи романа *Пут око света за 80 дана* Жила Верна

Колекција *Вишејезични Верн* је паралелни вишејезични корпус романа Жила Верна (Jules Verne) *Пут око света за 80 дана* изграђен у оквиру самосталног пројекта који је реализован на Математичком и Филолошком факултету Универзитета у Београду. Жил

[71] SpiderLing, http://corpus.tools/wiki/SpiderLing
[72] hrWaC, http://nlp.ffzg.hr/resources/corpora/hrwac/
[73] SETimes.HR, http://nlp.ffzg.hr/resources/corpora/setimes-hr/
[74] CST's Lemmatiser, https://cst.dk/online/lemmatiser/uk/
[75] HunPos, https://code.google.com/archive/p/hunpos/
[76] mate-tools, https://code.google.com/archive/p/mate-tools/

Верн је најпревођенији француски аутор и други најпревођенији аутор на свету[77] те су његова дела доступна у електронском облику на многим језицима, што их чини одличним кандидатима за израду паралелних корпуса (Vitas and Krstev 2012a). Ово је један од разлога зашто је одабран Жил Верн и његов роман *Пут око света за 80 дана.* Други разлог јесте специфичност садржаја самог текста који је погодан за различите врсте анализа, посебно за поступак препознавања и обраде именованих ентитета (Vitas et al. 2008, 250). Корпус је иницијално сачињен од паралелних превода романа на шеснаест језика: бугарски, хрватски, енглески, француски, немачки, грчки, мађарски, италијански, македонски, пољски, португалски, румунски, руски, српски, словеначки, шпански. Корпус данас садржи двадесет превода поравнатих са оригиналном француском верзијом текста, а у припреми су и преводи на холандском и кинеском, као и још по једна верзија превода на енглеском и немачком језику.

Поступак припреме текстова и паралелизација урађена је на исти начин као и за остале корпусе, а постигнуто је 1-1 упаривање сегмената оригиналног текста на француском и његових превода на пројектне језике. За сваки текст појединачно је урађена аутоматска токенизација, морфосинтаксичка анотација и лематизација уз помоћ алата који су оригинално произведени за одређени језик или који су јавно доступни и прилагођени датом језику. Ознаке коришћене у процесу анотације текстова (осим за бугарски и немачки језик) у складу су са спецификацијама из пројекта MULTEXT, највећим делом са спецификацијама 3. верзије из пројекта Multext-East (Tufiş et al. 2009, 45). Урађена анотација додата је XML коду паралелног корпуса. Разрешавање морфолошке и лексичке вишезначности је урађено делимично аутоматски док је коначна провера урађена потпуно ручно.

Овај корпус је постао веома користан језички ресурс и коришћен је у многим апликацијама за анализу садржаја текстова, највише у погледу бројчаних израза, властитих имена и временских израза. Садржај текстова одличан је за препознавање и обраду именованих ентитета те је корпус искоришћен за тестирање система NERosetta[78],

[77] Статистички подаци о фреквентности превођења аутора у свету доступни су на: http://www.unesco.org/xtrans/bsstatexp.aspx?crit1L=5&nTyp=min&topN=50

[78] NERosetta, http://www.korpus.matf.bg.ac.rs/nerosetta/

веб апликације која може да пореди различите шеме анотирања именованих ентитета и стратегије за њихово означавање што је приказано у (Krstev et al. 2013). Експеримент је извршен само за оне језике за које је систем за препознавање именованих ентитета био доступан: француски, енглески, српски, хрватски и грчки.

Постављањем јединствених упита за екстракцију информација из датог корпуса детаљно су анализирани преводи текстова при чему су утврђене битне разлике и то на местима где се тако нешто не очекује (Vitas et al. 2008, 255-256). Полилексемске јединице (MWU expressions) које су ручно анотиране, а који су постале део система Unitex (детаљније у поглављу 3 одељак 3.3.2), представљају одличан извор за поређење анотација у електронским речницима различитих језика (Vitas and Krstev 2012a).

Орвелова *1984* за српски језик

Српски језик није био део Multext-East пројекта на самом почетку његове реализације. Захваљујући пројекту TELRI у којем су учествовали истраживачи са Математичког факултета Универзитета у Београду урађен је структурно анотирани корпус српског превода романа *1984*. Текст је, као и за све остале језике, био означен SGML етикетама до нивоа реченице и поравнат са енглеским оригиналом. MSDs спецификације за српски језик усклађене су са ознакама које се користе у електронским речницима за српски језик. У првом тренутку произведен је минимални речник који је садржао само облике речи које се појављују у анотираном корпусу 1984. Сам процес анотације корпуса прошао је кроз неколико фаза (Krstev, Vitas and Erjavec 2004, 435):

1. Сви лексички ресурси на српском језику везани за *1984* су морфолошки означени коришћењем система Intex[79]. Као резултат добијена је текстуална датотека која је садржала коначно анотиран текст представљен у форми регуларних израза.
2. Свим непрепознатим речима ручно су додељене леме и морфосинтаксичке категорије, за вишезначно препознате речи којима је додељено више лема или више скупова морфосинтаксичких категорија ручно је одабрана права лема и

[79] Претеча система Unitex, доступно на: http://www.nyu.edu/pages/linguistics/intex/

скуп морфосинтаксичких категорија, док су леме и морфосинтаксичке категорије за једнозначно препознате речи проверене. Овај корак је рађен итеративно што је омогућило корекцију постојећих речника и лексичких ресурса и њихову допуну.

3. Трећи корак подразумевао је израду скрипта у програму Perl који је конвертовао текст анотиран српским е-речницима у текст анотиран према Multext-East спецификацијама.

У верзији 4.0 српски морфосинтаксички лексикон је ажуриран и допуњен тако да он сада садржи надскуп свих лема и њихових облика које се јављају у роману *1984*. Овај текст је коришћен за развој алата за препознавање полилексемских израза (MWE), допуну српског е-речника полилексемских речи са крајњим циљем израде нове верзије овог текста који би био анотиран и полилексемским речима (Krstev, Vitas and Trtovac 2011, 573).

Са овом докторском дисертацијом започиње се рад на изградњи српско-немачког паралелног корпуса - СрпНемКор. Целокупна процедура изградње СрпНемКор-а у погледу припреме текстова и паралелизације је скоро идентична са постојећим паралелним корпусима, са мало унапређеним системом и програмом за паралелизацију, и прати фазе изградње једног паралелног корпуса објашњене у одељку 2.3. СрпНемКор представљен је у поглављу 6 заједно са свим резултатима који су постигнути радом на њему.

3 Израда паралелних корпуса у Србији

У претходном поглављу представили смо већину паралелних корпуса на којима је последњих година радила Група за језичке технологије Универзитета у Београду анализирајући поступак њиховог настајања, њихову структуру и величину у погледу текстуалног материјала који садрже, као и могућности претраге и изглед добијених резултата. У овом поглављу детаљније ћемо представити фазе израде кроз које су прошли већ описани паралелни корпуси и језичке ресурсе и алате који су том приликом коришћени за обраду материјала са посебним освртом на правну регулативу за регулисање ауторских права која се са једне стране односе на материјал који је део корпуса, а са друге стране на права приступа и коришћења самог корпуса који као ауторско дело, такође, подлеже ауторским правима. У току израде корпуса који је предмет ове дисертације прошли смо кроз све наведене фазе и користили већину приказаних језичких ресурса и алата за обраду текстова што је детаљније приказано у поглављу 6 одељак 6.2.

3.1 Ауторска права, права приступа и коришћења садржаја корпуса

Ауторско право представља законску заштиту аутору књижевног, научног и уметничког дела, гарантујући му ексклузивно право да контролише продукцију и коришћење свог дела (Брзуловић Станисављевић 2010, 73). Ауторско дело је оригинална духовна творевина аутора, изражена у одређеној форми, без обзира на његову уметничку, научну или другу вредност, његову намену, величину, садржину и начин испољавања, као и допуштеност јавног саопштавања његове садржине (Завод за интелектуалну својину Републике Србије 2009, 1). Ауторска права регулишу се националним законима које свака земља доноси у складу са међународним нормама и представљају економска права власништва која се могу пренети на издаваче, агенције или нека друга лица. Према Бернској конвенцији о ауторским правима (Bernska konvencija 1971) коју је потписала већина земаља, трајање ауторског права је доживотно, плус најмање 50 година од смрти аутора. У многим земљама, па и у Србији, то је данас

продужено на 70 година од смрти аутора, односно ако има више аутора, последњег коаутора (Брзуловић Станисављевић 2010, 76).

Ауторским и сродним правима ауторима се гарантује сигурност да су њихова дела заштићена од неовлашћеног копирања или пиратерије, а са друге стране им се даје подстицај у облику признања или новчане накнаде (Завод за интелектуалну својину Републике Србије 2009, 5-6). Термин „ауторско право" карактеристичан је за европско право, док се у англосаксонском праву користи термин „копирајт" (copyright). Основна разлика између ова два термина је у томе што је ауторско право у суштини лично право аутора засновано на вези између аутора и његовог дела, док се копирајт искључиво односи на дело као такво (Prlja i dr. 2012, 10).

Развој информационих технологија, интернета и дигиталног света донео је велике изазове у многим областима права па и у погледу ауторских. Интернет представља највећи ресурс информација на свету који је широко отворен за представљање и преузимање информација. Свако ауторско дело на интернету заштићено је ауторским правом на основу Бернске конвенције о ауторским правима и на основу националних закона о ауторским правима без обзира да ли је то на самом делу назначено или није (Prlja i dr. 2012, 8).

Сами корпуси, као ресурси на интернету, такође подлежу ауторским правима. Према неким мишљењима приликом креирања корпуса креатори су у обавези да пре прикупљања и обраде текстова утврде да ли одређени текст у тренутку коришћења подлеже ауторским правима или не. Већина текстова која се уврштава у корпус (новински чланци, научни и књижевно-уметнички текстови) заштићена је на неки начин ауторским и сродним правима што подразумева да се без дозволе носиоца ауторских права дело не може умножавати или репродуковати односно у случају креирања корпуса дигитализовати, конвертовати у дигиталне формате, анотирати, дистрибуирати и стављати на увид истраживачима. У случају да текст који се уврштава у корпус подлеже ауторским правима први корак је утврдити да ли постоји издање текста чије је ауторско право истекло, у супротном неопходно је постићи споразум са носиоцима ауторских права односно добити дозволу за употребу текста као дела електронског корпуса. Том приликом

је врло битно јасно спецификовати за коју врсту истраживања је потребан текст, треба дати опис медијума и формата који ће се користити за складиштење текста, листу свих публикација које ће се произвести на основу коришћења текста и формулацију захвалности носиоцу ауторског права које укључује и референце на његово ауторско дело као саставни део произведених публикација (Barnbrook 1996, 33). Са друге стране, корпус као готов производ такође је ауторско дело те се и његово умножавање, дистрибуција и експлоатација уређују на неки начин ауторским правима, односно одговарајућим лиценцама.

Када је реч о корпусима, постоји неколико алтернатива у процесу обезбеђивања свих потребних дозвола од носиоца ауторских права на корпусним текстовима (McEnery and Hardie 2012, 59):

1. укључити у корпус само оне текстове који су јавно добро или користе лиценцу која омогућава слободно умножавање, конверзију и дистрибуирање;
2. креирати корпус на основу текстова са веба, не дистрибуирати сам корпус већ листу адреса корпусних текстова на вебу;
3. креирати корпус не тражећи дозволу од носилаца ауторских права на корпусним текстовима, а потом не дистрибуирати корпус већ омогућити корисницима ограничен приступ који не нарушава закон о ауторским правима.

Приликом креирања некомерцијалних корпуса, као што је реч у случају корпуса који је предмет ове дисертације, креатори корпуса позивају се на 3. алтернативу која је имплементирана преко конкорданцера 4. генерације и заснива се на такозваној поштеној употреби из следећих разлога:

1. Корпус се не дистрибуира.
2. Корисници преко веб читача и веб сумеђе корпуса постављају и прослеђују упит корпусу, а као резултат добијају генерисане конкорданце ограниченог контекста.
3. У пракси се конкорданце у веб читачу показују као део реченице, реченица или пар реченица, али не и као цео текст. На основу конкорданци корисници не могу на једноставан начин да реконструишу цео текст.

Поред закона о ауторским и сродним правима, креатори корпуса морају да воде рачуна и о закону о заштити приватности из следећих разлога (Utvić 2013, 65-66):

1. (ЗП1) приликом претраге корпуса у текстовима се могу наћи информације о особама и организацијама које задиру у њихову приватност.
2. (ЗП2) информације придружене текстовима корпуса (метаподаци) попут идентитета особа које су извори говорних текстова, звучни записи њиховог говора и транскрипти тих записа такође могу нарушити приватност, како самих говорника, тако и особа и организација о којима говоре.
3. (ЗП3) уколико се од корисника корпуса очекује да се региструје како би могао да добије корисничко име и лозинку, односно да приступи корпусу преко система ауторизације, постоји потенцијална опасност да се лични идентификациони подаци који се захтевају од корисника приликом регистрације злоупотребе.

Приликом израде паралелног корпуса који је предмет дисертације поштоване су норме дефинисане за израду некомерцијалних корпуса, као и лиценце којима се регулише употреба одређених ресурса и језичких алата који су коришћени приликом обраде текстова одабраних за корпус: ЗП1 јер се у романима јављају само фиктивни ликови, ЗП2 јер није у питању говорни корпус и ЗП3 јер сам систем води рачуна о регистрацији корисника.

3.2 Прикупљање и дигитализација текстова

Прикупљање текстова представља први корак у процесу израде корпуса. Када није реч о опортунистичким корпусима (опортунистички корпуси детаљније су објашњени у поглављу 2 одељак 2.2.2), аутори унапред утврде критеријуме по којима ће бити бирани текстови за садржај корпуса и приступају њиховом прикупљању. Независно од врсте корпуса који се креира, аутори корпуса морају имати законско право да користе прикупљене текстове што је објашњено у претходном одељку. Приликом прикупљања текстови могу бити у различитим форматима (рукопис, штампани текст, дигитална слика текста, електронски текст и многи други). Први и основни критеријум за креирање једног

електронског корпуса јесте постојање текстова у електронском формату. Ако су одабрани текстови већ у електронском формату они се као такви користе у даљој обради, док се текстови у другим форматима конвертују у формат електронског текста, односно ради се дигитализација.

Дигитализација подразумева конверзију аналогних (недигиталних) објеката (штампаних књига, рукописа, аудио и видео материјала) у дигитални облик. Дигитализација штампаних или писаних текстова најчешће подразумева сканирање материјала при чему се недигитални текст најпре трансформише у неки од формата дигиталних слика (GIF, JPEG, PNG, TIFF, BMP), а затим се дигитална слика оптичким препознавањем карактера конвертује у електронски читљив текст. Квалитетна дигитализација штампаног или рукописног текста или сликовне грађе подразумева следеће фазе које су део претходног процесирања материјала након кога следи оптичко препознавање карактера (Ikonomov and Dobreva 2008, 2):

1. *Сканирање штампаног материјала*: подразумева превођење у стандардне формате - препоручује се TIFF формат јер JPEG и PNG често као излаз дају мутне и недовољно јасне слике.
2. *Обрада слике*: раздвајање страница, исецање делова и поравнавање маргина; уклањање закривљења насталих процесом сканирања, прилагођавање контраста, прилагођавање величине и резолуције слике.
3. *Креирање излазног документа који ће се користити у дигиталним библиотекама*: формат у коме излазни документ треба да се појави је најчешће PDF на коме се касније примењују технологије оптичког препознавања карактера.

Већина текстова одабрана за корпус који је предмет дисертације дигитализована је за потребе израде корпуса и примењен је софтвер за оптичко препознавање карактера. Процес оптичког препознавања карактера детаљније је објашњен у поглављу 4 одељак 4.2.1, док је у поглављу 6 одељак 6.2.1 приказано који текстови одабрани за корпус су дигитализовани, а који су већ преузети у електронском формату, као и како је урађена дигитализација.

3.3 Доступни алати и језички ресурси за обраду текстова

3.3.1 IMS OCWB

Два најзначајнија софтверска окружења која је Група за језичке технологије до сада користила за развој паралелних корпуса су IMS OCWB и систем Unitex[80]. IMS OCWB је бесплатан софтвер отвореног кода развијен за рад са текстуалним корпусима. Оригинални IMS CWB развијен је средином деведесетих година 20. века у Институту за (аутоматску) обраду природних језика у Штутгарту (Institut für Maschinelle Sprachverarbeitung - IMS)[81], по коме је програм и добио назив (Proisl and Uhrig 2012, 2750). Софтвер данас носи назив IMS OCWB и још увек се активно развија. Тренутно доступна верзија 3.4 се може користити под условима лиценце GNU GPL (GNU General Public License)[82], док је 3.5 у фази бета-тестирања. Софтвер садржи алат за постављање упита за претрагу CQP (Corpus Query Processor) заснован на регуларним изразима који користе синтаксу усклађену са спецификацијом стандарда POSIX (објашњено детаљно у одељку 3.6). CQP дозвољава да се упит задаје у форми логичког (буловског) израза. Уз употребу овог софтвера и његовог процесора на Математичком факултету у Београду развијена је онлајн сумећа за претрагу развијених корпуса. О алату CQP и поступку постављања упита помоћу њега више се може сазнати у (Evert 2016). Од 2011. године аутори софтвера развијају веб сумећу CQP Web са циљем да се омогући рад са произвољним корпусом креираним помоћу IMS OCWB алата, као и претраживање на интернету. Сумећа CQP Web је развијена по угледу на веб сумећу за IMS OCWB алате – BNCweb, коју користи Британски национални корпус.

3.3.2 Unitex

Unitex представља слободан софтвер отвореног кода развијен за обраду и анализу корпусних текстова на природним језицима применом језичких ресурса и алата као што су електронски морфолошки речници, локалне граматике у форми графова и табеле лексикон-граматика (Paumier 2011, 9). Софтвер се дистрибуира под лиценцом LGPL (Lesser

[80]Unitex, http://unitexgramlab.org/
[81] Institut für Maschinelle Sprachverarbeitung, http://www.ims.uni-stuttgart.de/
[82] GNU GPL лиценца је доступна на: http://www.gnu.org/licenses/gpl-3.0.en.html

General Public License)[83], док се пратећи лингвистички ресурси могу користити под лиценцом LGPLLR (Lesser General Public License For Linguistic Resources)[84]. Систем је развијен у Лабораторији за аутоматску документацију и лингвистику (Laboratoire d'Automatique Documentaire et Linguistique - LADL) као компатибилна замена и надградња алата INTEX који је првобитно развио Макс Силберштајн (Max Silberztein) из исте лабораторије за обраду корпуса помоћу морфолошких електронских речника у форматима DELA (Silberztein 1993). Након интеграције лабораторије LADL и Универзитета Марн-ла-Вале Париз Исток (Université Paris-Est Marne-la-Vallée) Институт Гаспар-Монж (Institut d'électronique et d'informatique Gaspard-Monge) развијен је алат Unitex са подршком за Unicode кодну шему чиме се отворила могућност коришћења алата за широк спектар језика. Актуелна верзија Unitex-а је 3.2 и може се преузети са званичног сајта Unitex-а на коме је доступан и детаљан приручник о коришћењу самог програма[85]. У званичној дистрибуцији Unitex-а налазе се основни језички ресурси за 22 језика, укључујући и српски.

Unitex је имплементиран коришћењем програмских језика C/C++ и Java, и може се компилирати на платформама Linux, Windows и MACOS. Систем обраде корпусних текстова заснива се на коришћењу лексичких ресурса који се имплементирају у виду коначних аутомата и трансдуктора, а који се користе приликом претходне обраде и претраге текста. Претходна обрада текста или претпроцесирање, приликом које улазни текст остаје неизмењен, а сва обрада се односи на његову копију, подразумева токенизацију и нормализацију текста, његову сегментацију на реченице, као и морфосинтаксичко означавање применом електронских морфолошких речника монолексемских и полилексемских речи.

Приликом покретања система Unitex први корак је одабир језика са којим ће се радити. Одабиром језика систему се указује са којим писмом ради те се на основу тога и повезује са одговарајућим датотекама језика које постоје у радном простору приватног

[83] LGPL лиценца је доступна на: http://www.gnu.org/licenses/lgpl-3.0.html
[84] LGPLLR лиценца је доступна на: http://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/lgpllr.html
[85] Приручник је доступан на: https://unitexgramlab.org/

каталога корисника. Сваки језик представља поткаталог једног приватног каталога који је одабран приликом првог покретања програма (на пример, MojUnitex/Serbian_Latin). Један поткаталог језика садржи додатне поткаталоге у које се смештају сви текстови који ће бити анализирани, компилирани речници за обраду текстова, графови и подграфови за претрагу текстова, графови који се користе за отклањање вишезначних анотација у текстовима, као и поткаталози у које се смештају изворни облици речника лема и флективни графови за производњу речника облика. Поред поменутих поткаталога садржај поткаталога језика су и конфигурационе датотеке којима се дефинише алфабет једног језика, као и конфигурација система Unitex.

Након покретања програма и одабира језика приступа се процесу обраде текста. Текст који је улазни и који се анализира мора бити такозвани чист, обичан, односно сирови текст (plain или raw text) који систем препознаје преко проширења имена датотеке .txt или .xml. Такође, подразумева се да је текст већ у Unicode-у, а у противном систем омогућава аутоматско конвертовање у складу са Unicode кодном шемом одабраног језика. Друга текстуална датотека са којом ради Unitex јесте датотека са проширењем имена .snt. Ова датотека добија се као резултат фазе претпроцесирања односно претходне обраде која подразумева неколико корака: нормализацију сепаратора, поделу у реченице, нормализацију недвосмислених облика, токенизацију и примену речника. Сви ови кораци су детаљно објашњени у Приручнику за Unitex (Unitex 3.2 User Manual) Себастијана Помијеа (Sébastien Paumier) из 2018. године.[86]

Систем Unitex користи локалне граматике у форми правила (регуларних израза) или у форми графова као форму за задавање упита. Локалне граматике су алати за описивање значајних лингвистичких феномена. Приликом креирања локалних граматика најпре се описују једноставне, опште конструкције које се често појављују у тексту како би се те граматике, касније, могле искористити за опис сложенијих конструкција. Основне предности локалних граматика су што се оне могу независно конструисати, одржавати и примењивати и што могу реферисати на информације из електронских морфолошких речника.

[86] Оригинални приручник је написао Себастијан Помије који је дуго радио на развоју система; сада се приручник допуњује прилозима многих корисника.

Локалне граматике у форми регуларних израза заснивају су на регуларним операцијама које описују разноврсне лингвистичке феномене (морфолошке, синтаксичке, семантичке). Специфичност примене регуларних израза у систему Unitex огледа се у коришћењу лексичких информација приликом задавања упита као што су лема, врста речи, семантички маркери, морфолошке категорије (ово је све детаљније објашњено у следећем одељку). Регуларни изрази могу бити (Paumier 2011, 51):

1. токени или лексичке маске
2. конкатенација, односно спајање два регуларна израза
3. унија два регуларна израза
4. Клинијева звезда.

Ове појмове ћемо сада детаљније објаснити примерима. Под токенима се подразумевају појединачни несловни карактери (интерпункцијски знаци, цифре и слично), ниске слова[87], сепаратор реченица {S}, ознака за заустављање {STOP}, лексичке маске. Они се могу користити уз помоћ морфолошких филтера за регулисање упита у форми регуларних израза у POSIX формату, али велику предност Unitex-а представљају лексичке маске које користе информације из речника. Оне представљају упитни образац који се сравњује са једним или више токена. Постоји више врста лексичких маски (Krstev 2016b):

a. *Предефинисане лексичке маске*. Уграђене су у систем Unitex и не зависе од језика. Пример <E> празно слово, указује да на одређеном месту у тексту нечега не мора бити; <TOKEN> - сравњује се са сваким токеном осим са бланко карактером и друге.[88]
b. *Лексичке маске које реферишу на речник текста*. Ове маске су зависне од језика, односно зависе од тога којим речницима је текст обрађен. Најједноставније маске реферишу на лему или врсту речи. Пример <ljubav.N> реферише на све облике речи чији је канонски облик *љубав,* а врста речи именица.

[87] Шта је „Слово“ је дефинисано избором језика.
[88] Све предефинисане лексичке маске су детаљно представљене у Приручнику за Unitex Себастијана Помијеа.

c. *Лексичке маске које користе синтаксичке и семантичке маркере*. Код ових израза маркери се одвајају знаком +, а из речника текста се издвајају само они улази који имају све наведене синтаксичке или семантичке маркере. Пример <N+Bot> издваја из речника улазе који су именице (код за врсту речи N) и који означавају биљке (семантички маркер Bot).

d. *Лексичке маске са искључивањем семантичких маркера*. Код ових маски семантички и синтаксички маркери се раздвајају знаком тилда **~**, а из речника текста издвајају се само они улази који немају ниједан од наведених синтаксичких или семантичких маркера. Пример <N+NProp~Hum> издваја из речника улазе који у коду имају N+NProp што означава властите именице (код N+NProp), али који у коду немају маркер Hum (људско биће)**.** Односно, издваја примере из речника текста који се односе на властита имена људи која не означавају људе. На пример, регуларни израз издваја из речника Dunavom, Dunav.N+NProp+Top+Hyd+River:ms6q**.**

e. *Лексичке маске са флективним условима*. Код ових маски ограничења се постављају коришћењем флективних кодова који се користе и у примењеним речницима. Да би се ова ограничења користила неопходно је да им претходи бар ознака речи или један маркер. Пример <N:ms4> препознаје у тексту све облике који су према речнику именице мушког рода у једнини у акузативу.

Такође, постоји и могућност негације лексичких маски уз помоћ знака „!” који се умеће одмах после знака за мање.

Конкатенација подразумева спајање регуларних израза на неколико начина. Пример <Prep+p2><N:2> проналази предлоге који траже генитив (+p2) иза кога следи именица у генитиву. Приликом формирања сложенијих регуларних израза са различитим операторима користе се округле заграде. Пример (<PRD+ProA:ms1>+<A:ms1>)<N:ms1> проналази именице у мушком роду у номинативу једнине којима претходи или заменица у мушком роду у номинативу једнине или придева у мушком роду у номинативу једнине.

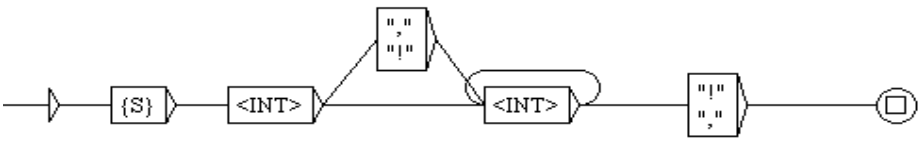
Унија регуларних израза постиже се коришћењем оператора + и сравњује се са једним од операнада односно означава избор. Пример <PRD+ProA:ms1>+<A:ms1>

проналази или заменице у мушком роду у номинативу једнине или придеве у мушком роду у номинативу једнине.

Клинијева звезда (користи се оператор *) омогућава да се у тексту сравни нула, једно или више појављивања регуларног израза на који се примењује. На пример, <A>* <N> препознаје секвенце у тексту од једне, две, три и више речи од којих је последња именица, а могу да јој претходе придеви. Ако је потребно препознати у тексту бар једно појављивање израза на који се примењује Клинијева звезда довољно је да се тај регуларни израз понови. Пример <A><A>*<N> препознаје секвенце у тексту које имају бар један придев испред именице.

За моделирање сложенијих лингвистичких феномена регуларни изрази нису адекватни јер брзо постају сувише комплексни те се уз њих примењују и колекције графова и подграфова. Графови представљају погодну визуелизацију регуларних израза. Графови за претрагу и анотацију називају се синтаксичким графовима (Syntactic graphs). Сваки чвор у графу садржи једноставне регуларне изразе који садрже ниске, лексичке ознаке или позиве других графова, другим речима све оно што садрже и регуларни изрази.

Пример 1:



Пример 1 представља граф који препознаје на почетку реченице (<S>) знак узвика (<INT>) иза кога може, али и не мора да следи интерпункцијски знак за узвик или зарез (“,”+”!”), а иза тога може се појавити произвољан број узвика (петља), а све се завршава опет интерпункцијским знаком за узвик или зарез. За претпроцесирање и обраду текстова одабраних за корпус који је предмет дисертације користили смо систем Unitex у који су уграђени језички ресурси за обраду текстова на српском језику. Поступак је објашњен у поглављу 6 одељак 6.2.3.

3.3.3 Е-речници

Електронски морфолошки речници заједно са локалним граматикама представљају неопходне језичке ресурсе у процесу аутоматске обраде текстова. Под електронским речником подразумева се речник намењен обради и анализи текста који садржи неопходне информације за лакши процес сегментације и морфолошке анотације. Процес стварања електронских речника је спор јер се највећи део њихове конструкције обавља ручно, али веза ових речника са корпусима омогућава напредну обраду текста која не би била могућа без њихове подршке (Vitas and Krstev 2012b, 279). Аутоматска обрада текста увек почиње од појединачних речи које представљају основне јединице од којих је неки текст сачињен. Међутим, како појединачне речи нису увек природна јединица обраде, речници монолексемских речи допуњују се другим врстама речника међу којима су и речници полилексемских речи[89] (Тртовац и Андоновски 2014, 231).

Међународна мрежа лабораторија која се бави рачунарском лингвистиком RELEX (Laporte 2003), направила је модел изградње електронских морфолошких речника који су усвојиле бројне организације које се баве обрадом природних језика. Систем Unitex ради са електронским морфолошким речницима развијеним према овом моделу. Реч је о речницима у формату DELA. Како би се направила разлика између монолексемских и полилексемских јединица овај електронски речник је организован у два одвојена подсистема: речник монолексемских (DELAS и DELAF) и речник полилексемских јединица (DELAC и DELACF). На основу ових модела, у оквиру Групе за језичке технологије Универзитета у Београду, изграђени су електронски морфолошки речници српског језика на латиничном и ћириличном писму чији су аутори др Цветана Крстев, професор Филолошког факултета Универзитета у Београду, и др Душко Витас, професор Математичког факултета Универзитета у Београду. Систем речника монолексемских и полилексемских речи допуњен је и речником који садржи коначне аутомате и трансдукторе за препознавање непознатих речи (Vitas and Krstev 2012b, 280).

[89] Термини „монолексемске речи“ и „полилексемске речи“ користе се према усвојеним енглеским терминима “simple words” и “multiword units”

Да би се електронски морфолошки речници српског језика могли примењивати у анализи и обради најразличитијих текстова пре почетка њихове израде било је потребно водити рачуна о следећим специфичностима (Krstev et al. 2006a, 552-553) (Vitas and Krstev 2005, 140):

1. *Употреба два алфабета:* у српском језику у равноправној су употреби два алфабета, ћирилица и латиница, и текстови се појављују на оба.
2. *Правопис заснован на фонологији:* употреба различитих варијанти српског језика (екавски и ијекавски говор) због чега постоје дублети.
3. *Транскрипција:* страна властита имена транскрибују се у складу са српским правописом.
4. *Богат морфолошки систем:* огледа на нивоима флексије и деривације.
5. *Слободан ред речи у реченици.* Реченице дозвољавају слободан редослед погледу субјекта, предиката, објекта и осталих делова реченице, као и употреба енклитика.

Речник монолексемских речи састоји се из три дела (Krstev, Vitas and Pavlović-Lažetić 2008, 3):

1. речник лема DELAS,
2. низ трансдуктора којима се описују својства флективних парадигми и
3. речник флективних облика DELAF.

DELAS је речник канонских облика (лема) и користи се за генерисање речника DELAF. Свака јединица у речнику DELAS састоји се из: канонског облика речи (леме), везе са подређеним облицима (ако постоје) и одговарајућих маркера који описују својства леме (синтаксичка, семантичка, дијалекатска, употребна и друга).

Формат одреднице у речнику DELAS:

lemma,Knnn[+SinSem]*

lemma – проста реч у облику који је обично идентичан оном у традиционалним речницима (за именице је то номинатив једнине, за глаголе инфинитив)
K – ознака за врсту речи
nnn – алфанумеричка ознака која одређује класу лема које деле иста флективна својства, а описују се одговарајућим трансдуктором *Knnn*
+SinSem – маркер који описује синтаксичка, семантичка и друга својства леме

Пример одреднице у речнику DELAS:

bibliotekarka,N661+Hum+Prof+GM

bibliotekarka	канонски облик (лема)
N	врста речи што је у овом примеру именица
661	флективна класа која генерише све флективне облике
+Hum	семантички маркер који указује да се ради о особи
+Prof	семантички маркер који указује да се ради о професији
+GM	семантички маркер који указује да је у питању именица женског рода изведена моцијом рода из одговарајуће именице мушког рода

Структура речника DELAS омогућава прецизно и аутоматско генерисање свих облика одредница које су садржај речника DELAF. DELAF је речник подређених облика или речник флективних облика и користи се у аутоматској обради текста. Одредницу у DELAF речнику прате следеће информације: канонски облик речи (лема); ознака за врсту речи; синтаксички, семантички и други маркери који су преузети од леме; скуп кодова који указују на вредности граматичких категорија канонског облика речи (једномесни кодови састављени од великог или малог слова или цифре) (Krstev, Vitas and Pavlović-Lažetić 2008, 4-5). На основу садржаја речника DELAF могуће је извршити аутоматску сегментацију текста у речи, као и морфолошку анализу применом метода лексичког препознавања (Silberztein 1993).

Формат одреднице из речника DELAF:

form,lemma[+SinSem]*[:categories]*

form – подређени (реализовани) облик односно облик просте речи препознат у тексту
lemma - канонски облик (лема) преузет из речника DELAS
+SinSem – маркер који описује синтаксичка, семантичка и друга својства леме
:categories – могуће граматичке категорије датог облика просте речи у виду јединствених карактерских кодова

Пример одреднице из речника DELAF:

bibliotekarkom,bibliotekarka.N+Hum+Prof+GM:fs6v

bibliotekarkom	подређени (реализовани) облик
bibliotekarka	канонски облик (лема)
N	врста речи што је у овом примеру именица (податак преузет од канонског облика)
+Hum+Prof	семантички маркери који су преузети као податак од канонског облика и у овом примеру указују да је у питању професија људи

+GM	семантички маркер који указује да је у питању именица женског рода изведена моцијом рода из одговарајуће именице мушког рода
f	категорија род што је у овом примеру женски
s	категорија број што је у овом примеру једнина
6	категорија падеж што је у овом примеру инструментал
v	категорија аниматности што је у овом примеру живо

Српски речник DELAS тренутно садржи више од 192.000 улаза од којих је генерисано око око 7 милиона одредница у речнику DELAF, од тога око 2,5 милиона различитих облика.[90]

Полилексемске речи (Multi-Word Units - MWUs) представљају секвенце појединачних речи организоване у одговарајуће језичке структуре са прецизно одређеним значењем. Такође, могу се дефинисати и као секвенце појединачних речи које формално представљају ниске алфабетских карактера одређеног језика показујући одређен степен морфолошке, дистрибутивне, синтаксичке или семантичке некомпозитности (Krstev et al. 2010, 226). Структура вишечланих речи је врло стриктна и обично се редослед речи и додатних неалфабетских карактера не може мењати како се не би нарушило значење полилексемске јединице.

Последњих деценија групе за језичке технологије почеле су интензивно да развијају речнике и граматике полилексемских речи сматрајући их једним од битних алата за аутоматску обраду природних језика (Savary 2008, 2-3). Јединице које чине структуру једне полилексемске речи углавном се могу појединачно морфолошки анализирати што омогућава њихову исправну и исцрпну флективну анализу за шта се могу искористити већ развијени речници монолексемских речи. Па ипак, израда електронских морфолошких речника полилексемских речи сложенији је посао у односу на израду електронских морфолошких речника простих речи.

Речник полилексемских јединица састоји се, као и речник монолексемских речи, из два речника: DELAC и DELACF. Одредницу у речнику DELAC прате леме свих делова дате полилексемске јединице које су дефинисане према речнику DELAS заједно са флективним обликом и кодовима граматичких категорија из речника DELAF коју прати флективни

[90] Подаци о броју улаза и одредница у електронским морфолошким речницима српског језика су из марта 2019. године.

класни код и одговарајући синтаксички и семантички маркери који се односе на полилексемску лему (Krstev and Vitas 2009, 205).

Пример одреднице из речника DELAC:

javna(javan.A7:aefs1g) biblioteka(biblioteka.N612:fs1q), NC_AXN+Org

javan, biblioteka		канонски облици (леме) компонената полилексемске јединице дефинисане према речнику DELAS
A7, N612		флективни класни кодови дефинисани према речнику DELAS за претходно наведене леме редом
aefs1g		граматичке категорије дефинисане према речнику DELAS
	a	категорија степен поређења што је у овом примеру позитив
	e	категорија одређеност што је у овом примеру без значења
	f	категорија род што је у овом примеру женски
	s	категорија број што је у овом примеру једнина
	1	категорија падеж што је у овом примеру номинатив
	g	категорија аниматност што је у овом примеру без значаја
f s1q		граматичке категорије дефинисане према речнику DELAS
	f	категорија род што је у овом примеру женски
	s	категорија број што је у овом примеру једнина
	1	категорија падеж што је у овом примеру номинатив
	q	категорија аниматности што је у овом примеру неживо
NC_AXN		флективни класни код полилексемске јединице, AXN означава да се испред именице налази придев који се са именицом слаже у роду, броју, падежу и аниматности
+Org		семантички маркер што је у овом примеру организација

Флективни облици у речнику DELACF генеришу се аутоматски на основу информација које стоје уз одредницу у речнику DELAC и на основу података у речницима DELAS/DELAF. Одредницу у DELACF речнику прате следеће информације: канонски облик (лема) полилексемске јединице, одређен флективни код класе и скуп кодова који указују на вредности кодова граматичких категорија облика полилексемске речи. За генерисање одредница за речник DELACF потребна су два типа трансдуктора: трансдуктори који производе флективне облике монолексемских речи и други који регулишу однос између конституената у одређеној полилексемској речи.

Пример одреднице из речника DELACF:

javnoj biblioteci,javna biblioteka.NC_AXN:fs7q

javnoj biblioteci		подређени (реализовани) облик
javna biblioteka		канонски облик (лема)
NC_AXN		флективни код класе AXN означава да се испред именице налази придев који се са именицом слаже у роду, броју, падежу и аниматности
fs7q		вредност граматичке категорије облика полилексемске јединице:
	f	категорија род што је у овом примеру женски

s	категорија број што је у овом примеру једнина
7	категорија падеж што је у овом примеру локатив
q	категорија аниматности што је у овом примеру неживо

Српски речник DELAC тренутно садржи више од 18.000 улаза од којих је генерисано око 320.000 одредница у речнику DELACF, међу којима је око 160.000 различитих облика.

Поред општег речника у оквиру Групе за језичке технологије развијају се и специјални речници као што су речник топонима и других геополитичких имена, затим речник српских личних имена DELAS-PERS, а од посебног значаја су терминолошки речници из различитих домена као што су библиотекарство и информатика, рударство и геологија и други (Тртовац 2016).

Применом припремљених електронских морфолошких речника монолексемских и полилексемских јединица на текст који се анализира систем производи следеће три датотеке: *dlf* – сортирана листа монолексемских речи из текста са свим могућим интерпретацијама из речника монолексемских речи, *dlc* – сортирана листа полилексемских речи из текста са свим могућим интерпретацијама из речника вишечланих речи и *err* – сортирана листа непознатих речи, односно простих речи које речници нису препознали у тексту који се анализира. Посматрано скуповно, *dlf* и *dlc* представљају пресек текста и речника, док *err* представља разлику текста и речника. Датотека *err* садржи сортирану листу непрепознатих речи у тексту који се обрађује. Део листе су речи које примењени речници нису препознали као део свог садржаја, а које могу бити потенцијални кандидати за речнике. Обрада непрепознатих речи подразумева ручну обраду и разврставање на оне које припадају општем и оне које улазе у састав специјалног терминолошког речника српског језика или у посебан речник књижевног дела или аутора. Такође, део сортиране листе су често и речи за које се након обраде може утврди да су грешке у самом тексту које се током евалуације ове листе и поправљају.

Текстови одабрани за корпус који је предмет дисертације обрађени су коришћењем ових речника у оквиру система Unitex. Поступак обраде детаљно је објашњен у поглављу 6 одељак 6.2.3.

3.3.4 Ворднет

Ворднет је лексичко-семантичка мрежа првобитно развијена за енглески језик на Принстонском универзитету као лексикон намењен истраживачима из области психолингвистике. Први Ворднет, такозвани Принстонски Ворднет (Princeton WordNet), почео је да развија професор Џорџ Милер (George Miller) 1985. године са групом психолингвиста и лингвиста у лабораторији за когнитивне науке са циљем да се створи ресурс који омогућава увезивање рачунарских могућности са традиционалним лексикографским начином представљања информација (Miller et al. 1990, 236).

Ворднет је замишљен као помоћно средство за концептуално претраживање речника и да се користи у блиској вези са конвенционалним онлајн речником. Међутим, „у Ворднету речнички појмови нису представљени алфабетским редом, као у традиционалним речницима, већ концептуално – на основу семантичке меморије и поимања“ (Митровић 2018, 45). Структура Ворднета (Fellbaum 1998) заснива се на концептима који су груписани у синсетове, когнитивне парове синонима (synset или synonymous set), који су повезани концептуално-семантичким везама као што су хипонимија и меронимија. Изградња Ворднета се данас креће у правцу глобалног Ворднета који омогућава компаративну језичку анализу, затим управљање знањем, управљање садржајем, екстракцију и проналажење информација и машинско превођење који се заснивају на вишејезичности (Krstev et al. 2006c, 114).

Рад на Српском Ворднету започет је у оквиру пројекта BalkaNet (BalkaNet Multilingual Balkan Wordnet)[91] у периоду од 2001. до 2004. године. Пројекат BalkaNet имао је за циљ проширење вишејезичне базе Ворднет успостављене у оквиру пројекта Euro WordNet (EWN)[92] балканским језицима (Митровић 2018, 54-58). BalkaNet је омогућио развој поравнатих Ворднетова за грчки, турски, бугарски, румунски и српски језик, док је за чешки језик настављена изградња започета у оквиру Euro WordNet. Структура Ворднета на балканским језицима заснована је на структури Принстонског Ворднета, али су

[91] BalkaNet, http://www.dblab.upatras.gr/balkanet/
[92] Euro WordNet (EWN), http://projects.illc.uva.nl/EuroWordNet/

укључени и нови концепти специфични за сваки балкански језик у погледу историјских, друштвених, географских и других прилика (Krstev 2006, 275-285).

Структура српског Ворднета заснива се на структури Принстонског Ворднета из кога је на самом почетку рада преузет највећи део појмова. Након завршетка пројекта BalkaNet настављен је рад на српском Ворднету као волонтерски рад сарадника. Од 2006. године кооперативан рад на доградњи српског Ворднета настављен је на постдипломским студијама на Катедри за библиотекарство и информатику на Филолошком факултету Универзитета у Београду (Крстев и др. 2008). Развој српског Ворднета ослањао се на постојеће ресурсе српског језика, а највише на Речник Матице српске. Због непостојања адекватног електронског енглеско-српског речника синсетови су превођени ручно, уз очување семантичке структуре Ворднета, док су значења литерала (делови синсетова) преузимана из Речника Матице српске колико год је то било могуће. Морфолошке, синтаксичке и семантичке особине тих литерала, кодови њихових флективних класа, синтаксички и семантички маркери уведени су из електронског морфолошког речника простих речи српског језика.

3.4 Анотација корпуса

3.4.1 Структурна анотација

Након дигитализације и потпуне обраде уз помоћ система Unitex приступа се структурној анотацији текстова. Доступни алати за паралелизацију (објашњено у следећем одељку овог поглавља) захтевају да текстови на улазном и циљном језику буду добро формирана XML[93] документа која су у исто време и валидна у односу на одговарајућу дефиницију типа документа, DTD.

Под добро формираним XML документом подразумева се да су задовољени одређени формални критеријуми (Krstev 2016a), као на пример, да свака почетна етикета, осим етикета празних елемената, мора да има завршну етикету, и да се садржаји

[93] Детаљније о XML-у може се погледати на званичној веб страни World Wide Web конзорцијума (https://www.w3schools.com/xml/) и курсу проф. др Цветана Крстев који је припремљен за студенте Катедре за библиотекарство и информатику Филолошког факултета Универзитета у Београду (http://poincare.matf.bg.ac.rs/~cvetana/kurs-xml/).

елемената не преклапају. Грешке у XML документима која нису добро формирана открива XML парсер који истовремено обавештава аутора о њима. XML парсер је уграђен у различите XML едиторе који се широко користе за креирање XML документа, а њихов задатак је да утврде да ли одређен XML документ задовољава све неопходне формалне критеријуме. Обавештење о грешкама у формирању XML документа појављују се у једном делу едитора дајући аутору приближну информацију о томе у ком делу документа је утврђена грешка, а на аутору је да на основу њих направљене пропусте и поправи. Слика 10 илуструје пример једног добро формираног XML документа који се реферише на спољашњи DTD.

```
<?xml version="1.0" encoding="UTF-8"?>      <!-- XML декларација-->
<!DOCTYPE body SYSTEM "body.dtd">           <!-- DTD декларација-->
<body>                                      <!-- Почетна етикета кореног елемента-->
<div>
<p><seg>Tomas Bernhard MOJE NAGRADE</seg></p>
<p><seg>Nagrada Franc Grilparcer</seg></p>
<p><seg>Povodom dodele nagrade Akademije nauka u Beču, Grilparcerove nagrade,
morao sam sebi da kupim odelo budući da sam iznenada, samo dva sata pre svečanog prijema,
shvatio da na ovoj, nesumnjivo izuzetnoj, ceremoniji, ne mogu da se pojavim u pantalonama i džemperu,
te sam na takozvanom Grabenu zaista odlučio da odem do Kolmarkta i da se obučem svečano, kako i priliči,
te sam pomenutim povodom tragao za radnjom muške garderobe pod nazivom Ser Entoni koju sam znao
budući da sam u dotičnoj već toliko puta kupovao čarape, i ako me sećanje dobro služi,
bilo je petnaest do deset kad sam stupio u salon Ser Entoni,
a dodela Grilparcerove nagrade bila je zakazana za jedanaest, imao sam, dakle, dovoljno vremena.</seg>
<seg> Kad je već... </seg></p>
</div>
</body>                                     <!-- Завршна етикета кореног елемента-->
```

Слика 10. Пример једног дела XML документа за дело Томаса Бернхарда „Моје награде”

Под валидним XML документом подразумева се да је он записан односно креиран у складу са DTD-ем који се користи. DTD преко формалне синтаксе прецизно описује који се елементи могу појављивати у документу, који је њихов тачан редослед, шта је могућ садржај свих дозвољених елемената, шта су њихови атрибути и њихове дозвољене вредности. Валидност XML документа проверава, као и у случају формираности, XML парсер. Валидација није условљена добро формираним XML документом односно добро формирани XML документ не мора да буде и валидан. Да ли ће се грешка у валидацији одређеног XML документа сматрати фаталном или не зависи од апликације која ће га користи. Предност DTD-а је та што на прецизан начин указује који се све елементи и

атрибути користе за анотацију одређеног документа те је немогуће испустити оне који су обавезни или користити оне који нису предвиђени.

Валидни документ реферише на DTD у односу на који се његова валидност проверава. Референца се задаје у јединственој *дефиницији типа документа* која говори шта је коренски елемент документа и URL адреса на којој се DTD може наћи (ако је у питању такозвани спољашњи DTD) односно директно назив DTD датотеке која се смешта у исти каталог у коме се налази и XML документ за који се ради валидација. Такође, постоји и могућност да DTD буде наведен у оквиру самог документа. У том случају, DTD декларација се смешта у пролог XML документа између XML декларације и почетне етикете кореног елемента. Структурна анотација текстова за корпус који је предмет дисертације заједно са DTD-ем који је коришћен за њихову валидацију детаљно су објашњени у поглављу 6 одељак 6.2.4.

3.4.2 Морфосинтаксичка анотација

Морфосинтаксичка анотација, у ширем смислу, подразумева лингвистичку анотацију којим се сваком токену у корпусу придружују информације као што су врста речи (етикетирање врстом речи), канонски облик или лема (лематизација) и вредност морфолошких категорија (род, број, падеж, лице и слично). У ужем смислу, морфосинтаксичком анотацијом се токену придружује само нека од наведених информација.

Етикетирање врстом речи подразумева додељивање морфолошких дескриптора односно етикета (tag) којима се дефинише врста речи (именица, глагол, придев и слично). Уз сваку врсту речи може стајати одређени скуп морфолошких категорија, које прецизније описују дату етикету, и могу бити карактеристичне за више врста речи (на пример, падеж - карактеристично за именице, придеве, заменице) или само за поједине врсте речи (на пример, степен за поређења придева или глаголски облик за глаголе). Приликом етикетирања врстом речи на почетку је потребно дефинисати скуп етикета које ће бити коришћене за анотацију, као и њихово прецизно значење. Величина скупа етикета који се користи за морфосинтаксичку анотацију зависи од природног језика који се обрађује, али

и од намене анотираног корпуса. Такође, за исти природни језик могуће је креирати скупове етикета различите величине што све зависи од предвиђене примене и расположивих алата.

Поред етикетирања врстом речи и доделе морфолошких категорија, сваком издвојеном облику се може доделити и канонски облик речи или лема. Лема представља парадигму лексеме у стандардном речнику која представља скуп облика који имају исту основу, припадају истој врсти речи и имају исто значење. На пример, лема именица је облик у номинативу једнине (ако постоји), док су глаголи у речнику представљени инфинитивом (ако постоји) као лемом.

3.5 Паралелизација

Када је реч о софтверу за паралелизацију текстова Група за језичке технологије определила се за програмске пакете развијене на програмском језику JAVA у оквиру Лоренске лабораторије за информатичка истраживања и примене (Laboratoire Lorrain de Recherche en Informatique et ses Applications - LORIA)[94] у Француској, XAlign и Concordancier. Пакет XAlign обавља аутоматску паралелизацију текстова и креирање битекста у формату заснованом на спецификацији TEI и направљен је за коришћење из командне линије, док се пакет Concordancier, са графичком корисничком сумеђом, користи за претрагу паралелних конкорданци уз помоћ регуларних израза и контролу и корекцију погрешно упарених варијанти јединица превођења која се обавља ручно.

Потреба за добрим окружењем са графичком корисничком сумеђом, која би у себи садржала све потребне компоненте које се користе у разним фазама припреме паралелних текстова за паралелизоване корпусе, мотивисала је Групу за језичке технологије да развије сопствено интегрисано развојно окружење за паралелизацију корпуса (Utvić, Stanković i Obradović 2008, 567). Програмски пакет ACIDE састоји се из два модула, Alignment и TMX. Поступак паралелизације одвија се преко модула Alignment који интегрише програмске пакете XAlign и Concordancier креираних коришћењем програмског

[94] Laboratoire Lorrain de Recherche en Informatique et ses Applications, приступљено 26.3.2019, http://www.loria.fr/loria-news

језика JAVA што омогућава коришћење у свим оперативним системима. Подразумевани улаз јесу добро формиране XML датотеке изворног и циљног језика које су уједно и валидне у односу на прописани DTD. Као излаз добија се нова XML датотека са упареним варијантама јединица превођења.

Кроз други модул, модул TMX, обављају се кораци генерисања датотеке у формату TMX и њено разлагање на датотеке појединачних текстова у формату XML. Модул је добио име по стандарду за складиштење такозваних преводилачких меморија TMX (детаљније у поглављу 2 одељак 2.3). За процес вертикализације добијених TMX формата развијена су два модула: први, омогућава разлагање добијених TMX документа на XML документе за сваки појединачни језик са сачуваним информацијама о јединицама превођења; други, врши конверзију добијених докумената у вертикализовани текст.

Модул TMX омогућава да се на основу битекста у формату XAlign генеришу верзије истог битекста у формату TMX, али и у другим форматима као што су Vanilla и HTML. Произведени TMX документ састоји се из заглавља и тела документа. У заглављу документа смештени су метаподаци који описују паралелизоване текстове док је тело документа састављено од јединица превођења које обухватају две или више семантички еквивалентних јединица превођења. Овај модул омогућава разлагање текстова у формату TMX на појединачне текстове у формату XML, а као резултат добијају се две датотеке: једна садржи анотиране варијанте јединица превођења на изворном, а друга на циљном језику. Процес разлагања обавља се уз помоћ XSL-трансформација[95]. Овако добијене датотеке могу представљати улазне текстове за два једнојезична корпуса. У последњем кораку који подразумева креирање корпуса са морфолошком и структурном анотацијом, индексирање текстова и њихову ефикасну претрагу коришћењем регуларних израза користи се пакет IMS OCWB објашњен у одељку 3.3.1 овог поглавља.

У поглављу 6 одељак 6.2.5 описан је и анализиран поступак паралелизације српских и немачких текстова одабраних за корпус који је предмет дисертације заједно са датотекама које су добијене као крајњи производ разлагања.

[95] Extensible Stylesheet Language (Прошириви језик стилских листова) – омогућава трансформацију XML докумената било у други XML документ или у (X)HTML документ

3.6 Претрага паралелних корпуса у Корпусу савременог српског језика

Претрага паралелних корпуса у Корпусу савременог српског језика врши се коришћењем регуларних израза. У области рачунарства и информатике регуларни изрази се дефинишу као ниска која описује, мења или упарује скуп ниски према одређеним синтаксним правилима. Синтакса и семантика регуларних израза стандардизована је на неки начин појавом стандарда Преносива сумећа оперативног система (Portable Operating System Interface - POSIX) када су регуларни изрази разврстани у две класе: основни и проширени регуларни изрази. Већина упитних језика данас, као део алата за претрагу и обраду корпуса, користи проширене регуларне изразе.

Синтакса POSIX[96] регуларних израза дефинише две групе елемената: обичне карактере или литерале и метакарактере. Обични карактери представљају сами себе, док су метакарактери карактери са специјалним значењем чији је задатак да ближе одреде оно што је кориснику потребно да добије као резултат претраге. Значење метакарактера доста зависи од типа обрасца у оквиру кога се користи, а њихова употреба знатно олакшава претрагу и даје боље резултате.[97]

Синтакса регуларних израза која се користи изгледа овако:

1. . –„џокерски знак”, метакарактер који проналази било који карактер, другим речима замењује било који карактер осим карактера за нови ред.
2. [] – класа карактера, проналази било који карактер који се налази у угластим заградама. На пример, [abc] проналази a, b и c, али не и d.
3. [...-...] – класа карактера са цртицом, проналази било који карактер који је у опсегу или интервалу карактера задат у угластим заградама. На пример, [a-z] проналази било које од малих слова енглеског алфабета.
4. [^] – негативна класа карактера, проналази било који карактер осим оних који се налазе у угластим заградама. На пример, [^abc] проналази све карактере осим a, b и c, јер знак ^ унутар угластих заграда значи негацију.

[96] POSIX, http://pubs.opengroup.org/onlinepubs/000095399/basedefs/xbd_chap09.html
[97] О регуларним изразима више се може погледати на http://www.regular-expressions.info/xml.html, приступљено 26.3.2019.

5. ^ - метакарактер који се може налазити на почетку регуларног израза, изван угластих заграда. Тада се као резултат претраге добијају све ниске које почињу задатим упитом. Регуларни израз ^abc проналази све ниске које почињу са abc, на пример, abcdefg.
6. $ - метакарактер који се пише на крају регуларног израза и означава да израз мора бити пронађен на крају ниске. На пример, регуларни израз abc$ проналази све ниске које се завршавају са abc.
7. * - метакарактер којим се означава Клинијево затворење, проналази ниске у којима се карактер или израз који му претходи може понављати произвољан број пута, а може бити и изостављен. На пример, ab* проналази a, ab, abb и тако редом.
8. + - метакарактер који означава позитивно Клинијево затворење и проналази ниске у којима се карактер или израз који му претходи појављује једном или више пута. На пример, ab+ ће пронаћи ab и abb, али неће само a као што је то био у случају ab*.
9. ? – метакарактер којим се означава опционо појављивање карактера или израза који му претходе односно проналази претходни карактер или израз једном или ниједном. На пример, ab? ће пронаћи a и ab.
10. {n}, {n,} и {n,m} - квантификатори који проналазе претходни карактер или израз тачно n пута, бар n пута односно најмање n, а највише m пута (n и m представљају позитивне целе бројеве односно бројеве од 0 до 9). На пример, a{2} проналази само aa, a{2,} проналази aa, aaa, aaaa, и тако редом, док a{2,3} проналази aa и aaa.
11. ( ) – проналази било коју секвенцу унутар заграда. На пример, ако га комбинујемо са знаком + имаћемо: (ab)+ проналази ab, abab, ababab и тако даље. Заграде служе за груписање знакова, тако да квантификатори *, +,? и { } на њих гледају као на једну целину.
12. | - логички ИЛИ оператор (алтернација) спаја два регуларна израза и проналази један од њих. На пример, P(1|2) проналази P1 или P2.

13. \ - обрнута коса црта третира специјалне карактере буквално, односно као обичне карактере. На пример, знак +, који означава „пронађи претходни карактер” једном или више пута, записан као \+ проналази сам карактер +. Сви карактери * ? + . ( ) { } [ ] ^ \ $ су синтаксни карактери регуларних израза, а постају буквални ако се испред њих стави знак „\”. Такође, појединим карактерима даје специјално значење те \n и \t представљају знак за нови ред и табулатор.

Све наведене операције у регуларним изразима имају свој приоритет како би се регуларни изрази једнозначно тумачили. Операције су наведене према опадајућем приоритету, [...-...], [^], \, (), *, +, ?, { }, ^, $, |, и представљају метакарактере из којих се као посебна подгрупа могу издвојити квантификатори: *, +, ?, {n}, {n,}, {n,m}.

Након примене регуларних израза приликом претраге корпуса као резултат добијају се конкорданце које представљају листу појављивања токена који су задовољили упит за претрагу, са позицијама у тексту и цитатима или изводима из контекста. Ако је корпус анотиран, уз сваку лексему која је резултат претраге стоје и додатне информације у виду врсте речи или леме (канонски облик речи). У том случају се и упити за претрагу могу постављати коришћењем тих додатних информација. Неки примери овакве претраге паралелних корпуса дати су у претходном поглављу одељак 2.5.3, док су поступак претраге корпуса који је предмет ове дисертације и анализа добијених резултата детаљније је приказани у поглављу 6 одељак 6.3.3.

4 Стварање предуслова за ефикасно проналажење информација у паралелним корпусима

Интензиван развој технологије од друге половине 20. века, а посебно последњих деценија, условио је појаву новог начина похрањивања информација и појаву електронских публикација. Електронске публикације, објављене данас углавном на интернету, доступне су корисницима различитог узраста и интересовања 24 сата са различитих удаљених места те су постале незаобилазни ресурс у свим доменима рада. Такође, постале су предуслов и за појаву дигиталних библиотека као специјалних колекција „дигиталних докумената који се у виду различитих дигиталних података (текст, слика, звук, видео, анимација) или њихових комбинација (мултимедија) похрањују на мрежи, описују различитим метаподацима и повезују са другим информационим системима" (Тртовац 2017, 23). Да бисмо могли да претражујемо и проналазимо информације у различитим базама података неопходно је да документи буду описани на одговарајући начин. Један начин описа односи се на одређивање формалних особина објекта и доделу метаподатака у складу са усвојеним стандардима и каталошким правилима. Други начин описа односи се на индексирање садржаја документа које може бити ручно и аутоматско применом одговарајућих технологија. Паралелни корпуси, као један вид дигиталне колекције, такође, могу постати део комплексне дигиталне библиотеке која захтева и адекватан опис уз помоћ релевантних метаподатака који корисницима омогућавају да их претражују на одређени начин и проналазе информације у њима.

Корпус паралелних текстова који се у дисертацији анализира смештен је у окружење које нуди систем претраге преко метаподатака и претраге садржаја текста формулисањем упита у форми кључних речи. Из тог разлога, колекције текстова морају бити описане одговарајућим метаподацима, а садржај текста индексиран на одређени начин или подржан језичким ресурсима.

У овом поглављу је дат преглед настанка појма „метаподатак", његове улоге и значаја у опису дигиталних објеката и колекција, као и формата и стандарда који се на

међународном нивоу користе за њихову израду, а који се примењују и у Србији. Поред система израде метаподатака, на крају поглавља дат је кратак преглед технологија за оптичко препознавања карактера и препознавање именованих ентитета.

4.1 Метаподаци – појам, дефиниције, врсте, значај

Термин „метаподатак" 1969. године је сковао Џек Е. Мејерс (Jack E. Meyers) са циљем да на најбољи могући начин опише све производе који су били у вези са мета моделом (Meta Model) који је развио, али и да означи компанију која их производи и продаје (Greenberg 2005, 19). Овако дефинисани термин се у литератури први пут појавио године 1973, и то у Мејерсовом каталогу производа, и врло брзо су га усвојили припадници различитих научних заједница, а међу њима и рачунарска и библиотечко-информациона, као термин који може релевантно да опише средства за приказивање карактеристика и својства објеката и података.

Данас се метаподаци у најширем смислу дефинишу као „подаци о подацима" односно „информације о информацијама". Према Карен Којл (Caren Coyle), метаподаци су „конструисане информације, што значи да их је створио човек и да нису настали у природи... Метаподатке су развили људи за неку потребу или функцију... Они нису свет, то је начин на који видимо свет у неком тренутку за неке специфичне потребе" (Coyle 2005, 160-163). Према Гејл Хоџ, чија је дефиниција и најшире прихваћена, метаподаци су структуриране информације које описују, објашњавају, лоцирају или на други начин чине лакшим проналажење, коришћење или управљање неким извором информација (Hodge 2004, 157). У савременом дигиталном свету метаподаци се често посматрају и као структуриране информације које пружају податке о начину креирања, интелектуалном садржају или контексту појединачних ресурса или колекција (Gill et al. 2008, 1-19), док творац глобалне светске мреже Тим Бернерс-Ли (Tim Berners-Lee) дефинише метаподатак као машински читљиву информацију о ресурсима на вебу или другим ресурсима са јасно дефинисаном семантиком и структуром (Berners-Lee 2016).

Термин „метаподатак" је најшире прихваћен у области библиотекарства као средство погодно за опис различитих врста садржаја. Све библиотеке имају метаподатке,

или бар неки њихов облик који се користи за формалну структуру описа, било аналогних, било дигиталних докумената. Вранеш и Марковић сматрају да је метаподатак „машински читљив податак који се односи на друге податке, на информације о изворима и њиховим ауторима, на библиографске и каталошке информације" (Вранеш и Марковић 2008, 205), а и Крстев тврди да се у пракси „метаподаци најчешће користе за опис дигиталних објеката па имају сличну улогу као и записи библиотечких каталога" (Крстев 2002, 9). На овај начин је јасно дефинисано да је метаподатак каталошки податак о физичком или дигиталном објекту и да је у том смислу близак каталогизацији у библиотекарству.

Традиционална библиотечка каталогизација представља један облик метаподатака. Каталошки листићи, библиографски опис, стандардни бројеви (ISBN[98], ISSN[99] и слично), стручна и предметна класификација, анотиране библиографије, стандардно цитирање као и предметни индекси представљају, у ствари, различите врсте метаподатака који су се у библиотекарству користили много пре коришћења самог термина. Широко прихваћени формати за каталошки опис UNIMARC[100] и MARC21[101], са правилима која прате каталогизацију (ISBD[102], AACR2[103]) представљају стандарде који су временом постали основа за развој других стандарда и модела података за израду метаподатака за опис разноврсних текстуалних и нетекстуалних објеката (архивски материјал, визуелни материјал, скупови података из друштвених наука) у дигиталном облику. На основу овога се може закључити да иако је каталогизација нешто што припада физичком објекту, а метаподатак се превасходно везује за електронске изворе, метаподаци који описују дигиталне објекте се у суштини не разликују од метаподатака

[98] Међународни стандардни број за књигу (International Standard Book Number - ISBN), доступно на: http://www.isbn.org/

[99] Међународни стандардни број за серијске публикације (International Standard Serial Number - ISSN), доступно на: https://www.issn.org/

[100] Универзални машински читљив каталог (Universal Machine Readable Catalogue - UNIMARC). Приручник је доступан на: http://archive.ifla.org/VI/3/p1996-1/sec-uni.htm

[101] Машински читљив каталог (Machine Readable Catalogue - MARC). Приручник је доступан на на: https://www.loc.gov/marc/bibliographic/

[102] Међународни стандард за библиографски опис (International Standard Bibliographic Description - ISBN). Приручник је доступан на: http://www.ifla.org/files/assets/cataloguing/isbd/isbd-cons_2007-en.pdf

[103]Англо-америчка каталошка правила (Anglo-American Cataloguing Rules - AACR2). Приручник је доступан на: http://www.aacr2.org

који описују аналогне облике (Greenberg 2003, 1877). Разлика постоји само у начину и правилима описа који су прилагођени потребама и захтевима савременог друштва.

Дигитална револуција променила је многе аспекте људског живота, од приватног до професионалног, утичући, између осталог, и на то како се представљају и добијају информације. Дигитализација која представља основу концепта „друштво знања" коме теже све земље у 21. веку има два основна циља: омогућити ширем кругу корисника приступ различитим материјалима и сачувати их и заштитити од уништавања. Стварање дигиталне форме културног и научног наслеђа и њихово постављање у отворени приступ представља богат извор материјала и садржаја за јавно коришћење и разнолика научна и стручна истраживања. „Откривање, идентификација и претраживање дигиталних извора чине се једноставним сваком кориснику, који ће у томе убеђењу и остати уколико су они добро припремљени за дигитализацију, уредно и квалитетно пренесени у електронски облик и опремљени одговарајућим индексима и конзистентним метаподацима који омогућавају да се њихов садржај вишеструко претражује" (Вранеш 2014, 9).

Метаподаци се креирају из различитих разлога и за различите потребе, али се најчешће користе као алат који омогућава бољу видљивост и доступност дигиталног објекта преко веба (Shukair et al. 2013, 10). Поред основне улоге да омогуће и олакшају проналажење извора избором различитих критеријума, метаподаци имају и задатак да организују електронске изворе на најбољи могући начин, олакшају размену података између различитих система (репозиторијума, база података и слично), олакшају дигиталну идентификацију преко трајних идентификатора као што су URI (Unifrom Resource Identifier) (детаљније у поглављу 5 одељак 5.1.2) адресе или DOI (Digital Object Identifier)[104] бројеви и омогуће одговарајуће архивирање и заштиту дигиталних информација. Они, такође, контролишу начин на који су подаци приказани и могу да укажу на везе између различитих информационих јединица.

Постоји неколико врста метаподатака који омогућавају детаљан опис свих особина једног дигиталног извора. У зависности од врсте података који се уносе, они се могу

[104] Digital Object Identifier, https://www.doi.org/

поделити у неколико група. Општеприхваћена подела је на три основна типа (Hodge 2004, 157-158):

1. *Описни (библиографски) метаподаци.* Ова врста метаподатака описује извор, укључујући најчешће следеће елементе: наслов, аутор, кључне речи, годину издавања, садржај, апстракт и слично. Овакав вид описа неког извора постојао је још и у лисним каталозима библиотека деценијама уназад, а касније и у електронским каталозима али под другим називом – библиографски опис. За овај опис коришћени су одређени формати, као што су различите варијанте MARC формата који прате прихваћени стандард за библиографски опис ISBD, али и концептуални модели података, као што је FRBR[105], које су прописале водеће организације из ове области ради унифицираног описа библиотечког материјала. На њима су изграђени бројни други стандарди и модели података који се данас широко користе за креирање метаподатака.
2. *Структурални метаподаци*. Ова врста метаподатака даје информацију о томе како је објекат структурисан односно организован. Томови, поглавља, стране једне књиге описују се овим метаподацима, при чему се добија одређен хијерархијски низ који указује на структуру објекта и омогућава његов адекватан приказ, као и кретање читаоца кроз њега.
3. *Административни метаподаци*. Ова врста метаподатака даје информације о томе како објекат треба чувати, који су услови коришћења, како су регулисана ауторска права, порекло и власништво објекта. Они се могу даље поделити у две мање групе:
    a. *метаподаци за управљање правима*: ова група укључује податке о праву својине, интелектуалном праву, праву копирајта (разлика између ауторског права и копирајта објашњена је у 3. поглављу одељак 3.1), као и праву приступа и коришћења описиваног објекта,
    b. *метаподаци за чување и заштиту*: дају информације о условима и начинима заштите и трајног похрањивања.

[105] Функционални захтеви за библиографске записе (Functional Requirements for Bibliographic Records - FRBR). Приручник је доступан на: https://www.ifla.org/publications/functional-requirements-for-bibliographic-records

4.1.1 Стандарди за израду метаподатака и формати приказа

Како би се у контексту дигиталних репозиторијума и пројеката дигитализације омогућила јединствена размена метаподатака јавила се потреба за дефинисањем стандарда за опис кроз различите формате са прецизно дефинисаном синтаксом за опис објеката у дигиталним библиотекама и репозиторијумима (Klarin 2007, 44). Запис метаподатака дигиталних објеката, поред стандардног библиографског описа, садржи и податке о датуму креирања записа, датуму последње измене, име креатора записа, идентификациони број записа и слично, као и детаљан опис начина приступа дигиталном објекту.

Метаподаци омогућавају аутоматско проналажење и управљање информацијама јер представљају скуп елемената са атрибутима који се користе за опис одређеног дигиталног објекта. Скуп елемента дефинисаних за опис одређене врсте информационог извора назива се схема. За сваку схему прописан је одређен број елемената, од којих сваки има своје значење односно дефиницију (познато још као и семантика схеме) и вредност (врло често дефинисана из прописаног скупа вредности) (Sarić, Magdić i Essert 2011, 137). „Схеме метаподатака у начелу одређују називе елемената и њихову семантику. По избору, оне могу да одреде правила садржаја за начин на који садржај мора бити формулисан (на пример, како идентификовати главни наслов) и/или правила представљања за начин на који садржај мора бити представљен (на пример, правила о писању великих слова). Такође, могу постојати правила синтаксе за начин на који би елементи и њихов садржај могли бити кодирани. Синтаксно независне схеме метаподатака су оне схеме које немају унапред задата правила синтаксе” (Hodge 2004, 158)

Структурирана група елемената чини јединствени запис који описује одређени дигитални објекат. Записи метаподатака могу да се креирају одвојено (такозвани експлицитни метаподаци), а затим се поставља веза ка дигиталном објекту који се описује. У пракси, као веза најчешће се користи Униформни локатор ресурса (Uniform Resource Locator - URL) који представља адресу објекта који се описује, или URL адреса

репозиторијума односно дигиталне библиотеке у коме је објекат складиштен. Одвојено чување метаподатака олакшава њихово управљање, претраживање и проналажење.

Други начин је да се креирају комплексне схеме које ће омогућити опис објеката на различитом нивоу у оквиру једног записа (такозвани имплицитни метаподаци). То значи да су метаподаци уграђени у, на пример, HTML документе или заглавља других докумената, чиме се обезбеђује да се они неће изгубити, избегавају се могући проблеми који могу настати приликом повезивања објеката и метаподатака и омогућава се истовремено ажурирање објеката и метаподатака. Трећи начин чувања метаподатака је у оквиру одређене базе података. Записи се могу директно креирати у њој, или се могу преузети из других извора (на пример, са веб-страна). Ако су метаподаци уграђени у базе података постављају се везе ка одговарајућим дигиталним објектима.

Једна стандардизована схема за израду метаподатака подразумева да је она једноставна за употребу како за креатора тако и за корисника, да може де се користити у различитим репозиторијумима, да може да се мапира са другим схемама и да институција која је одржава може редовно да је ажурира чиме се обезбеђује трајност записа. Схеме за израду метаподатака направљене су тако да корисници могу брзо и лако да доделе неопходне метаподатке, док од њихове вештине зависи колико ће детаљно један дигитални објекат бити описан, а самим тим и какав ће бити квалитет претраживања и проналажења информација о конкретном дигиталном објекту (Тртовац 2016, 58).

За већину схема за израду метаподатака данас користи се XML синтакса. Са развојем семантичког веба и иницијативе „Отворени повезани подаци” (Linked (Open) Data) (детаљније у поглављу 5 одељак 5.2) јавила се и потреба за повезивањем података на нивоу њиховог значења како би се омогућила јединствена претрага која групише све сродне јединице повезане са релевантним изворима на вебу. За потребе овако комплексног умрежавања података развијен је Оквир за описивање ресурса (Resource Description Framework - RDF). О самом RDF-у детаљније ће бити речи у поглављу 5 одељак 5.1.2, а структура и начин умрежавања података детаљније су описани поглављу 6 одељак 6.5.

Приликом одабира схеме за израду метаподатака треба имати у виду њену функционалност, погодност за опис самог садржаја, али и једноставност за употребу како у поступку креирања метаподатака тако и у процесу даље употребе. Захваљујући искуству у раду са системима за каталогизацију библиотекари су у стању да врло брзо савладају начин уноса метаподатака и особине формата за унос. Међутим, јако је важно да и корисници, односно креатори дигиталних објеката, буду у стању да сами уносе метаподатке и постављају дигиталне објекте, јер поједини пројекти израде дигиталних библиотека укључују волонтерско ангажовање корисника.

У наставку овог одељка представљамо неке од међународних стандарда за израду метаподатака који користе XML синтаксу (Даблинско језгро, Стандард за кодирање и преношење метаподатака, Шема за опис метаподатака објеката и заглавље Иницијативе за кодирање текста). На крају поглавља на једном примеру приказана је структура метаподатака у сваком од наведених стандарда понаособ и њихов упоредни приказ. Структура метаподатака у паралелном корпусу који је предмет дисертације описана у поглављу 6 одељак 6.3.1.

Даблинско језгро

Даблинско језгро (Dublin Core – DC) је стандард за израду метаподатка који се састоји од сета елемената за опис широког спектра извора на мрежи. Назив је добио по месту Даблин у Охају где је 1995. године одржана радионица OCLC/NCSA Metadata Workshop у организацији Онлајн рачунарског библиотечког центра (Online Computer Library Center – OCLC)[106] и Националног центра за суперрачунарске апликације (National Center for Supercomputing Application – NCSA).[107] Реч *језгро* у називу указује на то да се у његовој основи налази одређени број елемената за опис метаподатака који чине језгро односно основу, али се то језгро односно основа може проширити.

Даблинско језгро је првобитно створен како би аутори најразличитијих извора на вебу самостално могли да их опишу. Тако су постављени и следећи циљеви:

[106] Online Computer Library Center, http://www.oclc.org/uk/en/default.htm
[107] National Center for Supercomputing Application, http://www.ncsa.illinois.edu/

1. *Једноставност у креирању и одржавању:* овим се различитим ауторима дигиталних извора информација омогућава да их опишу на једноставан начин.
2. *Универзално разумљива семантика и стандардизована синтакса:* постиже се применом формата XML или, у скорије време, RDF.
3. *Међународно учешће и локализација:* стандард је изворно написан на енглеском, али је до данас локализован и на друге језике као што су фински, норвешки, јапански, француски, португалски, немачки, грчки, шпански и други.
4. *Прилагодљивост:* предвиђено је да се Даблинско језгро може прилагодити разним новим дигиталним изворима информација које је потребно описати, проширити се зависно од локалних потреба и користити у новим ситуацијама и прилагодити им се.

Временом, организације као што су библиотеке, музеји и архиви почеле су да користе Даблинско језгро како би своје садржаје у дигиталном облику описале на једноставнији начин у односу на потпуне MARC каталошке записе (Wiebel 1997, 9-11). Међународна и мултидисциплинарна група професионалаца, запослених у овим типовима организација, су заједно са стручњацима из области рачунарства, кодирања текста и сродних дисциплина, установили семантику Даблинског језгра. Даљи развој Даблинског језгра преузела је организација која се бави израдом стандарда и речника за опис информација чији је фокус опис семантике стандарда за метаподатке, Иницијатива за израду метаподатка у Даблинском језгру (Dublin Core Metadata Initiative – DCMI)[108].

Основни скуп елемената метаподатака Даблинског језгра (Dublin Core Metadata Element Set – DCMES), састоји се од петнаест елемената који се могу поделити у три групе:

1. *Елементи који описују садржај ресурса*: наслов (dc:title), предмет (dc:subject), опис (dc:description), извор (dc:source), језик (dc:language), однос (dc:relation), покривеност (dc.coverage).
2. *Елементи који пружају податке о интелектуалној својини ресурса*: креатор (dc:creator), издавач (dc:publisher), сарадник (dc:contributor), права (dc:rights).

[108] Dublin Core Metadata Initiative, http://dublincore.org

3. *Елементи који описују физичке карактеристике ресурса*: датум (dc:date), врста (dc:type), формат (dc:format), идентификатор (dc.identifier).

Сви елементи који карактеришу Даблинско језгро лако се идентификују, необавезни су и поновљиви, не прописује се посебан редослед њиховог навођења, препоручује се употреба контролисаних речника и шифрарника за садржај елемената, а правила за унос елемената дефинише систем који их користи (библиотека, музеј, архив). Ово је управо и негативна страна овог стандарда. Чињеница да садржај поља није уједначен отежава размену података између система, а његова превелика општост не одговара захтевима уско специјализованих институција при опису специфичних дигиталних ресурса.

Иако је Даблинско језгро замишљен као једноставан и концизан стандард за опис дигиталних објеката на мрежи, у пракси је почео да се користи и за опис других врста садржаја, али и за сложеније примене те је временом шема проширена и установљен је квалификовани опис Даблинског језгра. Квалификовани опис подразумева коришћење додатних елемената и квалификатора као што су: публика којој је објекат намењен (audience), порекло (provenance), власници права (rights holder), метод примене (instructional method), метод прираста (accrual method), периодичност прираста (accrual periodicity) и политика прираста (accrual policy).

Многи стандарди за израду метаподатака које се данас широко користе засновани су управо на Даблинском језгру. Различите базе података, најчешће електронски каталози библиотека и дигитални репозиторијуми, нуде могућност аутоматског извоза метаподатака у стандарду Даблинско језгро који је даље погодан за размену са другим системима и апликацијама, али и за мапирање са другим схемама за метаподатке.

Основних петнаест елемената део су ширег вокабуларног система и техничких спецификација које развија DCMI, а који укључују класе објеката, вокабулар шема за кодирање података и шеме за кодирање синтаксе. Термини из вокабулара DCMI користе се у комбинацији са другим одговарајућим вокабуларима што је најбоље описано кроз

DCMI апстрактни модел (DCMI Abstract Model – DCAM).[109] DCAM представља RDF граф модел (детаљније описано у поглављу 5, одељак 5.1.2) који дефинише компоненте које се користе у Даблинском језгру приликом креирања метаподатака и описује како су оне повезане да чине јединствену информациону структуру. Овакав модел омогућава боље разумевање описа који се кодира и олакшава мапирање и преношење података између информационих система. Структура DCAM модела може да се примени на било који стандард за израду метаподатака. DCAM се састоји из три дела: DCMI модел ресурса (DCMI resource model) који дефинише компоненте за опис ресурса, DCMI модел скупова описа (DCMI description set model) који дефинише компоненте за опис структуре записа једног ресурса и DCMI модел вокабулара (DCMI vocabulary model) који дефинише просторе имена и синтаксу за опис ресурса. Пример записа у Даблинском језгру за дело „Моје награде“ аутора Томаса Бернхарда дат је као Прилог 3 - Пример записа у формату Даблинско језгро / XML синтакса.

Стандард за кодирање и пренос метаподатака

Рад на Стандарду за кодирање и пренос метаподатака (Metadata Encoding and Transmission Standard – METS) започет је касних деведесетих година 20. века реализацијом пројекта *Making of America II (MoA II)* у координацији Беркли универзитета (University of California, Berkeley) и Федерације дигиталних библиотека (Digital Library Federation – DLF). Пројекат је покренут са циљем да се креира стандард за кодирање метаподатака који описује дигиталне објекте (Hurley et al. 1999, 3). У оквиру пројекта је развијена дефиниција типа документа (DTD) која спецификује елементе за опис података и начин кодирања ограниченог скупа објеката (објекти који садрже текст и сликовне датотеке). Даљи развој шеме разматран је на „DTD радионици MoA II” (Making of America II DTD Workshop) одржаној фебруара 2001. године на којој је постигнут договор да се постојећи DTD преведе у XML Scheme у коју је имплементиран вокабулар за опис разноврсних метаподатака. Тако је настао стандард METS који је формулисао Џером Макдона (Jerom McDonough) са Њујоршког универзитета (Cundiff 2004, 52-53). Даљи рад

[109]Цео модел, заједно са свим његовим деловима, детаљно је описан на веб страни Иницијативе за израду метаподатака у Даблинском језгру (DCMI). Доступно на: http://dublincore.org/documents/abstract-model/

на стандарду, изради веб-странице и званичне документације, као и одржавање и ажурирање саме шеме, наставила је Конгресна библиотека. Године 2004. шему је регистровала Национална организација за информатичке стандарде (National Information Standards Organization – NISO).[110]

Стандард METS је креиран како би се олакшало управљање, чување, коришћење и размена метаподатака о дигиталним објектима између репозиторијума, али и како би лакше могле да се опишу и организују компоненте сложених дигиталних објеката. Стандард омогућава кодирање дигиталног библиотечког материјала пружајући механизам за повезивање различитих делова садржаја, као и садржаја и метаподатака који чине један дигитални објекат (Digital Library Federation 2016). Структура шеме која је у основи стандарда је прилично флексибилна и једноставна, састављена од модула који садрже различите елементе за опис различитих врста метаподатака неопходних за опис дигиталних објеката (Guenther and McCallum 2003, 14). У оквиру овог стандарда дефинисана су три основна типа метаподатака који описују дигитални објекат: описни, административни и структурални. Наведени метаподаци су интегрисани у једну датотеку, у оквиру које је сваки од три типа метаподатака описан у одвојеном одељку, а повезани су преко интерних идентификатора (Gartner 2002).

Један METS запис састоји се од седам могућих одељака који описују различите делове дигиталног објекта од којих је један обавезан – *Структурна мапа (Structural Map)* (Cantra 2005, 240-250) (Gartner 2002):

1. *METS заглавље (METS Header - metsHdr)*: садржи информације о самом METS документу као што су креатор, сарадник и слично.
2. *Описни метаподаци (Descriptive Metadata Section- dmdSec)*: садржи описне метаподатке, који могу бити преузети из већ постојећих записа (нпр. MARC записа), независно се креирати, или представљати обе врсте.
3. *Административни метаподаци (Administrative Metadata Section - amdSec)*: садржи административне метаподатке као што су: како је документ настао и у

[110] National Information Standards Organization, http://www.niso.org/home/

ком дигиталном репозиторијуму се налази, ко полаже ауторска права, из ког оригиналног изворног објекта је настао, које је порекло објекта и многе друге.

4. *Група датотека (File Group Section - fileSec)*: садржи листу свих датотека чији садржај у целини представља дигитални документ који се описује (на пример, ако се описује књига која није у PDF формату већ је свака страна књиге посебна JPEG датотека свака JPEG датотека наводи се посебно у логичном редоследу са свим својим техничким карактеристикама (величина датотеке, идентификациони број у дигиталном репозиторијуму, линк на дигитални репозиторијум и тако даље).
5. *Структурна мапа (Structural Map - structMap)*: представља структуру дигиталног објекта који се описује. За сваку наведену датотеку појединачно се раде описни и административни метаподаци у одговарајућим сегментима METS записа при чему сваки сегмент добија своју идентификациону ознаку приликом креирања. У сегменту *Структурна мапа* се управо преко ових идентификационих ознака повезују одговарајући сегменти. Ово доста олакшава проналажене тражених метаподатака о одређеној датотеци посебно када је у питању велики запис који је већ сам по себи тежак за сналажење.
6. *Повезиване структурне мапе (Structural Map Linking Section - structLink)*: овај сегмент омогућава де се преко линкова повежу сегменти METS записа на основу хијерархијске структуре дате у сегменту Структурална мапа.
7. *Понашање (Behaviour Section - behaviourSec)*: садржи информације о томе како би метаподаци о дигиталном објекту који се описује требало да буду технички обликовани да би били прилагођени корисницима. Овде могу бити укључене информације као што су: подаци о специфичним софтверским пакетима који се користе приликом обраде метаподатака или одређени параметри који се користе за приказивање датотека метаподатака.

Пример METS записа за дело „Моје награде“ аутора Томаса Бернхарда дат је као Прилог 4 - Пример записа у формату METS / XML синтакса.

Схема за опис метаподатака објеката

Са израдом METS стандарда проблем стварања стандардизованог оквира за метаподатке у дигиталним библиотекама делимично је превазиђен. Овај формат је по улози у свету дигиталних библиотека еквивалентан са MARC форматом у свету традиционалне библиотечке каталогизације. Међутим, METS стандард не прописује форму садржаја што омогућава тек делимичну размену метаподатака између њега и других стандарда (Library of Congress 2016). Канцеларија Конгресне библиотеке за развој мреже и MARC стандарде (Library of Congress' Network Development and MARC Standards Office)[111] је 2002. године развила Шему за опис метаподатака објеката (*Metadata Object Description Schema* – MODS) са циљем да допуни друге стандарде за метаподатке и буде алтернатива између оних једноставних који садрже мало елемената и не тако комплексну структуру као што је Даблинско језгро, и веома детаљних са јако пуно елемената уређених у сложену структуру као што је MARC21.

Шема MODS је у великој мери компатибилна са MARC записима јер је преузела семантику еквивалентних елемената података из MARC21 за библиографске податке, односно представља његов подскуп, док је у основи формат XML са кључним MARC21 елементима, који су груписани у логичке компоненте (McCallum 2004, 83). Уместо троцифрених ознака и кодова за поља и потпоља који постоје у формату који се традиционално користи MARC библиографске податке, MODS шема, као и већина шема за израду метаподатака заснованих на формату XML, садржи језичке ознаке за њихово кодирање. Стога је MODS нашао широку примену у библиотекама, користећи се превасходно у следећим случајевима:

1. као спецификовани формат за претрагу и проналажење преко URL адреса;
2. као проширење METS шеме;
3. за представљање метаподатака означених за преузимање (harvesting);
4. за оригинални опис ресурса коришћењем XML синтаксе;
5. за представљање једноставног MARC записа у XML формату;

[111] Library of Congress' Network Development and MARC Standards Office, https://www.loc.gov/marc/ndmso.html

6. за унос метаподатака у XML-у који могу бити упаковани заједно са електронским изворима.

Шема се састоји од 20 основних елемената (titleInfo, name, typeOfResource, genre, originInfo, language, physicalDescription, abstract, tableOfContent, targetAudience, note, subject, classification, relateditem, identifier, location, accessCondition, part, extenstion, recordInfo) са припадајућим атрибутима, а може се по потреби проширити додатним елементима.[112] Такође, MODS шема нуди и могућност употребе јединствених идентификатора на нивоу елемената, чиме је омогућено њихово повезивање.

Иако MODS шема поседује неоспорне предности у односу на друге шеме метаподатака важно је напоменути да пошто су заступљени и елементи којих нема у формату MARC21 приликом мапирања може доћи до губитка података. Такође, следећи принцип Даблинског језгра, редослед елемената у шеми не подразумева и редослед приказа података, ниједан елемент у шеми није обавезан, а сви основни елементи су поновљиви.

Ради олакшавања преузимања и повезивања метаподатака представљених различитим стандардима, припремљен је MODS „Lite", који представља скраћену верзију MODS шеме састављену од елемената који могу да се мапирају са 15 основних елемената из Даблинског језгра, а развијен је и Стандард за опис метаподатака у нормативним записима (Metadata Authority Description Standard – MADS)[113] како би се омогућила нормативна контрола појединих ентитета унетих у MODS записе, а која је у складу са нормативном контролом у записима у MARC21. Пример MODS записа за дело „Моје награде" аутора Томаса Бернхарда дат је као Прилог 5 - Пример записа у формату MODS / XML синтакса.

Иницијатива за кодирање текста и TEI заглавље

Иницијатива за кодирање текста (Text Encoding Initiative – TEI) је међународни пројекат започет 1987. са циљем развијања, одржавања и објављивања хардверски и

[112] Наведених 20 основних елемената шеме MODS описани су детаљно у приручнику који је доступан на http://www.loc.gov/standards/mods/mods-outline-3-6.htm
[113] Metadata Authority Description Standard, http://www.loc.gov/standards/mads/

софтверски независних правила за обележавање електронских текстова (TEI: History 2016). Као резултат рада на пројекту објављене су *Смернице за кодирање и размену електронских текстова* (Guidelines for Electronic Text Encoding and Interchange) које су временом изашле у пет верзија (TEI Guidlines 2016). Први предлог *Смерница* (TEI P1) објављен је 1990, док је последњи предлог (TEI P5) објављен 2007. године. Године 1999/2000. формиран је и посебан непрофитни конзорцијум (TEI Consortium)[114] са циљем да промовише, одржава и развија TEI стандард.

За основу TEI стандарда је на почетку одабран SGML стандард, са утврђеним етикетама и правилима који описују структуру и елементе документа. Са развојем формата XML новије верзије *Смерница* засноване су на тој спецификацији.

Основни ентитети које се појављују у свим документима могу да се поделе у неколико категорија (Bernard and Ide 1997, 626):

1. скуп карактера који ће бити коришћени у документу,
2. ознаке за дефинисање TEI заглавља,
3. ознаке за дефинисање одређених целина текста (параграфа, навода, листи напомена, библиографија) и
4. једноставне структуре текста (карактеристике фонта, могући интегрисани делови).

Како већину корисника занима само један или само неколико типова електронског текста, дефиниције TEI-елемената су организоване модуларно са хијерархијским преузимањем елемената и атрибута, што омогућава корисницима да одаберу подскуп језика потребан за кодирање одређеног документа, дела документа, односно скупа докумената. Након одабира модула за кодирање садржаја, корисник доступне шаблоне прилагођава додавањем или брисањем додатних модула, елемената, атрибута и тако даље. Поред обавезног модула **tei**, којим се описује инфраструктура за кодирање схеме која је описана TEI Смерницама, шаблони могу да садрже и следеће модуле, од којих су прва три обавезна (TEI Consortium 2016):

[114] TEI Consortium, http://www.tei-c.org/index.xml

1. *Core*. Садржи елементе који су заједнички свим TEI-документима и који могу да се користе за опис произвољног типа текста;
2. *Header*. Садржи елементе којима се спецификује TEI заглавље (описано у наставку овог одељка);
3. *Textstructure***.** Садржи елементе којима се описује подразумевана структура која је заступљена код већине врста текстова;
4. *Analysis***.** Садржи елементе за представљање семантичких или синтаксичких интерпретација којима се може означити цео текст или неке његове делове;
5. *Certainty*. Садржи елементе којима се указује на проблеме или недоумице у погледу структурног означавања текста;
6. *Corpus*. Садржи елементе за опис језичког корпуса као структуре која се кодира TEI стандардом;
7. *Dictionaries*. Садржи елементе за кодирање свих врста лексичких ресурса као што су једнојезични и вишејезични речници, лексикони и слично;
8. *Drama*. Садржи елементе за кодирање драмских текстова, текстова играних филмова и серија, текстова радио емисија и сличне садржаје;
9. *Figures*. Садржи елементе за кодирање слика које су саставни део текста;
10. *Gaiji*. Садржи елементе за представљање напомена и језичких идентификација и репрезентацију карактера у описиваном TEI документу;
11. *Iso-fs*. Модул који омогућава опис карактеристика структуре текста која се кодира, на пример, „word structure“ – „структура речи“, „agreement structure“ – „структура сагласности“;
12. *Linking*. Садржи елементе за повезивања различитих делова текста који нису обавезно у линеарном или хијерархијском односу;
13. *Msdescription*. Садржи елементе за детаљан опис рукописног изворног текста. Модул је развијен како би каталогизатори и научници који раде са старим рукописима могли да опишу и кодирају специфичности овакве грађе које се не могу описати другим модулима овог стандарда;

14. *Namesdates.* Садржи елементе за кодирање и детаљнији опис имена и других фраза које описују људе, места или организације и датум и време;
15. *Nets.* Садржи елементе за графичко приказивање веза између делова текста који се кодира овим стандардом како би читаоци могли лакше да их визуелно сагледају и разумеју;
16. *Spoken*. Садржи елементе за транскрипцију говорног материјала;
17. *Tagdocs*. Модул који може да се користи за документовање XML елеменат и класа елемената који чине једну схему за означавања, посебно ону која је описана TEI Смерницама. Модул се такође може користити и за аутоматско генерисање схема или DTD-а на који се те схеме позивају;
18. *Textcrit*. Садржи елементе за кодирање критичких издања или критичких напомена унетих у текст;
19. *Transcr*. Модул који се користи за репрезентацију изворног ресурса као што су рукописи или други писани материјали на основу којих је урађен дигитални материјал који се кодира TEI стандардом;
20. *Verse***.** Садржи елементе за кодирање текста који је у облику стиха.

Са једне стране TEI смернице се користе за стварање дигиталних библиотека које омогућавају приступ великој количини текстуалног материјала при чему се акценат ставља на редак и осетљив материјал који једини дигитално може да постане широко доступан, док се са друге стране користе и у специјализованим истраживачким пројектима за представљање мањих колекција текстова које покривају ограничену тему (Ерјавец 2010, 5). Цео TEI стандард заснован је на XML синтакси. За потребе израде TEI записа развијен је посебан алат Roma[115] који омогућава корисницима да конструишу свој језик за обележавање заснован на постојећем TEI-језику, додавањем или брисањем модула, појединачних елемената и атрибута уз могућност преименовања елемената и атрибута, редефинисања њиховог садржаја, као и применом многих других опција.

Како је првобитна структура TEI стандарда била опширна и компликована за ширу популацију корисника, временом је развијена једноставнија „TEI lite“ верзија или „лака

[115] Roma, http://www.tei-c.org/Roma/

верзија TEI" како би се почетницима омогућило да се на једноставан начин упознају са основама *Смерница*, да би у што краћем року могли и практично да их примене. Верзија „TEI lite" се данас широко примењује у библиотекама.

Поред детаља о томе како треба кодирати сам текст документа, TEI општа правила дефинишу и део уграђеног заглавља који садржи метаподатке о делу, TEI заглавље (TEI Header). Оно може да се користи за унос библиографског описа како електронских, тако и неелектронских верзија текстова. Библиографски подаци се, у суштини, не разликују од оних који су унети приликом каталогизације документа, па се тако записи у неком од формата MARC могу користити за креирање TEI заглавља и обрнуто.

Заглавље TEI означава се етикетом *<teiHeader>* и састоји се из пет делова (TEI Consortium 2016, 18):

1. *Опис датотеке (file description - <fileDesc>).* Овај део садржи библиографске податке произведеног електронског текста. Из овог дела корисници текста могу да произведу одговарајуће библиографско цитирање, док библиотекари или архивисти овај део заглавља могу да искористе за креирање одговарајућих каталошких записа. Овај део заглавља садржи, такође, и информације о извору или изворима из којих је добијени електронски документ изведен.
2. *Опис кодирања (encoding description – <encodingDesc>).* Овај део описује везе између произведеног електронског текста и једног или више извора из којих је изведен. Овде могу бити дати детаљни подаци о томе да ли је текст коригован и нормализован током транскрипције, који ниво кодирања и анализе текста је примењен и слично.
3. *Профил текста (text profile - <profileDesc>).* Овај део садржи класификационе и контекстуалне информације о електронском тексту који се описује, као што су предмет текста, како је настао, особе које су учествовале у његовом стварању и слично. Подаци из овог дела заглавља често се користе у корпусима или језичким колекцијама како би се подстакло коришћење контролисаних описних речника, а у сврху организације информација и обезбеђивања терминологије

за каталогизацију (Harpring 2010) или претраживање садржаја текста. Овај део заглавља може да се користи у било ком кораку аутоматске обраде текстова.

4. *„Контејнер" део (xeno Data - <xenoData>).* Овај део омогућава увоз метаподатака насталих коришћењем других формата. На пример, MARC запис за кодирани документ може да се извезе у MARC/XML или MODS формат и као такав увезе преко овог дела у заглавље. Сет метаподатака који се увози може бити и у формату Даблинско језгро.
5. *Историја ревизија (revision history).* Овај део садржи податке о свим променама насталим од тренутка стварања електронског текста који се описује. Метаподаци из овог дела заглавља битни су за контролу верзије.

Метаподаци у форми TEI заглавља су за неки српски текст највероватније први пут урађени за српски превод дела *1984* Џорџа Орвела у оквиру пројекта TELRI када је и урађен структурно анотирани корпус српског превода романа *1984* и додат вишејезичном паралелном корпусу Орвелова 1984 (пројекат Орвелова 1984 је детаљније објашњен у поглављу 2, одељци Орвелова 1984 (Multext-East): паралелни преводи романа 1984 Џорџа Орвела и Орвелова 1984 за српски језик). Пример TEI заглавља за дело „Моје награде" аутора Томаса Бернхарда дат је као Прилог 8 – Пример записа у формату TEI заглавље / XML синтакса.

4.1.2 Израда метаподатака у Србији и значај библиотека

Библиотеке у Србији имају највише искуства у процесу дигитализације у односу на остале установе културе у земљи па самим тим и у процесу израде метаподатака. Велике библиотеке као што су Народна библиотека Србије, Универзитетска библиотека „Светозар Марковић" и Библиотека града Београда су искуство у овој области стекли учешћем у великим европским пројектима у којима учествују више година уназад. Највише пројеката у којима су ове библиотеке учествовале реализовано је у оквиру иницијативе Еуропеана[116] која преко развијеног портала омогућава приступ до преко 50 милиона дигиталних објеката из библиотека, архива, музеја и аудио-визуелних колекција.

[116] Europeana, https://www.europeana.eu/portal/en

Поступак израде метаподатака углавном зависи од захтева и потреба пројеката односно иницијатива које реализују процес дигитализације, али и од самих дигиталних објеката, њихове структуре и карактеристика. За израду метаподатака користе се стандарди у чијој је основи углавном XML синтакса, а најчешће коришћени стандарди описани су у одељку 4.1.1 овог поглавља. Развијени стандарди се преко различитих формата интегришу у системе и репозиторијуме са којима организације раде и које развијају омогућавајући тако лакшу размену метаподатака са другим системима. Са друге стране, неке организације дефинишу своје стандарде за израду метаподатака заснивајући их на већ постојећим стандардима, а у складу са специфичностима информационих извора који се описују.

Поред библиотекара стручно обучених за овај посао, креатори метаподатака могу бити и мање обучени библиотекари који недовољно или уопште не познају структуру формата који се примењује или чак сами корисници. Системи и апликације најчешће омогућавају опис објеката односно унос метаподатака у једном формату, али омогућавају аутоматски извоз (експорт) у друге формате захваљујући процесима мапирања који су програмски решени у оквиру датог система. Такав облик аутоматског извоза метаподатака нуди и систем COBISS.

Кооперативни онлајн библиографски систем и сервиси (Cooperative Online Bibliographic System and Services - COBISS)[117] је систем који представља платформу националних библиотечко-информационих система Србије, Словеније, Босне и Херцеговине, Црне Горе, Македоније, Бугарске и Албаније повезаних у регионалну мрежу COBISS.Net[118]. Систем COBISS настао је, развија се и одржава у Институту информацијских знаности (IZUM)[119] у Марибору још од 1987. године када је успостављен као систем за узајамну каталогизацију библиотека Југославије. На принципима COBISS система изграђен је систем COBISS.SR као интегрисани библиотечки систем Србије. У систему се за размену података користи формат COMARC, варијанта MARC формата израђена према UNIMARC-у у IZUM-у за потребе система COBISS, а у оквиру COMARC-а постоје: COMARC/B за

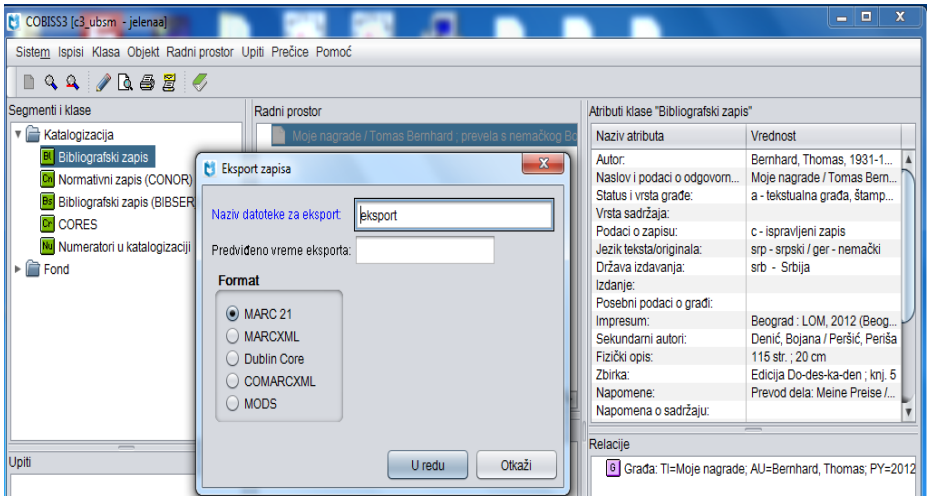
[117] Cooperative Online Bibliographic System and Services, http://www.cobiss.net/platforma_cobiss-SR.htm
[118] COBISS.Net, http://www.cobiss.net/default-sr.asp
[119] Institut informacijskih znanosti, http://www.izum.si/

библиографске податке и COMARC/A за нормативне податке, базирани на UNIMARC формату, као и COMARC/H за податке о стању фонда који је развио IZUM. Када је реч о међународној размени библиографских података поред формата COMARC/B и COMARC/A користи се и формат MARC21.

Записи у систему COBISS су доступни у форматима MARC21 и COMARC према стандарду ISO2709 (ISO2709:2008)[120] односно у форматима Даблинско језгро/XML, MODS/ XML, MARC21/XML и COMARC/XML. Сви ови формати су у складу са међународним стандардима па омогућавају комуникацију и размену података уз минималне губитке у погледу садржаја и функционалности између различитих система, односно различитих дигиталних библиотека (Тртовац 2016, 60). На овај начин је олакшано слање метаподатака о дигиталним објектима другим системима и учешће у бројним пројектима путем којих се представљају и повезују подаци о различитим дигиталним библиотекама. Овако структурирани метаподаци коришћени за опис дигиталних објеката омогућавају квалитетније проналажење информација о њима.

Библиотеке у Србији које раде у системском окружењу COBISS имају могућност да креиране записе аутоматски извезу у форматима који су претходно наведени (Слика 11). На овај начин се за кратко време може добити велики број записа у траженом формату који су спремни за даљу обраду и размену са другим системима и различитим дигиталним библиотекама.

[120] ISO2709 (ISO2709:2008), https://www.iso.org/standard/41319.html



Слика 11. Радни простор у систему COBISS за извоз метаподатака

Сви одељку Прилози приказан је пример романа „Моје награде” (Meine Preise) аустријског писца Томаса Бернхарда (Thomas Bernhard) у свим поменутим форматима и дат је упоредни приказ и анализа метаподатака (Прилог 9 – Упоредни приказ метаподатака). Поређење је урађено према запису у COMARC/B формату који је најобимнији од свих приказаних формата. Формати COMARC/XML, MARC/XML, Даблинско језгро, MODS и MARC21 који су предвиђени за извоз у оквиру система мапирани су са форматом COMARC/B што је још један разлог зашто је управо запис у овом формату узет као основа за поређење. Ако се погледа табела највећи број метаподатака је адекватно пренет из COMARC/B у тражене формате што значи да је мапирање урађено уз минималне губитке података.

Формати COMARC/B и COMARC/XML су у потпуности компатибилни. Број метаподатака као и редослед њиховог навођења у потпуности је исти, а разлика је само у синтакси која је дефинисана за сваки од ових формата посебно. Главна разлика формата COMARC/B у односу на остале јесте постојање података о стању фонда (холдингу) и система упутница. Подаци о холдингу подразумевају локалне локацијске податке и податке о сигнатури и инвентарном броју који се додељују библиотечкој грађи, док упутнице представљају информациони систем којим се „перманентно уједначавају,

употпуњавају или разрешавају облици одреднице[121] и свих елемената који уз њих иду" (Јанчић 1991, 212). Упутни систем доприноси већој информативности и повезаности система каталога. Ови подаци уносе се у блок 9[122] који је намењен националној употреби јер се њима UNIMARC прилагођава домаћим потребама. Према препорукама подаци из овог блока искључују се из међународне размене информација те они нису пренети ни у један од понуђених формата, што се може и видети у табели. Поред података о холдингу податак о држави издавања се такође не преноси у понуђене формате, податак о преводу се не преноси у формат Даблинско језгро, а податак о УДК броју се не преноси у формат MODS. Када је реч о језику и писму публикације у форматима Даблинско језгро, MODS и MARC21 преноси се само податак о језику публикације. Подаци из системског поља (датум креирања и преузимања записа, податак о креатору записа, податак о ономе ко преузима запис, датум последње измене, податак о идентификационом броју у каталогу (COBISS ID)) не преносе се у формате Даблинско језгро и MODS. Поред метаподатака у овим форматима, у табели је анализирано и постојање метаподатака у TEI заглављу и формату METS иако они нису добијени аутоматским извозом из система COBISS већ су направљени ручно уз помоћ приручника и већ постојећих примера.

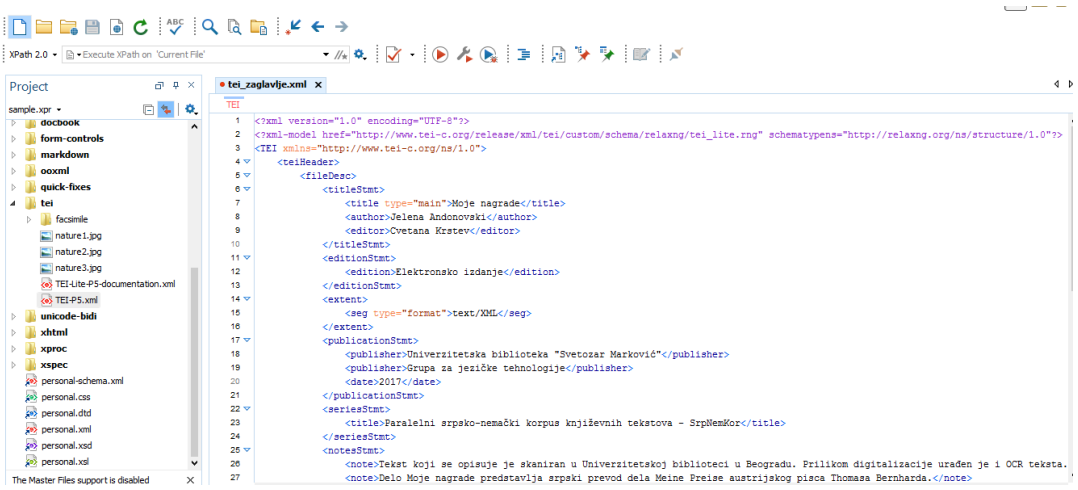
Запис у TEI заглављу направљен је ручно на основу Смерница TEI P5 у оквиру којих је детаљно дефинисана структура TEI заглавља са свим расположивим елементима и пропратним атрибутима за кодирање метаподатака и већ постојећих шаблона у едиторима за креирање XML докумената. За израду TEI заглавља коришћен је Oxygen XML Editor[123] у оквиру кога већ постоје шаблони за израду TEI записа, TEI-Lite-P5-documentation.xml и TEI-P5.xml (Слика 12). За израду TEI заглавља за дело „Моје награде" коришћен је шаблон TEI-Lite-P5-documentation.xml јер је структура TEI заглавља у оквиру овог шаблона у односу на исту структуру у шаблону TEI-P5.xml детаљнија. Поред

[121] Одредница представља скуп речи које одређују место каталошке јединице у библиотечком алфабетском каталогу (Јанчић 1991, 20).

[122] У оквиру формата UNIMARC као и његовим варијантама постоји десет блокова (0xx-9xx) података од којих сваки окупља податке са одређеним значењем.

[123] Oxygen XML Editor је софтвер који садржи у себи бројне алате за креирање, уређивање и објављивање XML докумената. Развијен је за различите платформе, све веће оперативне системе и као самостална апликација, односно Eclipse plug-in. Софтвер није бесплатно доступан али је могуће преузети едитор бесплатно на 30 дана као тестну верзију. Уз преузимање добија се и лиценцни кључ којим се покреће програм. Доступно на: https://www.oxygenxml.com/.

елемената који су већ дати ручно су додати још неки (на пример, <sourceDesc> за опис оригиналног извора на основу кога је настао електронски текст који се описује) како би запис садржао већи број метаподатака и самим тим боље описао представљени електронски текст.



Слика 12. Oxygen XML Editor – радно окружење

Запис садржи основне податке о електронској верзији текста „Моје награде" као што су: име аутора, податак о издању, формату, месту издавања, издавачу, години издавања, већој целини којој овај електронски текст припада, али и напомене у којима је наведено која је сврха и намена овако припремљеног документа. Поред ових метаподатака, представљено TEI заглавље садржи и метаподатке о физичком извору дела „Моје награде" на основу кога је и настао електронски извор. Ту су дати следећи библиографски подаци: наслов, оригинални наслов, податак о аутору, преводиоцу и аутору додатног текста, подаци о месту издавања, издавачу и години издавања, као и податак о едицији којој дело припада. TEI заглавље, такође, садржи и податке о креатору овог XML записа и датум последње измене што се може поистоветити са метаподацима који су код осталих формата у табелу унети у поље „Подаци о запису који се системски додељују". Оно што овде није унето од метаподатака јесу подаци о штампању, физичким карактеристикама (број страна и формат публикације) и УДК број што у већини

представљених формата постоји и подаци о стању фонда (холдингу) и инвентарном броју који постоје само у формату COMARC.

Запис у формату METS направљен је ручно на основу постојећих приручника и примера[124] који постоје на веб страни Конгресне библиотеке за различите врсте информационих извора. Према табели запис за дело „Моје награде" у овом формату садржи већину метаподатака као и записи у осталим форматима. Метаподаци о самом запису (креатор записа, датум креирања записа, идентификациони број записа) могу се поистоветити са метаподацима који су код осталих формата унети у поље „Подаци о запису који се системски додељују". Када је реч о описним метаподацима, METS омогућава да се у овај сегмент записа интегришу већ готови записи у другим форматима (MARC, DC, EAD и други) те је и у овом случају тако урађено. Интегрисан је већ постојећи запис у формату Даблинско језгро. Главна разлика записа у формату METS у односу на остале јесте постојање података о носиоцу ауторског права и података о техничким карактеристикама дигиталног објекта који се описује.

4.2 Припрема текста

Поред претраге преко формалних особина корпусног материјала, односно претраге преко метаподатака, проналажење информацију у текстуалним корпусима може бити и претрагом целог садржаја текста што доприноси прецизнијем и квалитетнијем одзиву резултата претраге одређеног корпуса. У наставку одељка представљамо оптичко препознавање карактера и препознавање именованих ентитета које су примењене и на текстовима који су одабрани за паралелни корпус који је предмет ове докторске дисертације.

4.2.1 Оптичко препознавање карактера

Оптичко препознавање карактера (Optical Character Recognition - OCR) представља конверзију, односно превођење, слика текста (руком писаног, куцаног на машини или штампаног текста) у машински кодирани текст. Процес сканирања, као први корак

[124] Примери су доступни на: http://www.loc.gov/standards/mets/mets-examples.html

дигитализације садржаја, производи „слике" документа у одређеном формату, које се могу листати, прегледати, читати и штампати, али напредна употреба таквих „слика", у смислу претраге целовитог текста и повезивања садржаја са другим изворима на вебу, није могућа. Применом методе OCR-а текст постаје претражив преко сваке појединачне речи, а могуће је идентификовати и слике, фотографије и разне мултимедијалне садржаје и њих обрадити на адекватан начин. Процес оптичког препознавања карактера који подржава томе намењен софтвер пролази кроз неколико корака (Тртовац 2016, 106-107):

1. *Учитавање слике извора који се обрађује.* Добар систем за оптичко препознавање карактера подржава различите формате за дигиталне фотографије сканираних материјала (TIFF, JPEG, PNG, као и PDF) који служе као основа, то јест, улаз за оптичко препознавање карактера.
2. *Утврђивање резолуције слике и врсте и величине фонтова који се јављају у тексту.*
3. *„Бинаризација".* Представља претварање слика у боји у црно-беле слике са добрим контрастом.
4. *Уклањање оквира и подвлака у тексту.*
5. *Препознавање и означавање речи и размака у тексту.*
6. *Препознавање појединачних знакова као главног дела оптичког препознавања карактера.* Слика сваког знака конвертује се у одговарајући код карактера према одабраним кодним табелама. Препознавање никада није потпуно успешно, па се код непрепознатих карактера поправке најчешће врше ручно консултовањем извора. Најбољи су они системи који на поправкама „уче" односно програмирани су тако да се „обучавају" на одређеном броју страница у току чега корисник даје разрешења за све недоумице које систем усваја и примењује их у наставку рада када наиђе на сличан проблем.
7. *Коришћење речника као подршке за боље препознавање карактера.* Овај корак може да побољша квалитет препознавања; наиме, неки карактери могу да изгледају слично, али појављивање одређене речи у електронском речнику

може да помогне код одлучивања. Ова техника је примењена за полуаутоматску корекцију корпуса који је изграђен за потребе ове дисертације.

8. *Чување резултата у различитим излазним форматима.* Софтвер обично нуди следеће излазне формате: PDF, .doc, .txt. Нарочито је пожељан PDF формат јер је најшире распрострањен у дигиталним библиотекама, као и .doc (Word) јер је текст у овом формату погодан за даљу обраду.

За оптичко препознавање карактера користи се данас разноврстан софтвер, а за потребе корпуса СрпНемКор цео процес сканирања материјала и оптичког препознавања описан је у поглављу 6 одељак 6.2.1.

4.2.2 Препознавања именованих ентитета

Препознавање именованих ентитета (Named Entity Recognition - NER) једна је од технологија за проналажење информација у самим документима и водећа тема у области обраде природних језика више од петнаест година. Технологија подразумева лоцирање специфичних појмова у тексту и њихово класификовање у одређене категорије. Проблем препознавања и класификације именованих ентитета је битан за разне научне дисциплине и области: то је први од пет подзадатака екстракције информација која представља специјализовану подобласт проналажење информација; с друге стране, у оквиру рачунарске лингвистике, тај проблем је од значаја за обраду природних језика и аутоматско превођење (Utvić 2008).

Потреба за препознавањем и класификовањем именованих ентитета истакнута је кроз бројне пројекте и конференције од краја осамдесетих година прошлог века па све до данас, али су прецизна дефиниција појма „именовани ентитети" као и задатак њиховог аутоматског препознавања постављени на шестој и седмој Конференцији о разумевању порука (MUC-6 1995, MUC-7 1998 - Message Understanding Conference) (Grishman 1996). Именовани ентитети су дефинисани као властита имена, или називи уопште, и одређени изрази за износе, а као задатак препознавања именованих ентитета дефинисано је препознавање имена (имена особа, имена организација и локација), израза којима се описују датуми и време на часовнику и бројчаних израза (процентуалних и новчаних

израза) у тексту (Chinchor and Robinson 1997) (Chinchor and Marsh 1998). У (Sekine et al. 2002, 1822-1824) дат је преглед преко 150 категорија именованих ентитета са примерима, док је у другом раду истих аутора (Sekine et al. 2004, 1977-1980) представљено преко 200 категорија организованих у хијерархијску структуру дрвета.

Данас се за препознавање именованих ентитета користе разноврсни приступи, као што су модели засновани на лингвистичким знањима и ресурсима и модели засновани на статистичким карактеристикама одређеног природног језика. Екстракција информација из дигиталног документа коришћењем технологије препознавања именованих ентитета од великог је значаја и за апликације семантичког веба које омогућавају њихово увезивање са сродним ентитетима у релевантним ресурсима на мрежи као што су нормативне датотеке личних имена, нормативне датотеке географских имена, различите врсте онтологија и слично, а које су данас део шире семантичке мреже „Отворени повезани подаци". Анотација именованих ентитета урађена је и на текстовима у паралелном корпусу СрпНемКор да бисмо омогућили што боље проширење упита за претрагу и испитали могућности умрежавања са релевантним ресурсима који су део шире мреже семантичког веба.

5 Семантички веб и мрежа отворених повезаних података

У последњој деценији интензивно се ради на стварању веба као глобалне базе података која садржи универзалну мрежу семантичких исказа. У првој јавној верзији веба који је у бити створен деведесетих година 20. века преовладавао је такозвани модел одозго надоле (top-down) изградње статичних страница које су биле намењене прегледању. Овакав приступ може се поистоветити са традиционалним приступом који користе библиотеке у изради статичних документа у облику каталошких записа у неком од MARC формата. Са новим трендом слободног приступа информационим системима и базама података много података смешта се коришћењем технологије облака (cloud technology), која се користи на захтев, при чему се од концепта „софтвер као услуга" SaaS (Software as a Service) створио концепт „подаци као услуга" DaaS (Data as a Service) (Truong et al. 2009). Са друге стране, с огромним бројем материјала на вебу који сваким даном расте поставио се проблем њихове организације, дуплирања података и непостојања обједињене претраге (Fargo et al. 2013, 28) велике количине информација.

Семантички веб пружа механизам за структурирање и увезивање података на нивоу њиховог значења и формирање скупова података који се могу претраживати преко система обједињене претраге. Тако је од „веба докумената" (Web of Documents) створен „веб података" (Web of Data) где се значењем података исказује садржај исказан на природном језику које се експлицитно анотира тако да га машине могу прочитати и коректно интерпретирати (разумети). На овај начин решавају се неки од проблема класичног веба као што су: немогућност претраге веба коришћењем сложених упита над више извора, све присутније обиље података на вебу уз немогућност њиховог тумачења, ограниченост рачунара на пренос и приказ података на вебу без стварне улоге у њиховој обради (Sikos 2015, 17).

Како су примењене неке од технологија семантичког веба, односно како је припремљен један скуп отворених повезаних података на основу паралелног корпуса који је предмет ове докторске дисертације детаљније је објашњено у поглављу 6 одељак 6.5.3, док је у овом поглављу приказан концепт „семантички веб", његово значење и

технологије на којима се заснива са посебним освртом на неке од иницијатива које су развијене на овом концепту, али и на значај библиотека и библиотечких ресурса у целом систему.

5.1 Семантички веб

5.1.1 О семантичком вебу

Семантички веб или Веб 3.0 дефинисан је као механизам за проналажење и повезивање података на вебу. Циљ семантичког веба је да веб уместо мреже докумената постане мрежа података којима је придружено значење и који су међусобно повезани. Визију семантичког веба изнео је 1994. године на првој World Wide Web (WWW - W3) конференцији творац WWW-а Тим Бернерс Ли (Tim Berners Lee) као „Веб применљивих информација – информације изведене из података кроз семантичку теорију тумачења симбола. Семантичка теорија указује на виђење „значења" у коме логичко повезивање термина успоставља интероперабилност између система."[125] (Shadbolt et al. 2006, 96). У (Berners-Lee et al. 2001), првом оригиналном чланку на тему семантичког веба, Тим Бернерс-Ли је заједно са сарадницима дефинисао семантички веб као „проширење постојећег веба у којем се информацијама даје јасно дефинисано значење, што омогућава људима и рачунарима да сарађују" истичући апстрактни ниво појединачних докумената односно проблем представљања знања (knowledge representation). Према (Coyle 2012, 10) израда семантичког веба заснива се на означавању и повезивању података у документима, са једне стране, и објављивању скупова повезаних података на вебу, са друге стране. У ширем смислу се под појмом „семантички веб" подразумева скуп стандарда W3 конзорцијума које обухватају темељне технологије за увезивање података на нивоу њиховог значења. Конзорцијум описује семантички веб једноставно као мрежу података (Web of Data) која се заснива на два основна принципа:

[125] The Semantic Web is a Web of actionable information—information derived from data through a semantic theory for interpreting the symbols. The semantic theory provides an account of "meaning" in which the logical connection of terms establishes interoperability between systems.

1. употреба заједничких формата за интеграцију и комбиновање података из различитих извора и
2. употреба заједничког језика за опис података.

Уз то W3 конзорцијум је дефинисао шест начела на којима се темељи семантички веб (Koivunen et al. 2011):

1. *Све може бити идентификовано URI идентификаторима.* Сви објекти на мрежи (људи, места, ствари из физичког света, апстракције) могу се идентификовати оваквом врстом идентификатора. URI идентификаторе могу додељивати сви они који имају надзор над просторима имена који и одређују на шта се они односе у стварном свету.
2. *Ресурси и везе могу имати типове.* Веб који смо познавали састојао се од ресурса и веза између њих. Такви извори нису садржали метаподатке за опис њиховог контекста и међусобних односа. Применом технологија и принципа семантичког веба ресурсима и њиховим везама додељују се типови тако да они постају разумљиви концепти. Додељивањем типова рачунари могу да препознају неке карактеристике ресурса које су до сада биле разумљиве искључиво људима. На пример, да је неко дело роман или научно-истраживачки рад, да је један ресурс верзија другог ресурса, да је ресурс производ неког аутора, или да је ресурс део софтвера који зависи од рада другог софтвера.
3. *Дозвољене су делимичне информације.* Семантички веб је, као и обичан веб, неограничен у смислу количине информација и веза између њих. На пример, неки од повезаних ресурса могу престати да постоје и њихове адресе се могу користити за означавање других ресурса. Алати семантичког веба треба да толеришу ове промене и функционишу упркос њима.
4. *Нема потребе за апсолутном истином.* На мрежи не постоји јединствена база података која је у потпуности истинита, односно, која садржи потпуно поуздане податке. Семантички веб то не мења и унутар њега свака апликација закључује шта је поузданије на основу доступних података.

5. *Развој се подржава.* Сличне концепте различите групе аутора могу да дефинишу на различитим местима или иста група аутора може дефинисати сличне концепте у различитим временским тренуцима. Став је да је окупљање таквих сличних концепата из различитих извора корисно. Циљ је да се опишу извори тако да додавање нових података не тражи измену већ постојећих, али је потребно да се назначе разлике и разреши двосмисленост и недоследност. Семантички веб користи описне конвенције које се могу проширити како се људско знање шири.
6. *Минималистичко обликовање.* Семантички веб чини једноставне ствари једноставнијим, а комплексне могућим. Циљ W3 конзорцијума је стандардизација најнеопходнијих технологија. Овај приступ омогућава имплементацију једноставних апликација које се заснивају на већ постојећим стандардима као што је, на пример, Даблинско језгро, али се истовремено ради на истраживању будућих сложенијих технологија.

По аналогији са светском мрежом, семантички веб захтева рачунарски мега систем са следећим карактеристикама (Janik et al. 2011, 470):

1. *Експлицитно и једноставно представљање података*. Уобичајени приказ података у мега систему треба да буде експлицитан и једноставан без приказивања технологија које су основи.
2. *Дистрибуирани систем*. Мега систем треба да омогући слободну дистрибуцију података без централизоване контроле о томе ко су носиоци ауторских права. Дистрибуирана контрола и поседовање, ако се раде правилно, олакшавају усвајање и повећање броја података.
3. *Унакрсно повезивање*. Да би се омогућило комплексно умрежавање компонената односно сетова података из различитих области подаци морају унакрсно да се повежу.
4. *Слободно повезивање коришћењем општих језика за описивање и структурирање података*. У мега систему компоненте односно сетови података и везе између њих морају бити описане стандардним језицима за

обележавање који имају велику флексибилност тако да се могу користити у различитим системима.

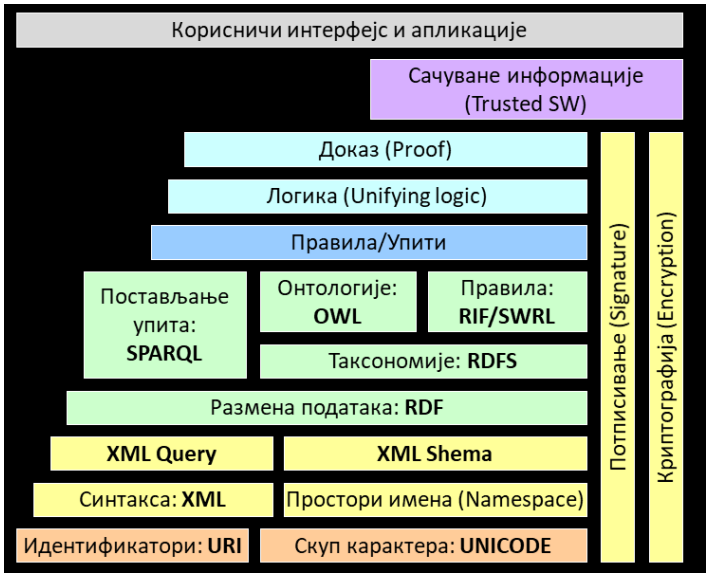
5. *Једноставно објављивање и употреба*. Мега систем треба да омогући једноставно објављивање и коришћење умрежених података.

Да би ове карактеристике биле реализоване, семантички веб користи разноврсне технологије и стандарде за означавање значења података и остваривање веза између њих.

5.1.2 Структура семантичког веба

Комплексну структуру семантичког веба графички је представио Тим Бернерс Ли преко шеме познатије и као вишеслојна „торта" семантичког веба (Semantic Web Layer Cake) односно стек семантичког веба (Semantic Web Stack)(

Слика 13). Преко „торте" су приказани статички и динамички делови односно слојеви семантичког веба од којих су у доњим слојевима стандардизовани концепти и модели док су у горњим слојевима они који су још у развоју.



Слика 13. Вишеслојна „торта" семантичког веба[126]

[126] Слика је настала на основу оригинала у Web 3.0 – beyond the Semantic Web, a way to global SOA, preuzeto 7.4.2019, https://gyires.inf.unideb.hu/GyBITT/08/ch02s03.html и објављена је у (Томашевић 2018, 75)

URI/Unicode. Основу „торте" чине статички делови као што су кодна шема Unicode и јединствени идентификатори ресурса URI. Синтакса URI идентификатора одређена је стандардом RFC 3986 (Berners-Lee et al. 2004) и дефинише се као компактна ниска знакова која идентификује апстрактни или физички извор, односно, као ниска из знакова ASCII карактерског скупа која подлеже синтаксичким правилима. Јединствени идентификатори ресурса користе се за именовање свих појмова и веза између њих. Ово је кључни елемент концепта „повезани подаци" и користи се аналогно, на пример, идентификаторима за нормативну контролу у традиционалном библиотекарству. URI идентификатори додељују се не само веб локацијама већ и свим ентитетима у семантичком вебу: именима аутора, именима издавача, називима места, насловима књига и слично (Antoniou and Van Harmelen 2008, 67-68). У овом случају недвосмисленост је од великог значаја и један URI може да се додели само једном ентитету. URI идентификатор може бити Универзални локатор ресурса (Uniform Resource Locator - URL), међународна идентификација ресурса (Internationalized Resource Identification - IRI) или нека друга врста јединственог идентификатора. URL је интернет локација преко које је омогућен приступ одређеном ресурсу на вебу. IRI је интернет протокол стандард који дефинише употребу карактера за дефинисање URI из Универзалног скупа карактера (Universal Character Set (Unicode/ISO 10646)[127]).

URI мора бити трајан и не треба да садржи делове који су подложни променама. Један од начина да се ово постигне јесте да се користи домен који је под директном контролом организације која генерише податке или да се користи трајни униформни локатор ресурса (Persistent Uniform Resource Locator - PURL). Иако је сама технологија URI идентификатора дефинисана стандардом, постоје смернице за њихово обликовање које је дефинисао Тим Бернерс-Ли под називом „Cool URIs", а који се могу сажето сврстати у три категорије (W3C Style 2019):

1. једноставност (simplicity) - препоручљиви су кратки „мнемонички" URI идентификатори који се лако памте и шаљу електронском поштом;

[127] ISO 10646, https://www.iso.org/standard/69119.html

2. стабилност (stability) - препоручљиви су трајни URI идентификатори односно почетна дефиниција URI идентификатора треба да опстане што дуже;
3. управљивост (manageability) – промена неких укључених елемената у структуру URI идентификатора не утиче на њихову стабилност; на пример, ако је година укључена у структуру URI идентификатора једне године, структура URI идентификатора који у себи има другу годину не би требало да се мења.

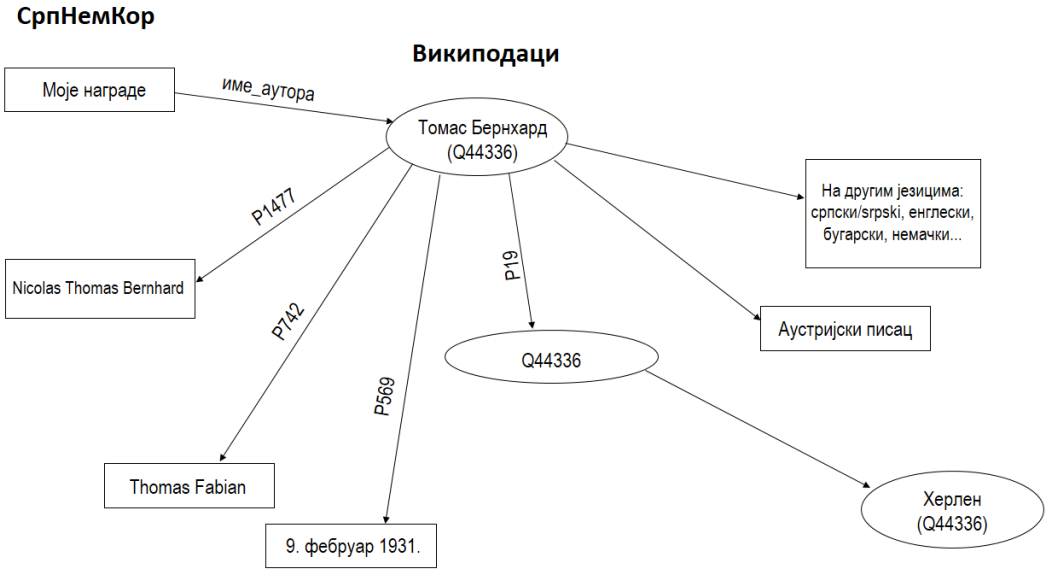
XML. XML синтакса је други ниво „торте". Језик XML означава постојање заједничке синтаксе за опис ентитета и веза између њих. XML је прошириви језик за обележавање (детаљније је објашњен у поглављу 3 одељак 3.4.1), али не и средство за опис семантичких веза између њих. За то се користи стандард Оквир за описивање ресурса (Resource Description Framework - RDF)[128].

RDF. RDF је стандард који је развио W3 конзорцијум као модел којим се подаци могу представити на различите начине и који омогућава моделирање метаподатака о ресурсима на вебу анотирањем њиховог значења или функција. Према (Miller 1998, 15) „RDF има инфраструктуру која омогућава кодирање, размену и употребу структурираних метаподатака. Ова инфраструктура омогућава интероперабилност метаподатака преко модела података и формата који подржавају уобичајене конвенције семантике, синтаксе и структуре. RDF не прописује конкретну семантику за опис ресурса, већ заједницама омогућава да саме дефинишу елементе за опис метаподатака који су им потребни".

У моделу података RDF разликујемо ресурсе, својства и изјаве (Workman 2016, 2). Ресурс представља метаподатак који се дефинише у RDF документу, а својство је атрибут којим се описују особине и карактеристике ресурса. Ресурси у RDF документу повезују се са својствима и формирају изјаве у форми уређене тројке: субјекат – предикат – објекат (*triple*) (RDF primer 2014). Субјекат је оно о чему се говори односно предмет изјаве, предикатом се представљају особине или карактеристике субјекта, док је објектом представљена вредност предиката. Субјекат једне изјаве може бити објекат друге изјаве, и обрнуто, а између њих се успостављају семантичке везе. Предикати, такође, могу бити субјекат друге изјаве. Другим речима, ова три саставна дела називају се ресурсима.

[128]Доступно на: http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/

Како изгледа једна изјава у форми уређене тројке приказаћемо на следећем примеру: „Моје награде” → „име_аутора” → „Томас Бернхард”. Субјекат ове изјаве је наслов романа „Моје награде” који је део веће колекције SrpNemKor (паралелни корпус који се у дисертацији представља) и он је ентитет који се описује. Аутор романа је писац Томас Бернхард који представља објекат у нашем примеру - „Томас Берхард”. Ентитети „Моје награде” и „Томас Бернхард” повезани су преко предиката који има вредност „име_аутора” чиме је дефинисана једна од особина субјекта. „Томас Бернхард”, даље, може бити део базе Википодаци и у њој представља субјекат који се описује. О бази Википодаци и структури записа за ентитет „Томас Бернхард” у њој детаљније ће бити речи у одељку 5.2.2. У том одељку приказани су примери неких изјава за ентитет „Томас Бернхард” у Википодацима и начин означавања субјеката, предиката и објеката у изјавама. Како скуп међусобно повезаних изјава формира граф где су субјекти и објекти представљени као чворови који су повезани предикатима (Workman 2016, 96), овде приказујемо пример графа за изјаве за ентитет „Томас Бернхард” које су детаљније објашњене у одељку 5.2.2 (Слика 14).



Слика 14. Пример RDF графа за ентитет „Томас Бернхард“ у бази Википодаци

Како би били машински читљиви, ресурсима, али и изјави у целини, додељују се URI идентификатори. Сви URI идентификатори који се користе долазе из речника односно URI колекција за представљање информација о одређеном домену. За обраду овако кодираних изјава и њихову размену између машина односно система користе се различити језици за обележавање. RDF најчешће користи XML формат као синтаксу за размену и обраду метаподатака. Користећи предности формата XML недвосмислено се представља значење ресурса који се описују и тако омогућава конзистентно кодирање, размена и машинска обрада стандардизованих метаподатака. Међутим, због потребе познавања појединачне структуре XML-а и RDF-а, као и њиховог међусобног прожимања у погледу описивања ресурса, RDF/XML је за већину потенцијалних корисника ипак био доста тежак за разумевање и имплементацију те су развијени формати N-Triples, Turtle и N3 као једноставније, текстуално засноване варијанте за записивање изјава у виду тројки у RDF-у чије су предности већа ефикасност у погледу моделирања изјава, брже и лакше читање и многе друге. Поред ових развијена је и JSON[129] варијанта за записивање RDF модела података за лакшу размену података између апликација (Auer et al. 2013, 10-12). Године 2009. W3 конзорцијум је стандардизовао RDFa (RDF in Attributes) са циљем да се поједностави интегрисање HTML-а и RDF-а и омогући заједничко представљање садржаја у оквиру једног HTML документа. Слика 15 илуструје пример изјава у RDF/Turtle формату за ентитет „Thomas Bernhard" у Нормативној датотеци Немачке националне библиотеке (GND).

[129] JavaScript Object Notation је текстуални формат који је потпуно језички независан премда користи конвенције из породице програмских C језика (C, C++, C#, Java, JavaScript, Perl, Python и слично). Универзалну структуру JSON-а данас подржавају углавном сви програмски језици, док је синтакса једноставна како људима за читање и разумевање, тако и машинама за парсирање и генерисање. О JSON-у више на https://www.json.org/.

```
<http://d-nb.info/gnd/118509861>
    wdrs:describedby <http://d-nb.info/gnd/118509861/about> .
<http://d-nb.info/gnd/118509861>
    a gndo:DifferentiatedPerson ;
    gndo:gndIdentifier "118509861" ;
    foaf:page <https://de.wikipedia.org/wiki/Thomas_Bernhard> ;
    owl:sameAs <http://dbpedia.org/resource/Thomas_Bernhard> ,
<http://d-nb.info/gnd/118509861>
    gndo:variantNameForThePerson "Tuomasi-Boenhade" ;
    gndo:variantNameEntityForThePerson_:node1d2kbdtnkx433561
_:node1d2kbdtnkx433561 gndo:personalName "Tuomasi-Boenhade" .

<http://d-nb.info/gnd/118509861>
    gndo:variantNameForThePerson "Boenhade, Tuomasi" ;
    gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433562 .
_:node1d2kbdtnkx433562 gndo:forename "Tuomasi" ; gndo:surname "Boenhade" .
<http://d-nb.info/gnd/118509861>
    gndo:dateOfBirth "1931-02-09"^^xsd:date ;
    gndo:dateOfDeath "1989-02-12"^^xsd:date .
```

Слика 15. Пример RDF записа у Turtle формату за ентитет „Thomas Bernhard" у бази GND

SPARQL. Још један битан део „колача" је и процес постављања упита над утврђеним повезаним скупом података према одређеним правилима упитног језика који је у позадини система. За претрагу се користи упитни језик заснован на Структурном језику за постављање упита (Structured Query Language - SQL), једноставни протокол и RDF упитни језик (Simple Protocol and RDF Query Language – SPARQL). Упитни језик SPARQL заснован је на упитном језику SQL и користи Turtle синтаксу за постављање упита (SPARQL 1.1 Query Language 2013). Поред основног система постављања упита, језик подржава и обликовање резултата претраге у смислу сортирања по одређеном параметру, одређивања максималног броја добијених резултата, а има и многе друге предности.

RDFS. RDF схема (RDF Schema - RDFS) је семантичко проширење стандарда RDF којим се обезбеђује механизам за опис комплексне структуре у погледу ресурса, њиховог значења и веза између њих. RDFS дефинише простор имена са хијерархијским концептом класа и њихових могућих својстава који заједно омогућавају креирање информатичких онтологија (Гардашевић 2013б, 32) (Volz et al. 2003).

Онтологије. Иако је реч о изворно филозофском појму, у контексту семантичког веба, под појмом "онтологија" сматра се сложени RDF речник чији елементи (класе и својства) имају јасно дефинисане типове и двосмерне логичке везе које омогућавају повезивање са

другим елементима (класама и својствима). За дефиницију онтологије у области информатике најчешће се узима (Gruber 1995): „Онтологија је експлицитна спецификација концептуализације. Термин је преузет из филозофије где је „онтологија" систематски исказ постојања. У системима вештачке интелигенције оно што „постоји" може бити представљено на одређен начин... У истом контексту можемо описати програмску онтологију која дефинише скуп репрезентативних термина. У таквој онтологији, дефиниције повезују имена ентитета у универзуму дискурса (класе, везе, функције или други објекти) са текстом који људи могу да прочитају описујући значење имена, односе између њих и ограничења њиховог коришћења".[130] У (Sure and Studer 2005, 192) се даље дефинишу неки појмови из претходне дефиниције: „*Концептуализација* представља апстрактни модел неке појаве у свету, који идентификује релевантне концепте те појаве. *Експлицитност* подразумева да су врсте коришћених концепата и ограничења њиховог коришћења експлицитно дефинисана. Ова дефиниција је врло често проширена са три додатна услова: Онтологија представља формалну, експлицитну спецификацију неке заједничке концептуализације одређене области интересовања. Појам формалности у овој дефиницији односи се на чињеницу да би онтологије требало да буду у машинама читљиве... Оне помажу да се знање представи на начин који машинама дозвољава да их обрађују".[131] Једном речју, онтологије описивањем концепата и веза између њих представљају формалну репрезентацију знања у некој посебној области. Могу се посматрати и као модел података који представља одређену област, односно домен, и користи се за закључивање над објектима у том домену и везама између њих. Захваљујући онтологијама могуће је логичко закључивање односно извођење транзитивних релација. Прецизније речено, није потребно задати све односе у једном

[130] An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an Ontology is a systematic account of Existence. For AI systems, what "exists" is that which can be represented... Thus, in the context of AI, we can describe the ontology of a program by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms.

[131] A conceptualization refers to an abstract model of some phenomenon in the world by identifying the relevant concept of that phenomenon. Explicit means that the types of concepts used and the constraints on their use are explicitly defined. This definition is often extended by three additional conditions: "An ontology is an explicit, formal specification of a shared conceptualization of a domain of interest". Formal refers to the fact that the ontology should be machine readable... Ontologies help to represent knowledge in a machine processable way…

скупу података већ их је могуће извести из аксиома речника који се користи. С растом семантичког веба расте и број различитих онтологија које описују поједина подручја. Неке од значајних онтологија у овом тренутку су: Пријатељ пријатеља (Friend of Friend - FOAF)[132] за опис људи и релација између њих, Даблинско језгро (Dublin Core - DC) за израду метаподатака, Једноставни систем за организацију знања (Simple Knowledge Organization System - SKOS)[133] за структурирање контролисаних речника појмова, Библиографска онтологија (Bibliographic Ontology - BIBO)[134] за опис библиографских ентитета, класификацију докумената, а може се сврстати и у цитатне онтологије.

OWL. Онтолошки језик (Ontology Web Language - OWL)[135] је један од језика за описивања онтологија на вебу. Подаци који се описују су хетерогени и потичу из различитих извора. RDFS није довољно комплексан и експресиван да ефикасно опише комплексну хијерархијску структуру веб онтологије. Због тога је развијен језик OWL који се заснива на стандарду RDF и надградња је RDFS. OWL користи XML синтаксу за опис, а графички приказ је базиран на Универзалном језику за моделирање (Universal Modelling Language - UML) (Antoniou and Van Harmelen 2004, 72).

Горњи слојеви „торте". Описана структура „торте" представља унифицирану логичку целину (unifying logic) која производи одређене резултате. Добијени резултати се првобитно анализирају као поуздани докази (Proof) који потврђују унапред постављени циљ овакве логичке структуре. Ако су подаци који се на почетку уносе поуздани добијени резултати су релевантни односно као крајњи циљ добијају се поуздани резултати (Trust). Такође, за поуздане улазе треба користити криптографска средства као што су, на пример, дигитални потписи за верификацију порекла извора. На основу целокупне структуре „торте" креирају се корисничка сумећа и различите врсте апликација (Obitko 2007)(Glimm and Stuckenschmidt 2016).

[132] Friend of Friend, http://www.foaf-project.org/
[133] Simple Knowledge Organization System, https://www.w3.org/2004/02/skos/
[134] Bibliographic Ontology, http://bibliontology.com/
[135] Ontology Web Language, https://www.w3.org/OWL/

5.2 Иницијатива „Отворени повезани подаци”

Највећи део идеје семантичког веба приказан је кроз иницијативу „повезани подаци” (Linked Data - LD) која је, са једне стране, омогућила да принципи и технологије семантичког веба постану приступачни и изводљиви за различите апликације, а са друге стране, увезивање великог броја скупова података (datasets) на вебу у глобалну базу знања. Иницијатива „повезани подаци“ представља кључну парадигму семантичког веба (Radulovic et al. 2015), а темељна начела иницијативе објаснио је Тим Бернерс Ли још 2006. године (Berners-Lee 2006), а представио га је на TED конференцији[136] 2009. године дефинишући га као природни развој семантичког веба која тежиште ставља на податке и на глобално повезивање скупова података користећи RDF. У (Bizer et al. 2011) концепт је дефинисан као „подаци који су објављени на вебу тако да су машински читљиви, њихов значење је експлицитно изражено, повезани су са другим скуповима података на вебу и могу се повезивати са новим скуповима података на вебу”. Поједине заједнице које делују у оквиру W3 конзорцијума описале су концепт на свој начин. Заједница повезаних података (Linked Data community) описује „повезане податке” као „скуп добрих пракси за објављивање и повезивање структурираних података на мрежи користећи URI, HTTP и RDF” (Linked Data 2019), док Инкубатор група библиотечких повезаних података (Library Linked Data Incubator Group) концепт описује као „податке који су објављени у складу са начелима дефинисаним тако да омогућавају повезивање података, скупова елемената и речничких вредности” (W3C Incubator Group 2011).

„Повезани подаци” су у отвореном приступу па се за назив користи још и „отворени повезани подаци” (Linked Open Data - LOD). Уз појам LOD често се додаје и именица *облак* (LOD cloud) и графички се представља као дијаграм међусобно повезаних скупова података (Гардашевић 2013б, 35). Рачунарство у облаку (cloud computing) је технологија у којој се преко мреже приступа ресурсима који су смештени на удаљеним серверима. То је систем који представља свеобухватан, практичан, дељиви извор ресурса у које спадају софтвер, базе података, хардвер и различите услуге (прорачуни, приступ

[136] Berners Lee, Tim. “Tim Berners-Lee on the web”, dostupno na https://www.ted.com/talks/tim_berners_lee_on_the_next_web?language=en

подацима и друго) при чему крајњи корисници не морају да знају физичку локацију и конфигурацију даваоца ресурса односно услуга (Mell and Grance 2011, 3) (Popović i Numović 2016, 818). Појам облак широко је распрострањен и у области семантичког веба. Додавање именице *облак* уз LOD указује на велику количину података која је толико обимна и сложена да се тешко обрађује уз помоћ постојећих алата и која се налази на различитим местима на вебу и позива преко различитих сервера. Овакво окружење посебно је значајно за библиотеке и друге институције културе јер је, са једне стране, финансијски повољније од одржавања појединачних система у појединачним институцијама, док с друге стране, мање библиотеке или појединачни пројекти на овај начин имају прилику да структурисане податке које поседују увежу са сродним ресурсима на вебу и постану видљивији и боље повезани.

Рад на облаку отворених повезаних података започет је 2007. године са 12 међусобно увезаних база података (Слика 16). Како се систем развијао, постао је корисно место за груписање ресурса у категорије према врсти података које чувају и заједници којој користе (Coyle 2012, 13). Сви ресурси из истог домена знања су у графичком облаку представљени истом бојом. У тренутку писања дисертације ресурси у облаку груписани су у девет категорија (Слика 17): спада у више више домена, географија, влада, науке везане за живот, лингвистика, медији, публикације, друштвене мреже, генерисано од стране корисника. За наше истраживање од посебног је значаја категорија *лингвистика* која је даље подељена на седам група (Слика 18): корпуси; лексикони и речници; терминологија, тезауруси и базе знања; метаподаци о лингвистичким ресурсима; категорије лингвистичких података; типолошке базе података; друго (Chiarcos et al. 2012).

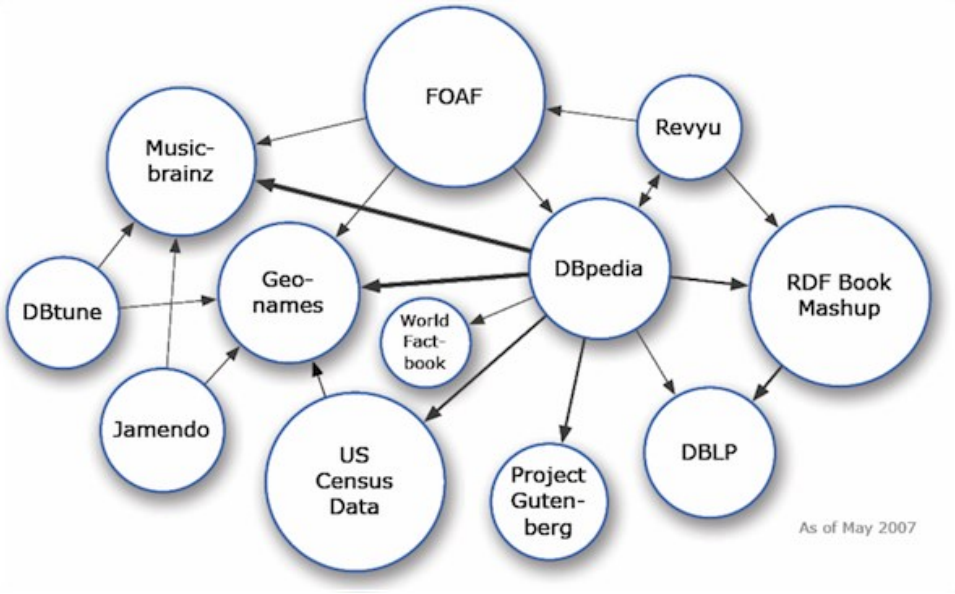
5.2.1 DBpedia

Базе знања заузимају највећи део облака, а једна таква база која тренутно (почетак 2019) представља и највећи део облака је DBpedia[137]. DBpedia је јавна иницијатива покренута са циљем извлачења структурираних информација са Википедије, како би се омогућило постављање комплексних упита и повезивање са другим скуповима података

[137] Dbpedia, http://wiki.dbpedia.org/

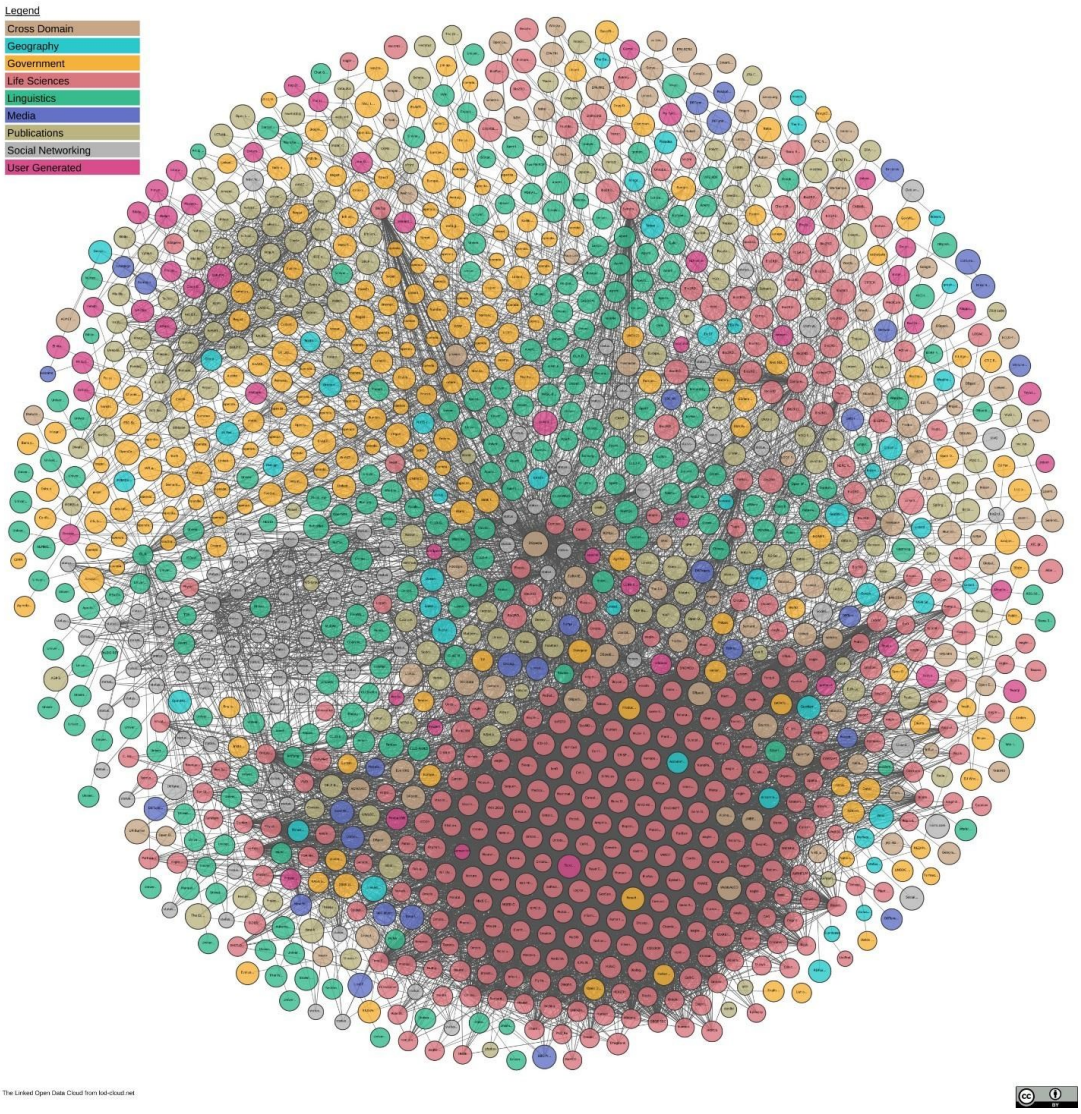
на вебу, односно примениле технологије семантичког веба. Ове информације су на вебу доступне под лиценцама Creative Commons Attribution-Share Alike 3.0 License[138] и GNU Free Documentation License[139]. Данас већина база знања покрива специфичан домен, па DBpedia у односу на њих има следеће предности (Bizer et al. 2009, 154):

1. покрива више домена,
2. аутоматски се мења и допуњује како се мењају и допуњују чланци на Википедији,
3. вишејезична је,
4. доступна је преко веба.



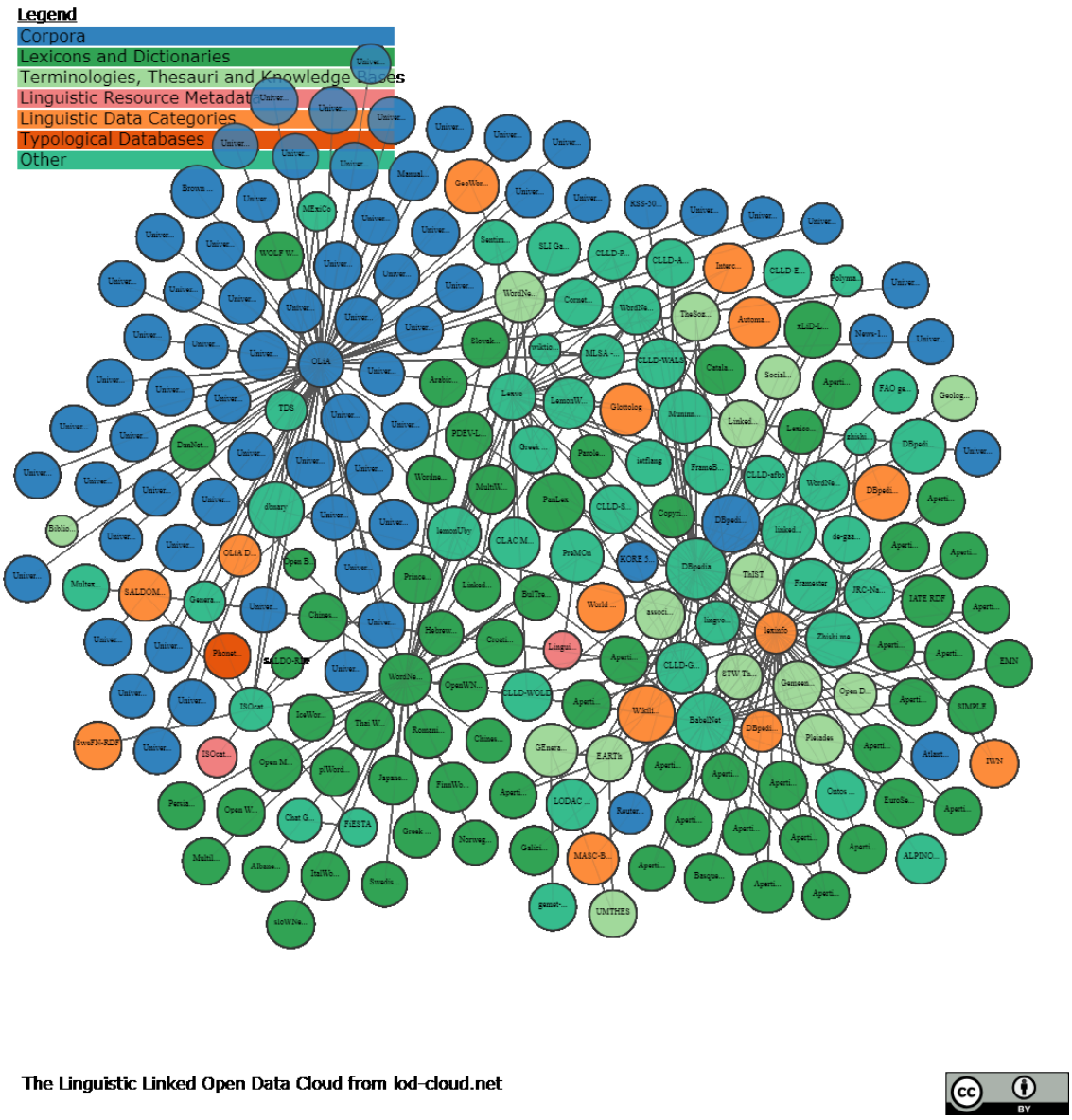
Слика 16. Први скупови података у LOD облаку[140]

[138] Creative Commons Attribution-Share Alike 3.0 License, http://creativecommons.org.rs/?page_id=74
[139] GNU Free Documentation License, https://www.gnu.org/licenses/fdl-1.3.html
[140] Дијаграм је преузет са званичног сајта Linking Open Data и доступан је на: http://lod-cloud.net/versions/2007-05-01/lod-cloud.png



Слика 17. Облак „Отворени повезани подаци”[141]

[141] Дијаграм је преузет са званичног сајта Linked Open Data http://lod-cloud.net/. Према подацима из јуна 2018. године облак тренутно садржи 1.234 скупа података и 16.136 веза.



Слика 18. Облак „Отворени повезани подаци из области лингвистике”[142]

За сваку страницу на Википедији DBpedia дефинише URI како би се идентификовао ентитет или концепт који се описује на датој страници. Како DBpedia покрива широк опсег домена који се великим делом поклапа са слободно доступним скуповима података на вебу, велики број издавача успоставља RDF везе између својих скупова података са DBpedia-ом, стварајући од DBpedia-е средиште „мреже података”. Током процеса екстракције, структуриране информације са Википедије као што су информативна поља,

[142] Дијаграм је преузет са званичног сајта Linguistic Linked Open Data http://linguistic-lod.org/llod-cloud/.

категорије и везе ка различитим страницама стварају уређене RDF тројке (RDF tripl) и додају се бази знања као својства одговарајућег URI идентификатора (Mendes, Jakob and Bizer 2012, 1813). За процес екстракције информација и креирање RDF изјава користи се Оквир за екстракцију информација DBpedia-е (DBpedia Information Extraction Framework – DIEF). У (Lehmann et al. 2015) детаљно је описан процес екстракције структурираних информација са Википедије и стварање RDF изјава. Како би се хомогенизовао опис информација у бази знања развијена је онтологија и шема која омогућава мапирање својстава из информационог дела Википедије са произведеном онтологијом. Онтологија DBpedia-је организује знање у 685 класа које формирају хијерархију и описане су са 2.795 различитих својстава.[143]

На званичној веб страни DBpedia-је[144], према подацима из 2016. године, стоји да база садржи око 13 милијарди RDF изјава од којих је око 1,7 милијарди екстраховано из енглеске DBpedia-е, 6,6 милијарди из DBpedia-ја на другим језицима [145]и око 4,8 милијарди из Википодатака и Викимедија заједнице. Према истој статистици у енглеској верзији DBpedia-е описано је преко 6,5 милиона ентитета односно ресурса од којих скоро 5 милиона има сажетке односно краће описе, скоро 2 милиона географске координате и око 1,7 милиона сликовне приказе. Негде око 5,5 милиона ентитета је класификовано према онтологији која се користи на: особе (1,5 милиона), места (840.000), ауторска дела (496.000), организације (286.000), врсте (306.000), биљке (58.000) и болести (6.000). Поред 6,6 милион ресурса енглеска DBpedia-ја садржи и 1,7 милиона категорија преузетих и SKOS онтологије, 7,7 преусмерених веб страна, 296.000 страна са вишезначним одредницама и 1,7 милиона унутрашњих чворова.

Када је реч о DBpedia-ји на немачком и српском језику постоје назнаке да је рад на њима започет. У (Hellmann et al. 2012) приказан је почетак рада на немачкој DBpedia-и који је започет 2012. године. Скуп RDF изјава немачке DBpedia-е који потиче из 2016. године може се преузети са званичне веб стране DBpedia-је[146]. Процес локализације

[143] Онтологија DBpedia-е доступна је на адреси: http://wiki.dbpedia.org/services-resources/ontology
[144] DBpedia version: Statistics, https://wiki.dbpedia.org/develop/datasets/dbpedia-version-2016-10
[145] DBpedia-ја је почела да се развија на другим језицима после енглеског али су сада све те DBpedia-је спојене у једну.
[146] Скуп RDF изјава немачке DBpedia-је доступан је на http://downloads.dbpedia.org/3.7/de/

DBpedia-е на српски започет је пре неколико година. Како је српски језик једини у Европи који користи два званична писма у том тренутку се наишло на проблем транслитерације са ћирилице на латиницу односно DIEF који се користи за екстракцију информација DBpedia-е не подржава овакву транслитерацију. За потребе транслитерације развијен је пост-процесор који прво врши транслитерацију, а затим добијене резултате прослеђује регуларном DIEF процесору. У (Milošević et al. 2014) представљена је иницијатива за почетак рада на српској DBpedia-и. Међутим, овај пројекат никада није у пракси реализован до краја.

5.2.2 Википодаци

Википодаци (Wikidata)[147] су поред DBpedia-е још једна база знања општег типа. Википодаци представља вишејезичну базу знања коју је развила Викимедија фондација (Wikimedia Foundation) са циљем да омогући преузимање и чување структурираних података на више језика са Википедије у интероперабилном машински читљивом формату (Vrandečić and Krötzsch 2014, 78-79). База Википодаци је званично објављена октобра 2012. године и данас има преко 50 милиона записа које могу да уређују сви регистровани корисници. Записи у бази су бесплатно доступни свим корисницима под лиценцом CC0 1.0 Universal (CC0 1.0) Public Domain Dedication[148] и данас садрже структуриране податке о ентитетима који се преузимају не само са Википедије већ и са других Викимедија сервиса (Википедија, Википутовања, Викиизвори и други).

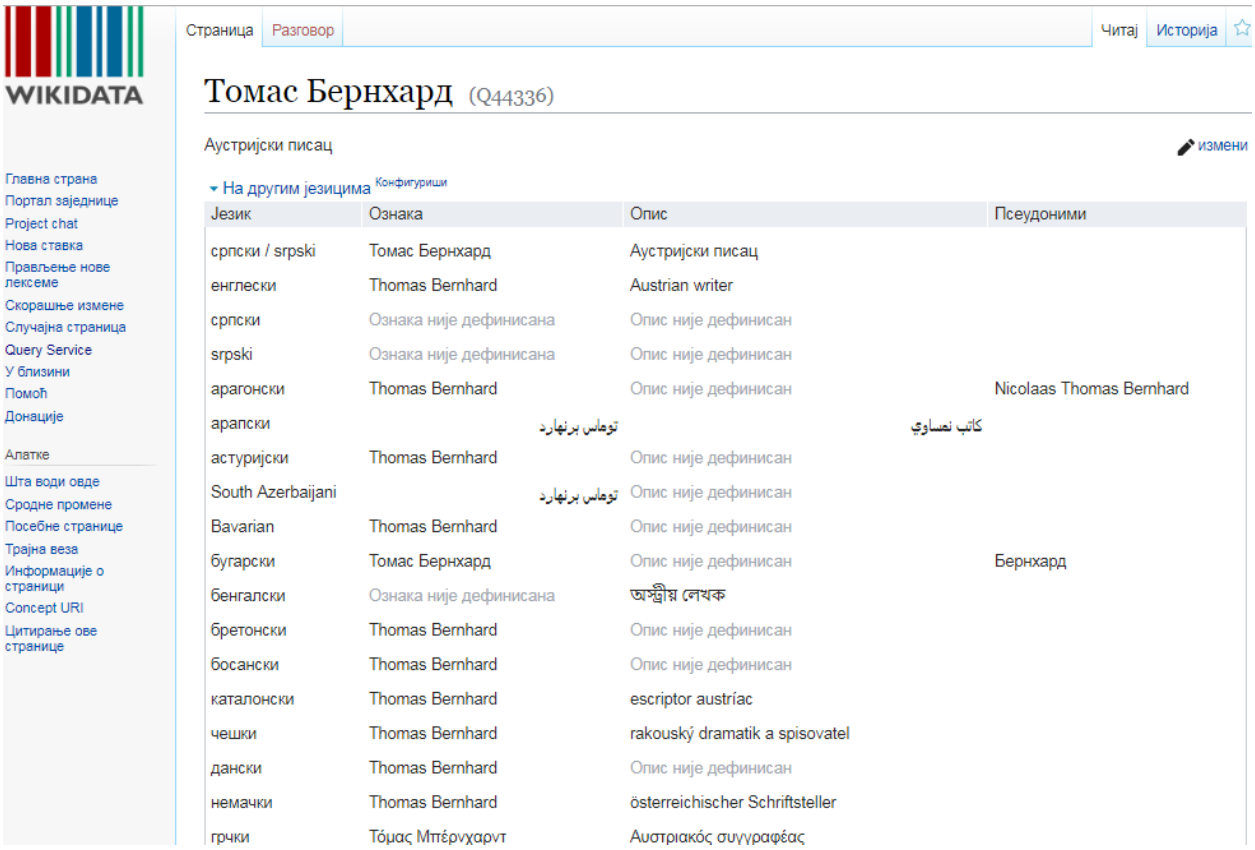
Све што се налази на Википодацима а садржи структуриране податке представља ентитет. Ентитети могу бити и субјекти (item), предикати односно својства (properties) и објекти (item). Једном речју ентитети могу бити сви делови изјаве па и сама изјава. Сваки ентитет има своју веб страну односно свој запис у Википодацима који садржи следеће делове: обележје (label), кратак опис (short description), листу варијантних облика имена на више језика (list of aliases), листу изјава (list of statements) и листу референтних веб страна (list of site links) (веб стране о субјекту који се описује на Википедији и другим интернет изворима). Прва три дела записа (обележје, кратак опис и листа варијантних

[147] Wikidata, https://www.wikidata.org/wiki/Wikidata:Main_Page
[148] CC0 1.0 Universal (CC0 1.0) Public Domain Dedication, https://creativecommons.org/publicdomain/zero/1.0/

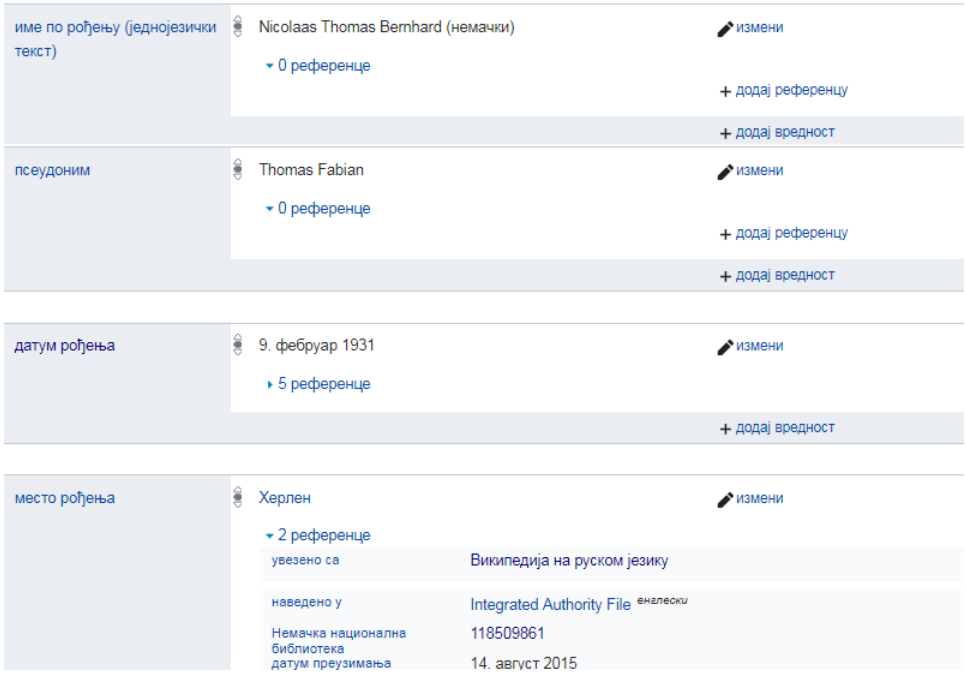
облика имена) познатији су и као *терми* (terms) односно текстуалне вредности које ближе описују ентитет и преко којих се најчешће врши претраживање. Сви ентитети имају јединствени идентификатор који се додељује аутоматски приликом креирања записа и не може се касније мењати, такозвани трајни идентификатор (permanent identifier). Субјекти и објекти имају јединствене идентификаторе који почињу са Q (на пример Q571 је *књига*), док предикати односно својства имају идентификаторе који почињу са P (на пример P366 је *намена, основна употреба субјекта*) (Erxleben et al. 2014, 51-52).

Структуру ентитета у Википодацима показаћемо на примеру ентитета „Thomas Barnhard". Ентитет „Thomas Barnhard" је у овом случају субјекат који се описује и запис има своју веб страну у бази Википодаци, https://www.wikidata.org/wiki/Q44336. Слика 19 приказује први део записа. Ознака ентитета на српском је „Томас Бернхард" уз коју стоји и идентификатор ентитета „Q44336". Испод ознаке стоји кратак опис „Аустријски писац" када је изабрана опција да подразумевани језик буде српски. Ако је изабрана опција да подразумевани језик буде енглески стоји кратак опис „Austrian writer" док за немачки стоји „österreichischer Schriftsteller". После кратког описа долази листа варијантних облика имена на различитим језицима (енглеском, немачком, бугарском, арапском, бенгалском и други).

Други део записа Википодатака чине изјаве. Слика 20 илуструје неке примере изјава за ентитет „Q44336" („Томас Бернхард"). Прва изјава односи се на „име по рођењу" и има следећу структуру: Q44336 → P1477 → „Nicolas Thomas Bernhard" у којој је субјекат Q44336 („Томас Бернхард"), предикат (својство) P1477 („име_по_рођењу"), а вредност предиката (својства) „Nicolas Thomas Bernhard". Друга изјава односи се на псеудоним: Q44336 → P742 → „Thomas Fabian", трећа на датум рођења: Q44336 → P569 → „9 фебруар 1931", а четврта на место рођења: Q44336 → P19 → Q9799. У овим примерима предикати (својства) имају своје јединствене идентификаторе и представљају ентитете на које се даље може референсати. На пример, предикат из прве изјаве у нашем примеру има следеће лексикализације: „име_по_рођењу"@sr, „birth name"@en, „Geburtsdatum"@de и тако даље**,** а све вредности се односе на ентитет који има идентификатор P1477.



Слика 19. Прва део записа „Томас Бернхард” у Википодацима



Слика 20. Примери изјава у Википодацима за Q44336 „Томас Бернхард”

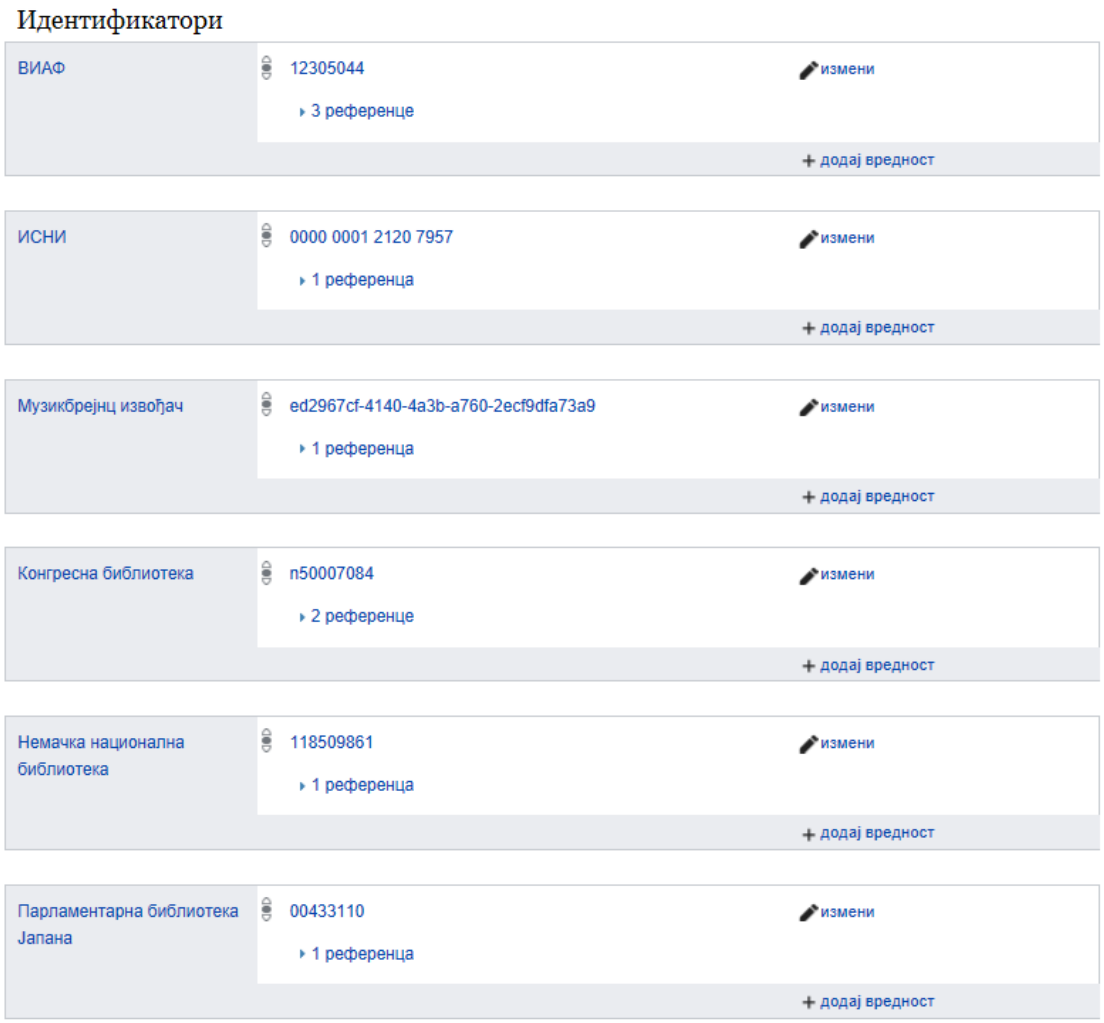
Вредности предиката (својстава) су објекти који могу бити текстуалне ознаке, као што су то „Nicolas Thomas Bernhard”, „Thomas Fabian”, „9 фебруар 1931” у првој, другој и трећој изјави, или могу бити други ентитети као што је то „Херпен” (Q9799) у четвртој изјави**.** Уз вредности објеката могу стајати и референце које, такође, имају форму својство-вредност. Референце могу имати облик класичног библиографског цитирања или могу бити веб адресе и скупови података на које се реферише од којих сваки може бити нови ентитет у Википодацима. У нашим примерима изјава референце видимо у четвртој изјави уз место рођења. Наведене су следеће референце: једна указује да је податак увезен са Википедијом на руском језику, друга указује да је податак наведен у Међународној нормативној датотеци VIAF и последња референца је на Нормативну датотеку Националне библиотеке Немачке GND. Уз референце стоји и датум преузимања података „14. август 2015”. У изјавама се такође дају информације да је он човек, да је писац, наводи се листа награда, образовање, занимања, чланство у политичкој странци, дела, место и датум смрти и слично.

Трећи део записа у Википодацима је листа референтних веб страна о субјекту. Слика 21 приказује неке примере из листе за ентитет Q44336 „Томас Бернхард”.

База Википодаци омогућава постављање SPARQL упита у кориснички оријентисаној сумеђи[149], а као помоћ корисницима су на располагању бројни типични примери упита који се могу искористити тако да корисници који не познају SPARQL језик могу да задовоље своје информационе потребе. Корисници имају могућност да поставе упит дефинишући параметре за претрагу у пољу „Filter”, док се у пољу „Show” дефинишу шта желе да добију као резултат претраге (листу веб страна, листу слика и друго). Током писања упита у левом оквиру веб стране аутоматски се исписују кључни параметри SPARQL упита.

Осим у табеларном облику, резултати се могу, у зависности од излазног скупа, приказати на више начина: као карта (map), линијски графикон (line chart), стубичасти дијаграм (bar chart), расути дијаграм (scatter chart), графикон површи (area chart), мапа дрвета (tree map), дрво (tree), временска скала (timeline), димензионе коцке (dimensions) или као RDF граф. Пример комплетног записа за Томаса Бернхарда дат је као Прилог 13 - Пример записа за Томаса Бернхарда у Википодацима / корисничко окружење .

[149] Wikidata Query Service, https://query.wikidata.org/



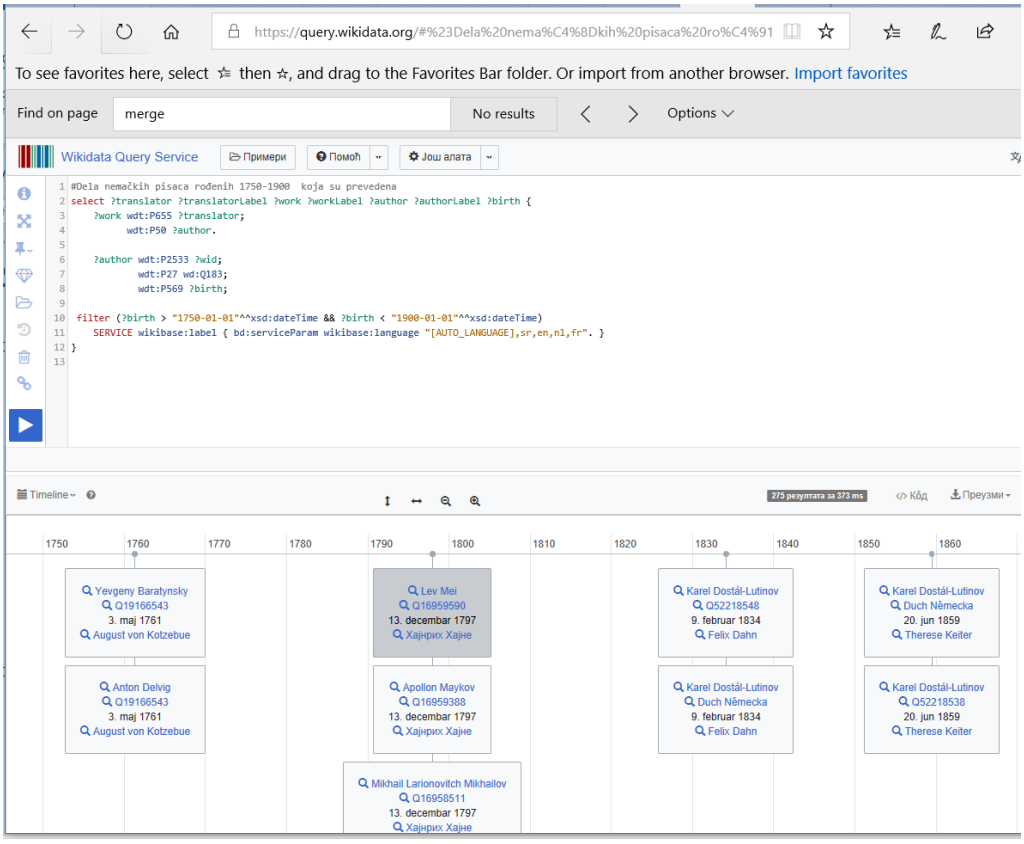
Слика 21. Примери из листе референтних веб страна за ентитет „Томас Бернхард”

Пример SPARQL упита у бази Википодаци за „Дела немачких писаца рођених 1750-1990 која су преведена” приказан је у наставку, а Слика 22 илуструје овај пример као временску скалу:

```
#Dela nemačkih pisaca rođenih 1750-1900  koja su prevedena
select ?translator ?translatorLabel ?work ?workLabel ?author ?authorLabel ?birth {
   ?work wdt:P655 ?translator;
         wdt:P50 ?author.

   ?author wdt:P2533 ?wid;
           wdt:P27 wd:Q183;
           wdt:P569 ?birth;

 filter (?birth > “1750-01-01”^^xsd:dateTime && ?birth < “1900-01-01”^^xsd:dateTime)
   SERVICE wikibase:label { bd:serviceParam wikibase:language “[AUTO_LANGUAGE],sr,en,nl,fr”. }}
```



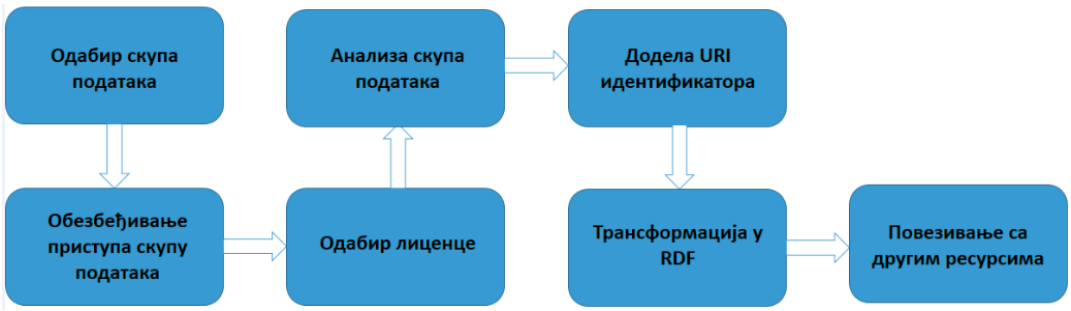
Слика 22. Упит у бази Википодаци „Дела немачких писаца рођених 1750-1990 која су преведена" са резултатима исписа на временској скали

5.2.3 Поступак повезивања података

Повезани подаци могу да садрже било који тип података, а сам LOD омогућава да свако може да постави и повеже своје податке са већ постојећим. На почетку развоја LODa припремљени су скупови података који су постављени у отворени приступ. Подаци су конвертовани у формат RDF и постављени на мрежу. Временом су се у пројекат укључиле велике институције као што су BBC[150], Thomson Reuters[151] и Конгресна библиотека (Library of Congress)[152] те су и библиотеке спознале предности ове иницијативе. Библиотеке садрже велики број структурираних метаподатака који су углавном у стандардним форматима који се затим лако конвертују, повезују и преносе. Бројне библиотечке

[150] British Broadcasting Corporation, http://www.bbc.com/
[151] Thomson Reuters, http://thomsonreuters.com/en.html
[152] Library of Congress, http://www.loc.gov/

организације у Европи и свету започеле су процесе повезивања података кроз различите пројекте (неке од њих су приказане у одељку 5.4 овог поглавља). Велики број електронских каталога, нормативних база података, база географских имена и слично постало је део овог пројекта, омогућавајући тако корисницима да са једног места приступе информацијама из различитих система, а институцијама да се међусобно повежу. Поступак повезивања података у оквиру „повезаних података" има неколико корака (Слика 23).



Слика 23. Процес повезивања података у систему „Отворени повезани подаци"

Први корак јесте одабир скупа података који ће постати део система „Отворени повезани подаци". На основу претходно дефинисаних захтева бира се један или више извора података. Извори података који се повезују у систем углавном припадају организацији која започиње рад на том пројекту и бирају се на основу потреба организације или на основу циља који жели да се постигне пројектом. Такође, организација може бити заинтересована да своје податке допуни изворима података које поседује нека друга организација што се разрешава у следећем кораку.

У другом кораку треба омогућити приступ скупу података. Подаци који се увезују у систем „Отворени повезани подаци" морају бити у јавном домену (public domain) односно јавно доступни. Ако су извори података у власништву организације која реализује пројекат на самој организацији је да их учини јавно доступним те да им тако омогући отворен приступ. Ако се извори података допуњују изворима података које поседује друга организација, а који су јавно доступни, њима је лако приступити. Међутим, ако извори података нису у јавном домену мора се тражити се одобрење од одговорне особе

односно установе или организације прво да би се приступило изворима података, а затим и да би се они ставили у јавни домен.

Трећи корак подразумева одабир лиценце под чијим ће условима бити омогућен приступ скуповима података и њихово коришћење. Како би се избегли правни конфликти неопходно је прво одредити носиоце ауторских права, а затим се одлучити за лиценцу којом ће се регулисати права приступа и коришћења података.

Четврти корак подразумева анализу скупа података у смислу њихове структуре и организације. Прво се анализирају карактеристике података као што су квантитет или распон вредности, а затим се анализира њихова структура и идентификују везе између њих. У овом кораку утврђује се схема за опис података. У неким случајевима схема за опис датог скупа података већ постоји и довољно је само анализирати је, док је у другим случајевима неопходно дефинисати нову. Овако припремљена схема за опис касније је погодна за валидацију у процесу трансформације у модел података RDF.

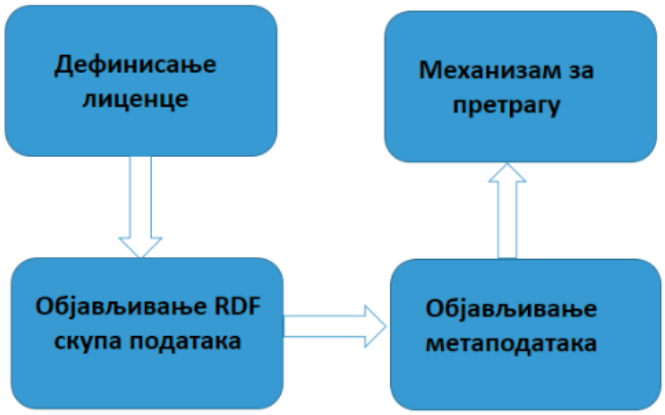
У петом кораку одређују се URI идентификатори за одабране скупове података. Прво се бира форма URI идентификатора, затим се бирају домен и путања која ће бити у њиховој основи, док трећи корак подразумева одабир обрасца URI идентификатора за класе и својства у оквиру онтологије.

Шести корак у процесу припреме скупа података за LOD је трансформација података у модел података RDF. Прво је важно да се одабере начин записивања RDF модела података (детаљније је објашњено у одељку 5.1.2 овог поглавља), а након тога се бира алат за трансформацију података на основу формата у коме се подаци налазе и потреба самог процеса трансформације што подразумева дефинисање процеса мапирања између података и онтологије. На крају се ради евалуација добијеног RDF записа.

Последњи корак је повезивање добијеног скупа података са релевантним скуповима података. Прво се дефинишу класе чије инстанце могу бити предмети повезивања, док је други корак идентификовање скупа података са којима је могуће успоставити повезивање.

5.2.4 Поступак објављивања повезаних података на веб

Поред повезивања скупова података неопходно је да су ти подаци објављени и доступни преко веба. Процес објављивања података, такође, подразумева неколико корака (Слика 24).



Слика 24. Процес објављивања скупа повезаних података

Први корак подразумева утврђивање правне сагласности односно лиценце којом се регулишу права приступа и коришћење скупа података у облаку. У случају да је претходно дефинисана лиценца непотпуна (у трећем кораку процеса повезивања података) или је дошло до неких промена у погледу лиценцирања бира се нова лиценца која по својим критеријумима може да регулише права приступа и коришћења у складу за захтевима организације која укључује податке у отворене повезане податке.

Други корак подразумева објављивање RDF скупа података на вебу према принципима иницијативе „отворени повезани подаци". Подаци у RDF-у се смештају у трајни репозиторијум који је доступан корисницима. Преко овог репозиторијума корисници имају могућност да приступе подацима и да их претражују постављањем релевантних упита за претрагу. Такође, у овом кораку се формира механизам за приступ подацима преко веба, утврђује се HTTP протокол који тумачи дефинисане URI идентификаторе и SPARQL упите за претрагу.

У следећем кораку објављују се метаподаци који описују добијени скуп података у RDF-у и добијену онтологију. Метаподаци се објављују у машински разумљивом облику, док се пропратна документација припрема и објављује за кориснике.

У последњем кораку припрема се механизам за претрагу добијеног скупа података прилагођен и рачунарима и корисницима.

5.2.5 Системи за организацију знања

Системи за организацију знања (Knowledge Organisation System - KOS) су ресурси развијени ради лакше организације колекција и докумената у различитим областима људског знања и проналажења информација у њима, али и ради дефинисања унифицираних облика појмова или термина који се користе у различитим областима људског знања. Ови системи садрже попише појмова или термина који могу бити организовани у комплексне хијерархијске структуре са развијеним системом упутница којима се указује на варијантни или сродни облик појма, као и могућношћу детаљног описа појмова. Према (Davenport and Prusak 1998, 5) „Знање је флуидна мешавина уоквиреног искуства, вредности, контекстуалних информација и експертског увида, која обезбеђује оквир за процену и укључивање нових искустава и информација. Изворно настаје и користи се у главама зналаца. Знање је у организацијама похрањено не само у документима или репозиторијумима већ исто тако и у организационим рутинама, процесима, праксама и нормама”[153]. Једном речју знање представља скуп чињеница, информација и вештина стечених учењем или искуством у циљу теоретског или писменог разумевања и решавања проблема које се похрањује у системе за организацију знања које развијају и одржавају различите институције. У зависности од садржаја, структуре и начина организације људског знања системи за организацију знања се могу поделити у три веће групе:

[153] Knowledge is a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information. It originates and is applied in the minds of knowers. In organizations, it often becomes embedded not only in documents or repositories but also in organizational routines, processes, practices, and norms.

1. терминолошке листе (нормативне датотеке као што су, на пример, VIAF, GND, LCNAF (детаљније објашњени у поглављу 6 одељак 6.5.1), речници и слично),
2. класификације и категорије (систем предметних одредница, шеме за класификацију и категоризацију и таксономије),
3. релационе листе (тезауруси, семантичке мреже, на пример, Ворднет (детаљније објашњен у поглављу 3 одељак 3.3.4) и онтологије).

Да би се омогућила јединствена употреба термина и појмова из различитих области људског знања, многи системи за организацију знања прерастају у контролисане речнике. За контролисане речнике врло често се данас користи и термин „таксономија". У савременом свету управљања информацијама термин „таксономија" се у ужем смислу користи да означи хијерархијски систем за класификацију и категоризацију, док се у ширем смислу односи на било који облик организације концепта знања (Hedden 2016, 34). Иако стручњаци из многих области знања нису у почетку прихватали термин сматрајући га сувише двосмисленим, он је ипак почео широко да се користи и у ове сврхе. Како су контролисани речници развијени за различите научне области и области људског знања, на пример, нормативне датотеке у библиотекама, речници у лингвистици, тезауруси у различитим научним областима и друго, тако су и у семантичком вебу развијени контролисани речници који дефинишу употребу концепата и термина за анотацију структурираних података. Један такав речник је и schema.org[154]. Контролисани речник Schema.org, који су 2011. године представили су Google, Yahoo! и Bing, садржи око 300 машински-читљивих дефиниција најфреквентнијих концепата као што су „особа", „дело", „догађај", „организација", „место" и многе друге (Sikos 2015, 32). На пример, термин „Library"[155] може да се опише уз помоћ својстава којима се дефинише радно време библиотеке у смислу колико дана у недељи је библиотека отворена и која је сатница по дану, услуге библиотеке и начин плаћања услуга, услови коришћења услуга, административни подаци као што су адреса, телефон, e-mail адреса и многе друге.

За описивање структуре система за организацију знања развијен је стандард Једноставни систем за организацију знања (Simple Knowledge Organization System - SKOS).

[154] schema.org, https://schema.org/
[155] schema.org: Library, https://schema.org/Library

SKOS се интензивно развија од 2005. године са циљем лакшег објављивања и коришћења система за организацију знања у концепту „повезани подаци”. Поред основне структуре, SKOS омогућава и опис садржаја система за организацију знања укључујући и концепте „надређени термин”, „подређени термин”, као и „повезани термин” и део је породице стандарда на којима се заснива семантички веб. Заснован је на RDF и RDFS који му омогућавају да структуру система за организацију знања представи у форми графа (Miles, Matthews and Wilson 2005, 3).

Библиотеке располажу великом количином добро структурираних и богатих података. Највећи део система за организацију знања, а пре свега контролисани речници са великим бројем структурираних података, развија се вековима у библиотечкој заједници како би се направио неки унифицирани систем за индексирања људског знања које би уједно било претраживо са једног места. Према препорукама W3 конзорцијума из 2009. године, SKOS је одмах примењен на неке од контролисаних речника у библиотекама као што су Француска нормативна датотека и Нормативна датотека предметних одредница Конгресне библиотеке. Интензивно коришћење контролисаних речника у библиотекама представља круцијалну ствар у синергији веба и библиотечког културног наслеђа. Године 2010. на 76. IFLA-ином конгресу представљене су смернице за укључивање контролисаних речника у семантички веб које су описане у (Vatant 2010).

5.3 Библиотеке и семантички веб

Примена технологија семантичког веба и „отворени повезани подаци” отварају врата библиотекама да метаподатке и информације које у њиховим системима постоје организују и објаве на потпуно нови начин и повежу их са свим релевантним ресурсима на вебу. Библиотеке су идеално место за примену технологија семантичког веба због следећих предности које имају у односу на друге заједнице:

1. *Квалитет записа у смислу конзистентности и потпуности описа.* Записи се раде по добро утврђеним међународно прихваћеним стандардима за каталогизацију и класификацију чиме се постиже доследност у опису свих врста грађе.

2. *Бројност записа.* Велики је број записа који су у јединственом формату и аутоматски се могу, уз помоћ различитих алата, конвертовати у RDF и интегрисати у окружење „ отворени повезани подаци”.
3. *Постојање великог броја контролисаних речника.* У библиотекама постоји добро развијен систем контролисаних речника као што су нормативне датотеке којима се регулише унифицирана употреба различитих ентитета (имена аутора, имена корпоративних тела, догађаја и слично).

Потреба да се опишу материјали у библиотекама који се вековима чувају створила је библиотечке каталоге који су кроз историју мењали облик и форму. Каталози које данас познајемо налазе се на вебу и засновани су на машински-читљивим форматима као што је MARC, а метаподаци се у каталошким записима наводе према тачно утврђеним правилима и стандардима као што су ISBD или AACR2. Појавом MARC формата шездесетих година 20. века омогућено је да каталошки листићи једне библиотеке постану машински читљиви, а развојем јавно доступних каталога на мрежи (Online Public Access Catalog - OPAC) библиотечки каталози постали су приступачни корисницима широм света. Са становишта приступа они су стално доступни – 24 сата, 7 дана у недељи - што је представљало велику технолошку револуцију у односу на форму лисних каталога који су се могли прегледати само физичким одласком у библиотеку. Такође, временом је развијен механизам за размену записа између институција што је довело до смањења дупликата. Прецизније речено, библиотеке које данас раде са електронским каталозима имају могућност да кроз систем за каталогизацију који користе преузимају записе које су креирале друге библиотеке како на националном тако и на међународном нивоу. На пример, библиотеке у Србији које користе систем за каталогизацију COBISS (детаљније објашњено у поглављу 4 одељак 4.1.2) имају могућност да међусобно преузимају записе на националном нивоу (кроз узајамну базу COBISS.SR), али и записе које је креирало преко хиљаду библиотека из Словеније, Босне и Херцеговине, Македоније, Црне Горе, Албаније и Бугарске (кроз узајамну базу COBISS.Net). Поред тога, кроз систем COBISS библиотекама у Србији је омогућено преузимање записа из каталога Конгресне библиотеке и

јединственог светског каталога WorldCat што доста олакшава посао када је у питању каталогизација стране књиге.

И за библиотечке каталоге у електронском облику израда каталошких записа подразумева креирање статичних документе који садрже структуриране метаподатке који се креирају у складу са националним каталошким правилницима појединачних држава. Једном речју они личе на лисне каталоге у којима записи представљају документе саме за себе. Са друге стране, претрага каталога је могућа од стране свих заинтересованих корисника у било које доба дана, али приступом појединачним националним библиографским центрима и то у форми предефинисаних упита предвиђених системом који се користи. Према (Coyle 2012, 56-57) „усмеравање библиотека на концепт 'отворени повезани подаци' не представља само потребу да се библиотечки каталози модернизују већ и потребу да се библиотечки каталози трансформишу из одвојене, затворене базе података у системе са комплексне технологијама које корисници користе у истраживањима".

Концептуално гледано овакви формати, иако међународно прихваћени, нису предвиђени да истакну виши ниво података односно њихово значење, везе са другим записима и ресурсима на вебу и омогуће интеракцију са корисником што су основни постулати семантичког веба. Такође, постављање сложених упита попут оних који се користе у релационим базама података постављањем вишеструких логичких услова није могућа. „Метаподаци обогаћени контекстуалним и релевантним везама омогућавају корисницима да се слободно „крећу" између библиотечких база података и екстерних информационих добављача као што су друге библиотеке и машине за претраживање. Глобалним и јединственим идентификовањем ентитета (дела, људи, места, догађаји), елемената за израду метаподатака и њихових својстава (аутор, наслов, предмет, везе) и одговарајућих вредности (инстанце), „отворени повезани подаци" нуде мноштво начина за обогаћивање информационих објеката метаподацима који могу да олакшају приступ информацијама и побољшају искуство корисника у коришћењу дигиталних библиотека" (Alemu et al. 2012, 562).

Године 2005. у (Dunsire 2005) представљени су главни проблеми објављивања библиографских података као „отворени повезани подаци”:

1. *Непостојање одговарајућег RDF речника*. Формати за метаподатке које користе библиотеке немају своју RDF репрезентацију.
2. *Некомпатибилност са мрежном технологијом и мањкавост података.* Метаподаци нису означени URI идентификаторима, као ни записи у целини што онемогућава да се као такве повлаче у систем „отворених повезаних података”.
3. *Организациони проблеми.* Недостатак отворених лиценци које регулишу употребу метаподатака у “повезане отворене податке”. Различите заједнице још увек немају стабилна мапирања својих метаподатака са моделом података RDF. Не постоји сагласност око преузимања и коришћења употребљивих метаподатака у RDF.
4. *Технички проблеми.* Недостатак отворених лиценци које би обухватиле сумеђе за програмирање апликација (Application Programming Interface - API), стандарде за метаподатке и клијентски софтвер у окружењу повезаних података.

Каталошки записи имају добру основу да постану део семантичког веба у погледу броја и структуре података, али су потребне техничке и концептуалне промене у теорији и начину израде самих записа. Из угла значења, за конверзију постојећих записа у „повезане отворене податке” главни проблем је одабир речника (онтологије) за опис. За структурирање и објаву неког скупа података као „отворени повезани подаци” постоје две могућности: конструисати сопствене речнике (онтологије) или користити постојеће. Библиотеке имају више могућности. Једна могућност је конверзија постојећих MARC записа у неке од стандарда за метаподатке поменуте у четвртом поглављу који користе XML синтаксу за опис. Системи за каталогизацију данас имају могућност да већ урађене записе у неком од MARC формата аутоматски извезу у неке од стандарда за метаподатке што је велика предност за даљу трансформацију у RDF онтологију. Друга могућност је да се записи у неком од MARC формата конвертују у UNIMARC који може да се мапира са ISBD стандардом у RDF репрезентацији која се налази у Отвореном регистру RDF речника

и елемената метаподатака (Open Metadata Registry - OMR)[156]. Трећа могућност је да се уместо MARC формата за каталогизацију користе концептуални модели који су развијени у библиотечким заједницама као што су Функционални захтев за израду библиографских записа (Functional Requirements for Bibliographic Records - FRBR) или Опис и приступ ресурсима (Resource Description and Access - RDA)[157] јер се могу моделирати онтологијама.

Од постављања главних проблема у (Dunsire 2005) до данас неке од већих библиотека у свету су своје каталошке базе података трансформисале у складу са принципима семантичког веба конвертујући записе из MARC формата у RDF онтологију. На пример, Конгресна библиотека је развила Оквир за библиографски опис (Bibliographic Framework - BIBFRAME) као нови модел података заснован на технологијама семантичког веба који омогућава израду, преузимање и размену библиографских метаподатака, при чему је омогућена конверзија записа из формата MARC21. Модел BIBFRAME је детаљније представљен у следећем одељку.

Поред додељивања такозваних формалних, односно библиографских метаподатака, у библиотечком свету постоји добро утврђена пракса предметизације и класификације, којом се одређена јединица описује предметним одредницама односно садржајно-описним кључним речима или се смешта у неку од класификационих шема. За предметизацију се данас широко користе нормативне датотеке предметних одредница израђене углавном према тезаурусима из одређене области. Системи нормативних датотека у библиотекама су „резултат нормативне контроле која доприноси Универзалној библиографској контроли. То је библиографска база јединствено унетих података, као што су имена аутора, чиме се омогућава да у узајамном каталогу, на пример, сва дела једног аутора буду окупљена око унифицирано унете одреднице, чиме се пружа прилика за мерење фреквенције коришћења, али и чувања ауторских права” (Вранеш и Марковић 2008, 235). Према (Reitz 2019) „нормативна датотека је списак усвојених облика одредница које се користе у каталогу библиотеке или датотеке библиографских записа, а одржавање датотеке омогућава да се одредница доследно примењује и да се нови облици додају колекцији. Посебне нормативне датотеке се воде за имена аутора,

[156] Open Metadata Registry, http://metadataregistry.org/
[157] Resource Description and Access, https://www.oclc.org/rda/about.en.html

јединствене наслове, наслове колекција и предметне одреднице”[158]. Неке од нормативних датотека које су данас део облака су: Међународна виртуелна нормативна датотека (Virtual International Authority File - VIAF), Нормативна датотека имена Конгресне библиотеке (Library of Congress Name Authority File - LCNAF), Нормативна датотека предметних одредница Конгресне библиотеке (Library of Congress Subject Headings - LCSH), Нормативна датотека Националне библиотеке Немачке (Gemeinsame Normdatei - GND), Нормативна датотеке Француске националне библиотеке (Répertoire d'autorité-matière encyclopédique et alphabétique unifié - RAMEAU)[159], Нормативна датотека геополитичких појмова (GeoNames)[160] и многе друге. Неке од нормативних датотека искоришћене су и за увезивање ентитета из корпуса који је предмет ове докторске дисертације што је детаљније објашњено у поглављу 6 одељак 6.5.1.

Када је реч о системима за класификацију, у библиотекама се користе неке од класификационих шема као што су Универзална децимална класификација (УДК) или Дјуијева децимална класификација (ДДК). Ове класификационе шеме представљају децималне нумеричке системе за изражавање предмета грађе која се обрађује и тиме се одређује њено место у систему људског знања. Нумерички систем у класификационим шемама погодан је за међународну размену података. Ови системи су, такође, погодни за примену технологија семантичког веба. Неке верзије Дјуијеве децималне класификације, као што је њено немачко издање, већ постоје у систему „отворени повезани подаци“. Универзална децимална класификација се полако припрема да постане део „отворених повезаних података”.

У погледу дигиталних библиотека велике су предности за примену технологија семантичког веба. Дигиталне библиотеке садрже различите врсте информација и различите ресурсе у дигиталном формату. Многе од њих представљају дигитализовану традиционалну библиотеку чији су аналогни извори информација дигитализовањем постали доступни ширем кругу корисника. Дигиталне библиотеке омогућавају библиотекарима обраду, дисеминацију и складиштење различитих врста информација у

[158] Превод дефиниције у (Тртовац 2016, 99)
[159] Répertoire d'autorité-matière encyclopédique et alphabétique unifié, http://rameau.bnf.fr/
[160] GeoNames, https://www.geonames.org/

дигиталном формату, а корисницима претрагу и анализу садржаја на више начина. Технологије семантичког веба омогућавају да се дигитални библиотечки репозиторијуми знања, применом утврђених стандарда и модела података, трансформишу у онтологије и повежу са сродним ресурсима у другим репозиторијумима на вебу. Такође, технологије семантичког веба примењују се и на корисничке сумеђе и интеракцију корисник-рачунар (приказивање информација, визуелизација и коришћење великог броја информационих колекција), профилисање корисника (узимајући у обзир свеукупну информациону сферу), персонализацију (балансирање између појединачних и колективних персонализација) и интеракцију корисника.

Коришћење концепта „отворени повезани подаци" логичан је след развоја у подручју интероперабилности. Примена технологија семантичког веба не подстиче нужно промену парадигме у подручју библиотечког пословања, већ нуди механизам за примену нових технолошких могућности које омогућавају размену информација и повезивање колекција са релевантним ресурсима било где на вебу, а корисници имају могућност да са једног места сагледају информације са више аспеката и у више расположивих ресурса. Данашњи корисници очекују да током рада на истраживању користе комплексне системе дигиталних колекција са могућношћу једноставног постављања упита за претрагу коришћењем библиотечки независног софтвера и хардвера (Sure and Studer 2005, 192). Један овакав систем представљен је и у овој дисертацији на примеру двојезичне колекције упарених текстова што је детаљније објашњено у поглављу 6.

До данас велики број библиотека успео је да своје ресурсе интегрише у концепт „отворени повезани подаци". Најбољи пример је Конгресна библиотека која је своје ресурсе структурирала према принципима семантичког веба и створила сервис ID.LOC.GOV одакле корисници могу да приступе свим онтологијама, контролисаним речницима и другим ресурсима које Конгресна библиотека развија у складу са овим принципима.

Када је реч о библиотекама у Србији не може се рећи да су добром положају. Са становишта библиографског описа, највећи број библиотека ради у систему COBISS који омогућава да се записи раде у формату COMARC и постоји могућност извоза записа у неке

формате за метаподатке (детаљније је објашњено у поглављу 4 одељак 4.1.2). Из овако припремљених записа даље се могу генерисати записи у RDF моделу. Међутим, записи урађени у систему COBISS представљају статичне документе и метаподацима нису додељена значења. Један од разлога је непостојање нормативних датотека било које врсте на националном нивоу. Нормативна датотека личних имена CONOR (Савић 2017) имплементирана је у рад система COBISS у априлу 2019. године. CONOR омогућава унифицирану контролу имена аутора у систему COBISS па се на овај начин постављају се темељи за будуће увезивање ентитета личних имена са неким од међународних нормативних датотека овог типа које су већ поменуте и налазе у систему отворених повезаних података. Када је реч о предметизацији односно додељивању предметних одредница нормативна датотека предметних одредница још увек није у изради. Највећи проблем је непостојање тезауруса из различитих области знања на основу којих би таква нормативна датотека могла да буде урађена. Што се тиче класификационих шема користи се систем УДК класификације која се на међународном нивоу интензивно припрема за облак „отворени повезани подаци". Са оваквим стањем библиотеке у Србији имају основа да размишљају о неком будућем раду на питању семантичког увезивања података из својих ресурса са релевантним ресурсима на вебу.

Међутим, иако још увек не постоје довољно развијени ресурси за примене технологија семантичког веба на националном нивоу, библиотеке у Србији могу да користе ресурсе које су већ у облаку и повежу податке из својих извора и своје ресурсе са њима. Поред тога, окружење „отворени повезани подаци" је отвореног типа са дефинисаним и објављеним упутствима за генерисање различитих скупова података што библиотекама даје отворен пут да саме припреме материјал који ће на овај начин постати део већег облака. На основу материјала корпуса који је предмет ове дисертације припремљен је двојезични речник општег типа који је генерисан као скуп отворених повезаних података и постављен у облак. Сам речник, упутства која су коришћена за генерисање и резултати генерисања детаљније су објашњени у поглављу 6 одељак 6.5.3.

5.4 Иницијативе и пројекти засновани на принципима семантичког веба

У овом одељку анализираћемо неке иницијативе и пројекте који се заснивају на принципима семантичког веба, њихове моделе података, начине претраге и приступ информацијама и испитаћемо које су предности овакве структуре за различите заједнице.

5.4.1 Европеана

Водећи пример семантичког веба у Европи јесте Европеана. Европеана је портал који омогућава приступ до преко 50 милиона дигиталних јединица из европских библиотека, музеја, архива и аудио-визуелних колекција. Прототип портала представљен је 20. новембра 2008. године, а пројекат је покренут са циљем да се корисницима омогући бесплатан приступ разноврсним дигиталним садржајима и представи Европи и свету европска културна баштина (Chambers and Schallier 2010, 106). Задатак Европеане је да преко агрегатора од сарадника прикупи метаподатке о дигиталним објектима који се чувају у њиховим дигиталним репозиторијумима. Цео концепт заснован је на отвореном приступу, али се свакако подржавају и поштују права интелектуалне својине. Претрага, коришћење метаподатака и преглед дигиталних објеката преко портала Европеана потпуно су бесплатни и доступни свим корисницима који су повезани на интернет, а приступ до дигиталних објеката регулисан је правним оквирима који важе у земљама где се дигитални објекти налазе, као и политиком институција које су власници објеката. Поред тога, „Европеана користи стандарде, интероперабилне и машински читљиве лиценце, како би омогућила повезивање података са другим апликацијама и сервисима. Лиценце на јасан начин одређују шта људи или роботи-претраживачи могу или не могу да раде са метаподацима и садржајима којима приступају" (Филипи-Матутиновић 2011, 4). Тако су, права и обавезе приликом коришћења садржаја доступних преко портала регулисани *Лиценцним оквиром пројекта Европеана* којим се регулише повезивање и слободно коришћење метаподатака и садржаја.[161]

[161] Лиценцни оквир Европеане је доступан на http://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20Licensing%20Framework.pdf.

Са Европеаном тренутно сарађује преко 3500 институција културе из 39 европских земаља које преко агрегатора достављају описе дигиталних објеката у виду метаподатака. Сви добављачи садржаја и агрегатори, приликом достављања података, усклађују се са техничким захтевима Европеане. Европеана прикупља, складишти и индексира метаподатке у централном индексу, док на веб страни добављачи или агрегатора постоји репозиторијум у коме су дигитални објекти смештени и где се они могу прегледати и преузети. За поступак прикупљања и индексирања користи се протокол OAI-PMH (Open Archives Initiative – Protocol for Metadata Harvesting)[162], што значи да је пожељно да партнери успоставе OAI-PMH репозиторијум који садржи одговарајуће метаподатке како би се олакшало њихово преузимање.

Метаподаци у оквиру Европеане описују се коришћењем Модела података Европеане (Europeana Data Model - EDM). EDM (Definition of Europeana Data Model elements 2016) (Андоновски и Недељков 2013) је модел заснован на принципима семантичког веба и обухвата стандарде који се користе за креирање метаподатака у библиотекама, музејима и архивима (METS (детаљније описано у поглављу 4 одељак 0), EAD[163], LIDO[164]), али и различите информатичке онтологије (DC, SKOS, FOAF) и нормативне базе података. Такође, у модел је уграђен и формат Семантички елементи Европеане (Europeana Semantic Elements - ESE) (Дакић и Андоновски 2012) који је на почетку развоја Европеане дефинисан за опис метаподатака.

Модел EDM је дизајниран тако да може да прихвати пун распон метаподатака који потичу из различитих домена, али и да подржи допуњавање постојећих података новим семантичким везама са другим културно-историјским подацима и ресурсима које је могуће пронаћи на интернету. Теоретски, модел EDM би требало да омогући да се овај процес аутоматизује и да се везе, које претходно нису биле експлицитно приказане, појаве пред очима истраживача који преко свог рачунара приступа једном једином

Превод Лиценцног оквира Европеане на српски објављен је у часопису Инфотека год. 12, бр.2 (2011) (Филипи Матутиновић 2011).

[162] Open Archives Initiative – Protocol for Metadata Harvesting, https://www.openarchives.org/pmh/

[163] Encoded Archival Description, XML формат за опис архивске грађе, https://www.loc.gov/ead/

[164] Lightweight Information Describing Objects, XML шема за опис музејских објеката, http://network.icom.museum/cidoc/working-groups/lido/lido-technical/specification/

порталу (Гардашевић 2013а, 89). У основи модела EDM је RDF који омогућава преузимање и комбиновање различитих вокабулара и чување оригиналних података. Како је структура модела прилагођена RDF-у она садржи одређен број елемената који су подељени на класе и својства, а који се међу собом повезују преко URI идентификатора (Dakić and Andonovski 2013, 12-13).

Систем Еуропеане, заснивајући се на принципима семантичког веба и LOD модела, поред шема за опис метаподатака, обједињује и различите системе контролисаних речника, а најзначајнији су записи нормативних података у системима VIAF и GND, затим нормативну датотеку геополитичких појмова на различитим језицима GeoNames, у којој су забележени подаци о надморским висинама, дубинама, становништву и слично, као и нормативну датотека предметних одредница Конгресне библиотеке (Library of Congress Subject Headings - LCSH)[165]. Пример једног записа у моделу EDM дат је као Прилог 14 - Запис у Еуропеани у EDM моделу / корисничко окружење.

5.4.2 Дигитална народна библиотека Америке

Дигитална народна библиотека Америке (Digital Public Library of America - DPLA)[166], покренута 2010. и званично представљена 18. априла 2013. године, је иницијатива која омогућава приступ до скоро 17 милиона дигиталних објеката из преко 3200 колекција из 18 институција-партнера у САД. Иницијатива је покренута са циљем формирања јединствене Дигиталне библиотеке Америке, слично Еуропеани у Европи. Окупља библиотеке, музеје и архиве Америке на једном месту, омогућавајући претрагу метаподатака дигиталних објеката који су смештени у дигиталне репозиторијуме матичних институција (Guthro 2013, 127).

Иницијатива DPLA има два главна производа. Први производ ове иницијативе је портал који омогућава претрагу и приступ дигиталним објектима из поменутих институција културе Америке. Други производ је платформа која са једне стране омогућава отворени приступ подацима и сервисима, а са друге програмерима, истраживачима и корисницима пружа окружење за развој нових платформи за учење,

[165] Library of Congress Subject Headings, http://id.loc.gov/authorities/subjects.html
[166] Digital Public Library of America, https://dp.la/

алата за истраживање и разноврсних апликација применом сумеђе за програмирање апликација API (Mitchell 2013, 34).

Позадинска структура DPLA заснована је на принципима семантичког веба. Метаподаци се, као и код Еуропеане, прикупљају по систему агрегације за шта се користе протоколи Иницијативе за отворене архиве OAI (Open Archives Initiative), OAI-ORE (Object Reuse and Exchange)[167] и OAI-PMH (Protocol for Metadata Harvesting). За креирање метаподатака изграђен је по угледу на EDM који се користи у Еуропеани модел DPLA профила за апликацију метаподатака (DPLA Metadata Application Profile - DPLA MAP). У основи модела су формати JSON-LD[168], који је неопходан за реализацију платформе API, и RDF, који омогућава преузимање и комбиновање различитих вокабулара и чување оригиналних података.

DPLA MAP је изграђен на принципима EDM модела у чијој је основи RDF. Следећи принципе RDF-а сам опис дигиталног објекта састоји се од изјава у форми уређених тројки које добијају одговарајуће URI идентификаторе и формирају структуру графа. За опис се користе елементи који су по принципу EDM-а разврстани на класе и својства који су између себе повезани преко датих URI идентификатора. У DPLA MAP интегрисан је формат Даблинско језгро за описне метаподатке, нека својства из EDM модела за опис поступака агрегације и вокабулара, различите информатичке онтологије као што су нормативне датотеке, али и стандарди за метаподатке као што су MODS, METS-wrapped MODS, MARC XML, VRA Core[169], CDWA[170], CIDOC CRM[171]. Достављени метаподаци у неком од ових формата лако се мапирају са DPLA MAP, док са друге стране заснованост на EDM-у значи

[167] Object Reuse and Exchange, https://www.openarchives.org/ore/
[168] Формат је заснован на већ постојећем JSON формату. Креиран је као формат за лакше реализације концепта Linked Data. Доступно на: http://json-ld.org/
[169] Стандард за опис слика и радова из уметности и културе. Доступно на: https://www.loc.gov/standards/vracore/
[170] Категорије за опис уметничких дела (Categories for the Description of Works of Art), стандард за опис радова из области уметности, архитектуре, групе и колекције радова, повезаних слика. Доступно на: http://www.getty.edu/research/publications/electronic_publications/cdwa/
[171] Концептуални референсни модел међународног комитета за документацију (International Committee for Documentation Conceptual Reference Model), међународни стандард којим се дефинише онтологија за опис концепата и информација и веза између њих из области културног наслеђа и музеологије. Доступно на: http://www.cidoc-crm.org/

да метаподаци који могу да се мапирају са EDM-ом могу да се мапирају и са DPLA MAP моделом (An introduction to DPLA Metadata Model 2015).

5.4.3 Оквир за библиографски опис

Оквир за библиографски опис (Bibliographic Framework - BIBFRAME)[172] је иницијатива коју развија Конгресна библиотека са циљем преузимања и размене библиографских метаподатака о библиотечкој грађи, применом технологије семантичког веба. Идеја о стварању новог модела у погледу библиографског описа библиотечке грађе потиче још из 2008. године када је представљена у званичном извештају библиотечке заједнице о будућности библиографске контроле (Mitchell 2013, 27). На основу овог извештаја Конгресна библиотека је у сарадњи са организацијом Zepheria 2012. године представила нови модел метаподатака BIBFRAME за библиографски опис библиотечке грађе и размену библиографских метаподатака преко веба настао ради побољшања постојећих механизама (Tharani 2015, 6).

Постојећи начин библиографске обраде библиотечке грађе, по некима назван још и традиционалан, заснован на постојећим стандардима и форматима доводи библиотеке у опасност да нису адекватно припремљене да се сусретну са потребама модерних корисника и убрзаним развојем веб технологија. Иако је изграђен велики број стандарда за израду метаподатака за библиографски опис библиотечке, архивске и музејске грађе ниједан се за сада није показао као довољно добра замена за постојеће формате за библиографски опис односно различите верзије MARC формата. Тако је започео рад на моделу BIBFRAME (Gonzales 2014, 10).

BIBFRAME се заснива на принципима семантичког веба и користи парадигму иницијативе „повезани подаци" за објављивање и размену података преко веба. Овај модел је представљен као потенцијална замена за MARC формат који се за обраду библиотечког материјала користи од шездесетих година 20. века. MARC формат је настао са аутоматизацијом библиотечких система и појавом електронских каталога као формат који омогућава обраду библиотечког материјала у машински читљивом облику. Иако је

[172] Bibliographic Framework, http://www.loc.gov/bibframe/docs/index.html

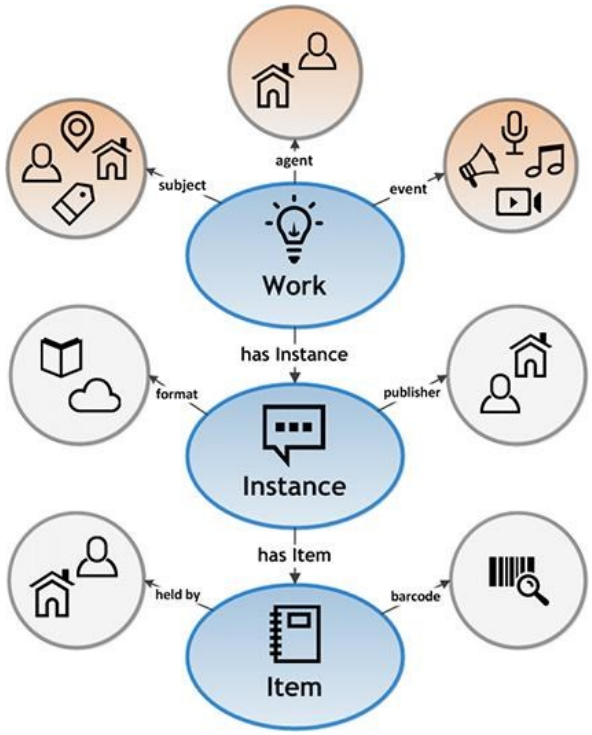
направљен за креирање машински читљивих података, формат је изграђен на принципима међународних библиотечких стандарда за опис физичких примерака библиотечке грађе (ISBD, AACR и други) те из тог разлога није увек адекватан за опис дигиталних, визуелних или мултимедијалних садржаја јер не садржи одговарајућа поља и потпоља за опис специфичних карактеристика овакве грађе. Поред тога, основни недостаци MARC формата су и постојање његових различитих варијанти (UNIMARC, UKMARC, MARC21 и многе друге), недостатак флексибилности која отежава размену података између локалних и обједињених каталошких система, немогућност да се искажу хијерархијске везе, као и његова превентивна оријентисаност ка библиотекама што отежава или чак онемогућава размену података са другим окружењима. Такође, формат не може да се користи у домену семантичког веб и његових технологија јер не дозвољава остваривање веза са другим ресурсима на вебу.

Оквири и принципи LD иницијативе интегрисани у модел BIBFRAME омогућавају машинама за претрагу индексирање библиографских метаподатака што је посебна предност у домену проналажења информација јер тако индексирани метаподаци олакшавају процес њиховог проналажења коришћењем различитих алата и програма за претрагу (Dean 2013). Иако је заснован на технологијама и принципима семантичког веба, BIBFRAME интегрише стандарде и моделе за библиографски опис библиотечке грађе као што су FRBR и RDA држећи се тако правила и принципа каталогизације.

Модел прави разлику између концептуалних садржаја и физичке(ких) манифестације(а) (Дело и Инстанца), омогућава недвосмислено идентификовање информационих ентитета (на пример, ентитети из записа у нормативним базама података) и омогућава успостављање и представљање веза између ентитета (Miller et al. 2012, 3). Тако је Конгресна библиотека развила посебан вокабулар за библиографску контролу који је интегрисан у BIBFRAME модел и који примењује правила за библиографски опис библиотечке грађе у складу са принципима LD модела. Структура модела прати структуру FRBR стандарда описујући ентитете преко три основне класе, Дело (Work), Инстанца (Instance) и Јединица (Item), које садрже одређен број

хијерархијски подређених атрибута који помажу у опису и повезивању основних класа (Слика 25).

У основи модела су RDF и XML. Као и у случају претходно поменутих модела развијених у оквиру иницијатива Еуропеана и DPLA, а следећи принципе RDF-а, сам опис дигиталног објекта састоји се од изјава у форми уређених тројки које добијају одговарајуће URI идентификаторе и формирају структуру графа. Сви елементи за опис груписани су у класе и својства који су између себе повезани преко додељених URI идентификатора. Систем обједињује и различите системе контролисаних речника од којих је најзначајнија нормативна датотека предметних одредница Конгресне библиотеке LCSH.



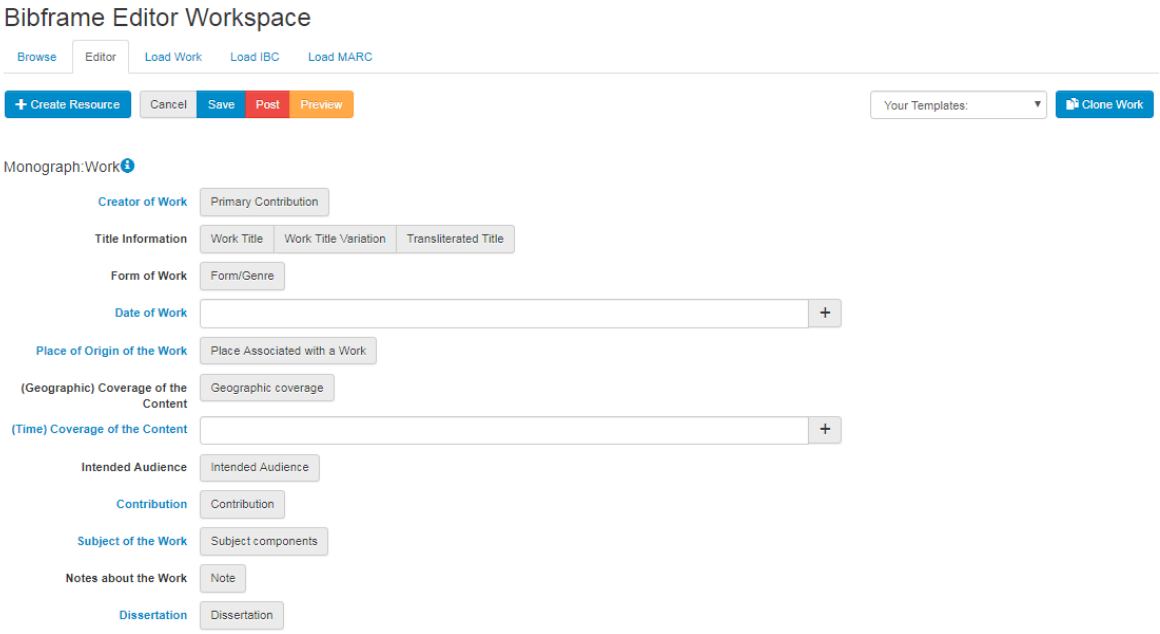
Слика 25. Структура BIBFRAME модела[173]

Временом је развијен алат који омогућава аутоматску конверзију записа у BIBFRAME модел[174]. Преко овог алата је могуће креирати и конвертовати записе у BIBFRAME на више начина. Један начин је да корисници сами креирају запис у BIBFRAME

[173] Слика је преузета са званичног сајта BIBFRAME модела: https://www.loc.gov/bibframe/docs/bibframe2-model.html

[174] Алат је доступан на: http://bibframe.org/bfe/index.html

моделу у уређивачком простору који је потпуно прилагођен корисницима, а након завршеног уноса систем сам генерише запис у форми RDF онтологије (Слика 26).



Слика 26. Сумећа алата за креирање записа у BIBFRAME модел

Други начин је да се учита запис преко неких идентификационих бројева као што је, на пример, LCCN (Library of Congress Catalog Number) преко кога се повлачи запис у MARC формату, а затим трансформише у BIBFRAME. Ми смо тестирали алат преко ове опције. За пример смо узели дело "Meine Preise" Томаса Бернхарда. У каталогу Конгресне библиотеке пронашли смо запис, учитали га у алат LCCN (Слика 27) и конвертовали га у BIBFRAME. Пример је дат као Прилог 15 - Пример записа за дело „Моје награде" Томаса Бернхарда у моделу BIBFRAME / корисничко окружење.

Слика 27. Сумећа алата за конверзију записа у BIBFRAME на основу идентификатора записа

6 Српско-немачки паралелни корпус - СрпНемКор

Предмет ове докторске дисертације је почетак израде српско-немачког паралелног корпуса књижевних текстова (СрпНемКор) те је у овом поглављу детаљно описан начин израде овог корпуса почев од прикупљања материјала до коначног постављања корпуса на веб и умрежавања са другим ресурсима доступним на интернету.

Као што је у другом поглављу ове дисертације објашњено Група за језичке технологије Универзитета у Београду је упоредо са развојем једнојезичних корпуса деведесетих година двадесетог века започела рад и на паралелним двојезичним односно вишејезичним корпусима. До сада је, у оквиру различитих пројеката, развијено неколико паралелних корпуса, а највише се радило на енглеско-српском и француско-српском паралелном корпусу који су детаљно описани у поглављу 2 одељак 2.5.3.

Немачки језик је званични језик Немачке, Аустрије и Лихтенштајна, затим један од три званична језика Луксембурга и један од четири званична језика Швајцарске, а говори се и у Источној Белгији, Северној Италији (Јужни Тирол) и области Алзас-Лотринген (Alsace-Lorraine) у Француској. Према званичним статистичким подацима[175] немачки језик представља матерњи језик великог дела становништва Европске Уније. Говори га више од 130 милиона људи као матерњи или други језик. Поред тога, према статистикама, немачки језик спада у један од три светска језика који се највише уче као страни, један је од десет језика који се највише говоре на свету и један од пет језика који се највише користе на интернету. Поред тога, координација рада центра „Аустријска библиотека" у оквиру Универзитетске библиотеке „Светозар Марковић" у Београду, затим координација сарадње Универзитетске библиотеке „Светозар Марковић" и Гете института у Београду, као и вишегодишња сарадња са Групом за језичке технологије у Београду подстакле су аутора ове дисертације да започне рад на изградњи овог корпуса. Координација центра „Аустријска библиотека" и стручна сарадња са Гете институтом представљали су добру

[175] Статистички подаци су преузети са: "Statista", preuzeto 25.3.2019, https://de.statista.com/statistik/daten/studie/918/umfrage/aussagen-ueber-die-deutsche-sprache/; "German Language: all about German language", преузето 25.3.2019, http://www.germanlanguageguide.com/german/facts/stats/; deutschland.de, преузето 25.3.2019, https://www.deutschland.de/de/topic/kultur/deutsche-sprache-ueberraschende-zahlen-und-fakten

основу за прикупљање материјала који је неопходан за креирање корпуса, док је сарадња са Групом за језичке технологије обезбедила неопходну техничку подршку у поступку израде корпуса. Претходно наведене чињенице о немачком језику су нас подстакле на размишљање о изради српско-немачког паралелног корпуса.

6.1 Садржај корпуса СрпНемКор

На почетку рада на дисертацији најважније је било одабрати почетни материјал за корпус, који би био довољно обиман и разнолик за примену развијених језичких алата, али и који би представљао добру основу за његову будућу надоградњу. Одлучили смо да корпус буде састављен од дела српских писаца који су превођени на немачки језик и дела писаца немачког говорног подручја (у овом случају се мисли на писце са територије Немачке и Аустрије) који су превођени на српски језик. На почетку је урађено истраживање и анализа како би се установило који су све немачки писци превођени на српски језик, а посебно који су српски писци превођени на немачки језик. Добру смерницу за почетак истраживања добили смо од библиотекара у Библиотеци Гете института у Београду и библиотекара библиотеке Катедре за германистику на Филолошком факултету Универзитета у Београду. На основу анализе добијених резултата одлучили смо да за корпус, за почетак, одаберемо дела писаца који су писали у другој половини 20. и почетком 21. века. Због једноставности у процесу припреме и обраде текстова на почетку рада смо се одлучили за прозне садржаје, и то романе.

Поред примарних критеријума „савремени писци" и „прозни текст, пре свега роман", приликом одабира материјала поставили смо и следеће критеријуме за одабир којима смо се руководили током прикупљања грађе:

1. *Награђиваност аутора*: одлучили смо да у корпус, за почетак, уђу дела и писци који су награђени неком од међународних или националних награда за књижевност као што су Нобелова награда, затим награда која се у Србији додељује за књижевно дело (НИН-ова награда, Андрићева награда, награда Меша Селимовић и друге) или неком од награда које се додељују у немачким

говорним подручјима за књижевно дело (Аустријска државна награда за књижевност, Бечка књижевна награда и друге).

2. *Доступност дела на оба језика у библиотекама у Србији*. Како библиотеке представљају полазну тачку у већини истраживања тако смо и ми састављајући листу писаца и њихових дела као потенцијалних кандидата за корпус покушали да утврдимо шта се од одабраног материјала налази у библиотечким фондовима првенствено у библиотекама у Србији, али и у фондовима библиотека у иностранству са којима Универзитетска библиотека „Светозар Марковић" има добро развијену мрежу међубиблиотечке позајмице.
3. *Популарност дела.* Поједина дела за корпус одабрана су на основу неких параметара популарности као што су: националне или међународне листе најчитанијих књига, велика популарност филма који је снимљен према мотивима одабране књиге, велика популарност позоришне представе чији је сценарио урађен по мотивима одабране књиге, или можда све заједно.
4. *Обимност дела*. Уз претходно наведене критеријуме из практичних разлога бирали смо дела која нису преобимна у погледу броја страна исписаног текста.

Сви наведени критеријуми су нам помогли да направимо прелиминарни списак писаца и њихових дела која би могла постати саставни део корпуса СрпНемКор. Са друге стране, било је потребно утврдити који је то број дела који би могао бити обрађен и смештен у овај корпус у временском периоду предвиђеном за израду ове докторске дисертације. Параметри као што је време потребно за припрему и обраду материјала, али и техничке могућности и расположивост сарадника на разним пословима које је потребно обавити у овим фазама рад утицали су на коначну одлуку о броју дела која ће постати део корпуса: одабрали смо седам романа различитих српских писаца и седам романа различитих писаца немачког говорног подручја.

Када је реч о српским писцима прво смо сагледали који су све српски писци из друге половине 20. века превођени на немачки језик. У овом кораку велику помоћ добили смо од запослених у Библиотеци Гете института који су нам обезбедили списак свих дела српских писаца која су преведена на немачки језик у прошлом и почетком овог века и која

се налазе у Библиотеци, што нам је дало одличне смернице за даље истраживање. Дела српских писаца углавном су бирана на основу критеријума „награђиваност аутора" и „доступност" те је на основу тога одабрано шест дела мушких писаца. Наша жеља је била да се у корпус поред дела мушких писаца укључе и дела женских писаца. Међутим, приликом истраживања утврдили смо да је јако мало српских женских писаца превођено на немачки језик, посебно када је реч о прози. Из тог разлога одабрано је једно дело. Списак дела српских писаца је следећи:

1. Срђан Ваљаревић и роман „Комо" (Como). За овај роман писац је 2006. године добио награду Културконтакт Аустрија, Беч. Роман је преведен на преко десет светских језика.
2. Давид Албахари и роман „Мамац" (Mutterland). За овај роман писац је 1996. године добио НИН-ову награду, као и награду Народне библиотеке Србије, Балканику и Мост-Берлин. Роман је преведен на шеснаест језика.
3. Данило Киш и роман „Пешчаник" (Sanduhr). За овај роман писац је 1972. године добио НИН-ову награду. Пешчаник представља најзагонетнију књигу Данила Киша. Преведен је, поред немачког, и на хебрејски, бугарски, македонски, словеначки, шпански језик и друге језике. Са друге стране, током прикупљања материјала утврђено је да је српски текст већ саставни део Корпуса савременог српског језика и у жељеном формату за корпус те је то искоришћено у припреми корпуса СрпНемКор.
4. Александар Тишма и роман „Употреба човека" (Der Gebrauch des Menschen). Александар Тишма један је од најпревођенијих српских писаца. Роман је преведен на скоро 30 светских језика. За овај роман писац је 1976. године добио НИН-ову награду.
5. Драган Великић и роман „Руски прозор" (Das russische Fenster). За овај роман писац је двоструко награђен 2007. године, НИН-овом наградом и Наградом Меша Селимовић, а у Аустрији је 2008. награђен наградом за допринос средњеевропској књижевности. Роман је, поред немачког, преведен и на бугарски, мађарски, италијански и грчки језик.

6. Гроздана Олујић и роман „Излет у небо” (Ein Ausflug in den Himmel). Роман је одабран на предлог ментора проф. др Цветане Крстев као једно од значајних дела овог писца и једно од дела које је изазвало велику пажњу јавности када је објављено. Роман је победио на конкурсу сарајевске „Народне просвјете” између 157 рукописа и преведен је све значајне светске језике (француски, енглески, немачки, шпански, норвешки, дански и друге). По мотивима романа урађена је позоришна представа „Чудна девојка” која је премијерно изведена 1959. године на сцени Београдског драмског позоришта, а снимљен је и истоимени филм 1962. године.
7. Владимир Арсенијевић и роман „У потпалубљу” (Cloaca maxima : eine Seifenoper). Роман „У потпалубљу” је први роман писца Владимира Арсенијевића и до сада једини деби роман награђен НИН-овом наградом. До сада је преведен на укупно двадесет светских језика, а по роману је у продукцији Југословенског драмског позоришта урађена и представа која је 1996. године награђена Стеријином наградом за најбољу представу. Уз све поменуто роман се уклопио и у критеријум „обимност дела” те је на основу свега поменутог постао саставни део корпуса СрпНемКор.

Када је реч о писцима немачког говорног подручја дела су примарно бирана на основу критеријума „доступност”, али и на основу осталих критеријума као што су „награђиваност” или „популарност” дела. Када је реч о делима женских писаца немачког говорног подручја које су превођене на српски избор је био знатно већи у односу на избор српских женских писаца које су превођене на немачки. Међутим, одабрано је једно дело како бе се направио равномеран однос мушких и женских писаца у обе групе. Као и код српских писаца и овде је одабрано седам дела различитих аутора:

1. Томас Бернхард (Thomas Bernhard) и роман „Моје награде” (Meine Preise). Томас Бернхард сматра се једним од највећих аутора савремене светске књижевности. Ово дело је одабрано и због доступности дела на оба језика на једном месту у „Аустријској библиотеци” Универзитетске библиотеке „Светозар

Марковић". Поред тога, дело је објављено постхумно, на двадесетогодишњицу смрти аутора и представља значајан рукопис из његове заоставштине.

2. Елфриде Јелинек (Elfride Jelinek) и роман „Пијанисткиња" (Die Klavierspielerin). Ово дело је, такође, примарно одабрано због доступности дела на оба језика на једном месту, а и зато што је ауторка добитница Нобелове награде за књижевност 2004. године. По овом роману је 2001. године снимљен и филм што је такође утицала да ово дело уђе у почетни садржај корпуса СрпНемКор.
3. Мило Дор (Milo Dor) и роман „Беч: јули 1999" (Wien: Juli 1999). Мило Дор је био аустријски писац српског порекла који је сва своја дела објавио под псеудонимом Мило Дор. Право име аутора било је Мирослав Дорословац. Добитник је бројних награда за књижевност, а за овај роман аутор је добио српску награду „Растко Петровић" која се додељује за најбољи роман у претходне четири године.
4. Гинтер Грас (Günter Grass) и роман „Ходом рака" (Im Krebsgang). Године 1999. писац је добио Нобелову награду за књижевност, а роман „Ходом рака" сматра се његовим најуспешнијим делом.
5. Патрик Зискинд (Patrick Süskind) и роман „Парфем: хронологија једног злочина" (Das Parfum: die Geschichte eines Mörders). Роман представља један од најчитанијих немачких послератних романа. Њујорк тајмс је 1986. прогласио „Парфем" за књигу године. Роман је преведен на преко 40 светских језика и на листи бестселера немачког магазина Шпигел био је скоро десет година. По мотивима романа је 2006. године снимљен и филм.
6. Кристоф Рансмајер (Christoph Ransmayr) и роман „Последњи свет" (Die letzte Welt). Писац је награђиван бројним наградама за књижевност. Роман је на предлог библиотекара библиотеке Катедре за германистику Филолошког факултета Универзитета у Београду одабран за садржај корпуса СрпНемКор.
7. Гинтер де Бројн (Günter de Bruyn) и роман „Буриданов магарац" (Buridans Esel). Писац је по професији библиотекар као и ликови у роману те се аутору тезе и ментору учинило интересантним да роман буде укључен у корпус.

Од одабраних аутора аустријски писци су Томас Бернхард, Елфриде Јелинек, Мило Дор и Кристоф Рансмајер, док су немачки писци Гинтер Грас, Патрик Зискинд и Гинтер де Бројн.

Највећи део материјала за припрему корпуса СрпНемКор пронађен је у фонду „Аустријске библиотеке" („Meine Preise"/„Моје награде", „Die letze Welt", „Die Klavierspielerin"/„Пијанисткиња", „Wien, Juli 1999", „Como") и у општем фонду („Буриданов магарац", „Беч, јули 1999", „Ходом рака", „Последњи свет", „Пешчаник", „Излет у небо", „Употреба човека", „Комо", „Руски прозор") , затим у фонду Библиотеке Гете института у Београду („Im Krebsgang", „Sandurh", „Der Gebrauch des Menschen"), али и фондовима Библиотеке Матице српске („Mutterland", „Das russische Fenster", „Мамац") и Националне библиотеке Аустрије („Buridans Esel", „Ein Ausflug in den Himmel") и то захваљујући добро развијеној међубиблиотечкој сарадњи и сусретљивости колега из Одељења за међубиблиотечку позајмицу Универзитетске библиотеке. Дело „Парфем" пронађено је на интернету у електронској верзији на оба језика и ове верзије су коришћене за даљу обраду.

6.2 Фазе у креирању корпуса СрпНемКор

6.2.1 Дигитализација текстова

Највећи број текстова који су тренутно саставни део корпуса СрпНемКор били су у штампаној форми. Текстови су прво сканирани, а затим је примењена метода оптичког препознавања карактера како би се добили текстови у машински читљивом облику чији се комплетан текст може претраживати, а неки његови делови се могу повезати са одговарајућим ресурсима на вебу. Метода оптичког препознавања карактера над текстовима одабраним за СрпНемКор дала је разноврсне резултате у зависности од квалитета „слике" која је добијена процесом сканирања. Сам квалитет „слике" доста је зависио од квалитета штампаног материјала који је сканиран. Ни код једног текста препознавање није било 100% успешно, али је код већине текстова успешност препознавања била доста добра. Сви текстови су сканирани у Универзитетској библиотеци „Светозар Марковић" у Одељењу за дигитализацију. За оптичко препознавање карактера коришћен је софтвер FineReader из пакета ABBYY.

6.2.2 Обрада немачких текстова

Текстови на немачком језику су након сканирања учитани у алат Транскрибус[176] који има уграђен софтвер ABBYY FineReader те је тада и примењена метода оптичког препознавања карактера. Транскрибус је платформа за аутоматско препознавање, транскрипцију и претраживање историјских докумената (Seaward and Kallio 2017). Састоји се од експертског алата Транскрибус, веб-сумеђе и неколико услуга у технологији облака. У оквиру Транскрибуса постоје алати за:

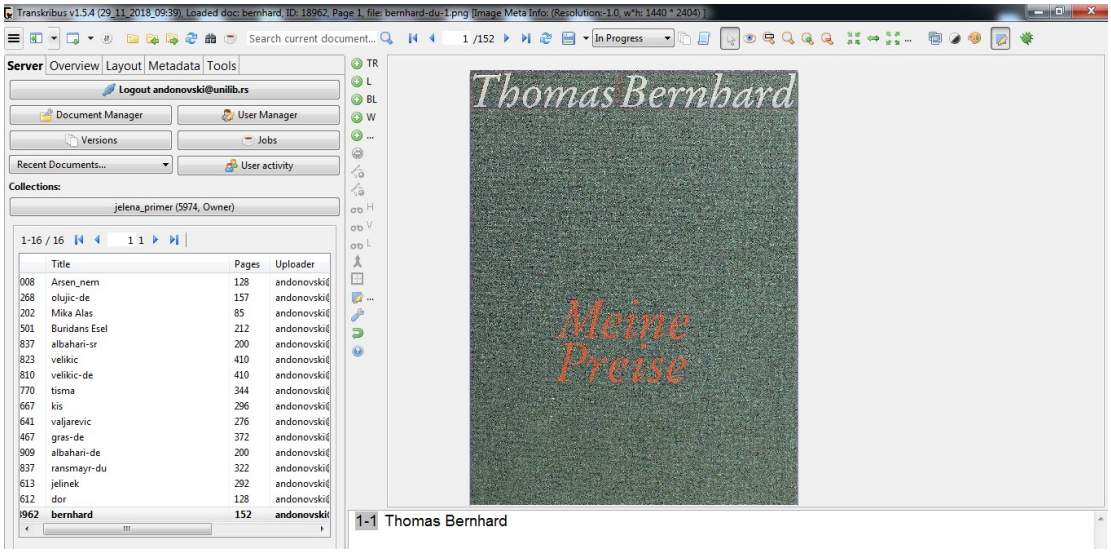
1. препознавање ручно писаних текстова (Handwritten Text Recognition - HTR),
2. анализу распореда елемената на страници (Layout Analysis),
3. анотацију структурних целина текста,
4. означавање именованих ентитета,
5. оптичко препознавање карактера коришћењем ABBYY Finereader Engine 11.

Транскрибус одржава Група за дигитализацију и дигиталну заштиту са Универзитета у Инсбруку, а његов развој финансира Европска комисија као део пројекта H2020 READ (Recognition and Enrichment of Archival Documents) (2016-2019)[177] (Mühlberger 2018), Пројекта за препознавање и обогаћивање архивских докумената. Намењен је истраживачима из области хуманистичких наука, архивистима, библиотекарима, као и стручњацима из области информационих технологија. Како је платформа још увек у фази развоја и тестирања све услуге које су на располагању су доступне потпуно бесплатно, а већи делови софтвера су у отвореном приступу. Делови програма реализовани су у Универзитетској библиотеци „Светозар Марковић" у Београду у оквиру пројеката које је финансирало и финансира Министарство културе и информисања Републике Србије: „Нови хоризонт дигитализације" за 2016. годину, „Рашчитана стара српска ћирилица: оживљена руком писана прошлост" за 2017. годину и „Рашчитаност старе српске ћирилице: историја и традиција на дохват руке" за 2018.

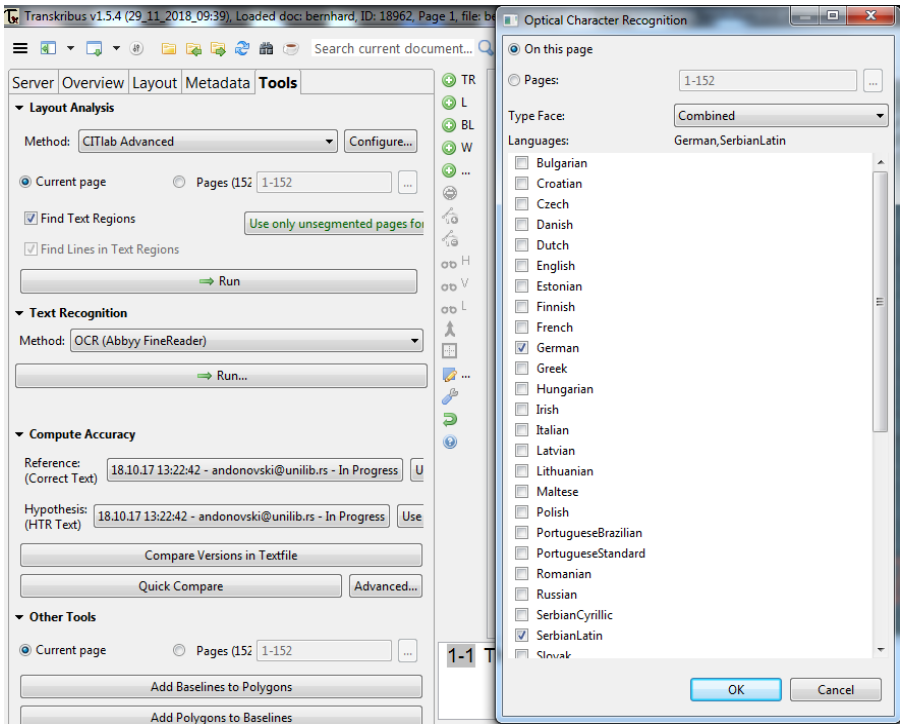
Транскрибус је иницијално изграђен за рашчитавање рукописног културног наслеђа које се налази у библиотекама и архивима и које се протеклих година ужурбано

[176] Transkribus, https://transkribus.eu/Transkribus/
[177] H2020 READ, https://cordis.europa.eu/project/rcn/198756_en.html

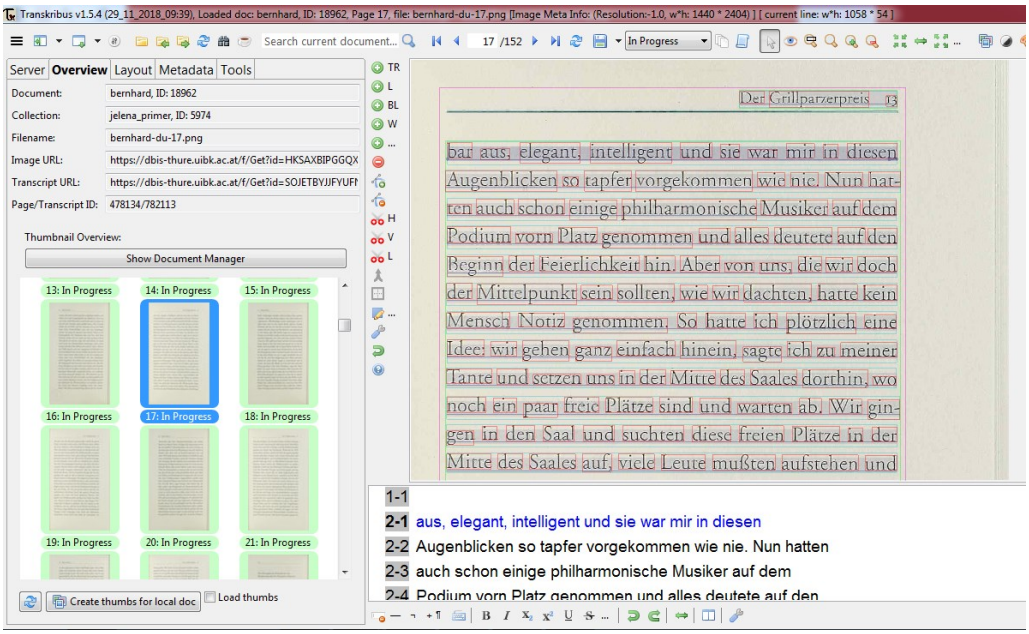
дигитализује како би се сачувало и учинило доступним корисницима. Међутим, Транскрибус омогућава и рад са штампаним текстовима те смо ми, уз сагласност и дозволу координатора пројекта Гинтера Милберга (Günter Mülhlberg), учитали немачке текстове које смо одабрали за корпус. Након учитавања текстови су смештени у одговарајућу датотеку за рад (Слика 28) и на њих је примењена технологија оптичког препознавања карактера (Слика 29). Након оптичког препознавања карактера одабрани текст је смештен у радно окружење које омогућава прегледање и корекцију произведених грешака. Радно окружење за прегледање и корекцију подељено је на три дела. Са леве стране смештене су обрађене странице текста у хронолошком редоследу. Сканирана слика одабране странице отвара се у горњем делу десне стране радног окружења, док се у доњем делу налази уређивачки простор, едитор у коме се налази рашчитани садржај странице (Слика 30).



Слика 28. Транскрибус - датотека за рад



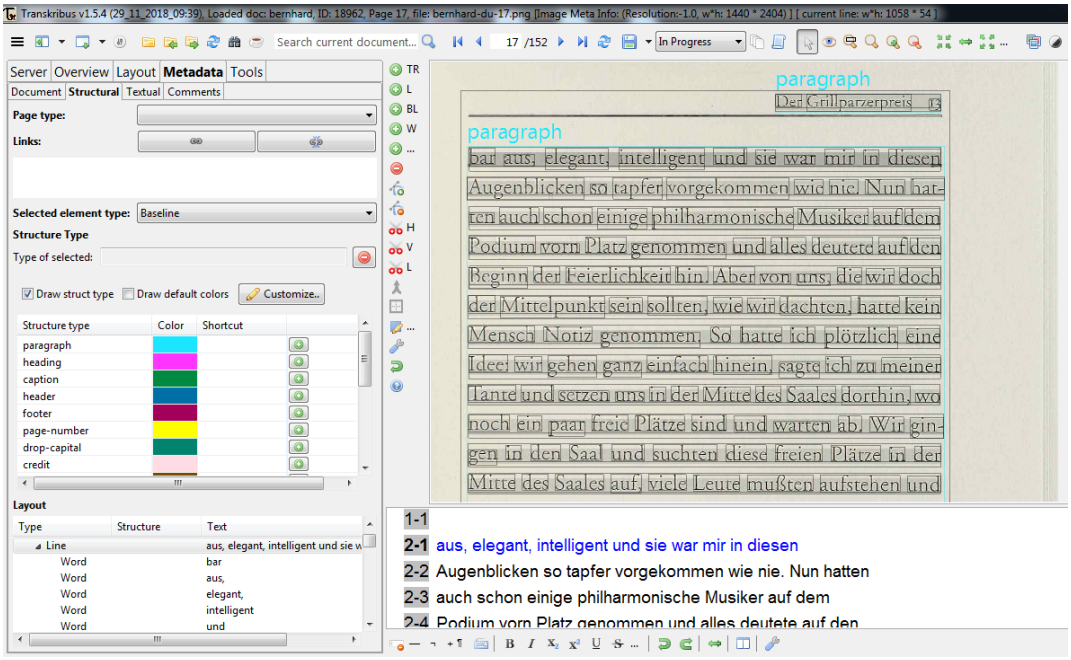
Слика 29. Транскрибус – радно окружење за оптичко препознавање карактера (подешавање параметара)



Слика 30. Транскрибус - радно окружење за прегледање и корекцију текста после оптичког препознавања карактера

Приликом поступка оптичког препознавања карактера урађена је и аутоматска сегментација текстова која подразумева сегментацију на текстуалне блокове (регионе) и линије текста у њима. Под текстуалним блоком подразумевају се непрекинуте текстуалне целине (наслови, поднаслови, пасуси и слично) док се под линијама текста подразумевају редови текста које је програм препознао у оквиру једног текстуалног блока. Овако урађена сегментација омогућава корисницима да током прегледања одабране странице у уређивачком простору прате текст по сегментима упоредо пратећи оригинал текста у горњем делу радног окружења.

Такође, програм омогућава и анотацију структурних целина одабраног текста (пасуса, фуснота, заглавља и слично) и форматирање текста (подебљање, подвлачење, промена величине фонта и слично) (Слика 31). Међутим, како је Транскрибус примарно намењен обради рукописног материјала интензивно се ради на развоју механизма који ће моћи аутоматски да обрађује руком писани текст. Након учитавања материјала у Транскрибус ручним прекуцавањем одређеног броја страна изабраног рукописног материјала припрема се транскрипт који представља машински читљив формат одабраног рукописа. На основу постојећег рукописа и његовог транскрипта софтвер „учи" како да прочита руком писани текст односно примењује технологију HTR (Handwritten Text Recognition – Препознавање руком писаног текста) (Sánchez et al. 2013). За разлику од технологије OCR која обрађује појединачне карактере штампаног текста, HTR технологија обрађује целе речи или пак целе линије, сканира их у различитим правцима и ставља у одговарајући низ. Транскрибус омогућава рад са различитим језицима и рукописним стиловима на основу чега се праве HTR модели за одређени језик (Дакић и Софронијевић 2018, 18). HTR модели се заснивају на алгоритмима за машинско учење, а технологија се обучава на основу најмање 25 страна транскрибованог материјала, што се назива „ground truth" (Gatos et al. 2014).

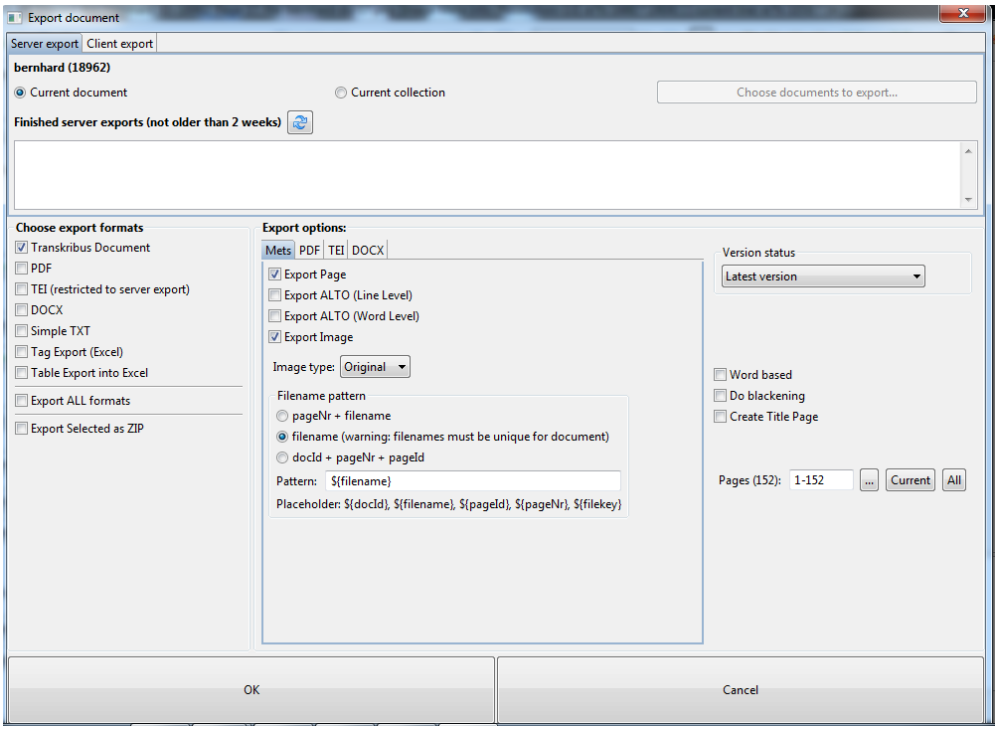


Слика 31. Транскрибус - радно окружење за анотацију структурних целина текста и форматирање текста

Библиотекари Универзитетске библиотеке „Светозар Марковић”, у сарадњи са волонтерима из различитих области, у оквиру поменутих пројеката развијају модел за српску ћирилицу. Модел за српску ћирилицу припрема се на основу транскрипата различитих рукописа који се као део српске културне баштине налазе се у фонду Универзитетске библиотеке „Светозар Марковић”, али и у фондовима различитих библиотека са којима Универзитетска библиотека сарађује. Сваки од припремљених транскрипата представља појединачни модел одабраног рукописа који се уграђује у кровни модел за српску ћирилицу. До сада су обрађени рукописи из легата Исидоре Секулић, из приватне збирке Михајла Петровића Аласа, легата Бранимира Ћосића, из Архива Суботица, из Архива Крушевца и многи други.

Након завршене обраде текста програм нуди извоз у више формата (Слика 32): ALTO shema[178], PDF, DOCX, TEI, TXT, Tag export (Excel), Table export into Excel. У оквиру свих ових формата постоје различити параметри и филтери којима се дефинише изглед документа после извоза. Корисник има могућност да одабере један од ових формата за извоз или да их одабере све одједном.

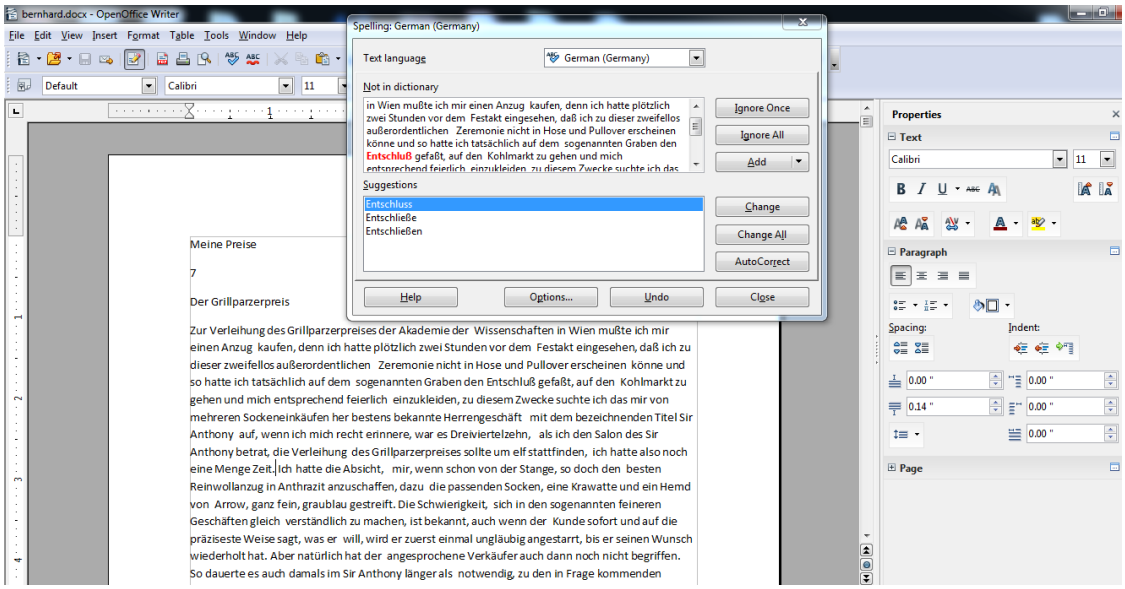
[178] ALTO shema, http://www.loc.gov/standards/alto/



Слика 32. Транскрибус – радно окружење за избор излазног формата

Иако алат нуди могућност извоза у .txt формат, који је нама био неопходан за даљи рад, ми смо текстове извозили у формат .docx како би такав текст прошао кроз даљу контролу. За контролу и исправку немачког текста коришћен је језички алат Hunspell[179] за типографску корекцију текстова. Овај алат за типографску корекцију је један од бољих када је немачки језик у питању, са уграђеним добрим немачким речником. Hunspell може да се користи у оквиру различитих окружења: LibreOffice, OpenOffice.org, Mozilla Firefox 3 & Thunderbird, Google Chrome, али и у оквиру софтверских пакета као што су macOS, InDesign, memoQ, Opera and SDL Trados. Ми смо овај алат успешно користили за типографску корекцију у оквиру алата OpenOffice (Слика 33). Сви текстови на немачком језику исправљени су на најбољи могући начин уз помоћ овог алата и припремљени за структурну анотацију.

[179] Hunspell, http://hunspell.github.io/



Слика 33. Hunspell окружење за контролу и корекцију текстова на немачком језику

На сликама 28-33 приказан је поступак обраде текста “Meine Preise” Томаса Бернхарда кроз алат Транскрибус, а затим и контрола применом речника за немачки језик у оквиру језичког алата Hunspell.

6.2.3 Обрада српских текстова коришћењем Unitex-а

Након дигитализације и оптичког препознавања карактера, за шта је као и код текстова на немачком језику коришћен софтвер Abby FineReader, текстови на српском језику су обрађени коришћењем система Unitex и припремљени за фазу структурне анотације. Обрада коришћењем система Unitex подразумевала је најпре примену електронских морфолошких речника српског језика и аутоматску поделу текстова у реченице. Применом електронских морфолошких речника утврђене су, са једне стране, грешке у текстовима које су затим ручно исправљане, а са друге стране добијене су и листе нових речи које су затим обрађене и уграђене у одговарајуће морфолошке речнике. За текстове чија сканирана слика није била квалитетна, па је самим тим и резултат оптичког препознавања био лош, примењивана је аутоматска процедура кориговања. Ова процедура се заснива на утврђивању парова који се највише мешају код препознавања (за ћирилицу, то су често п<=>н, п<=>и и и<=>н) и на е-речницима за српски. У зависности од

квалитета оптичког препознавања карактера одређени текстови су више пута обрађивани на овај начин како би се исправило што више грешака и како би се обрадило што је могуће више нових кандидата за речнике.

Речи за које је утврђено да су грешке, ручно су исправљане у самим текстовима, док су речима које су препознате као нови кандидати за речник додељене лема и врста речи и други подаци релевантни за српски е-речник. Друга обрада кроз систем представљала је проверу претходно исправљеног текста како би се утврдило да ли су све грешке у тексту поправљене и да ли су преостали још неки кандидати за унос у речнике. Обрадом текстова на српском језику за СрпНемКор електронски морфолошки речници за српски језик допуњени су са више од 2.500 нових речи (Табела 1).

Табела 1. Број нових речи у речницима DELAS опште лексеме и DELAS-PROP властитих имена према романима[180]

Немачки писци			**Српски писци**		
	Број нових речи у DELAS	**Број нових речи у DELAS-PROP**		**Број нових речи у DELAS**	**Број нових речи у DELAS-PROP**
Бернхард	135	87	**Албахари**	34	10
де Бројн	302	113	**Арсенијевић**	108	18
Дор	76	38	**Киш**		
Грас	234	138	**Великић**	218	97
Јелинек	259	55	**Ваљаревић**	49	23
Рансмајер	15		**Тишма**	359	40
Зискинд	147		**Олујић**	86	10
Укупно	1168	431	**Укупно**	854	198

Одабрани текстови обрађени су на писму оригинала односно штампане верзије, али су за потребе даљег рада односно производњу паралелизоване верзије и јединственог корпуса преведени у латиницу као што је то урађено и са другим текстуалним корпусима српског језика. Током обраде текстова кроз система Unitex примењен је и граф за обраду реченице, *Sentence-XML.grf,* чиме је извршена и аутоматска подела на реченице, односно сегментација текстова (детаљније у следећем одељку).

[180] За романе „Последњи свет“ аутора Кристофа Рансмајера и „Парфем“ аутора Патрика Зискинда остало је да се накнадноураде и додају нове речи у речник DELAS-PROP, док је за роман „Пешчаник“ аутора Данила Киша остало да се ураде и додају нове речи у оба речника, DELAS и DELAS-PROP

6.2.4 Структурна анотација

За поступак структурне анотације коришћене су следеће етикете: <body> - за анотацију тела текста, - за анотацију главних јединица текста (поглавља), <head> - за анотацију наслова и поднаслова, - за анотацију пасуса и <seg> - за анотацију реченица. Сегментација је рађена до нивоа реченице, односно, варијанте јединица превођења приликом паралелизације биле су реченице одабраних текстова. Поступак анотације текстова и на српском и немачком језику урађен је скоро у целини аутоматски. У првом кораку коришћен је систем Unitex и граф за обраду реченица *Sentence-XML.grf*. Овај граф примењен је на све текстове одабране за корпус, и на српском и на немачком језику. Применом графа *Sentence-XML.grf* текстови су анотирани XML ознакама за крај и почетак реченице „</seg><seg>“. На овај начин највећи део свих текстова аутоматски је означен до нивоа реченица. Међутим, ова аутоматска анотација није била потпуно успешна ни за један од текстова. Граф за обраду реченица није увек препознао интерпункцијски знак којим је означен крај реченице, односно карактер којим је означен почетак следеће реченице, а у неким случајевима ни једно ни друго. Као резултат, у појединим текстовима остајало је после овог корака доста неозначених сегмената. Највише проблема је било у текстовима који имају пуно дијалога.

Регуларни изрази коришћени су као други метод за аутоматску анотацију текстова. Њима су анотирани пасуси, сегменти који у претходном кораку нису обрађени, али и поглавља и пагинација када је то било могуће. На пример, регуларним изразом „\n</seg><seg>“ у тексту су проналажена сва места на којима се иза карактера за нову линију налазе XML етикете за крај и почетак реченице. На овим местима су применом текста замене „</seg> <seg>“ уметнуте XML ознаке за почетак и крај пасуса између ознака за крај и почетак реченица. За анотацију дијалога коришћен је регуларни израз „\n^»[A-ZÜÖÄ]”, за текстове на немачком, односно „\n^„[A-Š]”, за текстове на српском, који је препознавао сва места у тексту на којима иза карактера за нову линију почиње нови ред знаком навода и великим словом. На овим местима применом текста замене „</seg> <seg>»” уметнуте су XML ознаке за крај и почетак реченица и пасуса пре

знака навода. Кроз каснији поступак упаривања текстова анотирани су сви сегменти који су после примене регуларних израза остали необрађени.

За анотацију поглавља коришћени су, такође, регуларни изрази кад год је то било могуће односно ако су поглавља била означена. Регуларним изразом „\n([0-9]+)\.[\s]+" се, на пример, могу пронаћи су сва места у тексту која представљају наслов поглавља у засебном реду означен бројем записаним арапским цифрама иза кога следи тачка. Применом текста замене „\n<head>\1\.</head>\r\n" уметнуте су XML етикете за анотацију поглавља. За анотирање пагинације коришћен је регуларни израз „\n([0-9]+)[\s]+" којим су пронађена сва места у тексту која представљају пагинацију. Применом текста замене <!-- \1-->\r\n пагинација је означена као коментар у тексту у засебном реду. Анотирање пагинације рађено је на крају, након анотације поглавља, како не би дошло до забуне. За структурну анотацију коришћен је Notepad++ (Слика 34).

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE body SYSTEM "body.dtd">
<body>
<div>
<head>Tomas Bernhard
MOJE NAGRADE</head>
<head>Nagrada Franc Grilparcer</head>
<p><seg>Povodom dodele nagrade Akademije nauka u Beču, Grilparcerove nagrade,
morao sam sebi da kupim odelo budući da sam iznenada, samo dva sata pre svečanog prijema,
shvatio da na ovoj, nesumnjivo izuzetnoj, ceremoniji, ne mogu da se pojavim u pantalonama i džemperu,
te sam na takozvanom Grabenu zaista odlučio da odem do Kolmarkta i da se obučem svečano, kako i priliči,
te sam pomenutim povodom tragao za radnjom muške garderobe pod nazivom Ser Entoni koju sam znao budući da sam u
dotičnoj već toliko puta kupovao čarape, i ako me sećanje dobro služi,
bilo je petnaest do deset kad sam stupio u salon Ser Entoni,
a dodela Grilparcerove nagrade bila je zakazana za jedanaest,
imao sam, dakle, dovoljno vremena.</seg><seg> Kad je već
<!-- \07 -->
skupa konfekcija, onda nek bude najbolje odelo, od čiste vune, antracit sive boje,
a uz pomenuto - prikladne čarape, prikladna kravata i košulja marke erou, veoma fina, na tanke sivoplave pruge.</seg>
<seg> Poteškoće na koje kupac nailazi u ekskluzivnijim radnjama, u želji da ga smesta razumeju,
poznate su čak i kad kupac na najpreciznij mogući način izrazi šta zapravo hoće -
najpre se na njega podozrivo izbeče sve dok svoju želju ne ponovi.</seg>
<seg> I naravno da prodavac, kojem se čovek obrati, kupca ne razume čak ni tad.</seg>
<seg> Tako je i tad u SerEntoniju potrajalo duže nego što je bilo neophodno da se odaberu odela koja bi
eventualno dolazila u obzir.</seg>
<seg> Naravno, prilike u ovoj radnji bile su mi poznate iz mojih prethodnih kupovina čarapa,
a i sam sam, mnogo bolje od prodavca, uvek znao gde da pronađem odelo koje tražim.</seg>
<seg> Uputio sam se ka odelima koja su eventualno dolazila u obzir i
pokazao na sasvim određen model koji je prodavac skinuo sa štendera da bi mi ga pokazao.</seg>
<seg> Proverio sam kvalitet materijala i smesta u kabini za presvlačenje upriličio probu.</seg>
<seg> Nagnuo sam se napred, nekoliko puta napred-nazad, i shvatio da mi pantalone odgovaraju.</seg>
```

Слика 34. Едитор Notepad++: Томас Бернхард „Моје награде"

Припремљени XML документи за СрпНемКор су анотирани тако да буду валидни у односу на следећи једноставни DTD (значај дефиниције типа документа је детаљно објашњен у поглављу 3 одељак 3.4.1):

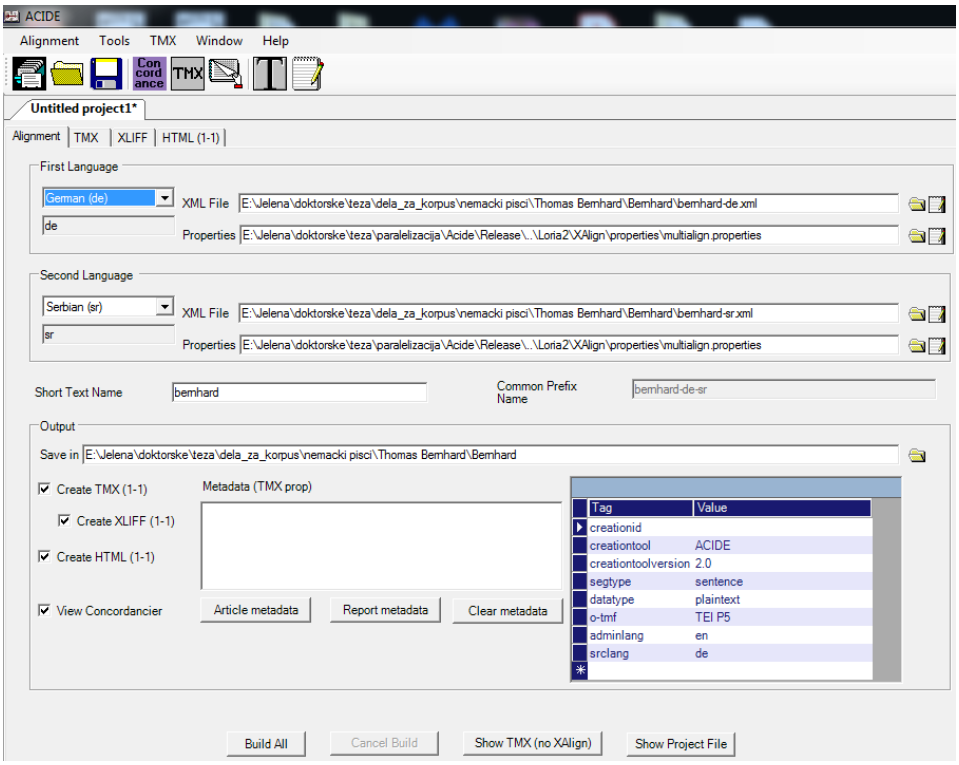
```
<!ELEMENT body (div+)>
<!ELEMENT div (head |p)+>
<!ELEMENT head (#PCDATA)>
<!ELEMENT p (seg+)>
<!ELEMENT seg (#PCDATA)>
```

Овај DTD дефинише да валидни XML документ садржи коренски елемент *body* који садржи бар један елемент *div* (*једно или више поглавља*). Сваки елемент *div* садржи или један или више елемената *head* (*један или више наслова*) и један или више елемената *p* (*један или више пасуса*) који могу бити произвољно измешани (*head* и *p*). Садржај сваког елемента *head* су парсирани карактерски подаци, односно текст са евентуалним карактерским ентитетима, док је садржај сваког елемента *p* један или више елемената *seg* (*једна или више реченица*). Према DTD, садржај сваког елемента *seg* су парсирани карактерски подаци односно текст.

6.2.5 Паралелизација

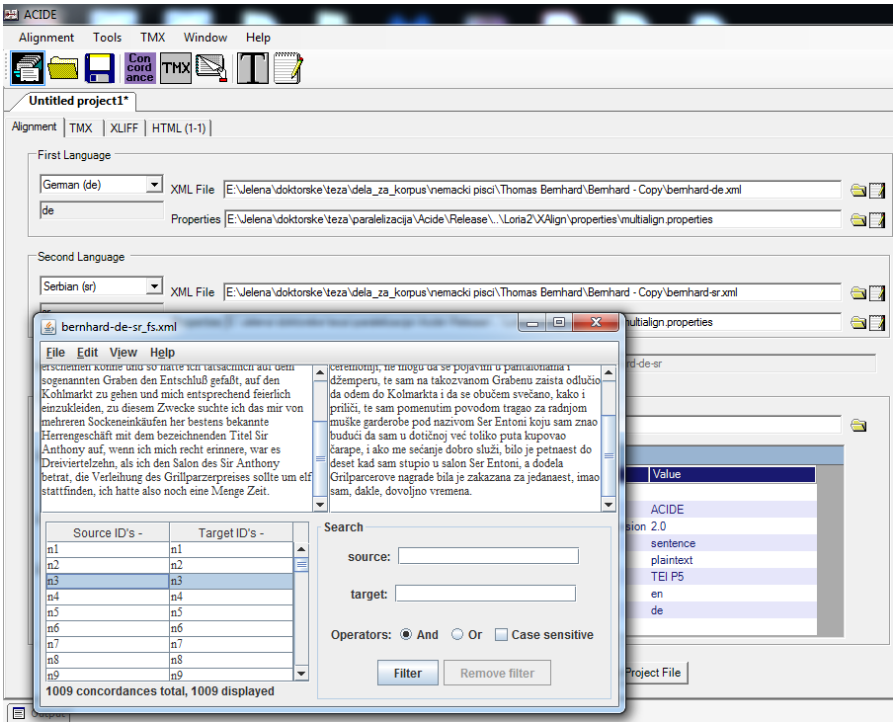
За паралелизацију коришћен је програмски пакет ACIDE. Овај програмски пакет омогућава да се сви кораци у припреми једног паралелног корпуса који су већ описани у поглављима 2 и 3 обаве кроз графичку корисничку сумећу (GUI). Сама структура програмског пакета ACIDE детаљно је објашњена у поглављу 3 одељак 3.5. Први корак у паралелизацији обављен је кроз мени Alignment (Слика 35). У оквиру овог менија одабрани су изворни и циљни језик из интегрисаних падајућих менија који се налазе у горњој левој страни радног окружења. Затим су задате улазне датотеке у XML формату уз сваки одабрани језик, као и адреса излазне датотеке. На основу одабраних језика и задатих XML датотека аутоматски је генерисан назив излазне датотеке, а након задавања свих неопходних параметара покренут је поступак аутоматске паралелизације преко Build All. Покретањем дугмета Build All програм прво проверава да ли су учитане XML датотеке добро формиране и валидне у односу на одабрани DTD. У случају да постоји грешка програм у доњем делу радног окружења исписује локацију указујући кориснику где се она налази у задатом или задатим XML датотекама.

Како је СрпНемКор састављен од преведених дела и српских писаца и писаца немачког говорног подручја тако су кроз рад мењани изворни и циљни језик. Када се радила паралелизација дела немачких писаца за изворни језик одабран је немачки, а за циљни српски језик. Када су паралелизована дела српских писаца изворни језик био је српски, а циљни немачки језик.

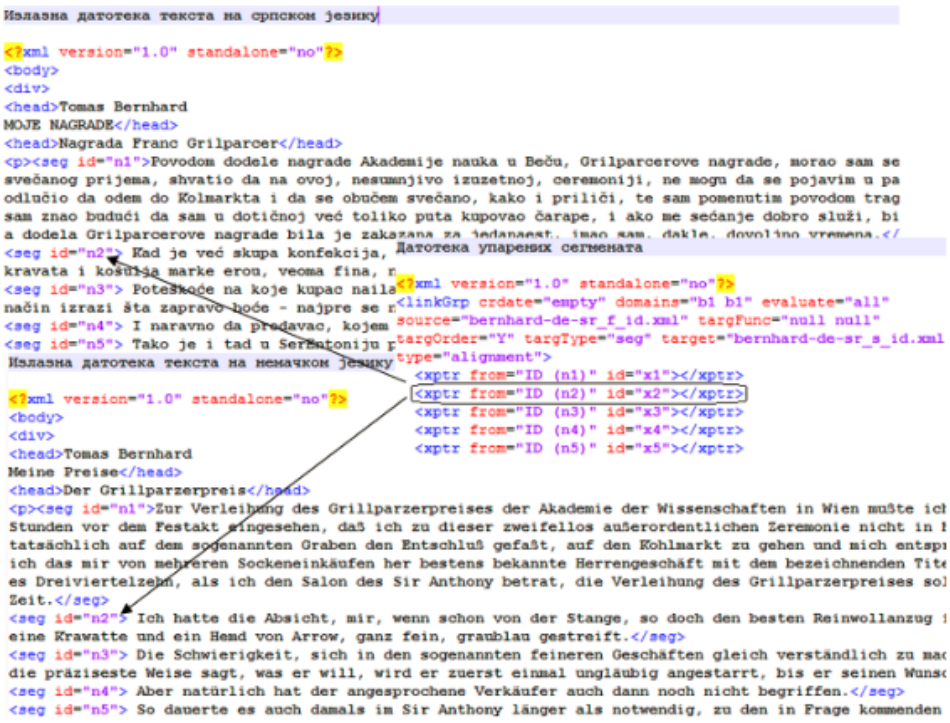


Слика 35. Мени Alignment – део програма ACIDE за аутоматско упаривање сегмената

Покретањем дугмета *Build All* генерисани резултати (битекстови) се одмах приказују у програму *Concordancier* (Слика 36). Као резултат добијене су три XML датотеке са информацијом о упаривању. Прве две представљају копије улазних датотека које су додатно анотиране идентификаторима, сваком сегменту, односно реченици, додељује се редни број који је означен XML атрибутом *id* уз сваку *seg* етикету, на пример, <seg id="n1">Tomas Bernhard</seg>. Трећа излазна датотека садржи кодиране информације о упареним сегментима користећи додељене идентификаторе. На пример, <xptr id="x1" from="ID (n1)"/> значи да је сегмент „1" из изворног текста упарен са сегментом „1" из циљног текста. Прецизније речено, програм идентификаторе који су додељени сегментима користи да би указао како су они повезани, то јест упарени (Слика 37).



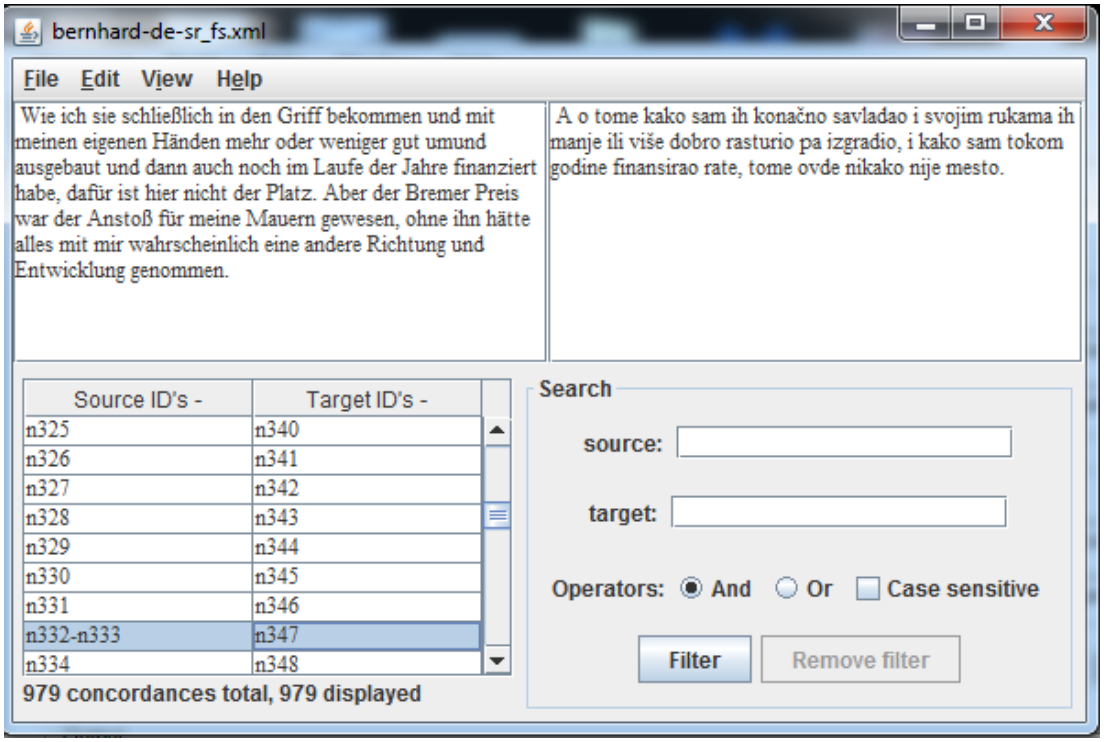
Слика 36. Concordancier – део програма ACIDE који приказује упарене сегменте у виду паралелних конкорданци



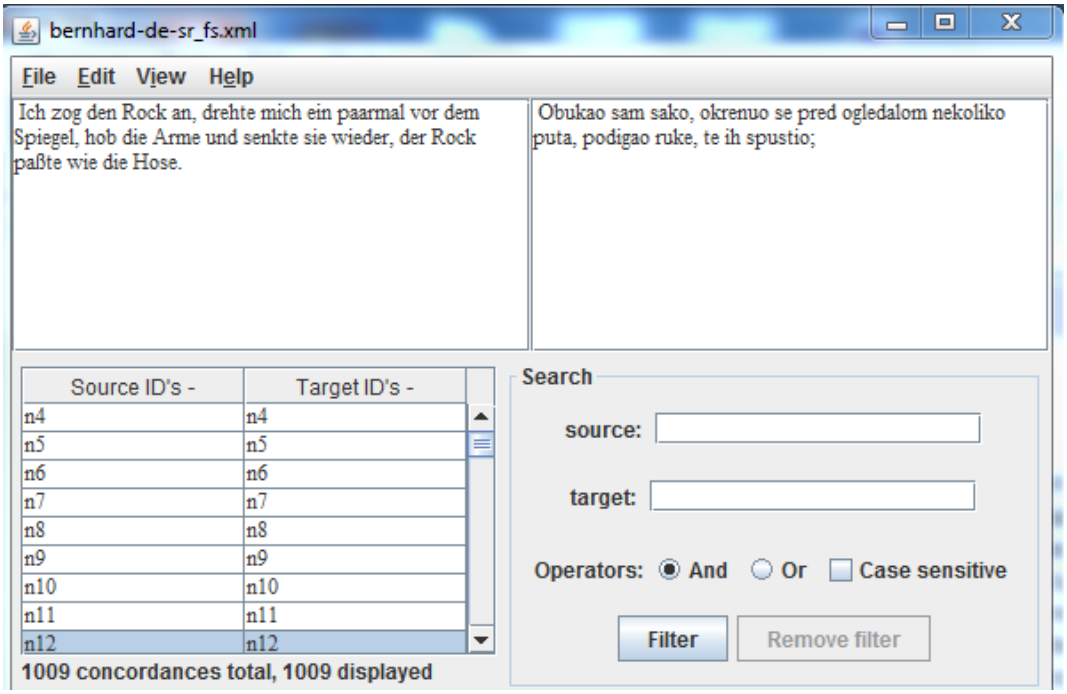
Слика 37. Садржај три излазне XML датотеке, идентификатори сегмената из прве две се користе у трећој

Као што је у поглављу 2 одељак 2.3 већ речено, приликом паралелизације тежи се 1-1 упаривању сегмената. Због различите структуре изворног и циљног текста у пракси често долази до погрешно упарених варијанти јединица превођења што захтева даљу, најчешће, ручну корекцију добијених битекста. Тако је био и случај приликом креирања корпуса СрпНемКор. Ручно су поправљене све погрешно упарене јединице превођења за све текстове који су паралелизовани. Сви разлози због којих може доћи до погрешно упарених сегмената који су наведени у поглављу 2 одељак 2.3, појавили су се и у паралелизацији текстова одабраних за овај корпус:

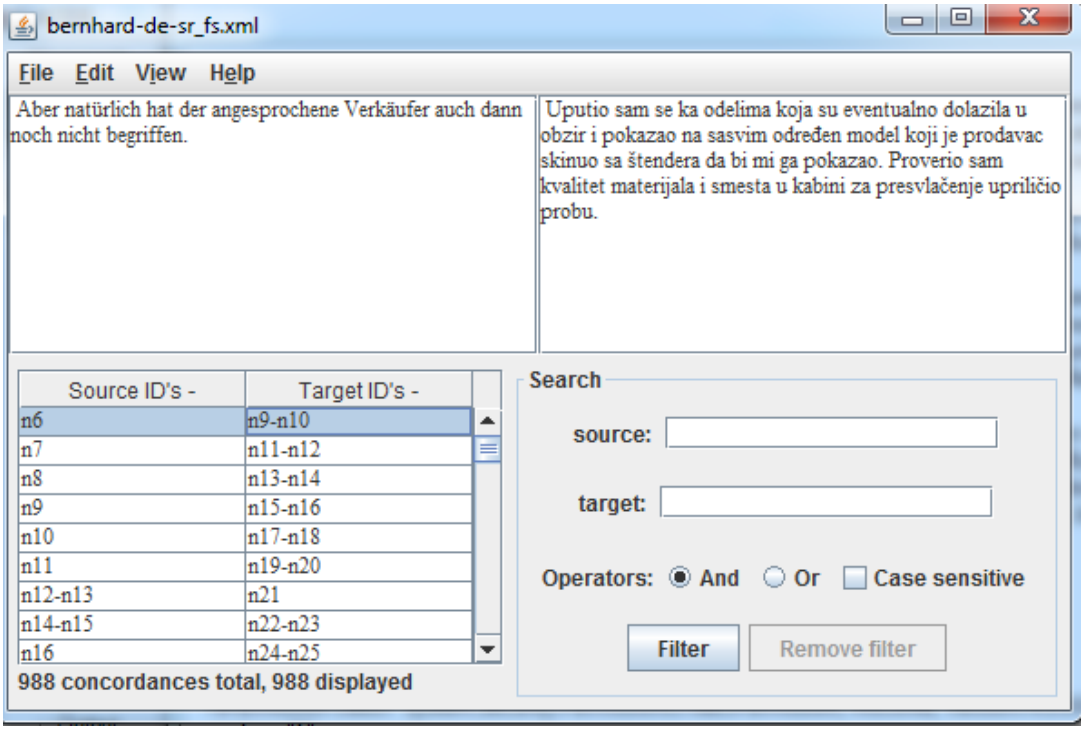
1. Постојале су разлике између оригиналног и преведеног текста у погледу броја реченица или пасуса. Овај проблем се појавио у скоро свим текстовима одабраним за овај корпус.
2. Изостављени су неки делови у преведеном тексту у односу на оригинал.
3. Постојале су разлике у означавању пасуса, па оригинални текст и преведени текст нису садржали исти број пасуса.
4. Постојале су разлике у сегментацији реченица и ово је био најчешћи разлог погрешно упарених сегмената. Приликом аутоматске сегментације програм је означавао сегменте, односно реченице, где год је препознао услов за то. Међутим, дешавало се да програм није увек препознао место за анотацију сегмента односно реченице јер за то нису били задовољени услови постављени програмом. Некада се на крају једног, односно почетак другог, сегмента налазио карактер који програмом није предвиђен као услов за анотирање. Приликом паралелизације јавили су се следећи проблеми:
    a. сегмент изворног језика је једна реченица, а одговарајући сегмент циљног језика су две или више реченица и обрнуто (Слика 38);
    b. сегмент изворног језика је једна реченица, а одговарајући сегмент циљног језика је део реченице и обрнуто (Слика 39);
    c. сегмент изворног и еквивалентан сегмент циљног језика се састоје од две или више реченица које нису у истом редоследу (Слика 40);
    d. сегмент изворног језика је једна реченица, а одговарајући сегмент циљног језика не постоји и обрнуто (Слика 41).



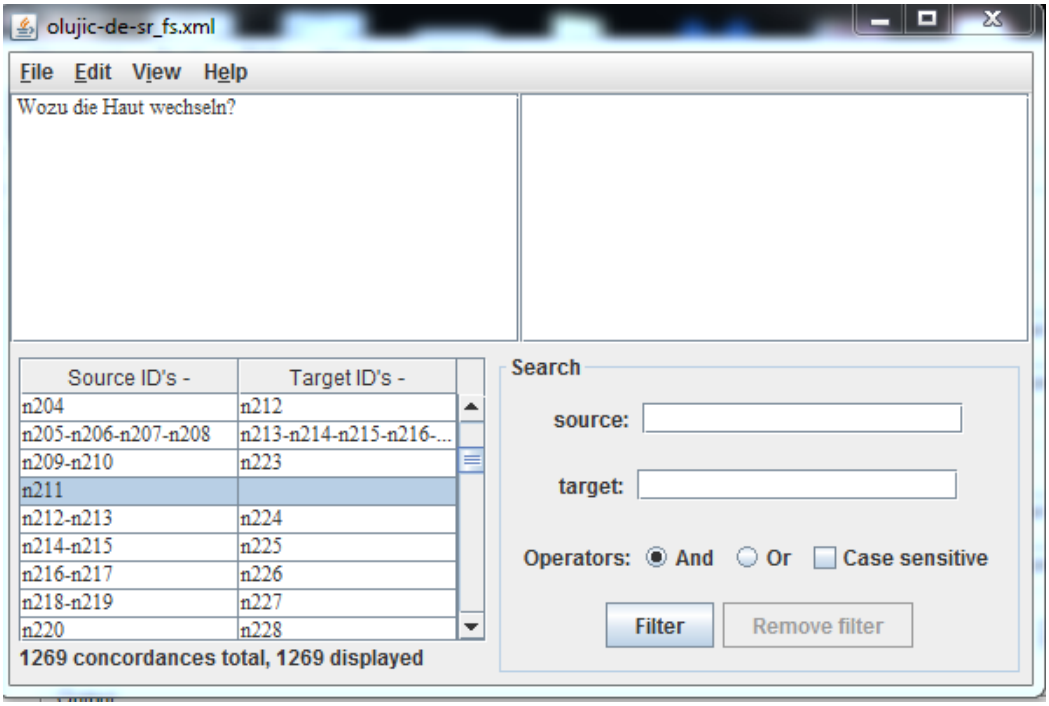
Слика 38. Concordancier - сегмент изворног језика (немачки) су две реченице, а одговарајући сегмент циљног језика (српски) је једна реченица



Слика 39. Concordancier - сегмент изворног језика (немачки) је једна реченица, а одговарајући сегмент циљног језика (српски) је део реченице



Слика 40. Concordancier - сегмент изворног (немачки) и еквивалентан сегмент циљног језика (српски) се састоје од две или више реченица које нису у истом редоследу



Слика 41. Concordancier - сегмент изворног језика (немачки) је једна реченица, а одговарајући сегмент циљног језика (српски) не постоји

Проблеми погрешно упарених сегмената које смо поменули појавили су се у скоро свим текстовима одабраним за овај корпус. Међутим, највише проблема је било са текстом „Излет у небо“ ауторке Гроздане Олујић и то посебно када је реч о проблемима који су наведени као прве три тачке, односно због великих разлика између оригиналног и преведеног текста[181]. Прецизније речено, сегментација српског и немачког текста овог романа није представљала велики проблем у техничком смислу. Текстови су обрађени без већих проблема и аутоматска сегментација оба текста је дала задовољавајуће резултате. Међутим, приликом упаривања појавила су се велика неслагања у погледу броја пасуса и броја реченица. Српски текст имао је много више реченица у односу на немачки, али је немачки текст имао много више пасуса у односу на српски. Ови проблеми су у осталим романима, у којима су се појавили, много лакше решени приликом ручне корекције погрешно упарених сегмената јер је преведени текст био доста усклађен са оригиналним. Међутим, приликом ручне корекције погрешно упарених сегмената романа „Излет у небо“ утврђено је да је немачки превод доста различит у односу на оригинални текст, односно можемо рећи да је преводилац потпуно слободно превео скоро цео роман. У току ручне корекције погрешно упарених сегмената аутор тезе је имао доста недоумица око спајања, односно раздвајања сегмената како би се постигло упаривање текстуалних целина које су смислене. Спајано је више реченица, па чак и пасуса у један сегмент како би се постигао неки вид конзистентности између оригинала и превода. Из тог разлога је било врло тешко успоставити 1-1 упаривање сегмената, а да они представљају логичке преводне парове. Након корекција урађена је паралелизација, али је остављен простор за даљи рад на овом тексту који би свакако подразумевао проналажење друге верзије превода на немачки језик на којој би се тестирало поновно упаривање са оригиналним текстом што би можда дало доста боље резултате у односу на тренутно доступну верзију паралелизованог текста. Слика 42 и Слика 43 илуструју неке примере упарених сегмената у роману „Излет у небо“ који су захтевали спајање више реченица и пасуса у један сегмент како би се на неки начин формирали преводни парови који имају смисла.

[181] Превод романа „Излет у небо“ је пронађен у Националној библиотеци Аустрије из које је позајмљен преко међубиблиотечке позајмице која је обављена у Универзитетској библиотеци „Светозар Марковић“.

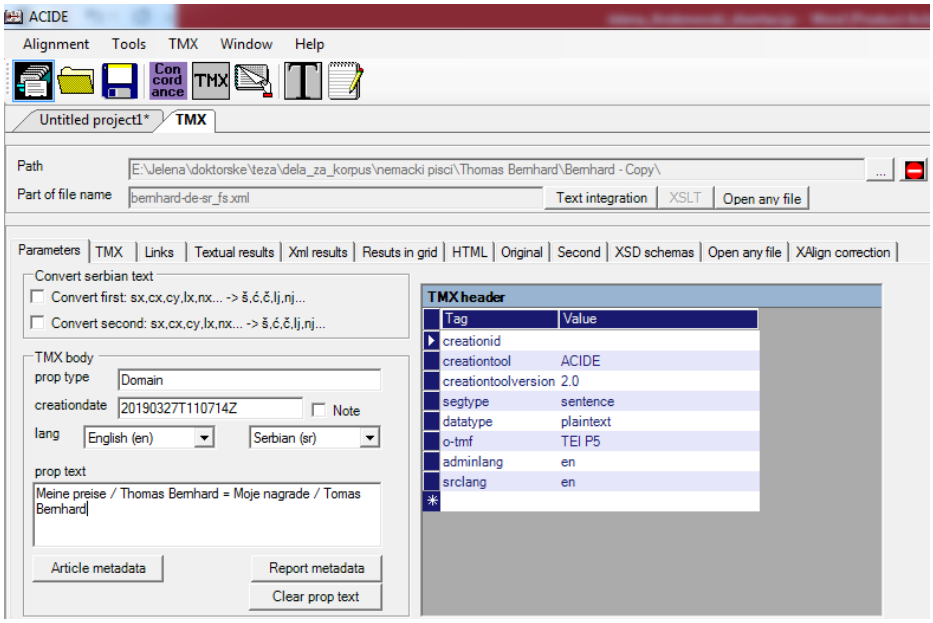
n622	Wer weiß, vielleicht schlafe ich, vielleicht habe ich keine Lust, midi zu unterhalten, vielleicht bereite ich irgendeine neue Rolle vor, mit der ich die Welt begeistern werde. Deswegen: halt ein, Hand! Störe die Ruhe nicht! Laß nur die blassen Gladiolen zurück mit wenigen Worten: der unvergeßlichen Lady Macbeth - eine „Ergebene Seele". Oder: der erhabenen Iphigenie auf Tauris — „Das Warten". Natürlich wäre ich auch manchmal generös, so daß ich in den Minuten zwischen zwei Aufnahmen oder zwei Auftritten, die von Beifall und Blumen angefüllt wären, dem „Warten" oder der „Ergebenen Seele" einige Male zuwinken würde. Dieses Winken wäre keinesfalls lang oder herzlich, ich hätte keine Zeit dafür . ..	n622	Ukras je i mesingana pločica s tvojim imenom na ulazu.
n623	Nein, es ist doch besser so — entschied ich — in die zwei dreisten Augen eines jungen Mannes versunken.	n623	I bez nje svi znaju da u beloj kući na vrhu Topčiderskog brda živi nezaboravna Lejdi Makbet, veličanstvena Antigona i još veličanstvenija Ana Karenjina. Ali, pazi, slava nije samo zadovoljstvo već i teret! Treba naučiti uloge, trčati s proba na snimanje, pa ponovo na probe. A onda, tu su i zavidnici, kritičari, sopstveni strah i sumnja da su u pravu one male zmije, koje tek pristižu, s tvrdnjom da si već predugo Lejdi Makbet, Antigona, Ana Karenjina: sjaši i ustupi mesto drugima, mladima, prodornijima, lepšima.

Слика 42. Пример упарених сегмената у роману „Излет у небо“ Гроздане Олујић

n879	— Warum hast du fast das ganze Jahr geschwiegen? fragt Mamas Mund, während der Blick uninteressiert von einem Gegenstand zum anderen gleitet. — Wera hat uns erzählt, daß es dir gut geht, aber daß du dich beinahe regelmäßig außerhalb des Heimes aufhältst. — Sie hat nicht gelogen. — Hat das etwas zu bedeuten?... — Mach dir keine Sorgen! — lache ich auf. — Es ist nicht das, was du meinst. Manchmal wird mir ihr Geschwätz über Kleider und Liebe lästig, dann gehe ich in die Bibliothek oder irre durch die Straßen. Das ist alles. Es gibt nichts Sorgenerregendes dabei. — Ja. Ja ... — murmelt sie, aber ich fühle, daß sie mir nicht glaubt. — Ihr habt einen neuen Geruch in der Wohnung? — sage ich. — Das ist auch alles, was wir an Neuem haben — sagt sie mit einer Stimme, in der weder Freude noch Trauer ist. — Bojka hat ihr Haar gefärbt, daher der Geruch. — Sie hofft sicher noch immer? — frage ich und fühle beinahe gleichzeitig die Unnötigkeit meiner Frage. — Der Mensch hofft, solange er lebt - sagt Mama und weicht einer direkten Antwort aus. — Sie ist noch immer ziemlich schön. Sie könnte einen andern finden. Stojan wird niemals zurückkehren. — Was weißt du darüber? — Mamas Stimme wird plötzlich schneidend und kalt. — Er wird wiederkommen. Sie glaubt fest daran. Sie hat einen Kredit für Möbel aufgenommen, und es besteht auch die Möglichkeit, daß sie zu einer Wohnung kommt. Sie hat sieben Jahre auf ihn gewartet, es dürfte nicht sein, daß er nicht wiederkommt. — Und dennoch: er wird nicht wiederkommen — sage ich. — Sie müßte das begreifen, wenn sie nicht ein Mädchen sein will, das bis zum Tode wartet. „Was kann ich dazu?" — meldet sich in meinem Ohr eine ungeduldige Stimme, und vor meinen Augen zeigt sich der fragende Blick eines hochgewachsenen jungen Mannes und das überraschte Gesicht einer Frau, die sich eng an ihn lehnt. „Ich habe sie nicht getrieben, daß sie sieben Jahre auf mich wartet!" „Du hast ihr geschrieben. Bist ab und zu gekommen" kehrt wie vom Magnetophonband meine Stimme zurück. Die Augen des jungen Mannes waren verlegen, aber kalt. „Anfangs. Später habe ich nur manchmal zu Neujahr	n879	Sa police na zidu, zapažam, nestale su moje knjige. Šta li su uradile s njima? Bacile ih? Prodale? Nemam snage da pitam. Nemam snage ni da proverim je li ono malo mojih prnja još uvek u ormanu. Ispod ormana izlazi bubašvaba, nekako mala, i kao ošamućena vrti se neko vreme po podu, zatim se ponovo vraća pod orman. Ubija li mama bubašvabe još uvek, na isti način, papučom? - Pa, eto, nemamo više ništa za večeru! — kaže mama, sklanjajući tanjire. - Znaš kako je s jednom platom... - ona postiđeno saginje glavu, a meni dolazi da ošamarim samu sebe. „Kog si đavola dolazila, budalo božja?" pitam se i s tugom razgledam mamino lice kao neko gradom opustošeno polje: oči upale, kosa pobelela, vrat uvučen u ramena kao da se nečega plaši, a na istanjenoj koži slepoočnice premreženoj borama podrhtava jedna modra žilica. Nekakvim usplahirenim pokretima ona, svaki čas, uzima nešto u ruke, tanjire, pletivo, bilo šta. Sklanja pogled. Oslabila je mnogo za ovih nekoliko meseci, ostarila, nekako se smanjila, neće dugo. A bila je žena kao grad: ćutljiva, uspravna, hrabra. Za vreme bombardovanja svi su trčali u naš podrum, uvereni da im se kraj Andrine Natke ništa ne može dogoditi. Nakon dedine smrti u očima joj se, prvi put, javila neka tama, kao neka nevidljiva crna paučina, ali još uvek je hodala uspravno, visoko uzdignute glave. Povila se tek kad je čula da je Marko poginuo, a nekako kao spuznula k zemlji kad smo skinuli tatu s konopca i zakopali ispod kruške divljake u dvorištu. Sada je manja od mene za pola glave, uznemirena, drhtavih ruku. A nekad je u njenim širokim, smeđim očima, bio nastanjen čitav moj svet. Hodala sam za njom kao psetance, čistila joj cipele, odnosila đubre i donosila vodu, sve dok se nije dogodilo ono s kupinama. Bilo mi je tada devet godina. Te godine, u vrbaku, kupine su rodile kao lude: sve se modrilo unaokolo od njih. Rešivši da obradujem majku, uzela sam vedricu i od jutra do večeri provlačila se kroz šibljak i kupinove vreže, skupljajući bobicu po bobicu. Još malo, još malo! Ako nakupim dovoljno, mama će napraviti kupinovo vino. U sebi sam je

Слика 43. Пример једног упареног сегмента у роману „Излет у небо“ Гроздане Олујић

После контроле, раздвајања сегмената и корекције погрешно упарених сегмената постигнуто је у потпуности 1-1 упаривање па је покретањем одговарајуће апликације у менију TMX (Слика 44) генерисана датотека у TMX формату на основу XML датотека са кодираним информацијама о упареним сегментима које су произведене у претходном кораку. Добијена датотека у TMX формату садржи заглавље (header) и „тело" текста (body) које чине преводне јединице означене XML етикетом <tu> са својим карактеристикама означеним XML етикетом <prop> (<prop type="Domain">Meine Preise = Moje nagrade / Thomas Bernhard</prop>) и унутар њих варијанте јединица превођења означене XML етикетом <tuv> која садржи XML атрибут xml:lang којим се дефинише језик варијанте јединице превођења (xml:lang="de" за немачки и xml:lang="sr" за српски) (Слика 45).



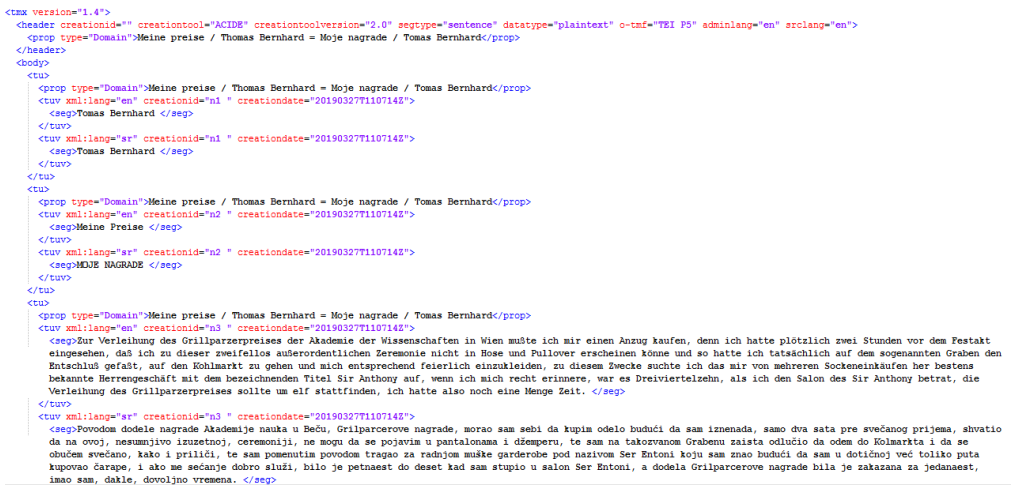
Слика 44. Мени TMX – део програма ACIDE за генерисање TMX документа

Добијени документ у формату TMX је валидан у односу на следећег DTD:

```
<! ELEMENT tmx (header, body)>
<! ELEMENT header (prop)>
<! ATTLIST header creationid ID #REQUIRED>
<! ATTLIST header creationtool CDATA #REQUIRED>
<! ATTLIST header creationtoolversion CDATA #REQUIRED>
<! ATTLIST header segtype CDATA #REQUIRED>
```

```
<! ATTLIST header datatype CDATA #REQUIRED>
<! ATTLIST header o-tmf CDATA #REQUIRED>
<! ATTLIST header adminlang CDATA #REQUIRED>
<! ATTLIST header srclang CDATA #REQUIRED>
<! ELEMENT prop (#PCDATA)>
<! ATTLIST prop type CDATA #REQUIRED>
<! ELEMENT body (tu)+>
<! ELEMENT tu (prop, tuv+)>
<! ELEMENT tuv (seg)+>
<! ATTLIST tuv xml:lang CDATA #REQUIRED>
<! ATTLIST tuv creationid ID #REQUIRED>
<! ATTLIST tuv creationdate CDATA #REQUIRED>
<! ELEMENT seg (#PCDATA)>
```

Овај DTD дефинише да валидни XML документ садржи коренски елемент *tmx* који садржи један елемент *header* и један елемент *body*. Елемент *header* садржи један елемент *prop* и више атрибута који дефинишу неке техничке карактеристике записа: идентификациони број записа (*creationid*), назив алата којим се ради упаривање сегмената (*creationtool*), верзију алата који се користи (*creationtoolversion*), тип сегмената (*segtype*), тип података (*datatype*), формат преводилачких меморија (*o-tmf*), језик сумеђе алата (*adminlang*) и језик текстуалних података (*srclang*). Вредност свих атрибута су карактерски подаци (*CDATA*), осим атрибута *creationid* чија је вредност јединствени идентификатор (*ID*). Сви поменути атрибути су обавезни (*#REQUIRED*). Садржај елемента *prop су* парсирани карактерски подаци (#PCDATA). Елемент *body* садржи један или више елемената *tu* (*једна или више јединица превођења*), а сваки елемент *tu* садржи један елемент *prop* и један или више елемената *tuv* (*једна или више варијанти јединица превођења*). Садржај сваког елемента *tuv* је један или више елемената *seg* (*једна или више реченица*). Према DTD, садржај сваког елемента *seg* су парсирани карактерски подаци односно текст.



Слика 45. TMX формат паралелизованог текста

Захваљујући развијеним модулима софтверског алата ACIDE обезбеђене су и додатне могућности визуелизације, али и кориговања грешака упаривања које кориснику знатно олакшавају успешно завршавање процеса паралелизације. Интегрисањем јединица паралелизованих текстова креира се интерна табеларна репрезентација (Слика 46). Поред табеларног приказа генерисан је и текстуални запис у такозваном „Vanila" формату (Слика 47), то јест формат који је производио један од првих програма за паралелизацију (Danielsson and Ridings 1997) и XML формат интегрисаних података (Слика 48). На крају, уз помоћ XSLT (Extensible Stylesheet Language Transformations) трансформација XML формат података је трансформисан у формат HTML (Слика 49) који је погодан уа објављивање на вебу.

	targets	source_Text	target_Text
▶	n1 x1	n1: Tomas Bernhard	n1: Tomas Bernhard
	n2 x2	n2: Meine Preise	n2: MOJE NAGRADE
	n3 x3	n3: Zur Verleihung des Grillparzerpreises der Akademie	n3: Povodom dodele nagrade Akademije nauka u Beču,
	n4 x4	n4: Ich hatte die Absicht, mir, wenn schon von der Stang	n4: Kad je već
	n5 x5	n5: Die Schwierigkeit, sich in den sogenannten feineren	n5: Poteškoće na koje kupac nailazi u ekskluzivnijim rad
	n6 x6	n6: Aber natürlich hat der angesprochene Verkäufer auc	n6: I naravno da prodavac, kojem se čovek obrati, kupca
	n7 x7	n7: So dauerte es auch damals im Sir Anthony länger al	n7: Tako je i tad u SerEntoniju potrajalo duže nego što j
	n8 x8	n8: Tatsächlich waren mir die Umstände in diesem Gesc	n8: Naravno, prilike u ovoj radnji bile su mi poznate iz m
	n9 x9	n9: Ich schritt auf das Regal mit den in Frage kommend	n9: Uputio sam se ka odelima koja su eventualno dolazil
	n10 x10	n10: Ich prüfte die Stoffqualität und machte sogleich in d	n10: Proverio sam kvalitet materijala i smesta u kabini z
	n11 x11	n11: Ich beugte mich ein paarmal vor und lehnte mich z	n11: Nagnuo sam se napred, nekoliko puta napred-naza
	n12 x12	n12: Ich zog den Rock an, drehte mich ein paarmal vor d	n12: Obukao sam sako, okrenuo se pred ogledalom nek
	n13 x13	n13: Ich ging ein paar Schritte mit dem Anzug durch das	n13: U odelu sam se malo prošetao radnjom, tragajući t
	n14 x14	n14: Schließlich sagte ich, daß ich den Anzug anbehalte	n14: Na kraju rekoh da bih ostao u odelu i dodatno obuk
	n15 x15	n15: Ich suchte mir eine Krawatte aus, band sie mir um,	n15: Pronašao sam kravatu, vezao je, pritegao čvor krav
	n16 x16	n16: Meine alte Hose und meinen Pullover hatten sie mi	n16: Stare pantalone i stari džemper zapakovali su mi u
	n17 x17	n17: Beim Gerstner wollten wir noch kurz vor der Feierli	n17: Nameravali smo da u Gerstneru pojedemo po send
	n18 x18	n18: Meine Tante war schon im Gerstner gewesen, sie h	n18: Tetka je već bila u Gerstneru, moju metamorfozu o
	n19 x19	n19: Ich selbst hatte bis zu diesem Zeitpunkt jahrelang k	n19: Inače, odelo nisam nosio godinama, sve do tog tre
	n20 x20	n20: In dieser Aufmachung war ich, erinnere ich mich, ei	n20: U ovom izdanju bio sam, sećam se, nekoliko puta i
	n21 x21	n21: Plötzlich, auf dem Graben wie gesagt und zwei Stu	n21: Iznenada, kao što rekoh, dva sata pre dodele nagra
	n22 x22	n22: Im Hinsetzen im Gerstner hatte ich aufeinmal das	n22: Prilikom zauzimanja mesta u Gerstneru, iznenada
	n23 x23	n23: Ich bestellte mir ein Sandwich und trank ein Glas Bi	n23: Naručio sam sendvič i popio čašu piva.

Слика 46. Табела интегрисаних јединица паралелизованог текста

```
*** Link: 1 - 1 ***
<source_Text id='n1'>
Tomas Bernhard
<target_Text id='n1'>
Tomas Bernhard

*** Link: 1 - 1 ***
<source_Text id='n2'>
Meine Preise
<target_Text id='n2'>
MOJE NAGRADE

*** Link: 1 - 1 ***
<source_Text id='n3'>
Zur Verleihung des Grillparzerpreises der Akademie der Wissenschaften in Wien mußte ich mir einen Anzug kaufen, denn ich hatte plötzlich zwei Stunden vor dem Festakt eingesehen, daß ich zu dieser zweifellos außerordentlichen Zeremonie nicht in Hose und Pullover erscheinen konnte und so hatte ich tatsächlich auf dem sogenannten Graben den Entschluß gefaßt, auf den Kohlmarkt zu gehen und mich entsprechend feierlich einzukleiden, zu diesem Zwecke suchte ich das mir von mehreren Sockeneinkäufen her bestens bekannte Herrengeschäft mit dem bezeichnenden Titel Sir Anthony auf, wenn ich mich recht erinnere, war es Dreiviertelzehn, als ich den Salon des Sir Anthony betrat, die Verleihung des Grillparzerpreises sollte um elf stattfinden, ich hatte also noch eine Menge Zeit.
<target_Text id='n3'>
Povodom dodele nagrade Akademije nauka u Beču, Grilparcerove nagrade, morao sam sebi da kupim odelo budući da sam iznenada, samo dva sata pre svečanog prijema, shvatio da na ovoj, neuporedivo izuzetnoj, ceremoniji, ne mogu da se pojavim u pantalonama i džemperu, te sam na takozvanom Grabenu zaista odlučio da odem do Kolmarkta i da se obučem svečano, kako i priliči, te sam pomenutim povodom tragao za radnjom muške garderobe pod nazivom Ser Entoni koju sam znao budući da sam u dotičnoj već toliko puta kupovao čarape, i ako me sećanje dobro služi, bilo je petnaest do deset kad sam stupio u salon Ser Entoni, a dodela Grilparcerove nagrade bila je zakazana za jedanaest, imao sam, dakle, dovoljno vremena.

*** Link: 1 - 1 ***
<source_Text id='n4'>
Ich hatte die Absicht, mir, wenn schon von der Stange, so doch den besten Reinwollanzug in Arthrazit anzuschaffen, dazu die passenden Socken, eine Krawatte und ein Hemd von Arrow, ganz fein, graublau gestreift.
<target_Text id='n4'>
Kad je već skupa konfekcija, onda nek bude najbolje odelo, od čiste vune, antracit sive boje, a uz pomenuto - prikladne čarape, prikladna kravata i košulja marke erou, veoma fina, na tanke sivoplave pruge.

*** Link: 1 - 1 ***
<source_Text id='n5'>
Die Schwierigkeit, sich in den sogenannten feineren Geschäften gleich verständlich zu machen, ist bekannt, auch wenn der Kunde sofort und auf die präziseste Weise sagt, was er will, wird er zuerst einmal ungläubig angestarrt, bis er seinen Wunsch wiederholt hat.
<target_Text id='n5'>
```

Слика 47. Текстуални запис у такозваном “Vanila” формату

```
- <link>
    <targets>n1 x1</targets>
    <source_Text>n1: Tomas Bernhard</source_Text>
    <target_Text>n1: Tomas Bernhard</target_Text>
  </link>
- <link>
    <targets>n2 x2</targets>
    <source_Text>n2: Meine Preise</source_Text>
    <target_Text>n2: MOJE NAGRADE</target_Text>
  </link>
- <link>
    <targets>n3 x3</targets>
    <source_Text>n3: Zur Verleihung des Grillparzerpreises der Akademie der Wissenschaften in Wien mußte ich mir einen Anzug kaufen, denn ich hatte plötzlich zwei
      Stunden vor dem Festakt eingesehen, daß ich zu dieser zweifellos außerordentlichen Zeremonie nicht in Hose und Pullover erscheinen könne und so hatte ich
      tatsächlich auf dem sogenannten Graben den Entschluß gefaßt, auf den Kohlmarkt zu gehen und mich entsprechend feierlich einzukleiden, zu diesem Zwecke suchte
      ich das mir von mehreren Sockeneinkäufen her bestens bekannte Herrengeschäft mit dem bezeichnenden Titel Sir Anthony auf, wenn ich mich recht erinnere, war es
      Dreiviertelzehn, als ich den Salon des Sir Anthony betrat, die Verleihung des Grillparzerpreises sollte um elf stattfinden, ich hatte also noch eine Menge
      Zeit.</source_Text>
    <target_Text>n3: Povodom dodele nagrade Akademije nauka u Beču, Grilparcerove nagrade, morao sam sebi da kupim odelo budući da sam iznenada, samo dva sata
      pre svečanog prijema, shvatio da na ovoj, nesumnjivo izuzetnoj, ceremoniji, ne mogu da se pojavim u pantalonama i džemperu, te sam na takozvanom Grabenu
      zaista odlučio da odem do Kolmarkta i da se obučem svečano, kako i priliči, te sam pomenutim povodom tragao za radnjom muške garderobe pod nazivom Ser Entoni
      koju sam znao budući da sam u dotičnoj već toliko puta kupovao čarape, i ako me sećanje dobro služi, bilo je petnaest do deset kad sam stupio u salon Ser Entoni, a
      dodela Grilparcerove nagrade bila je zakazana za jedanaest, imao sam, dakle, dovoljno vremena.</target_Text>
  </link>
- <link>
    <targets>n4 x4</targets>
```


Слика 48. XML формат интегрисаних података

Meine preise / Thomas Bernhard = Moje nagrade / Tomas Bernhard

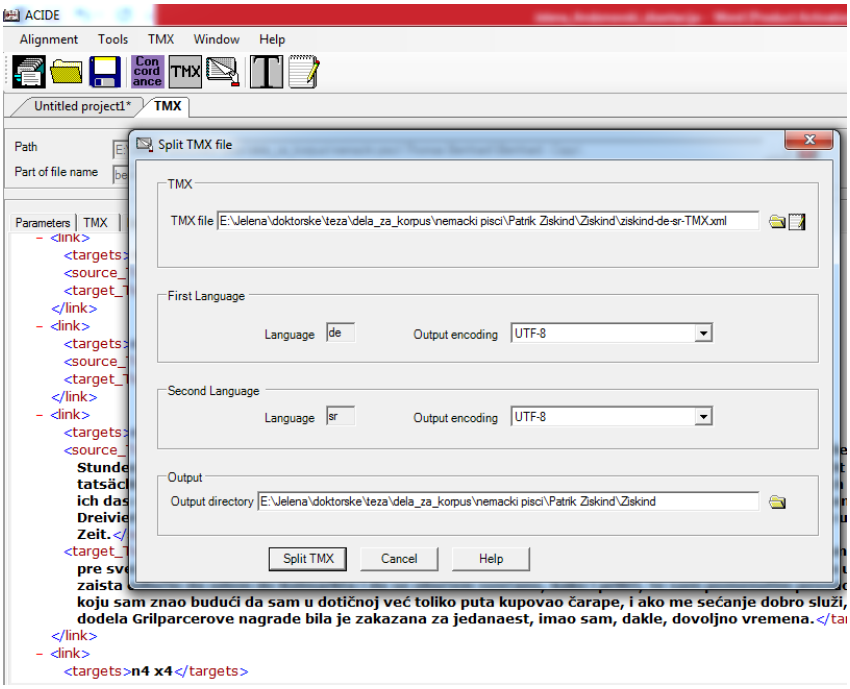
English (en)	Serbian (sr)
n1 : Tomas Bernhard	**n1 :** Tomas Bernhard
n2 : Meine Preise	**n2 :** MOJE NAGRADE
n3 : Zur Verleihung des Grillparzerpreises der Akademie der Wissenschaften in Wien mußte ich mir einen Anzug kaufen, denn ich hatte plötzlich zwei Stunden vor dem Festakt eingesehen, daß ich zu dieser zweifellos außerordentlichen Zeremonie nicht in Hose und Pullover erscheinen könne und so hatte ich tatsächlich auf dem sogenannten Graben den Entschluß gefaßt, auf den Kohlmarkt zu gehen und mich entsprechend feierlich einzukleiden, zu diesem Zwecke suchte ich das mir von mehreren Sockeneinkäufen her bestens bekannte Herrengeschäft mit dem bezeichnenden Titel Sir Anthony auf, wenn ich mich recht erinnere, war es Dreiviertelzehn, als ich den Salon des Sir Anthony betrat, die Verleihung des Grillparzerpreises sollte um elf stattfinden, ich hatte also noch eine Menge Zeit.	**n3 :** Povodom dodele nagrade Akademije nauka u Beču, Grilparcerove nagrade, morao sam sebi da kupim odelo budući da sam iznenada, samo dva sata pre svečanog prijema, shvatio da na ovoj, nesumnjivo izuzetnoj, ceremoniji, ne mogu da se pojavim u pantalonama i džemperu, te sam na takozvanom Grabenu zaista odlučio da odem do Kolmarkta i da se obučem svečano, kako i priliči, te sam pomenutim povodom tragao za radnjom muške garderobe pod nazivom Ser Entoni koju sam znao budući da sam u dotičnoj već toliko puta kupovao čarape, i ako me sećanje dobro služi, bilo je petnaest do deset kad sam stupio u salon Ser Entoni, a dodela Grilparcerove nagrade bila je zakazana za jedanaest, imao sam, dakle, dovoljno vremena.
n4 : Ich hatte die Absicht, mir, wenn schon von der Stange, so doch den besten Reinwollanzug in Anthrazit anzuschaffen, dazu die passenden Socken, eine Krawatte und ein Hemd von Arrow, ganz	**n4 :** Kad je već skupa konfekcija, onda nek bude najbolje odelo, od čiste vune, antracit sive boje, a uz pomenuto - prikladne čarape, prikladna kravata i košulja marke arow, veoma fina, na tanke sivoplave

Слика 49. Генерисани HTML формат погодан за приказивање на вебу

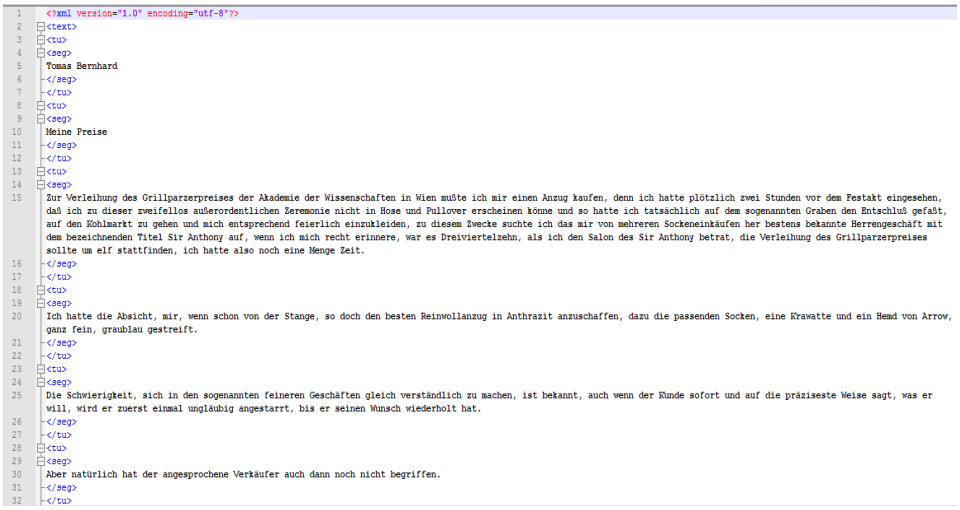
Овако генерисана датотека у TMX формату је у следећем кораку разложен на две XML датотеке која садрже преводне јединице. Разлагање датотеке у TMX формату на XML датотеке појединачних језика урађено је опцијом Split TMX уграђеном у програмски пакет ACIDE (Слика 50). Слика 51 и Слика 52 илуструју појединачне XML датотеке које су резултат разлагања TMX датотеке на њен немачки, односно српски део. Разлагањем датотеке у

TMX формату добијене су потпуно кориговане верзије улазних датотека појединачних језика спремне за креирање паралелног корпуса помоћу програмског пакета IMS CWB.

Као последњи корак урађена је вертикализација добијених коначних датотека (опција Verticalize у пакету ACIDE) (Слика 53), што је још један корак неопходан за укључивање паралелних текстова у корпус подржан системом IMS CWB.



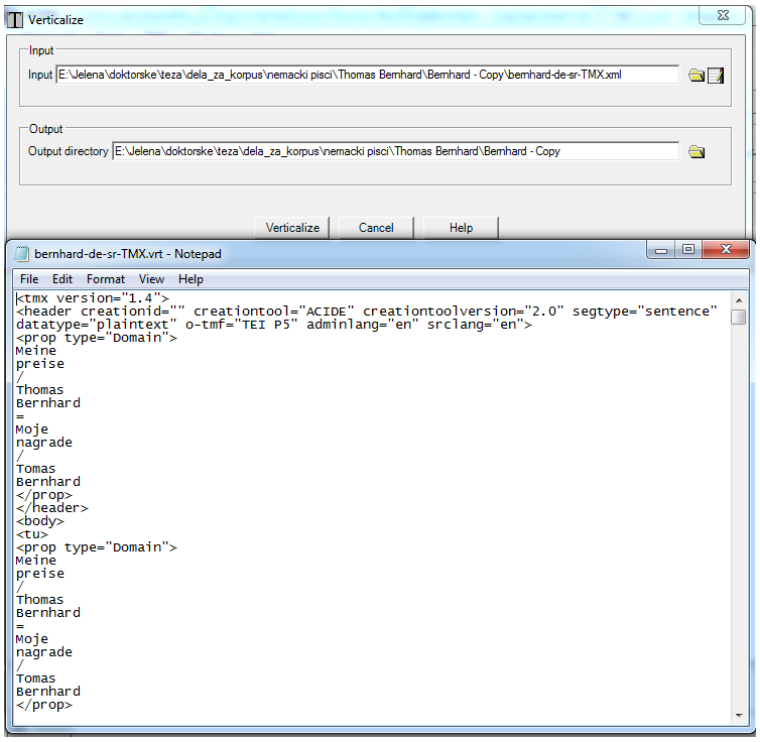
Слика 50. Разлагање TMX документа - Split TMX



Слика 51. Резултат разлагања TMX документа – немачки део паралелног текста

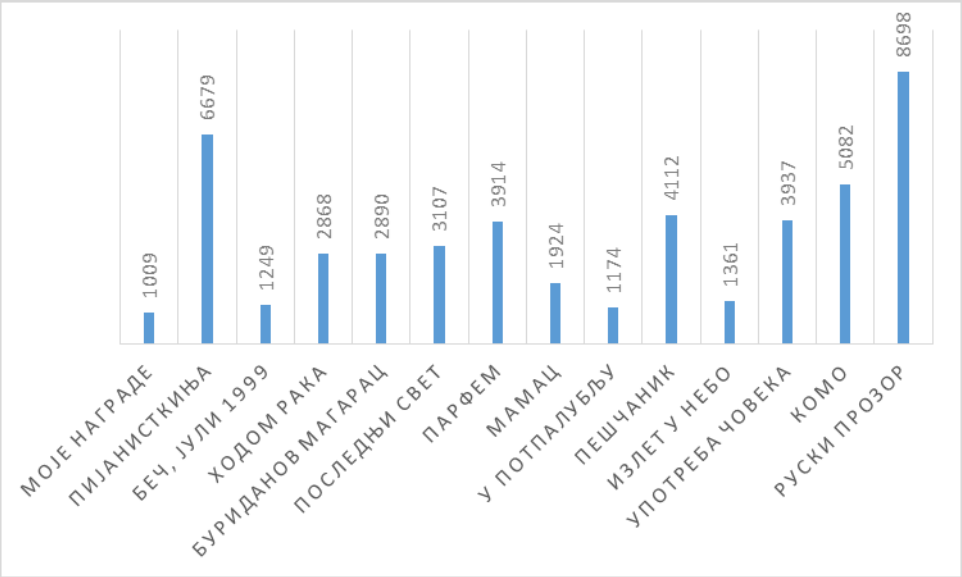
```
<?xml version="1.0" encoding="utf-8"?><text>
<tu>
<seg>
Tomas Bernhard
</seg>
</tu>
<tu>
<seg>
MOJE NAGRADE
</seg>
</tu>
<tu>
<seg>
Povodom dodele nagrade Akademije nauka u Beču, Grilparcerove nagrade, morao sam sebi da kupim odelo budući da sam iznenada, samo dva sata pre svečanog prijema, shvatio da na ovoj, nesumnjivo izuzetnoj, ceremoniji, ne mogu da se pojavim u pantalonama i džemperu, te sam na takozvanom Grabenu zaista odlučio da odem do Kolmarkta i da se obučem svečano, kako i priliči, te sam pomenutim povodom tragao za radnjom muške garderobe pod nazivom Ser Entoni koju sam znao budući da sam u dotičnoj već toliko puta kupovao čarape, i ako me sećanje dobro služi, bilo je petnaest do deset kad sam stupio u salon Ser Entoni, a dodela Grilparcerove nagrade bila je zakazana za jedanaest, imao sam, dakle, dovoljno vremena.
</seg>
</tu>
<tu>
<seg>
Kad je već skupa konfekcija, onda nek bude najbolje odelo, od čiste vune, antracit sive boje, a uz pomenuto - prikladne čarape, prikladna kravata i košulja marke erou, veoma fina, na tanke sivoplave pruge.
</seg>
</tu>
<tu>
<seg>
Poteškoće na koje kupac nailazi u ekskluzivnijim radnjama, u želji da ga smesta razumeju, poznate su čak i kad kupac na najprecizniji mogući način izrazi šta zapravo hoće - najpre se na njega podozrivo izbeče sve dok svoju želju ne ponovi.
</seg>
</tu>
<tu>
<seg>
I naravno da prodavac, kojem se čovek obrati, kupca ne razume čak ni tad.
</seg>
</tu>
```

Слика 52. Резултат разлагања TMX документа – српски део паралелног корпуса



Слика 53. Вертикализација текста

Након завршене паралелизације свих одабраних текстова за корпус произведено је укупно 48.004 преводних парова[182] (Слика 54), од тога у колекцији „Романи оригинално написани на немачком језику“ 21.716, а у колекцији „Романи оригинално написани на српском језику“ 26.288. На крају је припремљен корпус од 1.657.329 речи (Табела 2). Од романа оригинално написаних на немачком језику најкраћи је „Беч, јули 1999“ аутора Мила Дора са укупно 47.436 речи у српском (24.300) и немачком (23.136) делу корпуса, док је најдужи „Пијанисткиња“ ауторке Елфриде Јелинек са укупно 177.207 речи у српском (90.743) и немачком (86.464) делу корпуса. Од романа оригинално написаних на српском најкраћи је „У потпалубљу“ аутора Владимира Арсенијевића са укупно 49.523 речи у српском (24.329) и немачком (25.194) делу корпуса, док је најдужи „Руски прозор“ аутора Драгана Великића са укупно 213.675 речи у српском (96.952) и немачком (116.723) делу корпуса. најмање преводних парова има за роман „Моје награде“, укупно 1.009, док највише има за роман „Руски прозор“, укупно 8.698.



Слика 54. Број преводних парова у корпусу СрпНемКор према романима појединачно

[182] Укупан број произведених парова може се видети на http://jerteh.rs/biblisha/Statistika.aspx?tip=2&JID=11

Табела 2. Број речи у корпусу СрпНемКор према романима појединачно

Немачки писци			Српски писци		
	Немачки текст	Српски текст		Српски текст	Немачки текст
Бернхард	24.973	23.105	**Албахари**	39.698	43.706
де Бројн	72.014	79.283	**Арсенијевић**	24.329	25.194
Дор	23.136	24.300	**Киш**	64.571	68.131
Грас	51.376	52.099	**Великић**	96.952	116.723
Јелинек	90743	86464	**Ваљаревић**	59.810	63.352
Рансмајер	70562	67356	**Тишма**	91.217	97.116
Зискинд	72233	71559	**Олујић**	29.117	28.210
Укупно	405037	404166	**Укупно**	405.694	442.432
Укупан број речи корпусу	**1.657.329**				

6.3 СрпНемКор у Библиши

Паралелизовани романи за српско-немачки корпус постављени су у Библишу у оквиру посебне колекције „СрпНемКор" како би се тестирале расположиве могућности претраге које Библиша нуди, посебно у погледу претраге на немачком језику. Приликом креирања одлучили смо се да колекцију СрпНемКор поделимо на две потколекције: потколекцију романа српских аутора и потколекцију романа аутора немачког говорног подручја. Као и код већине дигиталних библиотека и репозиторијума, тако је и у Библиши било неопходно доделити идентификационе бројеве и колекцији и потколекцијама. Колекција СрпНемКор је добила идентификациони број 11, док су потколекције добиле идентификационе бројеве 11.1 (потколекција „Романи оригинално написани на немачком језику") и 11.2 (потколекција „Романи оригинално написани на српском језику").

Библиша је примарно дизајниран за складиштење чланака из часописа који излазе на српском и енглеском језику што значи да је структура метаподатака била прилагођена структури чланка, а претрага могућа само на српском и енглеском језику коришћењем расположивих лексичких ресурса на ова два језика. Складиштењем колекције СрпНемКор у Библишу ови сегменти система су ажурирани. Како колекција СрпНемКор садржи романе тако је и структура метаподатака прилагођена да на релевантан начин опише

монографске публикације (детаљније објашњено у одељку 6.3.1). Када је реч о претрази, у претраживачима и исписима резултата претраге додат је немачки језик, а када је реч о допуни лексичких ресурса у овој фази рада допуњена је база Терми како би се омогућило семантичко проширење упита и на немачком језику.

6.3.1 Структура метаподатака

Сви романи припремљени за СрпНемКор описани су одговарајућим метаподацима упоредо на два језика, у нашем случају на српском и немачком, и представљени у формату TMX (детаљније објашњено у поглављу 2 одељак 2.3) који је генерисан у поступку паралелизације. За сваки роман унети су: УДК број, име аутора, наслов, број страна (односи се на штампани примерак на основу којег је припремљен дигитални текст), сажетак (дат је кратак опис романа), место издавања, издавач, година издавања и име преводиоца. Метаподаци су у JSON формату са схемом која представља сет елемената дизајнираних да опишу колекцију у целини, све њене потколекције и објекте појединачно (Stanković et al. 2017). У наставку је дат пример метаподатака у JSON формату за роман „Моје награде / Meine Preise" аустријског писца Томаса Бернхарда. Запис је подељен на три сегмента у зависности од врсте метаподатака који се наводе (Слика 55).

Први део записа садржи податке за идентификацију објекта који се описује: идентификациони број објекта, УДК број, идентификациони број колекције и потколекције којој објекат припада, као и идентификационе бројеве записа из релевантних нормативних база података са којима је успостављена веза. Идентификациони број објекта додељује се приликом учитавања објекта у Библишу. Како су сви објекти у Библиши део одређене колекције, а у већини случајева и потколекције, тако је идентификациони број објекта састоји из три дела. Први део је идентификациони број колекције, други део је идентификациони број потколекције у оквиру колекције, а трећи део је идентификациони број објекта у потколекцији чиме се одређује његова позиција у дигиталној библиотеци.

```
{ "id" : "11.1.001",
   "UDC" : "821.112.2(436)-31",
   "JournalID" : "11.1",
   "CollectionID" : "11",
   "Refs" : ["VIAF:239728568", "GND:990847837", "WikiData:Q1287985", "LCNAF:n2010027085"],
    "Authors" : [
        {"OrdinalNo" : "1",
            "Name" : "Thomas Bernhard",
            "Mail" : "",
            "Institution" : [{"lang" : "de", "Institution" : ""}, {"lang" : "sr", "Institution" : ""}],
            "Refs" : [ "VIAF:12305044", "GND:118509861", "WikiData:Q44336", " LCNAF:n50007084"]}],
    "About" : [{ "lang" : "de",
              "Title" : "Meine Preise",
"Category" : "Roman",
"URL" : "",
"Abstract" : "\"Meine Preise\" wird zum 20. Todestag im Februar 2009 erstmals veröffentlicht. Bernhard hat
  sie 1980 fertiggestellt, zu Lebzeiten aber nie publiziert. Der Text gliedert sich in neun Kapitel und
  einen Anhang. Zornig Rückschau haltend, zieht Bernhard darin eine Bilanz der ihm verliehenen
  Literaturpreise.",
"Keywords" : "",
"Pages" : "139",
"PublishingPlace" : "Frankfurt am Mein",
"Publisher" : "Suhrkamp",
"PublishYear" : "2009",
"Translator" : ""},
 { "lang" : "sr",
"Title" : "Moje nagrade",
"Category" : "roman",
"URL" : "",
"Abstract" : "Knjiga \"Moje nagrade\" objavljena je 2009. godine na 20. godišnjicu smrti jednog od
  najznačajnijih pisaca nemačkog govornog područja, Tomasa Bernharda. Bernhard uzima u obzir književne
  nagrade koje je dobio i predstavlja govore koji su se uvek završavali skandalozno.",
"Keywords" : "","Pages" : "115",
              "PublishingPlace" : "Beograd",
               "Publisher" : "LOM",
             "PublishYear" : "2012",
               "Translator" : "prevela sa nemačkog Bojana Denić"}]}
```


Слика 55. Запис у Библиши за роман „Моје награде" аутора Томаса Бернхарда

У нашем примеру идентификациони број објекта је 11.1.001 на основу чега можемо утврдити да је у питању колекција са идентификационим бројем 11, потколекција је 11.1 и објекат са идентификационим бројем 001. На основу ових идентификационих бројева и података из претходног одељка о структурној организацији српско-немачког паралелног корпуса у Библиши можемо утврдити да роман "Моје награде / Meine Preise" припада колекцији СрпНемКор, потколекцији писаца немачког говорног подручја и да је први на листи романа у овој потколекцији.

На основу УДК броја 821.112.2(436)-31 може се утврдити да је у питању роман (роман се у УДК таблицама означава бројем 31 иза главног УДК) из аустријске књижевности (аустријска књижевност је у УДК таблицама означена бројем 821.112.2(436)). На крају овог првог дела дати су идентификациони бројеви записа о роману "Моје награде / Meine Preise" у нормативним базама VIAF, GND и LCNAF, као и у

бази података Википодаци и на основу чега је и направљена веза са њима. Значај ових база података, њихова намена, структура као и начин увезивања са релевантним записима у њима биће детаљније објашњено у одељку 6.5 овог поглавља.

У другом делу записа, „Authors”, наводе се подаци о аутору односно ауторима: име аутора, адреса електронске поште и афилијација на два језика. У нашем примеру име аутора је „Thomas Bernhard”. Како је структуром предвиђено да може постојати више аутора тако је Томасу Бернхарду аутоматски додељен идентификациони број „1”. Поред имена аутора, предвиђен је и унос адресе електронске поште аутора и назив институције у којој је аутор запослен на два језика што је у нашем случају остало непопуњено.

У трећем делу записа дати су библиографски метаподаци (детаљније објашњено у поглављу 4 одељак 4.1) о роману: наслов, категорија и сажетак на српском и немачком језику, као и број страна, место издавања, име издавача, година издавања и име преводиоца.

Корисници имају могућност да у Библиши сагледају листу свих романа у паралелном корпусу СрпНемКор у делу *Metadata browse/Projekti/SrpNemKor* (Слика 56). Сваки роман на листи има своју идентификацију која указује на колекцију и потколекцију којој роман припада као на његов редни број у припадајућој потколекцији, а корисници имају могућност да приступе детаљним метаподацима, роману у формату TMX, сажетом прегледу одабраног романа у формату PDF и записима у базама VIAF, GND, LCNAF и Википодаци везаним за одабраног аутора и одабрани наслов.

Number of documents: 14 Number of sentences: 48004

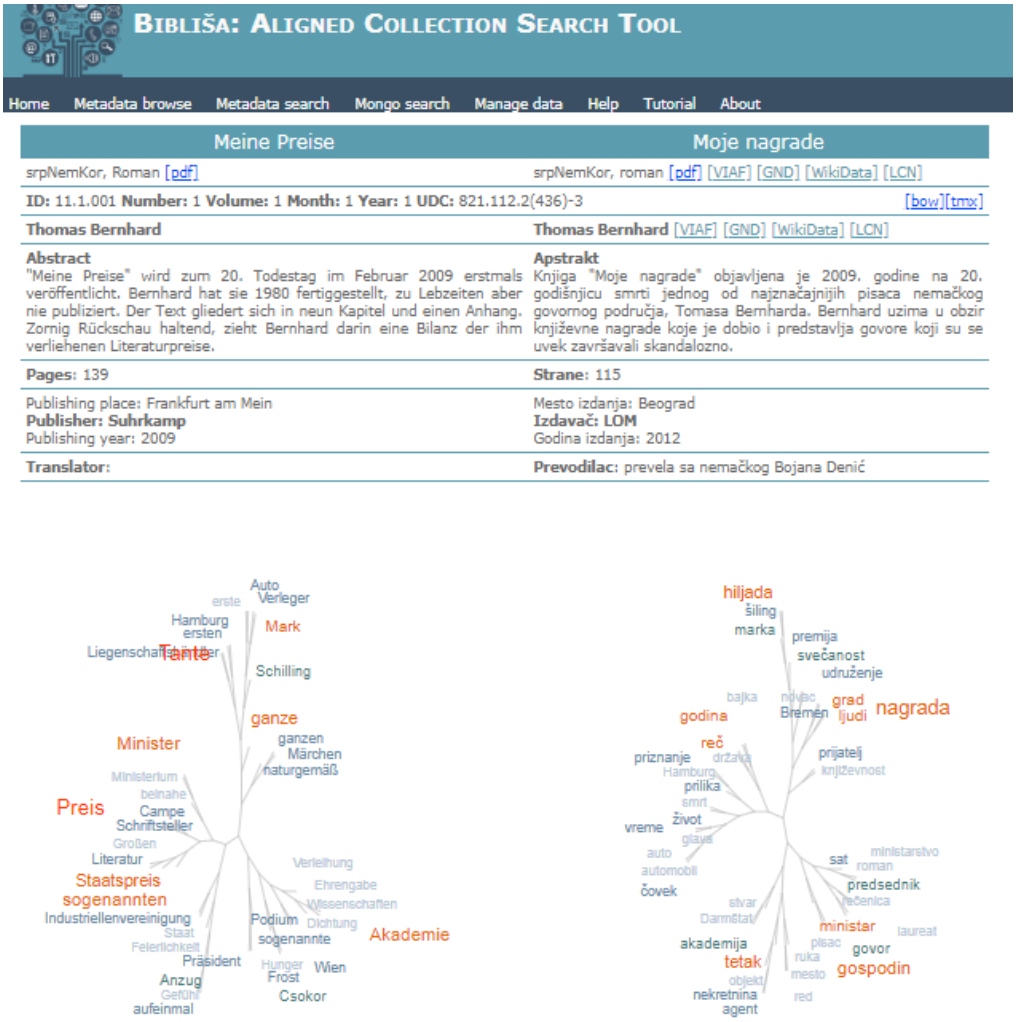
Journal Document	About Document
11.1.001 vol. 1 11.1	**Moje nagrade** [VIAF] [GND] [WikiData] [LCN] authors: **Thomas Bernhard** [VIAF] [GND] [WikiData] [LCN] [detaljnije] [tmx] [pdf]
11.1.002 vol. 1 11.1	**Pijanistkinja** [VIAF] [GND] [WikiData] [LCN] authors: **Elfride Jelinek** [VIAF] [GND] [WikiData] [LCN] [detaljnije] [tmx] [pdf]
11.1.003 vol. 1 11.1	**Beč, juli 1999** [VIAF] [GND] authors: **Milo Dor** [VIAF] [GND] [WikiData] [LCN] [detaljnije] [tmx] [pdf]
11.1.004 vol. 1 11.1	**Hodom raka** [VIAF] [GND] [WikiData] [LCN] authors: **Günter Grass** [VIAF] [GND] [WikiData] [LCN] [detaljnije] [tmx] [pdf]
11.1.005 vol. 1 11.1	**Buridanov magarac** [VIAF] [GND] authors: **Günter de Bruyn** [VIAF] [GND] [WikiData] [LCN] [detaljnije] [tmx] [pdf]
11.1.006 vol. 1 11.1	**Poslednji svet** [VIAF] [GND] [WikiData] authors: **Christoph Ransmayr** [VIAF] [GND] [WikiData] [LCN] [detaljnije] [tmx] [pdf]
11.1.007 vol. 1 11.1	**Parfem: hronologija jednog zločina** [VIAF] [GND] [WikiData] authors: **Patrik Süskind** [VIAF] [GND] [WikiData] [LCN] [detaljnije] [tmx] [pdf]
11.2.001 vol. 11.2	**Mamac** [VIAF] [GND] [LCN] authors: **David Albahari** [VIAF] [GND] [WikiData] [detaljnije] [tmx] [pdf]

Слика 56. Листа романа у потколекцији „Романи оригинално написани на немачком” у корпусу СрпНемКор

Преглед метаподатака за сваки роман појединачно доступан је на вези „детаљније”. Овде корисници имају могућност да на оба језика паралелно сагледају све доступне метаподатке и поново приступе роману у формату TMX, сажетом прегледу одабраног романа у формату PDF и записима у базама VIAF, GND, LCNAF и Википодаци везаним за одабраног аутора и одабрани наслов. Такође, за сваки роман генерисан је облак речи које су распоређене у облику стабла чиме су приказане семантичке везе између најчешће коришћених речи унутар текста (Gambette and Véronis, 2010). За све романе облак стабла је генерисан за све именице преко алата TreeCloud.org[183] који је доступан на вебу (Слика 57). У стаблу романа на српском најистакнутија је реч „награда“ као што је на немачком „Preis“. Ова реч се у српском везује за именице *држава,*

[183] TreeCloud.org, http://treecloud.univ-mlv.fr/cgi-bin/NuageArbore_10_EN.cgi#

књижевност, новац, град људи и друге, док се у немачком везује за именице *Staatpreis, Literatur, Schriftsteller, Campe, Ministerium* и друге.



Слика 57. Приказ метаподатака у Библиши за роман „Моје награде / Meine Preise" Томаса Бернхарда са облацима текста

Нерегистровани корисници имају могућност да виде само мање делове текста одабраног романа и првих девет поравнатих сегмената у формату TMX (Слика 58). Са ове странице корисници могу да приступе комплетним метаподацима о овом роману (веза About) и сажетом прегледу овог романа у формату PDF. Такође, у заглављу приказа је дат наслов романа упоредо на немачком и српском, податак о аутору и идентификациони број документа. У нашем примеру то је 11.1.001.

Слика 58. Роман „Моје награде / Meine Preise" Томаса Бернхарда у формату TMX

6.3.2 Допуна лексичких ресурса за двојезично претраживање

Смештањем колекције у Библишу омогућена је двојезична претрага пуног текста колекције. Међутим, да би се користиле главне могућности алата које подразумевају семантичко проширење упита за претрагу било је неопходно произвести одговарајућу листу преводних немачко-српских парова лексичких јединица која би се могла уградити у лексичке и терминолошке ресурсе који су саставни део Библише. За ове потребе користили смо алат BilTE (Bilingual Terminology Extraction)[184], који је развила Група за језичке технологије, који омогућава екстракцију двојезичне терминологије на основу паралелних двојезичних текстуалних колекција, терминолошких листи изворног језика и система за екстракцију полисемичних термина циљног језика.

Алат се састоји од неколико компонената које су развијене у програмским језицима C# и Python и ослања се на једнојезичну екстракцију полилексемских јединица (MWUs) за српски језик (Stanković et al. 2016) и поравнању на нивоу речи са листом термина изворног језика уз помоћ алата за машинско превођење GIZA++ (Och and Ney 2003) (Koehn et al. 2003). Да би се произвела жељена листа термина неопходно је да

[184] BilTE, http://bilte.jerteh.rs/

постоје следећи језички ресурси и алати: листа термина за изворни језик, паралелни корпус за изворни и циљни језик и екстрактор терминологије за циљни језик (Krstev et. al. 2018, 2487). На основу поменутих ресурса ради се обрада текста одабраног паралелног корпуса (чишћење, токенизација и утврђивање исправних величина слова) и производи листа упарених речи, врши се екстракција потенцијалних фраза и складиштење фраза уз помоћ алата GIZA++[185] и генерише листа потенцијалних преводних парова на изворном и циљном језику, такозвана „табела фраза".

Алат је до сада тестиран за екстракцију преводних енглеско-српских парова термина на примеру енглеско-српског паралелног корпуса научног часописа Инфотека коришћењем терминолошког речника из библиотекарства и информатике, а резултати су приказани у (Krstev et. al. 2018). Иако се у нашем случају не ради о екстракцији терминологије користили смо исти алат како бисмо добили листу најфреквентнијих фраза у нашем српско-немачком корпусу заједно са одговарајућим преводима и синонимима. У нашем случају за изворни језик одређен је немачки, а за циљни српски. За ове потребе користили смо следеће језичке ресурсе:

1. поравнати корпус књижевних текстова СрпНемКор са 48.004 упарених сегмената;
2. листу лексичких јединица за немачки језик као изворни. Листа лексичких јединица припремљена је коришћењем два извора. Као први извор коришћен је немачки Ворднет, Open-de-WordNet[186], из кога је добијено приближно 120.000 литерала. Као други извор коришћена је листа најфреквентнијих речи

[185] GIZA++, http://www.statmt.org/moses/giza/GIZA++.html

[186] Иницијатива Open-de-WordNet покренута је са идејом да се направи Ворднет на немачком језику који ће постати део ширег вишејезичног Ворднет окружења, бити у потпуно отвореном приступу и моћи да се користи у оквиру платформе Алати за обраду природних језика (Natural Language Toolkit - https://www.nltk.org/). Прва верзија Open-de-WordNet урађена је на основу лексикона синонима на немачком језику у отвореном приступу OpenThesaurus (OpenThesaurus German synonym lexicon) који је доступан на https://www.openthesaurus.de/ и енглеског Ворднета и објављена у пролеће 2017. године. О иницијативи Open-de-WordNet више се може погледати на https://ikum.mediencampus.h-da.de/projekt/open-de-wordnet-initiative/, док је Open-de-WordNet доступан на https://github.com/hdaSprachtechnologie/odenet и може се користити под лиценцом Creative Commons Attribution Share Alike 4.0 International која дозвољава комерцијалну употребу, измене, дистрибуцију и приватну употребу. Лиценца се може погледати на https://github.com/hdaSprachtechnologie/odenet/blob/master/LICENSE.

на немачком језику која је преузета из Wiki извора Викиречник[187] а која приближно броји око 10.000 речи. Ове две листе спојене су у једну и добијена је прелиминарна листа лексичких јединица за немачки језик. Како би се постигли бољи резултати урађена је лематизација речи (детаљније објашњено у поглављу 3 одељак 3.4.2) из дате листе лексичких јединица за шта је коришћен модел доступан на spaCy[188] који омогућава лематизацију користећи табелу од 355.354 монолексемских јединица. После лематизације и елиминације дупликата добијена је листа од 27.638 различитих лексичких јединица на немачком језику;

3. екстрактор полилексемске терминологије (MWT) за српски језик као циљни. За екстракцију полилексемских термина на српском делу корпуса коришћен је алат LeXimir (Stanković et al. 2011) заснован на електронским морфолошким речницима и локалним граматикама (детаљније објашњено у поглављу 3, одељак 3.3) за екстракцију полилексемских јединица за српски језик. Како се систем користи само за екстракцију полилексемских јединица произвели смо додатно и врећу речи[189] из српских текстова. Произведене листе садржале су 94.802 монолексемских и 48.159 полилексемских јединица. Након тога, урадили смо лематизацију применом електронских морфолошких речника за српски језик. После лематизације елиминисани су дупликати и добијена је листа од 77.297 различитих лексичких јединица на српском језику. Екстракција полилексемских јединица и лематизација детаљније је описана у (Stankovic et al. 2016).

Применом алата GIZA++ произведена је листа преводних немачко-српских парова полилексемских јединица. Пре добијања коначних резултата урађена су додатна филтрирања. Првом филтрацијом задржани су преводни парови у којима се немачки део

[187] Листа је доступна на https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists#German

[188] spaCy је бесплатна библиотека отвореног кода писана у програмским језицима Python и Cython која нуди модела за означавање именованих ентитета у енглеском, немачком, шпанском, португалском, француском, италијанском, холандском и у вишејезичним текстовима. Доступно на https://spacy.io/

[189] Врећа речи (Bag of Words) представља скуп речи у једном тексту са израчунатом фреквенцијом њиховог појављивања у том тексту независно од њихових граматичких облика и реда речи у реченици.а

преводног пара поклапа са неком ставком из првобитно припремљене листе немачких лексичких јединица. Другом филтрацијом елиминисани су преводни парови у којима се српски део преводног пара не поклапа са елементом из листе екстрахованих термина српског језика. Применом система BilTE добијена је листа од 14.142 потенцијална кандидата немачко-српских преводних парова лексичких јединица која је ручно евалуирана. Током процеса евалуације искључени су преводни парови због следећих разлога:

1. преводни парови који су били бесмислени, на пример „kap po kap”/“heraus” (*напоље*) или „karlovom primer”/“sehen” (видети);
2. преводни парови у којима српски део није имао никаквог смисла, на пример „deo mraz s naslov”/Frost (мраз);
3. преводни парови где је српски део имао веће значење од немачког, на пример „komad papira”/“papier” (*папир*)
4. преводни парови где је немачки део имао веће значење од српског, на пример „kafa”/“kaffe trinken” (*пити кафу*).

Након евалуације број је сведен на 3.984 исправна преводна пара. Табела 3 приказује неке примере из добијене табеле преводних парова. Добијена листа исправних преводних парова уграђена је у терминолошку базу Терми чиме је омогућено семантичко проширење упита на немачком језику.

Табела 3. Неки примери из немачко-српске листе преводних парова произведени на основу текстуалне колекције СрпНемКор

	Циљни језик	**Изворни језик**	**Циљни језик**	**Изворни језик**
Примери лексичких јединица из немачко-српске листе преводних парова	autobuskoj stanici	Haltestelle	starica	alt Frau
	baterijska lampa	Taschenlampe	omladinski dom	Jugendherberge
	beznadežan slučaj	Hoffnungslos Fall	pokretom ruke	Handbewegung
	centru grada	Stadtzentrum	boravišna dozvola	Aufenthaltsgenehmigung

Како би се омогућило семантичко проширење упита засновано на синонимима било је неопходно саставити одређену листу синонима која ће затим бити интегрисана у Терми. На основу претходно добијене табеле преводних парова лексичких јединица припремљене су две табеле кандидата за синониме. Једна табела садржала је кандидате за синониме на српском језику и на почетку је бројала 955 лексичких јединица, док је друга табела садржала кандидате за синониме на немачком језику и на почетку је бројала 906 лексичких јединица. Након ручне евалуације број кандидата за синониме на српском у табели је сведен на 864, док је број кандидата за синониме на немачком сведен на 791. Овако добијене табеле синонима уграђене су у базу Терми. Табела 4 приказује неке лексичке јединце на немачком и српском језику са примерима синонима на оба језика. Целокупна табела преводних немачко-српских парова лексичких јединица допуњена синонимима на оба језика употребљена је за припрему двојезичног електронског речника општег типа као скупа отворених повезаних података који је и објављен у окружењу „отворени повезани подаци". Овај поступак ће бити детаљно објашњен у одељку 6.5.3.

Табела 4. Примери синонима на немачком и српском језику

Лексичке јединице на српском	**Синоними на немачком**
potpun	absolut, gänzlich, total, völig, vollkommen, vollständig
bazen	Bassin, Becken, Schwimmbad, Schwimmbecken
majka	Mama, Mutter, Mutti
Лексичке јединице на немачком	**Синоними на српском**
Augenblick	tren, trenutak, moment, čas
fremd	tuđ, tuđinski, nepoznat
grenzenlos	bezgraničan, beskrajan

Након увоза добијена листа немачко-српских преводних парова додатно је анализирана, идентификовани су нови кандидати за синониме и додати у Терми. На пример, за појам „мајка" након учитавања листе преводних парова систем је као резултат

приказивао неколико термина на српском (“majka”, “mama”, “mati” и “mamica”) и немачком језику (“Mama” и “Mutter”). У добијеним резултатима појавили су се примери у са појмом „Mutti” на немачком који није био означен, а представља синоним за именицу „Mutter”. На основу тога је утврђено да појам „Mutti” у бази Терми не постоји и препознат је као нови кандидат за унос. Након верификације и ажурирања базе добијене резултате је могуће извести у више формата: TBX[190] (Слика 59), CSV[191], LMF, lemon или у виду Excel документа. LMF и lemon су детаљније објашњени у одељку 6.5.3.

```
<conceptEntry id="c113951">
    <langSec xml:lang="SR">
        <termSec><term>majka</term><termNote type="termType">entryTerm</termNote></termSec>
        <termSec><term>majka</term><termNote type="termType">synonym</termNote></termSec>
        <termSec><term>mama</term><termNote type="termType">synonym</termNote></termSec>
        <termSec><term>mati</term><termNote type="termType">synonym</termNote></termSec>
        <termSec><term>mamica</term><termNote type="termType">synonym</termNote></termSec>
    </langSec>
    <langSec xml:lang="DE">
        <termSec><term>Mutter</term><termNote type="termType">entryTerm</termNote></termSec>
        <termSec><term>Mutter</term><termNote type="termType">synonym</termNote></termSec>
        <termSec><term>Mama</term><termNote type="termType">synonym</termNote></termSec>
        <termSec><term>Mutti</term><termNote type="termType">synonym</termNote></termSec>
    </langSec>
</conceptEntry>
```

Слика 59. Лексичка јединица „мајка” у бази Терми у формату TBX са синонимима

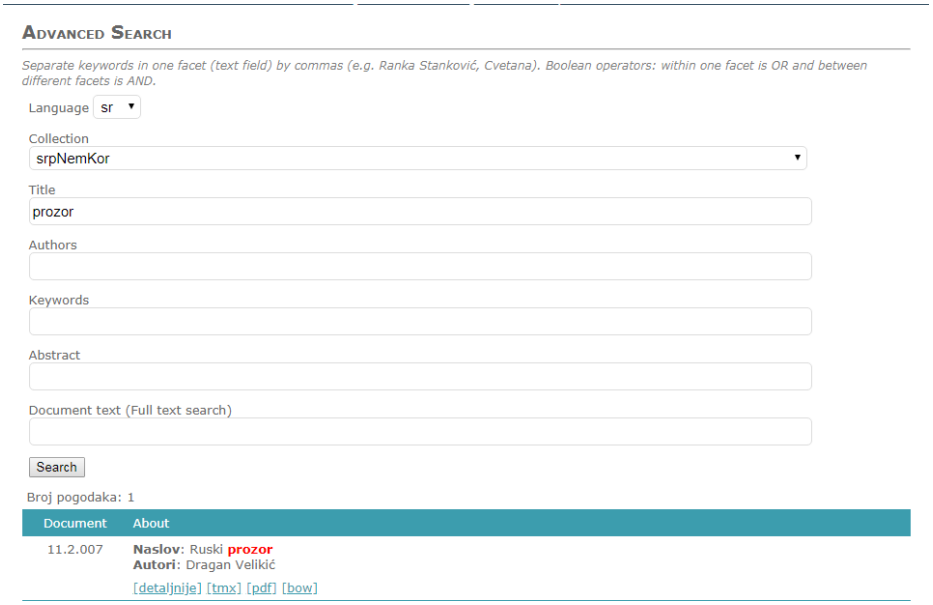
6.3.3 Претрага колекције СрпНемКор и анализа добијених резултата

Као што је већ поменуто у поглављу 2 одељак Библиша алат Библиша омогућава претрагу колекција на два начина: једнојезична претрага преко метаподатака и двојезична претрага пуног текста. У овом одељку показаћемо примере обе претраге на нашој корпусној колекцији паралелних текстова.

[190] TermBase eXchange је међународни стандард за представљање и размену информација о терминологији. Стандард је дефинисала Међународна организација за стандардизацију (International Organization for Standardization - ISO) у сарадњи са Удружењем за локализацију индустријских стандарда (Localization Industry Standards Association – LISA) 2008. године (ISO 30042:2008). Стандардом TBX дефинише се XML формат за размену терминолошких података. Терминолошка база података представљена TBX-ом мора да буде у складу са Оквиром за обележавање терминологије (Terminological Markup Framework - TMF), стандардом који прописује апстрактни модел података и смернице за опис и представљање терминологије у терминолошким базама података (ISO 16642:2017).

[191] Comma-separated values, вредности раздвојене зарезом, је формат за размену података између различитих апликација. Најчешће се користи за извоз односно увоз података који су представљени табеларно који су у датотеци у овом формату одвојени зарезом.

За претрагу преко метаподатака бирамо опцију „Metadata search” у окружењу Библиша и постављамо следећи упит: “Language: *SR* AND collection: *SrpNemKor* AND title: *prozor*”. У упиту за претрагу укрштен је језик претраге *српски*, колекција *СрпНемКор* и реч у наслову *прозор* што значи да се на основу упита за претрагу систем тражити све наслове у српском делу колекције *СрпНемКор* који у себи садрже реч *прозор.* Као резултат добија се роман „Руски прозор”. Резултат исписа садржи идентификациони број документа (у нашем примеру је то 11.2.007), податак о наслову, податак о аутору и везе на потпуне метаподатке (веза „детаљније”), роман у формату TMX, сажети преглед романа у формату PDF и врећу речи BOW (Слика 60).

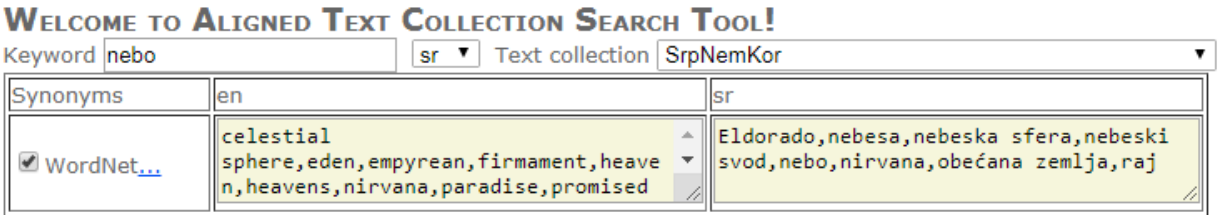


Слика 60. Окружење у Библиши за претрагу преко метаподатака и резултати исписа за упит “Language: SR AND collection: SrpNemKor AND title: prozor”

За претрагу пуног текста колекције СрпНемКор Библиша позива следеће лексичке ресурсе: електронске морфолошке речнике српског језика за морфолошко проширење упита на српском језику (детаљније у поглављу 3 одељак 3.3.3), српски Ворднет за семантичко проширење упита на српском језику (детаљније у поглављу 3 одељак 3.3.4) и терминолошку базу Терми за семантичко проширење упита на српском и немачком језику (детаљније у одељку 6.3.2 овог поглавља). Бирањем опције “Morphological query expansion” у окружењу за претрагу пуног текста добијају се конкорданце за све флективне

облике кључних речи из постављеног упита који су пронађени у одабраној колекцији. Ако поставимо као упит за претрагу „мама” и укључимо опцију “Morphological query expansion” као резултат добијамо упарене текстуалне сегменте једне колекције који садрже у српском делу неке од следећих облика једнине „маме”, „мами”, „маму”, „мамом” или облика множине „маме”, „мамама”.

За семантичко проширење упита Библиша позива семантичке мреже Ворднет. У случају претраживања колекције СрпНемКор могуће је извршити само семантичко проширење упита на српском језику позивањем српског Ворднета. На пример, за упит “Language: *SR* AND collection: *SrpNemKor* AND keyword: *nebo*” позивањем српског Ворднета упит се проширује следећим семантичким еквивалентима на српском: “Eldorado, nebesa, nebeska sfera, nebeski svod, nebo, nirvana, obećana zemlja, raj” (Слика 61). Постављени упит уједно је проширен и семантичким еквивалентима на енглеском језику позивањем енглеског Ворднета што за наше истраживање није од великог значаја. Како није пронађен ниједан немачки Ворднет у потпуно отвореном приступу који би нама послужио за истраживање, семантичко проширење упита на немачком на овај начин за сада није могуће.



Слика 61. Резултати семантичког проширења упита позивањем српског и енглеског Ворднета

За додатно проширење упита приликом претраге колекције СрпНемКор Библиша позива базу Терми. Као што је у претходном одељку наведено, база Терми је за потребе овог истраживања допуњена листом преводних немачко-српских парова лексичких јединица и новим синонимима. Након допуне базе урадили смо претрагу колекције СрпНемКор и анализирали добијене резултате на примерима монолексемских и полилексемских јединица. Као пример за монолексемску јединицу узели смо „мајка”: “Language: *SR* and collection: *SrpNemKor* and keyword: *majka*”. На основу добијених резултата претраге произведене су конкорданце поравнатих сегмената у којима се јавља

одговор на упит.Излазне конкорданце се могу произвести на три различита начина: “DE&SR”, “DE” и “SR”. Бирањем опције “DE&SR” (погодак се јавља у једном од језика или у оба) као резултат добили смо 1.413 конкорданци поравнатих сегмената са означеним одговором на упит у текстовима на немачком и српском језику. Табела 5 приказује неке примере генерисаних конкорданци поравнатих сегмената. Представљени примери су из романа „Руски прозор” Драгана Великића и „Употреба човека” Александра Тишме из потколекције „Романи оригинално написани на српском језику” и романа „Пијанисткиња” Елфриде Јелинек из потколекције „Романи оригинално написани на немачком језику”.

Табела 5. Примери генерисаних конкорданци поравнатих сегмената за упит “Language: SR and collection: SrpNemKor and keyword: majka” уз одабрану могућност приказа “DE&SR”

Ruski prozor = Das russische Fenster / Dragan Velikić, ID: 11.2.007 metadata	n1023 **Es kam vor,dass er ins Stadtzentrum ging, nur um zu überprüfen, ob noch immer der Film gezeigt wurde, den Mama und er so gerne sehen wollten.**	n1023 **Dešavalo se da ode u centar grada samo da bi proverio da li je i dalje na repertoaru film koji su majka i on želeli da vide.**
Ruski prozor = Das russische Fenster / Dragan Velikić, ID: 11.2.007 metadata	n2904 **Darüber dachte er nach, wenn er sich nach den Telefonaten mit seiner Mutter in sein Lieblingscafe in Budim begab.**	n2904 **O tome razmišlja kada posle subotnjih razgovora sa majkom krene u omiljeni kafe na Budimu.**
Upotreba čoveka = Der Brauch des Menschen / Aleksandar Tišma, ID: 11.2.005 metadata	n1686 **Die Mutter hat Vera schon am Tag nach ihrer Ankunft ermahnt: »Kein Wort über Vater und Gerd, das würde keiner verstehen.**	n1686 **Mati je Veru već sutradan po dolasku upozorila: „Nemoj im ništa govoriti o ocu i Gerdu, oni to ne bi razumeli.**
Die Klavierspielerin = Pijanistkinja / Elfride Jelinek, ID: 11.1.002 metadata	n210 **Die Mutter soll streng ihr Gewissen erforschen, ob sie ein ähnlich geschnittenes Kleid nicht in ihrer Jugend selbst getragen habe, Mutti?**	n210 **Neka majka strogo ispita svoju savest, nije li ona sama u svojoj mladosti nosila neku slično krojenu haljinu, mamice?**

Бирањем могућности “DE” (погодак се јавља у немачком делу поравнатог сегмената) као резултат добили смо 140 конкорданци поравнатих сегмената са означеним одговором на упит у немачком делу корпуса. Табела 6 приказује два примера генерисаних конкорданци поравнатих сегмената. Оба примера су из романа Давида Албахарија „Мамац” из потколекције „Романи оригинално написани на српском језику”. Термини на српском нису пронађени из различитих разлога. У првом примеру из табеле термин „мајчиног” није препознат као преводни еквивалент за термин „Mutter” зато што

је у питању придев изведен из именице „мајка” и као такав не представља њен синоним. У другом примеру из табеле аутор уопште није употребио термин „мајка” у овом сегменту текста за разлику од преводиоца који је употребио термин „Mutter”.

Табела 6. Генерисане конкорданце поравнатих сегмената за упит “Language: SR and collection: SrpNemKor and keyword: majka” уз одабрану могућност приказа “DE“

Mamac = Mutterland / David Albahari, ID: 11.2.001 metadata	n196 **Vielleicht ist meine Äußerung über die politische Einstellung des ersten Mannes meiner** Mutter **ungerecht.**	n196 **Možda grešim kada govorim o političkim opredeljenjima majčinog prvog muža.**
Mamac = Mutterland / David Albahari, ID: 11.2.001 metadata	n414 Mutter **war ein Traum gewesen, der in einem fremden Traum gelebt hatte.**	n414 **bila je san koji je živeo u tuđem snu.**

Бирањем могућности “SR” (погодак се јавља у српском делу поравнатог сегмената) као резултат добили смо 67 конкорданци поравнатих сегмената са означеним одговором на упит у српском делу корпуса. Табела 7 приказује три примера генерисаних конкорданци поравнатих сегмената. Представљени примери су из романа „У потпалубљу” Владимира Арсенијевића и „Пешчаник” Данила Киша из потколекције „Романи оригинално написани на српском језику” и романа „Буриданов магарац” из потколекције „Романи оригинално написани на немачком језику”. У прва два примера преводилац није користио еквивалент на немачком језику за лексичку јединицу „мајка” односно за облике „мајци” (датив једнине од „мајка”) и „мајке” (генитив једнине од „мајка”). У првом примеру, облик „мајци” појављује се у фрази на српском „оцу и мајци” за шта је у немачком тексту преводилац користио лексичку јединицу “Eltern”, док се у другом примеру облик „мајке” појављује се у фрази на српском „гола као од мајке” за шта је у немачком тексту преводилац користио лексичку јединицу “splitternackt”. У последњем примеру лексичка јединица “Mütter” у немачком делу текста, такође, није препозната. Лексичка јединица “Mütter” представља множину од “Mutter” који у бази Терми не постоји и овом приликом је препознат као нов кандидат за базу.

Табела 7. Генерисане конкорданце поравнатих сегмената за упит "Language: SR and collection: SrpNemKor and keyword: majka" уз одабрану могућност приказа "SR"

Cloaca Maxima = Cloaca Maximar / Vladimir Arsenijević, ID: 11.2.002 metadata	n661 **Für meine Eltern — wie für viele Städter — war das Landleben ein Quell exotischer Freuden.**	n661 **Kao mnogim ljudima koji su život proveli u gradu, mom ocu i majci selo predstavlja izvor egzotičnih uživanja.**
Peščanik = Sanduhr / Danilo Kiš, ID: 11.2.003 metadata	n3860 **Eva, splitternackt, hat mit der Rechten den untersten Ast ergriffen, während sie zwischen den Fingern der Linken den Apfel hält, den sie Adam anbietet.**	n3860 **Eva, gola kao od majke, uhvatila je desnom rukom najnižu granu, a u levoj ruci, između stisnutih prstiju drži jabuku, pružajući je Adamu.**
Burdans Esel = Buridanov magarac / Ginter de Brojn, ID: 11.1.005 metadata	n635 **Selbst feinsinnigste Mädchen lernen als Hausfrauen und Mütter rechnen;**	n635 **Čak i najprefinjenije devojke nauče da računaju kad postanu domaćice i majke;**

Као пример за полилексемску јединицу узели смо „брачни пар": "*Language: SR and collection: SrpNemKor and keyword: bračni par*". На основу добијених резултата претраге произведене су конкорданце поравнатих сегмената у којима се јавља одговор на упит. Бирањем опције "DE&SR" као резултат добили смо 16 конкорданци поравнатих сегмената са означеним одговором на упит у текстовима на немачком и српском језику. Табела 8 приказује неке примере генерисаних конкорданци поравнатих сегмената. Представљени примери су из романа „Беч, јули 1999" Мила Дора, „Пијанисткиња" Елфриде Јелинек и „Ходом рака" Гинтера Граса из потколекције „Романи оригинално написани на немачком језику". У првом примеру видимо да у немачкој страни корпуса лексичка јединица „Ehepaare" која означава множину није препозната и третира се као потенцијални кандидат за базу Терми.

Табела 8. Примери генерисаних конкорданци поравнатих сегмената за упит "Language: SR and collection: SrpNemKor and keyword: bračni par" уз одабрану могућност приказа "DE&SR"

Wien, Juli 1999 / Milo Dor = Becy, juli 1999 / Milo Dor, ID: 11.1.003 **metadata**	n75 **Die beiden Ehepaare lächelten freundlich vor sich hin und nickten verständnisvoll.**	n75 **Oba** bračna para **su se samo ljubazno smeškala i klimala glavama u znak potvrde da su ga razumeli.**
Wien, Juli 1999 / Milo Dor = Becy, juli 1999 / Milo Dor, ID: 11.1.003 **metadata**	n73 **Der rotgesichtige, dicknasige Kutscher trug einen Backenbart ä la Kaiser Franz Joseph sowie die unvermeidliche Melone auf dem rundlichen Schädel und erzählte seinen Gästen, zwei ältlichen amerikanischen Ehepaaren,[...]**	n73 **Kočijaš, crvenog lica i debelog nosa, s bakenbartom a la Franja Josif i obaveznim polucilindrom na okruglastoj glavi, objašnjavao je svojim mušterijama, bila su to dva američka** bračna para**,[...]**
Die Klavierspielerin / Elfride Jelinek = Pijanistkinja / Elfride Jelinek, ID: 11.1.002 **metadata**	n5983 **Für das** Ehepaar **Erika/Mutter.**	n5983 **Za** bračni par **Erika/majka.**
Im Krebsgang / Guinter Grass = Hodom raka / Ginter Grass, ID: 11.1.004 **metadata**	n2543 **Wohl deshalb ist das** Ehepaar **Stremplin noch vor der Verkündung des Urteils abgereist.**	n2543 **Verovatno je zato** bračni par **Štrempelin otputovao još pre izricanja presude.**

Бирањем могућности "DE" нисмо добили никакав резултат, док смо бирањем могућности "SR" као резултат добили 4 конкорданце поравнатих сегмената са означеним одговором на упит у српском делу корпуса. Табела 9 приказује сва четири примера генерисаних конкорданци поравнатих сегмената. Представљени примери су из романа „Парфем" Патрика Зискинда и „Буриданов магарац" Гинтера де Бројна из потколекције „Романи оригинално писани на немачком језику" и романа „Употреба човека" Александра Тишме и „Мамац" Давида Албахарија из потколекције „Романи оригинално писани на српском језику". У другом и трећем примеру где је у српском делу корпуса коришћена полилексемска јединица „bračni par" у немачком делу корпуса еквивалентне су лексичке јединице „Lehrerpaar" и „Paar", док су у првом и четвртом примеру у српском делу корпуса препознати синоними за „bračni par", „супружници" и „муж и жена", за шта су у немачком делу корпуса еквиваленти „Gatten" и „ein Mann und eine Frau".

Табела 9. Генерисане конкорданце поравнатих сегмената за упит "Language: SR and collection: SrpNemKor and keyword: bračni par" уз одабрану могућност приказа "SR"

Das Parfum / Patrick Süskin = Parfem / Patrik Ziskind, ID: 11.1.007 metadata	n3782 **Auf dem Cours, in hellstem Sonnenlicht, suchten biedere Bauern nach den Kleidern, diesie im Exzess der Orgie von sich geschleudert hatten, suchten sittsame Frauen nachihren Männern und Kindern, schälten sich wildfremde Menschen entsetzt ausintimster Umarmung, standen sich Bekannte, Nachbarn, Gatten plötzlich inpeinlichster öffentlicher Nacktheit gegenüber.**	n3782 **Po poljani pod blistavom sunčevom svetlošću čestiti seljaci tražili su odeću koju su u žaru orgije kidali sa sebe, primerene žene su tražile svoje muževe i decu, nepoznati ljudi su se užasnuto razdvajali iz najintimnijih zagrljaja, poznanici, susedi, supružnici odjedanput su stajali jedni preko puta drugih u najbolnijoj javnoj golotinji.**
Upotreba čoveka / Aleksandar Tišma = Der Brauch des Menschen / Aleksandar Tišma, ID: 11.2.005 metadata	n3726 **Da erschien eine Bedrohung von außen: das Lehrerpaar,[...]**	n3726 **Tada se pojavilo jedno uznemirenje sa strane: učiteljski bračni par, [...]**
Burdans Esel / Günter de Bruyn = Buridanov magarac / Ginter de Brojn, ID: 11.1.005 metadata	n1458 **Häßler organisiert eine Betriebsversammlung, auf der Erp Scheidung und Heirat bekanntmacht; Frau Broder-Erp wird in die Zentrale berufen, wo sie sich ganz ihrem Spezialgebiet, der Bibliothekssoziologie, widmen kann; das Paar bezieht eine Neubauwohnung [...]**	n1458 **Hasler organizuje zbor radnih ljudi, na kom Erp obznani jedan razvod i jednu ženidbu, gospođa Broder-Erp pređe u centralu, gde će u potpunosti moći da se posveti svojoj užoj struci, bibliotečkoj sociologiji, mladi bračni par se useljava u stan u novogradnji [...]**
Mamac / David Albahari = Mutterland / David Albahari, ID: 11.2.001 metadata	n1290 **Manchmal folge ich ihnen nur, zumeist sind das ein Mann und eine Frau mit zwei Kindern, im Winter [...]**	n1290 **katkad samo pođem za njima, obično su to muž i žena sa dvoje dece, obučeni, ako je zima, [...]**

6.4 Означавање именованих ентитета

На паралелној колекцији СрпНемКор тестирали смо расположиве алате за анотацију именованих ентитета. Именовани ентитети у српској страни корпуса анотирани су коришћење система за означавање именованих ентитета који је заснован на развијеним лексичким ресурсима за српски језик, електронским морфолошким речницима и локалним граматикама у форми коначних трансдуктора, описан у (Krstev et al. 2014). Означени именовани ентитети могу се сврстати су у пет категорија: нумерички изрази, лична имена, временски изрази, геополитичка имена и организације (Табела 10).

Табела 10. Листа ознака за именоване ентитете у српском језику

Нумерички изрази	Лична имена	Временски изрази	Геополитички појмови	Организације
amount.approx amount.exact measure.approx measure.exact measure.greaterThan measure.lessThan money.exact money.approx	persName.first persName.full persName.last persName.name persName.spec	time.date.abs time.date.period time.date.rel time.duration.abs time.duration.period time.duration.rel time.hour time.hour.abs time.hour.rel time.set	top.deoGr top.dr top.geo top.gr top.hyd top.reg top.supReg	org org.pol

Овај систем као резултат генерише се документ са анотираним именованим ентитетима одговарајућим XML ознакама. Пример анотације именованих ентитета показаћемо на тексту „Парфем" Патрика Зискинда из потколекције „Романи оригинално написани на немачком језику". Након анотације као резултат добија се датотека са XML ознакама именованих ентитета (Слика 62). У примеру су означена четири именована ентитета. Два именована ентитета спадају у категорију „лична имена", једно је име „Verhamona" (<persName.name>), а друго је презиме „Šenije" (<persName.last>). Именовани ентитет „nekoliko dana" (<time.duration.rel>) спада у категорију „временски изрази", док именовани ентитет „jedan posrednik" (<amount.exact>) спада у категорију „нумерички изрази".

```
On uopšte nije ni mislio na to da za grofa
<persName.name> Verhamona </persName.name> pronađe novi parfem.
On te večeri nije inače ni mislio da ga
<persName.last> Šenije </persName.last> ubedi da „Amor i Psihu" nabavi od
Pelisijea.
To je već uradio.
Stajao je tu na stolu ispred prozora u maloj staklenoj bočici sa izbrušenim
zapušačem.
Kupio ga je još pre <time.duration.rel> nekoliko dana </time.duration.rel>.
Naravno, ne lično.
Nije valjda mogao da ode sam kod Pelisijea i da kupi parfem!
To je uradio <amount.exact> jedan posrednik </amount.exact>, a i on je kupio
preko posrednika...
```

Слика 62. Пример неких означених именованих ентитета XML ознакама у српском тексту романа „Парфем"

Да би се олакшала претрага и омогућило да анотирани текст буде компатибилан са различитим софтвером Група за језичке технологије развила је алат за конверзију између формата који се најчешће користе за означавање именованих ентитета, NER&Beyond[192]. Коришћењем алата NER&Beyond документ са анотираним именованим ентитетима XML ознакама конвертује се у формате IOB (Inside-outside-beginning) и standoff који су погодни за визуелизацију именованих ентитета.

Формат IOB један је од најчешће коришћених формата за датотеке са означеним именованим ентитетима. Развијен је као формат за означавање токена у фразама (именичке фразе, глаголске фраза и слично) у процесу аутоматске синтаксичке анализе (chunking task) одабране реченице (Ramshaw and Marcus 1999). Приликом трансформације у формат IOB токенима се у означеној фрази додељују ознаке I, O или B у виду префикса. Префикс I означава да се токен налази унутар фразе, префикс B означава да се токен налази на почетку фразе, док префикс O означава да токен не припада фрази.

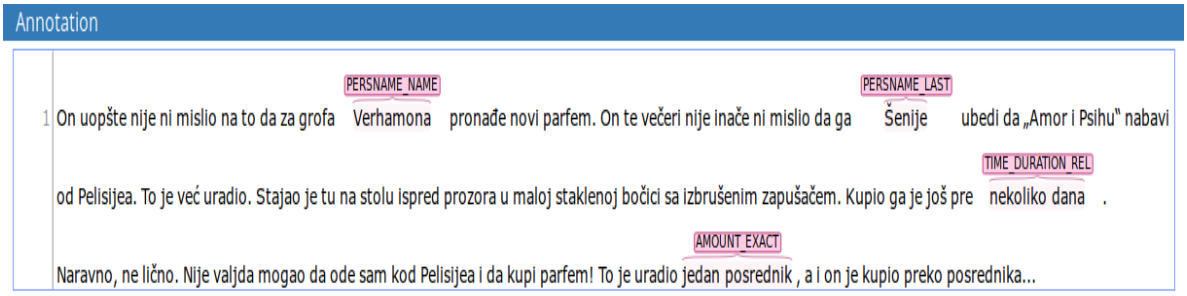
Пример трансформације у IOB формат приказаћемо, такође, на тексту „Парфем". Датотека са анотираним именованим ентитетима XML ознакама (Слика 62) трансформисана је у IOB формат (Табела 11). Реченица која садржи именовани ентитет „неколико дана" (временски израз) трансформацијом у формат IOB анотирана је на следећи начин: токен „неколико" означен је префиксом B, токен „дана" префиксом I, а остали токени у реченици означени су префиксом O. Реченица која садржи именовани ентитет „један посредник" (нумерички израз) трансформацијом у формат IOB анотирана је на следећи начин: токен „један" означен је префиксом B, токен „посредник" префиксом I, а остали токени у реченици означени су префиксом O. Остале реченице из примера нисмо приказали у табели пошто није препознат ниједан ентитет као фраза која се може трансформисати у формат IOB. Другим речима, токени у осталим реченицама из примера би добили префикс O као токени који не припадају ниједној целини, или префиксом B са значењем да са њима почиње фраза (која се истим токеном и завршава).

[192] Алат је доступан на http://nerbeyond.jerteh.rs/

Табела 11. Примери именованих ентитета у IOB формату

Kupio O ga O je O još O pre O nekoliko B-TIME_DURATION_REL dana I-TIME_DURATION_REL . O	To O je O uradio O jedan B-AMOUNT_EXACT posrednik I-AMOUNT_EXACT , O a O i O on O je O kupio O preko O posrednika... O

Текстови у IOB формату могу бити визуализовани коришћењем веб алата за анотацију WebAnno[193]. WebAnno је веб алат развијен за визуализацију и уређивање различитих врста анотација у језичким ресурсима (Yimam et al. 2013), између осталог и именованих ентитета. Слика 63 илуструје визуализацију уз помоћ овог алата. Именовани ентитети из нашег примера (Слика 62) (Табела 11) приказани су на следећи начин: Verhamona (PERSNAME_NAME), Šenije (PERSNAME_LAST), nekoliko dana (TIME_DURATION_REL), jedan posrednik (AMOUNT_EXACT).



Слика 63. Визуализација именованих ентитета применом алата за анотацију и визуализацију WebAnno

Још један формат у који смо конвертовали анотирани документ са XML ознакама именованих ентитета је такозвани standoff формат. Овај формат подразумева да се анотације ентитета смештају у посебну датотеку у односу на текст који се анотира. Пример

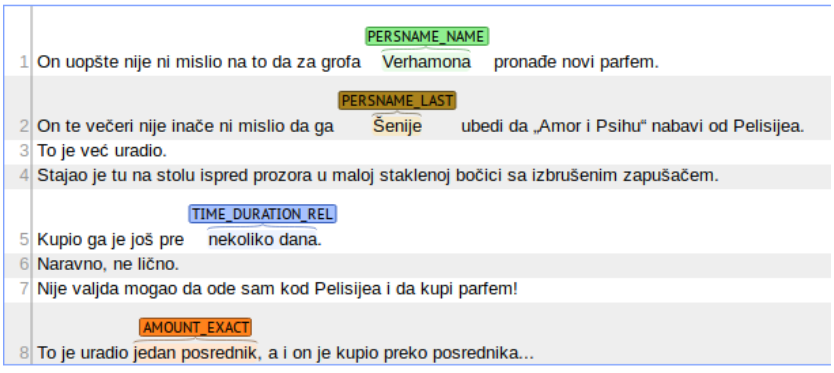
[193] WebAnno, https://webanno.github.io/webanno/ (аутори: Ранка Станковић, Бранислава Шандрих и Цветана Крстев)

из романа „Парфем” након конверзије у формат standoff производи датотеку са четири именована ентитета (Табела 12). Сваки ентитет смештен је у засебној линији и има свој ID који се налази на почетку линије којим је означен редослед њиховог појављивања у тексту. Уз додељен ID стоји ознака која означава категорију у коју спада препознат именовани ентитет и његову почетну и завршну позиција у тексту (то јест, редни број одговарајућег карактера) (brat standoff format 2019). На пример, уз именовани ентитет Verhamon” стоји редни број T1, ознака PERSNAME_NAME за лично име и бројеви 43 и 52 који означавају да овај ентитете почиње на 43, а завршава се на 52. позицији у тексту.

Табела 12. Примери именовани ентитети у standoff формату

T1	PERSNAME_NAME 43 52	Verhamona
T2	PERSNAME_LAST 114 120	Šenije
T3	TIME_DURATION_REL 291 304	nekoliko dana
T4	AMOUNT_EXACT 399 414	jedan posrednik

За визуализацију именованих ентитета у standoff формату може да се користи веб алат BRAT (Brat Rapid Annotation Tool)[194]. BRAT је први веб алат бесплатно доступан за структурну анотацију језичких ресурса заснован на клијент-сервер архитектури. Алат је детаљно описан у (Stenetorp et al. 2012). За разлику од резултата добијених применом алата WebAnno, овде је свака реченица у посебном реду са означеним именованим ентитетом стандардизованим ознакама (Слика 64).



Слика 64. Визуализација именованих ентитета применом веб алата BRAT

[194] BRAT, http://brat.nlplab.org/

Поред анотације именованих ентитета у српском делу корпуса за шта имамо већ развијене алате, покушали смо да анотирамо и именоване ентитете у немачком делу корпуса. За ову сврху користили смо алат за препознавање именованих ентитета развијен у оквиру пројекта „Европске новине" (Europeana Newspapers)[195] (Pekárek and Willems 2012). Пројекат „Европске новине" реализован је од фебруара 2012. до јануара 2015. године са циљем да се направи корпус историјских новина претражив преко метаподатака и пуног текста чланака. Том приликом је настао портал "Europeana Newspapers" који омогућава приступ до 18 милиона страница новинског текста. Поред дигитализације и анотације историјских новина и новинских чланака на основу чега је омогућена претрага пуног текста новина, за новинске колекције на појединим језицима развијена је и напредна претрага преко именованих ентитета. Претрага преко именованих ентитета развијена је за новинске колекције на француском, немачком и холандском језику. За ове потребе корпуси новинских колекција на холандском, немачком (обухвата колекције које су доставиле Национална библиотека Аустрије и Библиотека др Фридрих Тесман) и француском. У овим колекцијама ручно су анотирани именовани ентитети (особе, организације и локације односно геополитички појмови) како би се направили модели за тестирање и обучавање софтвера за препознавање именованих ентитета. За то је коришћен софтвер Stanford NER[196] (Neudecker 2016). Stanford NER је софтверски алат за препознавање именованих ентитета (личних имена, имена организација и геополитичких појмова) који је развила Група за обраду природних језика на Станфорд Универзитету (Natural Language Processing Group at Stanford University - Stanford NLP Group)[197]. На припремљеним корпусним колекцијама са означеним именованим ентитетима обучен је софтвера Stanford NER на основу чега су направљени и језички модели за NER за одабране језике.[198] Како су за немачки језик искоришћене две

[195] Europeana Newspapers, http://www.europeana-newspapers.eu/
[196] Stanford NER, https://nlp.stanford.edu/software/CRF-NER.html
[197] Stanford NLP Group, https://nlp.stanford.edu/
[198] Језички модели за NER за немачки, холандски и француски направљени у оквиру пројекта "Europeana Newspapers". Доступно на: https://github.com/EuropeanaNewspapers/ner-corpora

корпусне колекције новинских текстова за тестирање направљена су и два језичка модела у IOB формату: DE.lift[199] и DE.onb[200].

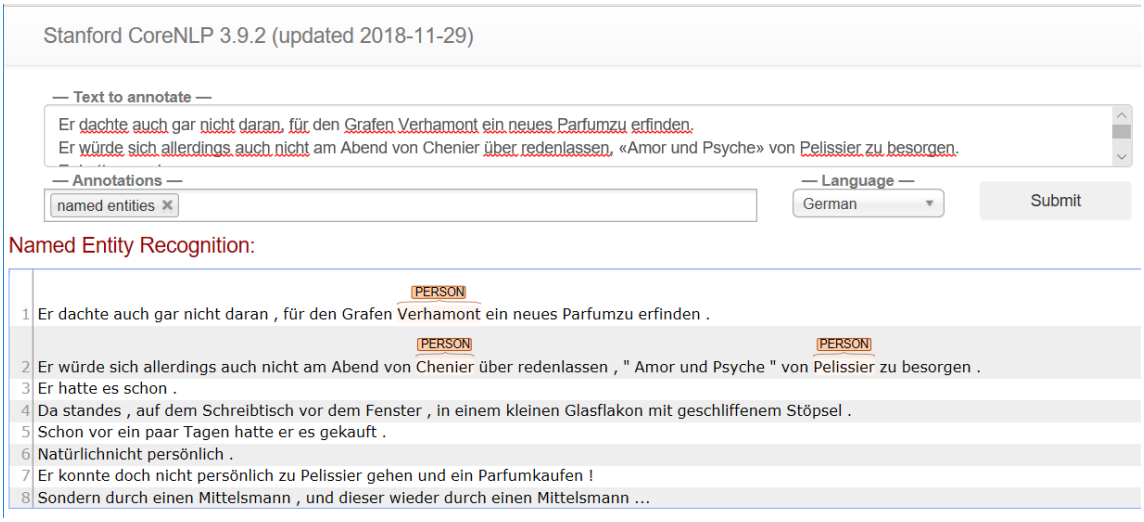
Одлучили смо да тестирамо Stanford NER и на нашем корпусу односно на немачком делу нашег паралелног корпуса и направимо нови језички модел коришћењем доступних текстуалних ресурса анотираних NE у IOB формату. За тестирање смо одабрали немачку верзију текста „Парфем". Станфорд NLP група развила је и веб алат за анотацију текстова.[201] Уз помоћ овог веб алата могуће је анотирати врсте речи, именоване ентитете и друго у текстовима на арапском, кинеском, енглеском, француском, немачком и шпанском. Веб алат је бесплатно доступан и развијен је у сумеђи која је прилагођена корисницима како би могли да тестирају језичке алате које развија Станфорд NLP група и виде резултате анотације. Алат се састоји из простора за унос текста који се анотира и из два падајућа менија којима се дефинишу врста анотације и језик текста (Слика 65). Добијени резултати анотације производе се у BRAT формату.

Слика 65. Stanford NER веб алат за анотацију именованих ентитета

Ми смо овај алат користили како бисмо тестирали анотацију именованих ентитета у немачком делу нашег паралелног корпуса. Као пример одабрали смо део текста "Parfum" на немачком који је превод примера на српском на коме смо илустровали анотацију именованих ентитета на српском раније у овом одељку. У овом примеру софтвер је анотирао особе Verhamont (PERSON), Cheier (PERSON) и Pelissier (PERSON) (Слика 66). Ако упоредимо са верзијом текста на српском видећемо да именовани ентитети нису подједнако анотирани. Два ентитета која означавају особе и која су анотирана у српском тексту анотирана су и у немачком тексту, али не на исти начин.

[199] DE.lift, https://github.com/EuropeanaNewspapers/ner-corpora/tree/master/enp_DE.lft.bio
[200] DE.onb, https://github.com/EuropeanaNewspapers/ner-corpora/tree/master/enp_DE.onb.bio
[201] Stanford NER веб алат за анотацију именованих ентитета на вебу доступан је на: http://corenlp.run/

Прецизније речено, у обе верзије текста означени су Verhamont и Šenije/Cheier ознакама за особе са том разликом што је у српском делу текста прецизно назначено да су у питању лично име и презиме. У немачкој верзији текста означен је још један ентитет као особа, Pelissier. Ако погледамо српску верзију текста видећемо да се тај ентитет, Pelisije, такође појављује, али га алат који је тамо примењен за анотацију није препознао као ентитет који треба означити. Већ смо видели да су у српској верзији текста означена још два ентитета „nekoliko dana" и „jedan posrednik" који представљају временски и нумерички израз. Ови изрази у немачкој верзији текста нису означени, а њихови еквивалентни би били "ein paar Tagen" и "einen Mittelsmann".

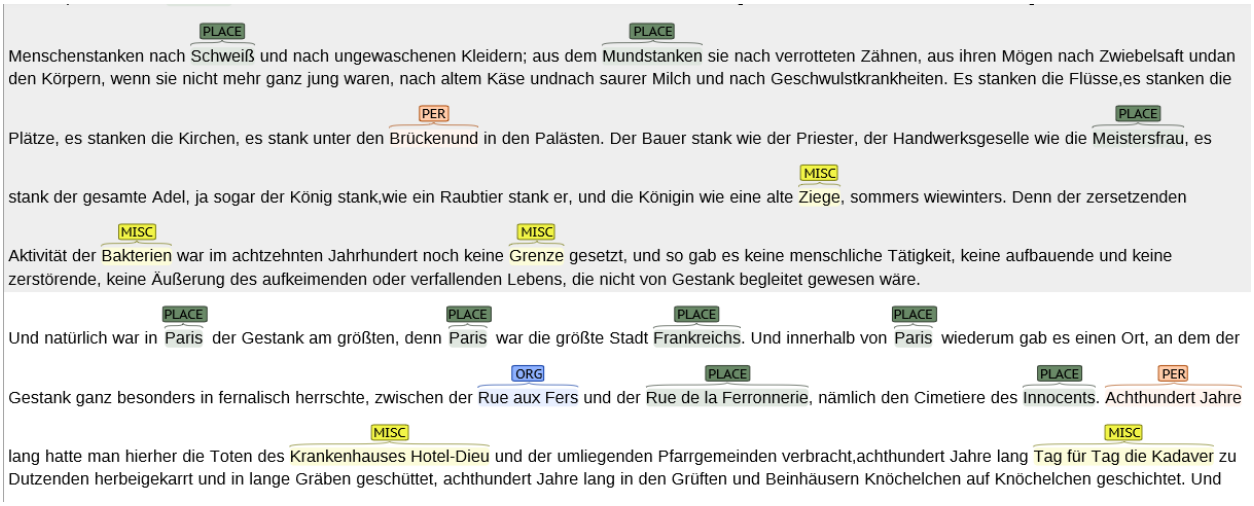


Слика 66. Резултати анотације именованих ентитета применом Stanford NER веб алата на примеру текста „Парфем" на немачком језику

Како добијени резултати не могу да се искористе за неки даљи рад јер алат не омогућава никакав извоз, приступили смо тестирању NER&Beyond алата како бисмо анотирали именоване ентитете у тексту "Parfum" у више формата. За анотацију именованих ентитета у тексту користили смо модел spaCy за који већ постоји модел за немачки, а резултате смо произвели у standoff формату (Табела 13). Визуализација резултата из standoff формата урађена је уз помоћ веб алата BRAT (Слика 67).

Табела 13. Примери именованих ентитета из текста “Parfum” у standoff формату

T21	PLACE 2360 2365	Paris
T22	PLACE 2395 2400	Paris
T23	PLACE 2422 2433	Frankreichs
T24	PLACE 2453 2458	Paris
T25	ORG 2558 2570	Rue aux Fers
T26	PLACE 2579 2600	Rue de la Ferronnerie
T27	PLACE 2628 2637	Innocents
T28	PER 2639 2656	Achthundert Jahre
T29	MISC 2694 2718	Krankenhauses Hotel-Dieu
T30	MISC 2787 2810	Tag für Tag die Kadaver
T31	MISC 2998 3022	Französischen Revolution
T32	MISC 3267 3276	Millionen
T33	ORG 3319 3340	Montmartregeschaufelt
T34	PLACE 3399 3409	Viktualien



Слика 67. Визуализација именованих ентитета из табеле 13 применом веб алата BRAT

Такође, тестирали смо Stanford NER модел за анотацију на овом тексту. За то смо користили два поменута модела за немачки језик која су произведене у оквиру пројекта “Europeana Newspapers”. Као резултат добили смо две датотеке у CoNLL02 формату. Формат CoNLL02 је дефинисан на Конференцији о учењу природних језика (Conference on Natural Language Learning - CoNLL) која је одржана 2002. године по којој је и добио назив. На конференцији је разматрана анотација четири врсте именованих ентитета у IOB формату: лична имена (PER), имена организација (ORG), локације (LOC) и остало (MISC). На

нашем тексту применили смо оба модела за немачки језик која су произведена у оквиру пројекта “Europeana Newspapers”. Применом језичког модела DE.lft и DE.onb добили смо резултате са мање анотираних именованих ентитета (Табела 14 и Табела 15) у односу на резултате добијене тестирањем модела spaCy у алату NER&Beyond. Резултате смо анализирали на две реченице из примера који је анализиран и након примене модела spaCy како бисмо упоредили резултате. Прегледом добијених резултата применом оба модела видимо да су анотирана места односно географски појмови, али се анотације не поклапају у потпуности у моделу DE.lft (испуштено је друго појављивање Paris, Rue de la Ferronnerie је означено као организација итд.). Применом оба модела анотирани су географски појмови, али након приме модела DE.lft анотирано је мање географских појмова у односу на примену модела DE.onb. Са друге стране, након примене модела DE.lft анотиране су организације за разлику од примене модела DE.onb где тих анотација нема односно ти именовани ентитети нису препознати. Ако упоредимо резултате добијене применом модела spaCy у алату NER&Beyond и резултате добијене применом модела DE.lft и DE.onb видећемо да модели DE.lft и DE.onb ипак нису довољно развијени за анотацију текстова као што су литерарни текстови који се разликују од историјских новинских текстова на којима су модели обучавани.

Табела 14. Примери анотације именованих ентитета применом модела DE.lft на роман „Parfum”

Und O natürlich O war O in O Paris B-LOC der O Gestank O am O größten O , O	denn O Paris O war O die O größte O Stadt O Frankreichs B-LOC . O	Und O innerhalb O von O Paris B-LOC wiederum O gab O es O einen O Ort O , O an O dem O der O	Gestank O ganz O besonders O in O fernalisch O herrschte O , O zwischen O der O Rue O aux O Fers O und O	der O Rue B-ORG de I-ORG la I-ORG Ferronnerie I-ORG , O nämlich O den O Cimetiere O des O Innocents O . O

Табела 15. Примери анотације именованих ентитета применом модела DE.onb на роман „Parfum”

Und O natürlich O war O in O Paris B-LOC der O Gestank O am O größten O	, O denn O Paris B-LOC war O die O größte O Stadt O Frankreichs B-LOC . O	Und O innerhalb O von O Paris B-LOC wiederum O gab O es O einen O Ort O , O an O dem O der O	Gestank O ganz O besonders O in O fernalisch O herrschte O , O zwischen O der O Rue O aux O Fers O und O	der O Rue O de O la O Ferronnerie O , O nämlich O den O Cimetiere O des O Innocents O . O

Спроведени експерименти су показали да се постојећи алати за обележавање именованих ентитета, како за српски тако и за немачки, могу применити на наш корпус само с делимичним успехом. Главни разлог је то што су сви коришћени алати, као махом и други алати за обележавање именованих ентитета, обучавани на новинским текстовима. У будуће треба уложити додатан напор за развој потребних алата за обележавање именованих ентитета литерарних текстова, за шта ће одлично послужити корпус ELTeC који се развија у оквиру COST акције Distant Reading[202], а који ће садржати литерарне текстове на немачком, српском и многим другим језицима из периода 1840-1920.

6.5 СрпНемКор и отворени повезани подаци

Посебна тема ове докторске дисертације јесте примена технологија семантичког веба на СрпНемКор. Са једне стране, увезали смо ентитете из добијеног паралелног корпуса са релевантним ресурсима који су део окружења „отворени повезани подаци”. У току рада на дисертацији одлучили смо се да повежемо имена писаца и наслове њихових романа које смо одабрали за садржај корпуса са релевантним записима који се налазе у три нормативне датотеке и једној општој бази знања. У следећем делу овог одељка

[202] Distant Reading COST Action, https://www.distant-reading.net/eltec/

приказани су ресурси које смо користили за повезивање ентитета из колекције СрпНемКор, њихова структура, као и начин повезивања.

Поред увезивања ентитета произвели смо и један сет података у форми отворених повезаних података. Одлучили смо да на основу паралелне српско-немачке колекције креирамо двојезични електронски речник општег типа као отворене повезане податке. За ово смо искористили већ генерисану табелу српско-немачких преводних парова лексичких јединица. Добијени сет података генерисан је у RDF и постављени су темељи за увезивање са другим сетовима података и објављивање у облаку „отворени повезани подаци". Како је изгледао процес генерисања и који алати су за то коришћени, као и како је добијени сет података објављен и увезан са другим сетовима података из облака анализирано је у одељку 6.5.3.

6.5.1 Коришћени ресурси

За повезивање ентитета из СрпНемКор-а користили смо четири ресурса из мреже отворених повезаних података: три нормативне датотеке (GND, VIAF и LCNAF) и базу знања Википодаци (детаљније представљено у поглављу 5, одељак 5.2.2). Повезали смо имена аутора и наслове романа са релевантним записима у поменутим базама података.

A. Нормативна датотека Националне библиотеке Немачке (Gemeinsame Normdatei - GND)[203]. Нормативна датотека Националне библиотеке Немачке (GND) је интегрисана датотека личних имена, имена корпоративних тела, конференција и догађаја, географских имена, предметних одредница и назива дела (Trunk 2019). GND развијају и одржавају Немачка национална библиотека, мрежа библиотека немачког говорног подручја, узајамни каталог Немачке за серијске публикације и бројне друге институције. До априла 2012. године постојале су четири одвојене нормативне датотеке: нормативна датотека личних имена (Normdateien Personennamendatei - PND), нормативна датотека корпоративних тела (Gemeinsame Körperschaftsdatei - GKD), нормативна датотека предметних одредница (Schlagwortnormdatei - SWD) и датотека унифицираних назива немачке музичке архиве (Einheitssachtitel-Dateides Deutschen Musikarchivs - DMA-EST). У

[203] GND, https://www.dnb.de/EN/Standardisierung/GND/gnd_node.html

мају 2012. године ове четири датотеке интегрисане су у јединствену нормативну датотеку GND (Hochstein 2013, 19). Записи у GND налазе се у отвореном приступу што је регулисано лиценцом CC0 1.0[204]. За библиографски опис нормативних јединица користе се следећи каталошки стандарди:

1. Resource Description and Access (RDA) – за опис личних имена која се појављују у библиографском опису и предметној каталогизацији,
2. Regelnfürden Schlagwortkatalog (RSWK) - за опис географских појмова који се користе у предметној каталогизацији у виду предметних одредница (Regeln für den Schlagwortkatalog 2010).

Немачка национална библиотека ради на сервисима повезаних података који ће заједници семантичког веба омогућити да користи националне библиографске податке Немачке, укључујући и нормативне записе. Да би GND била део шире семантичке мреже, од 2010. године Немачка национална библиотека, преко сервиса повезаних података, конвертује метаподатке у RDF/XML формат, што омогућава корисницима и корисничким групама да преузимају податке, а да притом не морају познавати библиографске формате. Поред стандардизованих и алтернативних облика имена успостављене су везе и са записима у другим нормативним датотекама на основу чега је GND укључена у мрежу отворених повезаних података. Поред формата RDF/XML постоји могућност експорта метаподатака и у форматима MARC21 Authority и MARC21/XML.

Од јануара 2014. године неки нормативни записи за географска имена садрже и информације преузете из међународне нормативне датотеке географских појмова GeoNames. Такође, нормативни записи GND базе део су Међународне виртуелне нормативне датотеке (Virtual International Authority File - VIAF) која је такође део облака „отворени повезани подаци". Пример записа у нормативној датотеци GND дат је као Прилог 10а - Пример записа за Томаса Бернхарда у бази GND / корисничко окружење и Прилог 10б - Пример записа за Томаса Бернхарда у бази GND / формат RDF/Turtle.

[204] CC0 1.0, https://creativecommons.org/publicdomain/zero/1.0/

Б. Међународна виртуелна нормативна датотека (The Virtual International Authority File - VIAF)[205]. Међународна виртуелна нормативна датотека (VIAF) је нормативна датотека имена (имена аутора, колективних тела, наслова дела) коју развија и одржава Рачунарски библиотечки центар на мрежи (Online Computer Library Center - OCLC)[206] у сарадњи са националним библиотекама и институцијама културе увезујући нормативне датотеке имена из националних библиотечких, музејских и архивских датотека у јединствену базу података. Иницијативу о оснивању VIAF-а покренули су Конгресна библиотека, Немачка национална библиотека и OCLC 1998. године (VIAF 2019) са циљем успостављања јединствене међународне нормативне датотеке личних имена која би увезала нормативне датотеке националних библиографских центара и била бесплатно доступна преко веба за све кориснике. Конзорцијум VIAF званично је основан на 69. IFLA конгресу 2003. године. Конгресна библиотека и Немачка национална библиотека уступиле су своје нормативне датотеке, а OCLC је развио алгоритам за увезивање записа из две поменуте базе и поставио сервер за складиштење записа (Benett et al. 2006, 3). До 2014. године база VIAF је садржала 38 милиона нормативних записа из 36 институција, заједно са 104 милиона библиографских записа који су у вези са датим именима (Hickey and Toves 2014). Тренутно VIAF остварује сарадњу са више од 40 организација из преко 30 земаља света.

Записи у VIAF-у структурирани су према принципима иницијативе „Отворени повезани подаци" (детаљније објашњено у поглављу 5 одељак 5.2). Сваки од њих има свој идентификациони број односно јединствени идентификатор URI и везе ка записима у другим нормативним датотекама. Сви записи у нормативној датотеци VIAF доступни су бесплатно што је регулисано лиценцом Open Data Commons Attribution License (ODC-BY)[207] у више формата: HTML, MARC21, XML, RDF и JSON. Пример записа у бази VIAF дат је као Прилог 11 - Пример записа за Томаса Бернхарда у бази VIAF / HTML.[208] VIAF садржи податке о готово свим нашим писцима који се заједно са осталим записима преузимају из

[205] VIAF, https://viaf.org/
[206] Online Computer Library Center, https://www.oclc.org/en/home.html
[207] Open Data Commons Attribution License (ODC-BY), https://opendatacommons.org/licenses/by/index.html
[208] Запис за Томаса Бернхарда у другим форматима: XML запис доступан је на https://viaf.org/viaf/12305044/viaf.xml; везе на друге записе у JSON формату доступан је на https://viaf.org/viaf/12305044/justlinks.json. Записи у форматима MARC21 и RDF могу се преузети.

нормативних датотека националних библиографских агенција широм света ако они тамо постоје. Када је реч о корпусу СрпНемКор, одабрани српски писци су аутори који су у свету доста превођени па самим тим и записи за њих постоје у нормативним датотекама великих светских библиографских агенција као што су Национална библиотека Немачке, Национална библиотека Француске, Национална библиотека Канаде, Национална библиотека Швајцарске и многе друге. Нормативна датотека личних имена у Србији до априла 2019. није постојала. Детаљније о нормативној датотеци у Србији и њеном значају било је у поглављу 4 одељак 4.1.2.

B. Нормативна датотека имена Конгресне библиотеке (Library of Congress Name Authority File - LCNAF). Конгресна библиотека има водећу улогу у развијању сервиса у складу са принципима семантичког веба и њиховом придруживању мрежи отворених повезаних података. Први сет података који је Библиотека објавила по овим принципима 2009. године била је Нормативна датотека предметних одредница Конгресне библиотеке (Library of Congress Subject Headings – LCSH) (Vatant 2010, 8). Остали сетови података објављени су 2010. године. Тако је настао Сервис повезаних података Конгресне библиотеке (Library of Congress Linked Data Service - ID.LOC.GOV)[209]. Сервис омогућава корисницима да преко кориснички оријентисане сумеђе, са једног места бесплатно приступе стандардима, контролисаним речницима и нормативним датотекама које Конгресна библиотека развија, бесплатно их претражују и преузимају записе за ентитете који су им потребни у више различитих формата. Један од укључених сервиса је и Нормативна датотека имена Конгресне библиотеке (LCNAF)[210]. LCNAF је нормативна датотека личних имена, имена колективних тела, наслова, конференција и догађаја. Сваки запис у LCNAF има свој јединствени идентификатор URI, варијантне облике имена, везе ка записима у другим сличним системима, као и могућност преузимања записа у више различитих формата: RDF/XML (MADS and SKOS), N-Triples (MADS and SKOS), JSON (MADS/RDF and SKOS/RDF), MADS - RDF/XML, MADS - N-Triples, MADS/RDF – JSON, SKOS - RDF/XML, SKOS - N-Triples, SKOS – JSON, MADS/XML, MARC/XML. Пример записа у бази LCNAF дат је као Прилог 12а - Пример записа за Томаса Бернхарда у бази LCNAF /

[209] ID.LOC.GOV, http://id.loc.gov/

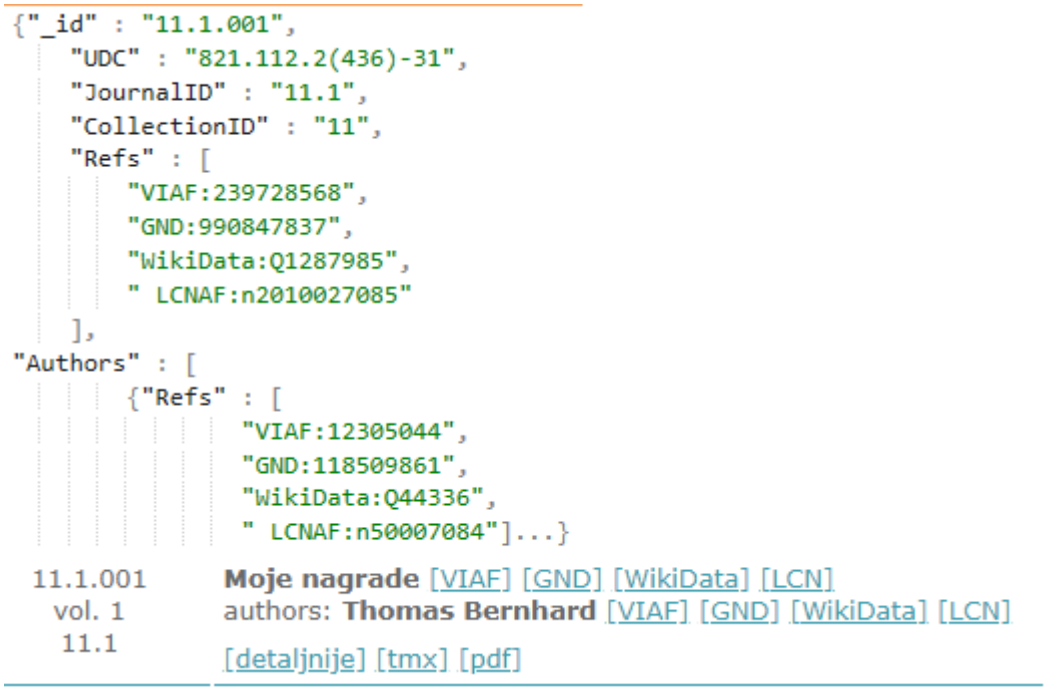
[210] Library of Congress Name Authority File, http://id.loc.gov/authorities/names.html

корисничко окружење и Прилог 12б - Пример записа за Томаса Бернхарда у бази LCNAF / формат RDF/XML.[211]

6.5.2 Поступак повезивања

Као што смо у претходном одељку видели сваки запис у поменутим базама података има свој јединствени идентификациони број који се користи за повезивање. У току рада на овој дисертацији увезали смо записе за имена писаца и наслове њихових романа који су обрађени за корпус. Име писца и наслов представљају и библиографске метаподатака о одређеном роману па је увезивање урађено у оквиру структурне шеме за метаподатке. Јединствени идентификациони бројеви записа наведени су у оквиру елемента "Refs" у одговарајућем делу шеме за метаподатке у зависности од тога да ли се повезују записи о писцу или наслову романа (целокупна шема за израду метаподатака детаљније је објашњена у одељку 6.3.1 овог поглавља). Елемент "Refs" групише податке о идентификационим бројевима записа у базама VIAF, GND, Википодаци (WikiData) и LCNAF на начин и у оном редоследу како се после појављују у корисничкој сумећи. Јединствени идентификатори преко којих смо увезали записе о наслову романа у базама VIAF, GND, Википодаци (WikiData) и LCNAF груписани су у првом делу записа за метаподатке, док су јединствени идентификатори преко којих смо увезали записе о писцу у базама VIAF, GND, Википодаци (WikiData) и LCNAF груписани у другом делу записа за метаподатке, "Authors". Слика 68 илуструје повезивање аутора Томаса Бернхарда и његов роман „Моје награде" са поменутим нормативним датотекама у структурној шеми за метаподатке и приказ корисничкој сумећи у оквиру Библише.

[211] Запис за аутора Томаса Бернхарда доступан је на http://id.loc.gov/authorities/names/n50007084.html



Слика 68. Библиша - везе ка базама VIAF, GND, Википодаци (Wikidata) и LCNAF за роман „Моје награде" и писца „Томаса Бернхарда"

6.5.3 Двојезични речник општег типа као отворени повезани подаци

Поступак креирања скупа података у форми отворених повезаних података са фазама израде је објашњен у поглављу 5 одељак 5.2.3. У овом делу дисертације приказујемо практичну примену ових корака на једном језичком ресурсу и анализирамо добијене резултате. Одлучили смо се да поступак применимо на двојезичном електронском речнику општег типа за који смо материјал припремили на основу паралелног корпуса СрпНемКор. Двојезични речници представљају језичке ресурсе који садрже алфабетски или абецедни попис лексичких јединица на једном језику и њене еквиваленте на другом језику уз које стоје лингвистичке анотације односно показатељи граматичких категорија који нам дају додатне податке о лексичкој јединици као што је врста речи (именица, глагол, прилог, предлог и тако даље), род ако су у питању именице (мушки, женски, средњи), глаголски вид ако су у питању глаголи (свршени и несвршени) и слично. Еквиваленти на другом језику најчешће представљају директне преводе лексичких јединица уз које, такође, стоје лингвистичке анотације, а врло често и дефиниције у виду описног дела који подразумева тумачење дате лексичке јединице или

дефиниције у виду синонима који представљају додатна објашњења. Двојезични речници се најчешће користе у процесима превођења са једног језика на други и могу бити једносмерни и двосмерни односно омогућавају превод на и са оба језика. У зависности од врсте лексичких јединица које су пописане једна подела речника може бити на опште и терминолошке. Општи речници пружају увид у лексику једног језика која се користи у књижевном говору, док терминолошки речници пружају увид у терминологију одређене струке или научне дисциплине.

Двојезични електронски речници по структури су скоро исти са двојезичним речницима у папирној форми: садрже алфабетски односно абецедни попис лексичких јединице, њихове еквиваленте на другом језику и ознаке лингвистичких анотација односно показатеље граматичких категорија. Примена технологија семантичког веба и трансформисање двојезичних електронских речника у облику скупа отворених повезаних података омогућава њихово увезивање са сродним ресурсима на вебу што је за кориснике значајно јер могу да сагледају и преводне парове лексичких јединица које у оригиналном речнику не постоје, а са друге стране, представљају значајан језички ресурс за системе за машинско превођење и допуну сродних лингвистичких ресурса који су део великог облака „отворени повезани подаци". Пример двојезичног електронског речника који ми овде анализирамо припремљен је на основу књижевних текстова тако да је реч о речнику општег типа, а како је припремљен на основу паралелне колекције немачко-српских текстова реч је о немачко-српском речнику. Лексичке јединице и њихови еквиваленти анотирану су врстом речи, а за велики број њих наведени су и синоними.

Немачко-српски електронски речник општег типа који ми овде анализирамо као скуп отворених повезаних података креиран је извозом и трансформацијом из базе Терми, у коју је складиштен резултат екстракције двојезичне паралелне листе немачко-српских преводних парова, са полуаутоматски додатим синонимима, о чему је било речи у одељку 6.3.2. Као пример трансформисања двојезичног електронског речника у форму отворених повезаних података користили смо принцип који је примењен на трансформацију двојезичних речника Apertium, а који је препоручен као пример добре праксе. Велики број двојезичних речника из Apertium базе трансформисан је у RDF што је

омогућило њихово међусобно увезивање и објављивање у облаку „отворени повезани подаци”. Цео поступак објашњен је у (Garcia and Vila-Suero 2015) и (Gracia et al. 2018). На крају, како двојезични речници представљају лингвистичке ресурсе наш двојезични електронски речник објавили смо у подоблаку „Отворени повезани подаци из области лингвистике” (Linguistic Linked Open Data Cloud - LLOD) (структура облака „отворени повезани подаци” детаљније је представљена у поглављу 5 одељак 5.2).

Према (Garcia and Vila-Suero 2015) и (Gracia et al. 2018) за моделирање речника и његову трансформацију у RDF коришћена су два модела података: Оквир за означавање лексике (Lexical markup framework - LMF) и Модел лексикона за онтологије (LExicon Model for ONtologies - lemon). LMF је ISO стандард (ISO 24613:2008) који дефинише оквир за означавање података у лексичким базама података и омогућава креирање лексикона и речника у електронском формату (Francopoulo 2013, 98). Креиран је са циљем да се обезбеди општи модел који омогућава описивање како једнојезичних тако и двојезичних и вишејезичних електронских лексичких ресурса са што мање лексичких информација о лексичким јединицама које се у њима налазе (Francopoulo et al. 2006, 233). Трансформацијом у LMF креирају се речници у формату који омогућава размену података и повезивање са релевантним ресурсима на вебу како би се омогућила њихова интероперабилност. Када су у питању двојезични ресурси, LMF модел омогућава формирање преводних парова лексичких јединица из речника који се описује. Овај модел смо и ми користили за наш двојезични немачко-српски речник, SrpNemLexDe-Sr, а целокупан припремљени скуп немачко-српских преводних парова извезен је из базе Терми у LMF/XML.

Структуру LMF/XML датотеке објаснићемо на примеру лексичких јединица на немачком “grenzenlos” и “unendlich” и њихових еквивалената на српском „bezgraničan” и „beskrajan” (Слика 69). Ове лексичке јединице из SrpNemLexDe-Sr речника само одабрали јер представљају синониме и на српском и на немачком да бисмо илустровали не само њихово увезивање у преводне парова већ и увезивање са синонимима. У примеру видимо да датотека LMF/XML има четири целине. Прва целина, <GlobalInformation>, је општег типа и у њој је наведен податак о стандарду, ISO 639-3, којим се дефинише

употреба кодова за језик у речнику. Друга два дела LMF/XML датотеке представљају појединачне речнике, немачки и српски, који су означени етикетом <Lexicon> коју прати етикета *feat* која кроз атрибуте att и val дефинише језик речника. Вредност атрибута att је "language" док је вредност атрибута val код за језик, "DE" односно "SR".

Свака лексичка јединица означена је етикетом <LexicalEntry> која добија свој идентификатор односно ID. Идентификатор је назначен атрибутом id који стоји уз етикету <LexicalEntry> и састоји се из три дела: први део је канонски облик лексичке јединице (лема) која се описује и коју прати код за врсту речи и код за језик. Идентификатор ћемо илустровати на примеру id="grenzenlos-A-DE". Значење кода A се разрешава у оквиру етикете *feat* која прати етикету <LexicalEntry>. Етикета *feat* кроз атрибут att чија је вредност "partOfSpeech" и атрибут val чија је вредност "A", у нашем примеру, дефинише да је лексичка јединица придев. Податак о канонском облику (леми) лексичке јединице разрешава се кроз етикету Lemma коју прати етикета feat са атрибутом att чија је вредност "writtenForm" и атрибутом val чија је вредност "grenzenlos" чиме се дефинише писани облик лексичке јединице у датом језику. Код "DE" указује да лексичка јединица припада речнику на немачком језику. Уз ове податке, лексичкој јединици је додељено и „значење" које омогућава њено увезивање са еквивалентом на српском језику у преводни пар. „Значење" се наводи у оквиру етикете <Sense> коју прати атрибут id са одговарајућом алфа-нумеричком вредношћу.

```xml
<?xml version="1.0" encoding="utf-8"?>
<LexicalResource>
  <GlobalInformation>
    <feat languageCoding="ISO 639-3" />
    <feat language="srp" />
  </GlobalInformation>
  <Lexicon>
    <feat att="language" val="DE" />
    <LexicalEntry id="grenzenlos-A-DE">
      <feat att="partOfSpeech" val="A" />
      <Lemma>
        <feat att="writtenForm" val="grenzenlos" />
      </Lemma>
      <Sense id="112175-DE">
        <Definition>
          <feat att="gloss" val="keine Grenze habend" />
        </Definition>
      </Sense>
    </LexicalEntry>
    <LexicalEntry id="unendlich-A-DE">
      <feat att="partOfSpeech" val="A" />
      <Lemma>
        <feat att="writtenForm" val="unendlich" />
      </Lemma>
      <Sense id="112131-DE">
        <Definition>
          <feat att="gloss" val="nicht begrenzt, von nicht absehbarem Ausmaß" />
        </Definition>
      </Sense>
    </LexicalEntry>
  </Lexicon>
  <Lexicon>
    <feat att="language" val="SR" />
    <LexicalEntry id="bezgraničan-A-SR">
      <feat att="partOfSpeech" val="A" />
      <Lemma>
        <feat att="writtenForm" val="bezgraničan" />
      </Lemma>
      <Sense id="112175-SR">
        <Definition>
          <feat att="gloss" val="Onaj koji je bez granice" />
        </Definition>
      </Sense>
    </LexicalEntry>
    <LexicalEntry id="beskrajan-A-SR">
      <feat att="partOfSpeech" val="A" />
      <Lemma>
        <feat att="writtenForm" val="beskrajan" />
      </Lemma>
      <Sense id="112131-SR">
        <Definition>
          <feat att="gloss" val="nije ograničen, nepredvidivog obima" />
        </Definition>
      </Sense>
      <Sense id="112175-SR">
        <Definition>
          <feat att="gloss" val="Onaj koji je bez granice" />
        </Definition>
      </Sense>
    </LexicalEntry>
  </Lexicon>
  <SenseAxis id="112131-SR-DE" senses="112131-SR 112131-DE" />
  <SenseAxis id="112175-SR-DE" senses="112175-SR 112175-DE" />
</LexicalResource>
```

Слика 69. Пример лексичких јединица „grenzenlos/unendlich” и „beskrajan/bezgraničan” у LMF/XML

У пракси лексичке јединице могу имати више значења односно више од једног еквивалента на другом језику, тако се и у моделу LMF лексичким јединицама може доделити више „значења” ако је то потребно. На пример, лексичка јединица „beskrajan”

може да се преведе са “grenzenlos” и “unendlich” јер има два „значења” у моделу LMF, „112131-SR” и „112175-SR”. Преко ових значења лексичка јединица „beskrajan” формира преводне парове са „grenzenlos” („112131-DE”) и „unendlich” („112175-DE”). Лексичке јединице „grenzenlos” и „unendlich” су у овом случају синоними. Са друге стране, „grenzenlos” може да се преведе и као „bezgraničan” и као “beskrajan” при чему су ове лексичке јединице у овом случају синоними. Поред значења у овом облику, лексичке јединице имају и описне текстуалне дефиниције (глосе) које објашњавају њихово значење, а које се наводе у оквиру етикете <Definition> коју прати етикета *feat* са атрибутима att и val. Атрибут att има вредност „gloss“, а вредност атрибута val је дефиниција значења. На пример, „beskrajan” има две дефиниције „nije ograničen, nepredvidivog obima” и „onaj koji je bez granice”.

Четврта целина датотеке LMF/XML представља преводне парове. Преводни парови су анотирани етикетом SenseAxis. Етикету SenseAxis прати атрибут *id* чија је вредност алфа-нумеричка која представља „значење” преводног пара. Да је у питању преводни пар видимо по кодовима за језик који стоје уз нумеричку ознаку. Атрибут id прати атрибут *senses* у коме се наводе идентификатори „значења” за лексичку јединицу на немачком и српском језику. Неки примери преводних парова су:

“unendlich=beskrajan” - <SenseAxis id=“112131-DE-SR” senses=“112131-DE 112131-SR” />
“grenzenlos=bezgraničan/beskrajan” - <SenseAxis id=“112175-DE-SR” senses=“112175-DE 112175-SR” />

Модел LMF омогућава моделирање речника, али само у формату XML. Да бисмо од речника креирали скуп повезаних података потребна је RDF серијализација. За ове потребе развијен модел *lemon* чија је структура заснована на моделу LMF. Модел *lemon* (Gracia et al. 2014) је RDF модел креиран за опис лингвистичких ресурса као што су лексикони и машински читљиви речници у складу са принципима семантичког веба како би постали део окружења „отворени повезани подаци”. Модел је дизајниран да омогући представљање речника у виду информатичке онтологије на вебу (Buitelaar et al. 2011) и следи принцип „семантика кроз референце” (semantics by reference). На овај начин је корисницима омогућено да на једном месту сагледају све лексичке јединице датог лексикона односно речника са што мање лингвистичких анотација и њихове еквиваленте на другим језицима ако су у питању вишејезични ресурси (McCrae et al. 2012, 703). За

додатне информације о о лингвистичким анотацијама које стоје уз лексичке јединице модел омогућава реферисање на релевантне онтологије. Модел *lemon* је конципиран као универзално средство за опис свих врста речника и лексикона независно од тога да ли је њихов садржај општег или доменског типа и да ли су у питању једнојезични, двојезични или вишејезични ресурси. Када је реч о двојезичним ресурсима, као што је случај у нашем примеру, и вишејезичним ресурсима, *lemon* предвиђа и коришћење модула за превођење односно модула који омогућавају увезивање преводних парова. Као и у поступку креирања речника по LMF моделу који ми овде анализирамо, целокупан припремљени скуп немачко-српских преводних парова извезен је из базе Терми у RDF. За креирање RDF графова коришћена је програмска библиотека dotNetRDF[212], а добијени резултати приказани су у Turtle формату. Како RDF подразумева формирање изјава у виду тројки (субјекат-предикат-објекат, објашњено у поглављу 5 одељак 5.1.2) у нашем примеру формиране су три одвојене датотеке: датотека речника изворног језика (у нашем случају немачки), датотека речника циљног језика (у нашем случају српски) и датотека преводних парова. Свакој од креираних датотека додељен је URI идентификатор, као и свакој лексичкој јединици у датотекама изворног и циљног језика, али и преводним паровима у датотеци преводних парова. За структуру URI идентификатора прихватили смо предлог из упутства (Gracia et al. 2018) да се користе препоруке (Archer et al. 2012) дефинисане у документу Интероперабилна решења за европску јавну администрацију (Interoperability Solutions for European Public Administrations) које прописују следећи формат: http://{domain}/{type}/{concept}/{reference}. URI идентификатори додељени датотекама имају фиксни и променљиви део. За све ентитете у речнику фиксни део се односи на домен на коме је речник објављен што је у нашем примеру *http://llod.jerteh.rs/* иза чега долази променљиви део. URI идентификатори датотека у нашем примеру су:

Датотека изворног језика - <http://llod.jerteh.rs/id/SrpNemLexDE>
Датотека циљног језика - <http://llod.jerteh.rs/id/SrpNemLexSR>
Датотека преводних парова - <http://llod.jerteh.rs/id/SrpNemLexDE-SR>

Што се тиче URI идентификатора лексичких јединица они се састоје од фиксног идентификатора датотеке, који је у случају датотеке изворног језика

[212] dotNetRDF, https://www.dotnetrdf.org/

http://llod.jerteh.rs/id/SrpNemLexDE, иза кога долази променљиви део који се односи на идентификатор лексичке јединице, у случају изворног језика идентификатор лексичке јединице на немачком језику. Примери URI идентификатора лексичких јединица „grenzenlos" и „unendlich":

"grenzenlos" - <http://llod.jerteh.rs/id/SrpNemLexDE/grenzenlos-A-DE
"unendlich" - <http://llod.jerteh.rs/id/SrpNemLexDE/unendlich-A-DE>

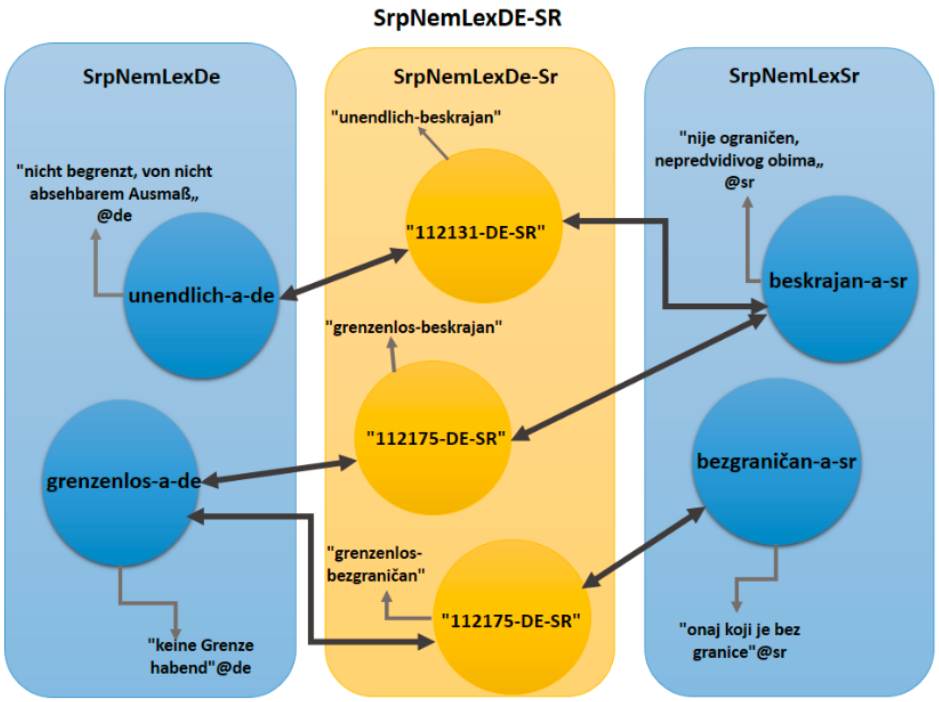
Датотека циљног језика има фиксни део у форми *http://llod.jerteh.rs/id/SrpNemLexSR* иза чега долази променљиви део који се односи на идентификатор лексичке јединице на српском језику. Примери URI идентификатора лексичких јединица „beskrajan" и „bezgraničan":

"beskrajan" - <http://llod.jerteh.rs/id/SrpNemLexSR/beskrajan-A-SR
"bezgraničan" - <http://llod.jerteh.rs/id/SrpNemLexSR/bezgrani%C4%8Dan-A-SR>

URI преводних парова има фиксни део у форми *http://llod.jerteh.rs/id/SrpNemLexDE-SR/,* што је URI датотеке преводних парова иза чега долази променљиви део који се односни на „значење" лексичких јединица преко кога је формиран преводни пар. Примери URI идентификатора преводних парова за претходно наведене лексичке јединице су:

"grenzenlos=beskrajan" - <http://llod.jerteh.rs/id/SrpNemLexDE-SR/112139-sr-sense-112139-de-sense-trans
"grenzenlos=bezgraničan" - <http://llod.jerteh.rs/id/SrpNemLexDE-SR/112175-sr-sense-112175-de-sense-trans
"unendlich=beskrajan" - <http://llod.jerteh.rs/id/SrpNemLexDE-SR/112131-sr-sense-112131-de-sense-trans>

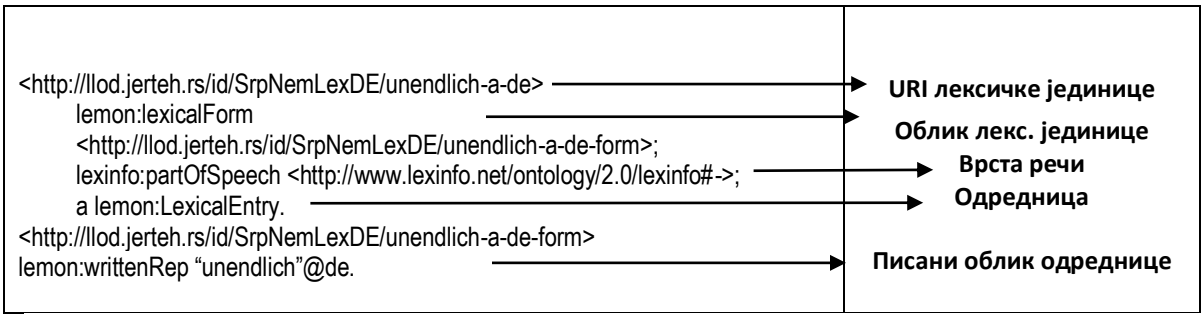
Као што је већ поменуто, трансформацијом скупа немачко-српских преводних парова из базе Терми у RDF произведене су три датотеке у Turtle формату (Слика 70). Структуру добијених датотека анализираћемо на примерима претходно одабраних лексичких јединица и њихових преводних парова. У датотекама језика структура описа лексичке јединице је следећа: на почетку се налази URI идентификатор лексичке јединице иза кога се преко ентитета lemon:lexicalForm указује на њен канонски облик (лему), затим се наводи податак о врсти речи преко ентитета *lexinfo*, док ентитет lemon:LexicalEntry указује да је у питању лексичка јединица која је део речника. На крају записа наведен је облик лексичке јединице који се користи у писању уз податак о језику који се наводи уз ознаку *@.*



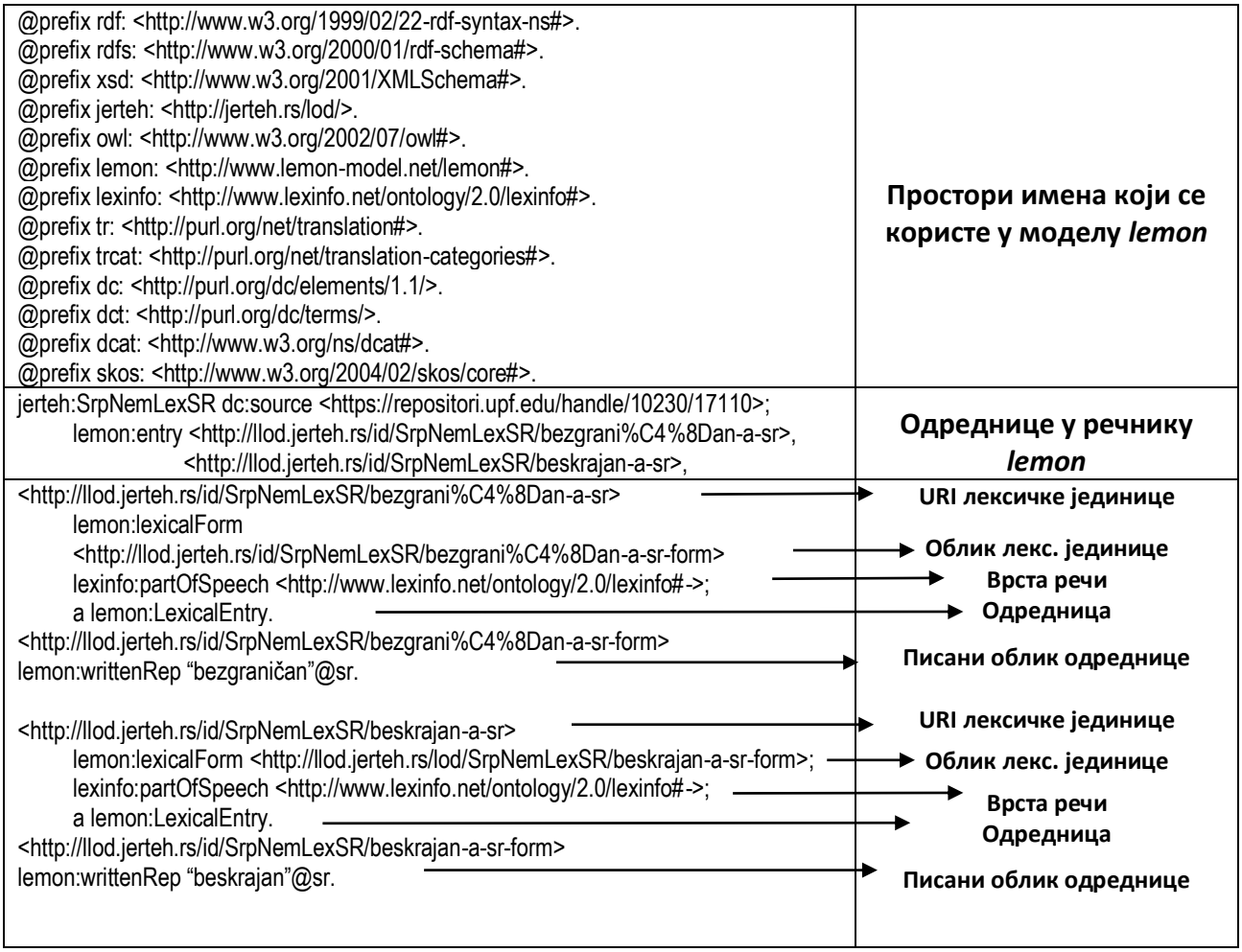
Слика 70. Графички приказ преводних парова "grenzenlos"="beskrajan", "grenzenlos"="bezgraničan", "unendlich"="beskrajan"

Пример лексичких јединица "grenzenlos" и "unendlich" у RDF/Turtle формату:

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>. @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>. @prefix xsd: <http://www.w3.org/2001/XMLSchema#>. @prefix jerteh: <http://jerteh.rs/lod/>. @prefix owl: <http://www.w3.org/2002/07/owl#>. @prefix lemon: <http://www.lemon-model.net/lemon#>. @prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>. @prefix tr: <http://purl.org/net/translation#>. @prefix trcat: <http://purl.org/net/translation-categories#>. @prefix dc: <http://purl.org/dc/elements/1.1/>. @prefix dct: <http://purl.org/dc/terms/>. @prefix dcat: <http://www.w3.org/ns/dcat#>. @prefix skos: <http://www.w3.org/2004/02/skos/core#>.	**Простори имена који се користе у моделу *lemon***
jerteh:SrpNemLexDE dc:source <https://repositori.upf.edu/handle/10230/17110>; lemon:entry <http://llod.jerteh.rs/id/SrpNemLexDE/grenzenlos-a-de>, <http://llod.jerteh.rs/id/SrpNemLexDE/unendlich-a-de>,	**Одреднице у речнику *lemon***
<http://llod.jerteh.rs/id/SrpNemLexDE/grenzenlos-a-de> → lemon:lexicalForm → <http://llod.jerteh.rs/id/SrpNemLexDE/grenzenlos-a-de-form>; lexinfo:partOfSpeech <http://www.lexinfo.net/ontology/2.0/lexinfo#->; → a lemon:LexicalEntry. → <http://llod.jerteh.rs/id/SrpNemLexDE/grenzenlos-a-de-form> lemon:writtenRep "grenzenlos"@de. →	**URI лексичке јединице** **Облик лекс. јединице** **Врста речи** **Одредница** **Писани облик одреднице**



Пример лексичких јединица "beskrajan" и "bezgraničan" RDF/Turtle формату:

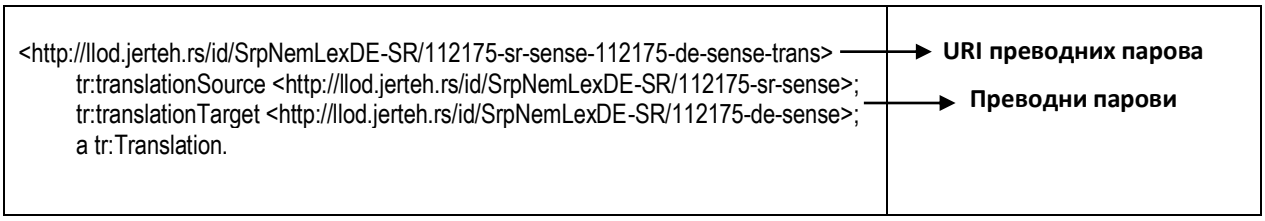


Датотека језика, датотека преводних парова има мало другачију структуру. За представљање преводних парова у *lemon* моделу користи се Модел за превођење (Translation Model). Модел за превођење (Gracia et al. 2014) користи две класе за опис које су преузете из језика OWL: Translation и TranslationSet. Класа TranslationSet групише све

преводне парове лексичких јединица који су произведени. Ова класа стоји на почетку датотеке преводних парова. Кроз класу Translation указује се на појединачне лексичке јединице које чине преводни пар. Елементом lemon:isSenseOf наводи се URI идентификатор који указује на одређену лексичку јединицу. Својствима translationSource и translationTarget указује се на датотеку изворног и датотеку циљног језика и на значење.

Пример преводних парова за лексичке јединице „grenzenlos" и „unendlich" и „beskrajan" и „bezgraničan" у RDF/Turtle формату:

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>. @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>. @prefix xsd: <http://www.w3.org/2001/XMLSchema#>. @prefix jerteh: <http://jerteh.rs/lod/>. @prefix owl: <http://www.w3.org/2002/07/owl#>. @prefix lemon: <http://www.lemon-model.net/lemon#>. @prefix lexinfo: <http://www.lexinfo.net/ontology/2.0/lexinfo#>. @prefix tr: <http://purl.org/net/translation#>. @prefix trcat: <http://purl.org/net/translation-categories#>. @prefix dc: <http://purl.org/dc/elements/1.1/>. @prefix dct: <http://purl.org/dc/terms/>. @prefix dcat: <http://www.w3.org/ns/dcat#>. @prefix skos: <http://www.w3.org/2004/02/skos/core#>.	**Простори имена који се користе у моделу *lemon***
jerteh:SrpNemLexDE-SR dc:source https://repositori.upf.edu/handle/10230/17110 tr:trans <http://llod.jerteh.rs/id/SrpNemLexDE-SR/112131-sr-sense-112131-de-sense-trans>, <http://llod.jerteh.rs/id/SrpNemLexDE-SR/112175-sr-sense-112175-de-sense-trans>,	**Преводни парови**
<http://llod.jerteh.rs/id/SrpNemLexDE-SR/112131-de-sense> → **URI значења (DE)** lemon:isSenseOf <http://llod.jerteh.rs/id/SrpNemLexDE-SR/unendlich-a-de>, → **URI лексичке јединице (DE)** a lemon:LexicalSense; → **Значење лексичке јединице** skos:definition "nicht begrenzt, von nicht absehbarem Ausmaß"@de. → **Текстуална дефиниција** <http://llod.jerteh.rs/id/SrpNemLexDE-SR/112131-sr-sense> → **URI значења (SR)** lemon:isSenseOf <http://llod.jerteh.rs/id/SrpNemLexDE-SR/%20beskrajan-a-sr>, → **URI лексичке јединице (SR)** a lemon:LexicalSense; → **Значење лексичке јединице** skos:definition "nije ograničen, nepredvidivog obima"@sr. → **Текстуална дефиниција** <http://llod.jerteh.rs/id/SrpNemLexDE-SR/112131-sr-sense-112131-de-sense-trans> → **URI преводног пара** tr:translationSource <http://llod.jerteh.rs/id/SrpNemLexDE-SR/112131-sr-sense>; → **Преводни парови** tr:translationTarget <http://llod.jerteh.rs/id/SrpNemLexDE-SR/112131-de-sense>; a tr:Translation. <http://llod.jerteh.rs/id/SrpNemLexDE-SR/112175-de-sense> → **URI значења (DE)** lemon:isSenseOf <http://llod.jerteh.rs/id/SrpNemLexDE-SR/grenzenlos-a-de>; → **URI лексичке јединице (DE)** a lemon:LexicalSense; → **Значење лексичке јединице** skos:definition "keine Grenze habend"@de. → **Текстуална дефиниција** <http://llod.jerteh.rs/id/SrpNemLexDE-SR/112175-sr-sense> → **URI значења (SR)** lemon:isSenseOf <http://llod.jerteh.rs/id/SrpNemLexDE-SR/beskrajan-a-sr>, → **URI лексичких јединица (SR)** <http://llod.jerteh.rs/id/SrpNemLexDE-SR/bezgrani%C4%8Dan-a-sr>; a lemon:LexicalSense; → **Значење лексичке јединице** skos:definition "onaj koji je bez granice"@sr. → **Текстуална дефиниција**



Приликом креирања модела *lemon* предвиђено је коришћење више простора имена за дефинисање различитих ентитета. Сви простори имена који су предвиђени да се користе у моделу наводе се на почетку сваке RDF/Turtle датотеке. Простори имена се у запису препознају према префиксу који стоји уз ентитет. Један од простора имена предвиђених *lemon* моделом је и Lexinfo. Lexinfo (Buitelaar et al. 2009) је онтологија која дефинише типове, вредности и својства за анотацију лингвистичких карактеристика лексичких јединица која су делимично изведена из регистра DatCatInfo. DatCatInfo је Репозиторијум категорије података (Data Category Repository - DCR) развијен према стандарду ISO 12620:2019 настао као замена за регистар метаподатака ISOcat (Kemps-Snijders et al. 2008) који је представљао званичан регистар информација о лингвистичким категоријама података на вебу и значајан ресурс за истраживања и развој различитих лингвистичких дисциплина. У нашем примеру простор имена Lexinfo је коришћен за дефинисање лингвистичке анотације лексичких јединица из речника која се односи на врсту речи. Ако погледамо лексичке јединице које смо до сада узимали за пример “grenzenlos”, “unendlich”, „bezgraničan” и „beskrajan” у датотекама језика видећемо да уз њих стоји анотација lexinfo:partOfSpeech <http://www.lexinfo.net/ontology/2.0/lexinfo#-> која дефинише да је за дефинисање врсте реч коришћена онтологија Lexinfo.

Други простор имена предвиђен *lemon* моделом је и SKOS. Овај простор имена искоришћен је за исказивање значења лексичких јединица у облику дефиниције. Уз дефиницију стоји анотација *skos:definition* коју прати дефиниција и код за језик на коме је дефиниција. На пример, skos:definition “onaj koji je bez granice”@sr приказује дефиницију на српском.

Како што смо већ поменули је за припрему речника SrpNemLexDe-Sr као скупа повезаних података коришћена је табела преводних парова са синонимима која је екстрахована из корпуса СрпНемКор и уграђена у Терми (детаљније објашњено у одељку

6.3.2), која је, затим, извезена у формате LMF/XML и lemon/RDF. Речник SrpNemLexDe-Sr, као и табела преводних парова, садржи 3.984 упарене лексичке јединице са 864 синонима на српском и 791 синонимом на немачком. Поред тога, за осам лексичких јединица додате су текстуалне дефиниције које објашњавају њихово значење. Поред трансформације речника SrpNemLexDe-Sr у формате LMF/XML и lemon/RDF припремљено је и заглавље које садржи следеће метаподатке: наслов пројекта и његов кратак опис, сарадници на пројекту, број лексичких јединица у речнику и податак о лиценци којом се регулишу права приступа и коришћења садржаја речника, GNU General Public License v3.0[213]. Овим су створени услови за објављивање речника као скуп повезаних података.

7 Постигнути резултати и будући рад

7.1 Постигнути резултати

У досадашњем раду на паралелним корпусима Група за језичке технологије при Универзитету у Београду израдила је неколико вишејезичних паралелних корпуса који су приказани у дисертацији у поглављу 2, одељак 2.4.2. Радом на овој докторској дисертацији започели смо израду новог паралелног корпуса. Нови паралелни корпус садржи материјал на српском и немачком језику, а у овој фази истраживања обухватио је рад на књижевним текстовима, романима. Према завршним резултатима произведени корпус садржи четрнаест паралелизованих романа, односно преко 1,6 милиона корпусних речи или 48.004 преводних парова.

Поступак израде српско-немачког паралелног корпуса није се много разликовао у погледу припреме, обраде и паралелизације одабраног материјала у односу на већ постојеће паралелне корпусе који обухватају српски језик, а који су приказани у поглављу 2, одељак 2.4.2. За припрему и корекцију текстова на српском језику коришћени су електронски морфолошки речници српског језика које развија Група за језичке технологије, док су за обраду немачких текстова коришћени алати у слободном приступу као што су Транскрибус и Hunspell, алат за правописну и морфолошку контролу.

[213] GNU General Public License v3.0, https://www.gnu.org/licenses/gpl-3.0.en.html

Електронски морфолошки речници српског језика, који се редовно ажурирају и допуњују новим речима, након обраде текстова за овај корпус допуњени са више од 2.500 нових одредница. За даљу израду паралелних текстова коришћени су алати које, такође, развија Група за језичке технологије: коначни трансдуктори у оквиру програма Unitex за аутоматску сегментацију текстова на реченице и програмски пакет ACIDE за паралелизацију.

Посебан акценат у дисертацији стављен је на претраживању и проналажењу информација у креираном српско-немачком паралелном корпусу. Паралелни српско-немачки корпус смештен је у дигиталну библиотеку Библиша која омогућава постављање упита за претраживање на више језика и њихово морфолошко и семантичко проширење позивањем различитих вишејезичних лексичких и терминолошких ресурса. Библиша је до сада садржала паралелне српско-енглеске колекције текстова. Из тог разлога претраживање је било могуће само на ова два језика позивањем лексичких и терминолошких ресурса који су омогућавали морфолошко и семантичко проширење упита само на српском и енглеском језику. Како су лексички ресурси за српски језик који се у овом окружењу користе за проширење упита у већој мери развијени, у дисертацији смо приказали и анализирали могућности допуне неких ресурса који су слободно доступни како би се омогућила претрага колекције на немачком језику и користиле могућности алата за семантичко проширење упита за претрагу засноване на синонимима. Из овог разлога упарени српско-немачки текстови додатно су обрађени применом алата за екстракцију двојезичне терминологије који је развила Група за језичке технологије. Алат и цео поступак екстракције, заједно са ресурсима који се користе, анализирани су у дисертацији, а екстракцијом и евалуацијом добијена је листа немачко-српских преводних парова лексичких јединица која је уграђена у терминолошку базу Терми. Произведеној листи немачко-српских преводних парова лексичких јединица додата је и листа синонима на немачком и српском језику. Тестирањем система за претрагу позивањем овог терминолошког ресурса и анализом добијених резултата у виду генерисаних конкорданци преводних парова пронађене су неке лексичке јединице на немачком језику које нису биле део добијене листе преводних парова па су као нови кандидати додати у базу

Терми. У дисертацији је приказана структура базе Терми, као и поступак њеног ажурирања и допуне новим кандидатима.

Поред алата за екстракцију терминологије на припремљеном паралелном корпусу примењени су и расположиви алати за анотацију именованих ентитета у текстовима на оба језика. За анотацију ентитета у српској страни корпуса коришћен је алат који развија Група за језичке технологије, док су за анотацију именованих ентитета у немачком делу корпуса коришћени алати и језички модели доступни на вебу. Добијени резултати трансформисани су у више расположивих формата за приказ анотираних именованих ентитета преко портала NER&Beyond и упоређени су резултати у произведеним различитим алатима и излазном форматима.

Поред система претраге, посебан акценат у дисертацији стављен је и на повезивању ентитета из корпуса са релевантним ресурсима на вебу у оквиру облака „Отворени повезани подаци" и креирању скупа података који је постао део облака „Отворени повезани подаци из области лингвистике". Повезивање ентитета, у нашем случају библиографских метаподатака који се односе на имена писаца и на наслове романа, извршено је са три међународне нормативне датотеке имена (VIAF, LCNAF и GND), као и са општом базом знања Википодаци. Како су нормативне датотеке производ библиотека, кроз дисертацију смо указали и на улогу и значај библиотека и њених ресурса за нове технолошке могућности које носи семантички веб односно веб 3.0, али и колико је битно да библиотеке увиде предности примене овакве технологије у области претраживања и проналажења информација. Поред увезивања поменутих библиографских метаподатака, на основу садржаја корпуса креиран је скуп повезаних података. Као скуп повезаних података припремили смо двојезични немачко-српски речник општег типа, SrpNemLexDE-SR. За садржај речника искористили смо већ припремљену листу немачко-српских преводних парова лексичких јединица са синонимима, а за креирање речника у облику скупа повезаних података користили смо предложена упутства и препоруке за моделирање отворених повезаних података из области лингвистике. На овај начин тестирали смо расположиве алате за трансформацију једног двојезичног речника у формат RDF и анализирали добијене резултате. Поред

припремљеног речника креирано је и заглавље са метаподацима о самом речнику у XML формату.

7.2 План за будући рад

Поред постигнутих резултата у дисертацију су анализиране и неке могућности за даљи рад на креираној немачко-српској колекцији текстова, а међу њима морфолошко и семантичко проширење упита за претрагу на немачком језику. Како Библиша омогућава семантичко проширење упита позивањем мреже Ворднет на српском и енглеском, у дисертацији је анализирана могућност семантичког проширења упита позивањем семантичке мреже Ворднет на немачком језику. Наше истраживање обухватило је анализу ресурса који је у отвореном приступу, али још увек у развоју, Open-de-WordNet. Садржај ове семантичке мреже искористили смо за евалуацију произведене листе лексичких јединица на немачком језику приликом припреме листе преводних парова и утврдили да постоји могућност да у будућем раду Библиша омогући позивање једне овакве семантичке мреже која би вршила семантичко проширење упита на немачком језику.

Када је реч о морфолошком проширењу упита, Библиша за сада омогућава само проширење упита на српском језику позивањем електронских морфолошких речник српског језика. Како је немачки језик, такође, богат флексијом разматрана је могућност морфолошког проширења упита на немачком. За морфолошко проширење упита на немачком потребно је утврдити расположивост језичких ресурса који се у ове сврхе могу користити, да ли су они у слободном приступу или је потребно тражити дозволу за њихово коришћење од стране аутора. Наше истраживање обухватило је тестирање и анализу алата STTS (Stuttgart-Tübingen Tagset) (Schiller et al. 1999) за који је, након евалуације добијених резултата, утврђено да даје добре резултате.

Увезивањем библиографских метаподатака из корпуса са ресурсима у облаку „отворени повезани подаци” остављен је простор за даљи рад у овом правцу. Једна од могућности је и увезивање анотираних именованих ентитета са ресурсима из облака, на пример, географских појмова са међународном нормативном датотеком географских

имена GeoNames. Такође, поред увезивања, предвиђено је и упаривање анотираних именованих ентитета у обе стране корпуса на основу чега се може саставити листа преводних парова која се, даље, може трансформисати у скуп отворених повезаних података.

Када је реч о речнику SrpNemLexDE-SR који је припремљен као скупу повезаних података, остало је да заједно са припремљеним заглављем постане доступан на вебу и да се објави у подоблаку „Отворени повезани подаци из области лингвистике“. Поступак објављивања скупа повезаних података анализиран је у поглављу 5 одељак 5.2.4 ове докторске дисертације. Предвиђено је да речник буде објављен на домену *http://llod.jerteh.rs/id/SrpNemLexDE-SR/.* У овом тренутку се још увек тестирају расположиви алати за објављивање.

Поред свих претходно наведених планова за будући рад на корпусу СрпНемКор предвиђа се и да корпус постане доступан и преко IMS CQP, као и други паралелни корпуси приказани у поглављу 2 одељак 2.5.3. Такође, очекује се да се добијени корпус и анализа његовог садржаја заједно са резултатима рада који су наведени у дисертацији искористе у истраживањима из области језика, али и да се произведени корпус искористи као добар пример у пракси у погледу моделирања ресурса као скупа повезаних података и увезивања ентитета са релевантним ресурсима на вебу, посебно у области библиотекарства и информатичких наука с обзиром на значајну примену библиотечко-информационих ресурса у целом процесу. Са друге стране, очекује се да докторска дисертација понуди решења за повећање видљивости дигиталне колекције која се претражује, представљајући модел комплексне организације података који доприноси релевантнијим резултатима претраге дигиталних колекција различитог садржаја и њихову бољу видљивост на вебу.

8 Библиографија

8.1 Библиографске референце публиковане на ћирилици

1. Андоновски, Јелена и Гордана Недељков. Метаподаци у пројекту „Библиотеке Европеане" – искуство двогодишњег пројекта. У *Културе у дијалогу. Књ. 3, Културна дипломатија и библиотеке*, ур. Александра Вранеш и Љиљана Марковић, 275-288. Београд: Филолошки факултет Универзитета, 2013.
2. Брзуловић Станисављевић, Татјана. Ауторско право и библиотеке. *Панчевачко читалиште 17* (2010): 73-77. Преузето 28.01.2014, http://citaliste.rs/casopis/br17/brzulovic_tatjana.pdf
3. Витас, Душко и Љубомир Поповић. „Конспект за изградњу референтног корпуса српског стандардног језика". У *Научни састанак слависта у Вукове дане, 12-16.9.2011.,* вол. 1 (2003): 221-227
4. Вранеш, Александра. Дигитална хуманистика и савремене библиотеке. *Инфотека* 15, 1(2014): 4-15.
5. Вранеш, Александра и Љиљана Марковић. *Од рукописа до библиотеке*. Београд: Филолошки факултет, 2008
6. Гардашевић, Станислава. „Европеана и њени модели метаподатака". *Читалиште* 22 (2013а): 85-93, преузето 12.06.2016, http://www.citaliste.rs/casopis/br22/gardasevic_stanislava.pdf
7. Гардашевић, Станислава. „Семантички веб и Linked (Open) Data: могућности и перспективе за библиотеке". У *Инфотека: часопис за дигиталну хуманистику* 14, 1 (2013б): 29-40. http://infoteka.bg.ac.rs/pdf/Srp/2013-1/INFOTHECA_XIV_1_2014_29-40.pdf
8. Дакић, Наташа и Јелена Андоновски. „Интегрисање метаподатака у Европеану коришћењем формата ESE". У *Хоризонти светског и европског библиотекарства у дигиталном добу: зборник радова са међународне научне конференције,*

Београд, 27-28. октобар 2011, уредници Весна Црногорац и Весна Иљац, 67-80. Београд: Библиотекарско друштво Србије, 2012

9. Дакић, Наташа и Адам Софронијевић. „Демократизација дигитализације у библиотекама". Београд: Универзитетска библиотека „Светозар Марковић", 2018
10. Ерјавец, Томаж. „Смернице Иницијативе за кодирање текста и њихова локализација". *Инфотека* год. 11, бр. 1(2010): 3-15, преузето 27.3.2019, http://infoteka.bg.ac.rs/pdf/Srp/2010-1/INFOTHECA_XI_1_April2010_3-15.pdf
11. Завод за интелектуалну својину Републике Србије. „Закон о ауторским и сродним правима". *Службени гласник РС, број 104/2009, 99/2011 и 29/2016*. Доступно на: http://web.archive.org/web/20100401015004/www.zis.gov.rs/sr/pdf_ap/autorsko_zakon.pdf
12. Јанчић, Светлана. *Алфабетски каталог монографских публикација.* Београд: Народна библиотека Србије, 1991
13. Костић, Александар. „Електронски корпус српског језика Ђорђа Костића". У *Зборник Матице српске* 64 (2003): 260-264
14. Крстев, Цветана. „Дигиталне библиотеке – разграничење појмова". *Инфотека* 3, 1-2(2002): 3-14.
15. Крстев, Цветана и Душко Витас. „Информатички поглед на библиографију". У *Српска библиографија данас*, ур. Александра Вранеш, 229-241. Нови Сад: Матица српска, 2008.
16. Крстев, Цветана, Бојана Ђорђевић, Сања Антонић, Невена Ивковић-Берчек, Зорица Зорица, Весна Црногорац и Љиљана Мацура. „Кооперативан рад на доградњи српског Wordnet-а". *Инфотека* год. 9, бр. 1-2 (2008): 57-75, преузето 04.02.2017, http://infoteka.bg.ac.rs/pdf/Srp/2008/INFOTHECA_IX_1-2_May2008_57-75.pdf
17. Митровић, Јелена. „Електронски језички ресурси и алати за обраду српског језика и њихово унапређивање путем модела групне расподеле рада". Докт. Дисертација, Филолошки факултет Универзитета у Београду, 2018, преузето 26.5.2018, http://phaidrabg.bg.ac.rs/o:19057

18. Ристовић, Зоран. „Од корпуса до учионице – примена паралелизованих текстова у настави енглеског језика у основној школи". *Инфотека* год. 13, бр. 2 (2012): 52-66, преузето 27.3.2019, http://infoteka.bg.ac.rs/pdf/Srp/2012-2/INFOTHECA_XIII_2_December2012_52-66.pdf
19. Ристовић, Зоран. „Кумулативни ефекти експлоатације вишејезичних корпуса у настави страних језика". Докт. Дисертација, Филолошки факултет Универзитета у Београду, 2016
20. Савић, Ана. "Нормативна контрола у Србији". *Инфотека* год. 17, бр. 1 (2017): 99-112, преузето 23.3.2019, https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/article/view/2017.17.1.5_sr
21. Станковић, Ранка, Бранислав Тривић, Оливера Китановић, Бранислав Благојевић и Велизар Николић. „Развој геолошког терминолошког речника GeolISSTerm". *Инфотека*, год. 12, бр. 1 (2011): 53-67
22. Томашевић, Александра. „Развој модела за управљање рударском пројектном документацијом". Докт. дисертација, Универзитет у Београду, Рударско-геолошки факултет, 2018.
23. Тртовац, Александра и Јелена Андоновски. „Улога дигиталног корпуса из библиотекарства и информатике у развоју речника српског језика". У *Србија између истока и запада: наука, образовање, култура, уметност: тематски зборник у 4 књиге*, ур. Александра Вранеш и Љиљана Марковић. Књ. 4, *Језици балкана у компаративном и интердисциплинарном контексту,* ур. Александра Вранеш, Љиљана Марковић, 227-241. Београд: Филолошки факултет Универзитета, 2014
24. Тртовац, Александра. „Дескриптори метаподатака и дескриптори садржаја у проналажењу информација у дигиталним библиотекама". Докт. дисертација, Универзитет у Београду, Филолошки факултет, 2016.
25. Тртовац, Александра. „Проналажење информација у дигиталним библиотекама". Београд: Универзитетска библиотека „Светозар Марковић", 2017
26. Утвић, Милош. „Анотација корпуса савременог српског језика". *Инфотека: часопис за библиотекарство и информатику* год. 12, бр. 2 (2011): 39-51, преузето

14.2.2016, http://infoteka.bg.ac.rs/pdf/Srp/2011-2/INFOTHECA_XII_2_Decembar_39-51.pdf

27. Утвић, Милош. „Листе учестаности Корпуса савременог српског језика”. У *Научни састанак слависта у Вукове дане - Српски језик и његови ресурси: теорија, опис и примене. научни састанак слависта у Вукове дане, Београд, 12-15. IX 2013*, вол. 43, бр. 3, 241-262. Београд: Међународни славистички центар, Филолошки факултет, 2014
28. Филипи-Матутиновић, Стела, прев. (2011). „Лиценцирање у пројекту Еуропеана”. *Инфотека* 12, 2 (2011), преузето 12.06.2016, http://infoteka.bg.ac.rs/pdf/Srp/2011-2/INFOTHECA_XII_2_Decembar_3-18.pdf

8.2 Библиографске референце публиковане на латиници

29. Alemu, Getaneh, Brett Stevens, Penny Ross, and Jane Chandler. “Linked Data for libraries: Benefits of a conceptual shift from library-specific record structures to RDF-based data models”. *New library world* Vol. 113, No. 11/12 (2012): 549-570, преузето 25.2.2019, https://www.emeraldinsight.com/doi/pdfplus/10.1108/03074801211282920
30. Allwood, Jens. “Multimodal Corpora”. In *Corpus Linguistics: an international handbook*, eds. Anke Lüdeling and Merja Kytö, 207-225. Berlin: Moution de Grutyer, 2008. Преузето 28.02.2016, http://gup.ub.gu.se/records/fulltext/79320/79320.pdf
31. Antoniou, Grigoris, and Frank Van Harmelen. “Web ontology language: Owl”. In *Handbook on ontologies*, 67-92. Berlin, Heidelberg: Springer, 2004, преузето 17.2.2019, https://www.math.vu.nl/~frankh/postscript/OntoHandbook03OWL.pdf
32. Antoniou, Grigoris and Frank van Harmelen. *A Semantic Web Primer.* Cambridge, Massachusetts. London: MIT Press, 2008, preuzeto 16.06.2016, https://wtlab.um.ac.ir/images/reports/The.MIT.Press.Semantic.Web.Primer.2nd.Edition.Mar.2008.eBook-DDU.pdf
33. Archer, Phil, Stijn Goedertier and Nikolaos Loutas. “Interoperability Solutions for European Public Administrations”. 2012, preuzeto 19.3.2019,

https://joinup.ec.europa.eu/sites/default/files/document/2013-02/D7.1.3%20-%20Study%20on%20persistent%20URIs.pdf

34. Aston, Guy. „Text categories and corpus users: a response to David Lee". *Language learning & technology* Vol. 5, No. 3 (2001): 73-76
35. Aston, Guy and Lou Burnard. *The BNC Handbook: exploring the British National Corpus with SARA*. Oxford: Oxford University Computing Service, 1998
36. Auer, Sören, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, and Amrapali Zaveri. "Introduction to linked data and its lifecycle on the web". In *Reasoning Web. Semantic Technologies for Intelligent Data Access*, 1-90. Berlin, Heidelberg: Springer, 2013, преузето 16.07.2017, https://www.researchgate.net/profile/Soeren_Auer/publication/221510762_Introduction_to_Linked_Data_and_Its_Lifecycle_on_the_Web/links/55aac28108ae481aa7fbca7a.pdf
37. Aziz, Wilker, Sheila Castilho Monteiro de Sousa, and Lucia Specia. "Cross-lingual Sentence Compression for Subtitles". In *The 16th Annual Conference of the European Association for Machine Translation*, 103-110. 2012, преузето 20.07.2016, https://www.researchgate.net/profile/Lucia_Specia/publication/268385243_Cross-lingual_Sentence_Compression_for_Subtitles/links/56179c2608aee2517b9d33a8.pdf
38. Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". *arXiv preprint arXiv:1409.0473* (2014).
39. Balvet, Antonio, Dejan Stosic and Aleksandra Miletic. "TALC-Sef a Manually-revised POS-Tagged Literary Corpus in Serbian, English and French". In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, eds. Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis, 26-31. Reykjavik: European Language Resources Association, 2014, preuzeto 04.02.2017, http://www.lrec-conf.org/proceedings/lrec2014/index.html
40. Bennett, Rick, Christina Hengel-Dittrich, Edward T. O'Neill, and Barbara B. Tillett. „Viaf (Virtual International Authority File): Linking die Deutsche Bibliothek and Library of

Congress Name Authority Files." In *World library and information congress: 72nd IFLA general conference and council*. 2006, преузето 2.2.2019, http://origin-archive.ifla.org/IV/ifla72/papers/123-Bennett-en.pdf

41. Barnbrook, Geoff. *Language and computers: A practical introduction to the computer analysis of language*. Edinburgh University Press, 1996.
42. Bernard, David and Nancy Ide. "The Text Encoding Initiative: Flexible and Extensible Document Encoding". *Journal of the American Society for Information Science*vol. 48, no. 7 (1997): 622–628, preuzeto 25. 2. 2016, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.30.9983&rep=rep1&type=pdf
43. Bernardini, Silvia, Dominic Stewart and Federico Zanettin. "Corpora in translator education: an introduction". In *Corpora in translator education*, eds. Federico Zanettin, Silvia Bernardini, Dominic Stewart, 1-13. Routledge: Taylor & Francis, 2003
44. Berners Lee, Tim. "Linked Data". *W3C website*, 27.06.2006, preuzeto 12.06.2016, http://www.w3.org/DesignIssues/LinkedData.html
45. Berners-Lee, Tim, James Hendler and Ora Lassila. "The Semantic Web". *Scientific American: Feature Article* (2001), преузето 16.2.2019, https://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf
46. Berners-Lee, Tim, Roy Fielding, and Larry Masinter. *Uniform resource identifier (URI): Generic syntax*. No. RFC 3986. 2004, preuzeto 20.2.2019, https://www.ietf.org/rfc/rfc3986.txt
47. Berners-Lee, Tim. "Metadata Architecture", preuzeto 09.08.2016, http://www.w3.org/DesignIssues/Metadata.html
48. Bernska konvencija o zaštiti književnih i umetničkih dela, преузето 21.02.2017, http://www.zis.gov.rs/upload/documents/pdf_en/pdf_ap/bern.pdf
49. Biber, Douglas. Representativeness in corpus design. *Literary and Linguistic Computing* Vol. 8, No. 4 (1993): 243-257

50. Bizer, Christian, Tom Heath, and Tim Berners-Lee. “Linked data: The story so far”. In *Semantic services, interoperability and web applications: emerging concepts*, 205-227. IGI Global, 2011, преузето 23.2.2019, http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf
51. Bizer, Christian, Jens Lehmannb, Georgi Kobilarova, Sören Auer, Christian Becker, Richard Cyganiakc and Sebastian Hellmannb. “DBpedia - A crystallization point for the Web of Data”. *Web Semantics: Science, Services and Agents on the World Wide Web* 7 (2009): 154-165.
52. „BNC Sampler: XML edition”, 2008, preuzeto 26.3.2019, http://www.natcorp.ox.ac.uk/corpus/sampler/sampler.pdf
53. „Brat are stored on disk in a standoff format”, преузето 15.2.2019, http://brat.nlplab.org/standoff.html
54. A brief SGML tutorial. *W3C,* приступљено 08.07.2016, https://www.w3.org/TR/WD-html40-970708/intro/sgmltut.html
55. British National Corpus. *Oxford Text Archive, IT Services, University of Oxford,* преузето 12.02.2016, http://www.natcorp.ox.ac.uk/corpus/index.xml
56. Buitelaar, Paul, Philipp Cimiano, Peter Haase, and Michael Sintek. “Towards linguistically grounded ontologies”. In *European Semantic Web Conference*, pp. 111-125. Berlin, Heidelberg: Springer, 2009, preuzeto 14.03.2019, https://link.springer.com/content/pdf/10.1007/978-3-642-02121-3_12.pdf
57. Buitelaar, Paul, Philipp Cimiano, John McCrae, Elena Montiel-Ponsoda, and Thierry Declerck. “Ontology lexicalisation: The lemon perspective”. In *WS 2 Workshop Extended Abstracts, 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, Paris, 10 November 2011,* 33–36. 2011, preuzeto 19.3.2019, http://oa.upm.es/9772/1/Ontology_Lexicalisation.pdf
58. Burnard, Lou. *Users reference guide for the British National Corpus*. Oxford: Oxford University Computing Service, 1995
59. Burnard, Lou, ed. *Reference Guide to BNC Baby.* 2008, preuzeto 26.3.2019, http://www.natcorp.ox.ac.uk/corpus/baby/manual.pdf

60. Burnard, Lou, ed. “Reference Guide for the British National Corpus (XML Edition)”. *Research Technologies Service at Oxford University Computing Services*, преузето 04.02.2016, http://www.natcorp.ox.ac.uk/docs/URG/
61. Callison-Burch, Chris, Colin Bannard, and Josh Schroeder. “Improving statistical translation through editing”. In *Proceedings of the Workshop of the European Association for Machine Translation,* 26-32. 2004a, preuzeto 12.07.2016, http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=D9FB6F9EAAE6145DB257BB4174BBB4A9?doi=10.1.1.561.338&rep=rep1&type=pdf
62. Callison-Burch, Chris, David Talbot, and Miles Osborne. “Statistical machine translation with word-and sentence-aligned parallel corpora”. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 175-183. Association for Computational Linguistics, 2004b, preuzeto 12.07.2016, http://www.aclweb.org/anthology/P04-1023
63. Camacho-Collados, José, Claudio Delli Bovi, Alessandro Raganato and Roberto Navigli. “A Large-Scale Multilingual Disambiguation of Glosses”. In *Proceedings of LREC 2016*, *Portorož, Slovenia, 23-28 May 2016,* 1701-1708, preuzeto 24.07.2016, http://lcl.uniroma1.it/disambiguated-glosses/files/A_Large-Scale_Multilingual_Disambiguation_of_Glosses.pdf
64. Cantra, Linda. “METS: The Metadata Encoding and Transmission Standard”. *Cataloguing and Classification Quarterly* vol. 40, iss. 3-4 (2005): 237–253, preuzeto 26. 2. 2016, http://www.columbia.edu/cu/libraries/inside/units/bibcontrol/osmc/cantara.pdf
65. Chambers, Sally and Schallier, Wouter. “Bringing Research Libraries into Europeana: establishing a Library-Domain Aggregator”. *Liber Quarterly* 20, 1 (2010): 105-118, preuzeto 02.06.2016, https://www.liberquarterly.eu/articles/10.18352/lq.7980/, DOI: http://doi.org/10.18352/lq.7980
66. Chiarcos, Christian, Sebastian Hellmann and Sebastian Nordhoff. “Linking linguistic resources: Examples from the Open Linguistics Working Group”. In *Linked Data in Linguistics. Representing Language Data and Metadata*, Christian Chiarcos, Sebastian Nordhoff and Sebastian Hellmann (eds.), 201-216. Heidelberg: Springer, 2012.

67. Chinchor, Nancy, and Patricia Robinson. “MUC-7 named entity task definition”. In *Proceedings of the 7th Conference on Message Understanding*, vol. 29. 1997.
68. Chinchor, Nancy, and Elaine Marsh. “Muc-7 information extraction task definition”. In *Proceeding of the seventh message understanding conference (MUC-7), Appendices*, 359-367. 1998.
69. Cho, Kyunghyun, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. “On the properties of neural machine translation: Encoder-decoder approaches”. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. 2014.
70. Christ, Oliver, Bruno Schulze, Anja Hofman and Esther König. “The IMS Corpus Workbanch: Corpus Query Processor (CQP): User’s Manual”. Stuttgart: University of Stuttgart, Institut für maschinelle Sprachverarbeitung, 1994
71. Coyle, Karen. “Understanding metadata and its purpose”. *Journal of Academic Librarianship* vol. 31, no. 2 (2005): 160–163, preuzeto 23. 2. 2016, http://www.kcoyle.net/jal-31-2.html
72. Coyle, Karen. “Semantic Web and Linked Data”. In *Library Technology Reports* 48, 4 (2012): 10-14, preuzeto 15.06.2016, http://search.proquest.com/openview/2e2d7bbf2b229725e1bea2b74a119500/1?pq-origsite=gscholar
73. Cribb M. V. “Machine Translation: The Alternative for the 21st Century?”. *TESOL Quaterly,* Vol. 34, No. 3(2000): 560-569
74. Cundiff, Morgan. “An introduction to the Metadata Encoding and Transmission Standard (METS)”. *Library Hi Tech* vol. 22, no. 1 (2004): 52–64, preuzeto 26. 2. 2016, http://polaris.gseis.ucla.edu/gleazer/260_readings/Cundiff.pdf.
75. Danielsson, Pernilla and Daniel Ridings. “Practical Presentation of a “Vanilla” Aligner. Technical report”. In *TELRI Workshop in Alignment and Exploitation of Texts*, *Ljubljana, 1-2 February*. Department of Swedish, Göteborg University, GU-ISS-97-2, Språkdata, 1997.

76. Dakić, Nataša and Jelena Andonovski. “Developement of a new format EDM for metadata ingestion in Europeana”. *Pregled NCD* 23 (2013): 11-21, preuzeto 12.06.2016, http://elib.mi.sanu.ac.rs/files/journals/ncd/23/ncd23011.pdf
77. Davenport, Thomas H., and Laurence Prusak. *Working knowledge: How organizations manage what they know*. Harvard Business Press, 1998.
78. „DCMI Abstract Model”. Dublin Core Metadata Initiative, preuzeto 1.4.2019, http://dublincore.org/specifications/dublin-core/abstract-model/
79. Dean, Jason W. “Charles A. Cutter and Edward Tufte: Coming to a Library Near You, via BIBFRAME”. In *The Library with the Lead Pipe* (4 December 2013), preuzeto 13.08.2016, http://www.inthelibrarywiththeleadpipe.org/2013/charles-a-cutter-and-edward-tufte-coming-to-a-library-near-you-via-bibframe/?format=pdf
80. “Definition of the Europeana Data Model elements: version 5.2.7”. *Europeana*, preuzeto 12.06.2016, http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation//EDM_Definition_v5.2.7_042016.pdf
81. Digital Library Federation. „METS: Metadata Encoding and Transmission Standard: primer and reference manual: version 1.6 revised”, preuzeto 26. 2. 2016, http://www.loc.gov/standards/mets/METSPrimerRevised.pdf.
82. Dimitrova, Ludmila, Nancy Ide, Vladimir Petkevic, Tomaz Erjavec, Heiki Jaan Kaaler and Dan Tufis. “Multext-east: Parallel and comparable corpora and lexicons for six central and Eastern European languages”. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 1998, преузето 13.02.2016, http://www.aclweb.org/anthology/P98-1050
83. Dobrić, Nikola. “Corpus linguistics – the basic form of linguistic analysis”. *Philologia*no. 7 (2009): 359-363, преузето 05.02.2016, http://www.uni-klu.ac.at/iaa/downloads/CORPUS_LINGUISTICS_-226_THE_BASIC_FORM_OF_LINGUISTIC_ANALYSIS_NDobric.pdf, http://philologia.org.rs/Files/broj_7.pdf#page=47

84. Dunsire, Gordon. “Cluster BibData”, preuzeto 23.2.2019, https://www.w3.org/2005/Incubator/lld/wiki/Cluster_BibData
85. Erjavec, Tomaž, Cvetana Krstev, Vladimír Petkevič, Kiril Simov, Marko Tadić and Duško Vitas. “The MULTEXT-East Morphosyntactic Specifications for Slavic Languages”. In *Proceedings of the Workshop on Morphological Processing of Slavic Languages: 10th Conference of the European Chapter, EACL 2003, Budapest, Hungary, April 13th, 2003*, eds. Tomaž Erjavec and Duško Vitas, 25-32, преузето 13.02.2016, http://poincare.matf.bg.ac.rs/~cvetana/biblio/04erjavec.pdf
86. Erjavec, Tomaž and Nancy Ide. “The MULTEXT-East Corpus”. In *Proceeding of First International Conference on Language Resources & Evaluation*, *Granada, Spain, 28-30 May 1998*, 971-974, преузето 14.02.2016, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.465.5846&rep=rep1&type=pdf
87. Erjavec, Tomaž. “MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora”. In *LREC* 2004, 2544-2547, преузето 13.02.2016, http://nl.ijs.si/et/teach/jsi07-hlt/Bib/Multext_LREC04.pdf
88. Erjavec, Tomaž. “MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora”. In *LREC* 2010, 1535-1538, преузето 13.02.2016, http://nl.ijs.si/ME/V4/doc/bib/mte-lrec2010.pdf
89. Erjavec, Tomaž, Ann Lawson and Laurent Romary. “East Meet West: Producing Multilingual Resources in a European Countext”. In *First International Language Resources and Evaluation Conference, Granada, Spain*, 1998
90. Erjavec, Tomaž, Camelia Ignat, Bruno Pouliquen and Ralf Steinberger. “Massive multilingual corpus compilation: Acquis Communautaire and totale”. In *Archives of Control Sciences* Vol. 15, No. 4 (2005): 529-540, преузето 06.05.2016, http://nl.ijs.si/et/Bib/2006_ACS-HLT_Special_issue_JRC-final.pdf
91. Erxleben, Fredo, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. “Introducing Wikidata to the linked data web”. In *International Semantic Web Conference*, pp. 50-65. Springer, Cham, 2014.

92. European Parliament Proceedings Parallel Corpus 1996-2011, преузето 14.07.2016, http://www.statmt.org/europarl/
93. Evert, Stefan and the CWB Development Team. „The IMS Open Corpus Workbench (CWB): CQP Query Language Tutorial", May 2016, преузето 01.04.2017, http://cwb.sourceforge.net/files/CQP_Tutorial.pdf
94. "Extensible Markup Language (XML)". *W3C. Information and knowledge domain*, преузето 28.01.2017, https://www.w3.org/XML/
95. Fargo, Filip, Boris Bosančić i Boris Badurina. "Povezani podaci i knjižnice". *Vjesnik biblioteke Hrvatske* 56, 4(2013): 25-52
96. Fellbaum, Christiane (Ed.). "WordNet: An Electronic Lexical Database". Cambridge, London: MIT Press, 1998
97. Francopoulo, Gil, ed. *LMF lexical markup framework*. London: John Wiley & Sons, 2013.
98. Francopoulo, Gil, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. "Lexical markup framework (LMF)". In *International Conference on Language Resources and Evaluation-LREC 2006*, 233-236. European Language Resources Association, 2006, preuzeto 15.03.2019, http://www.lrec-conf.org/proceedings/lrec2006/pdf/577_pdf.pdf
99. Francis, Winthrop Nelson and Henry Kučera. "Brown Corpus Manual: Manual of Information". Rhode Island: Department of Linguistics, brown University, 1964. Revised 1971, Revised and Amplified 1979, преузето 12.03.2016, http://www.hit.uib.no/icame/brown/bcm.html
100. Gale, William A. and Kenneth W. Church. "A program for aligning sentences in bilingual corpora". *Computational linguistics* Vol. 19, No. 1 (1993): 75-102.
101. Gambette, Philippe, and Jean Véronis. "Visualising a Text with a Tree Cloud". In *IFCS'09: International Federation of Classification Societies Conference, Mar 2009, Dresde, Germany*, 561-569. Berlin, Heidelberg: Springer, 2010
102. Garcia, Jorge and Daniel Vila-Suero. "Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries", preuzeto 13.03.2019, https://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/#bib-lmf

103. Gartner, Richard. “METS: Metadata Encoding and Transmission Standard”. *JISC Techwatch report TSW02-05*, 2002, preuzeto 26. 2.2016, http://www.academia.edu/1095658/METS_Metadata_Encoding_and_Transmission_Standard

104. Gatos, Basilis, Georgios Louloudis, Tim Causer, Kris Grint, Veronica Romero, Joan Andreu Sánchez, Alejandro H. Toselli, and Enrique Vidal. “Ground-truth production in the tranScriptorium project.” In *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on Document Analysis Systems*, 237-241. IEEE, 2014, preuzeto 2.2.2019, https://riunet.upv.es/bitstream/handle/10251/66452/paper-DAS-2013.pdf?sequence=1

105. Gavrilidou, Maria, Peny Labropoulou, Elina Desipri, Voula Giouli, Vasilis Antonopoulos, and Stelios Piperidis. “Building parallel corpora for eContent professionals”. In *MLR '04 Proceedings of the Workshop on Multilingual Linguistic Ressources*, 97-100. Stroudsburg: Association for Computational Linguistics, 2004, preuzeto 12.07.2016, http://delivery.acm.org/10.1145/1710000/1706253/p97-gavrilidou.pdf?ip=95.180.45.16&id=1706253&acc=OPEN&key=4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35%2E6D218144511F3437&CFID=642941053&CFTOKEN=69760177&__acm__=1468346807_241d38514d31f3cc2c3be5dc56f439c2

106. Gill, Tony, Anne J. Gilliland, Maureen Whalen, and Mary S. Woodley. *Introduction to Metadata*. Los Angeles: Getty Research Institute, 2008, preuzeto 23. 2. 2016, http://www.getty.edu/research/publications/electronic_publications/intrometadata/

107. Glimm, Birte and Heiner Stuckenschmidt. “15 Years of Semantic Web: An Incomplete Survey”. *Künstliche Intelligenz* Vol. 30, Issue 2 (2016): 117-130, preuzeto 1.4.52019, https://link.springer.com/article/10.1007%2Fs13218-016-0424-1

108. Gonzales, Brighid. “Linking libraries to the web: Linked Data and the future of the bibliographic record”. *Information Technology and Libraries* 33, 4 (2014): 10-22.

109. Gracia del Río, Jorge, Elena Montiel Ponsoda, Daniel Vila Suero, and Guadalupe Aguado de Cea. “Enabling language resources to expose translations as linked data on

the web". (2014): 409-413, preuzeto 14.03.2019, http://www.lrec-conf.org/proceedings/lrec2014/pdf/863_Paper.pdf

110. Gracia, Jorge, Marta Villegas, Asunción Gómez-Pérez, and Núria Bel. "The apertium bilingual dictionaries on the web of data". *Semantic Web* 9, no. 2 (2018): 231-240.

111. Greenberg, Jane. "Metadata and World Wide Web". In: *Encyclopedia of Library and Information Science*, 72, suppl. 35(2003), 1876-1888, preuzeto 01.06.2016, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.4528&rep=rep1&type=pdf

112. Greenberg, Jane. "Understanding metadata and metadata schemes". *Cataloging and Classification Quarterly* vol. 40, no. 3–4 (2005): 17–36.

113. Grishman R., Sundheim B. „Message Understanding Conference-6: A Brief History". In: *COLING*, vol. 1, 466-471. Stroudsburg: Association for Computational Linguistics, 1996, преузето 14.2.2019, http://www.aclweb.org/anthology/C96-1079

114. Gruber, Thomas R. "Toward principles for the design of ontologies used for knowledge sharing?". *International journal of human-computer studies* Vol. 43, no. 5-6 (1995): 907-928, preuzeto 16.06.2016, https://eecs.ceas.uc.edu/~mazlack/ECE.716.Sp2011/Semantic.Web.Ontology.Papers/Gruber.93b.pdf

115. Guenther, Rebecca and Sally McCallum. "New Metadata Standards for Digital Resources: MODS and METS". *Bulletin of the American Society for Information Science and Technology* Vol. 29, Issue 2 (2003): 12–15, preuzeto 26.2.2016, http://onlinelibrary.wiley.com/doi/10.1002/bult.268/epdf

116. Guthro, Clem. "Digital Public Library of America". *Maine Policy Review* 22, 1 (2013): 126 -129, preuzeto 11.08.2016, http://digitalcommons.library.umaine.edu/cgi/viewcontent.cgi?article=1600&context=mpr

117. van Halteren, Hans. “Source language markers in EUROPARL translations”. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1,* 937-944. Stroudsburg: Association for Computational Linguistics, 2008.

118. Harpring, Patricia. *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works*. Los Angeles: Getty Publication, 2010, preuzeto 25. 2. 2016, http://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/

119. Hedden, Heather. *The accidental taxonomist*. Medford, New Jersey: Information Today, Inc., 2016.

120. Hellmann, Sebastian, Claus Stadler, and Jens Lehmann. “The German DBpedia: A sense repository for linking entities”. In *Linked data in linguistics*, 181-190. Berlin, Heidelberg: Springer, 2012.

121. Hermann, Karl Moritz and Phil Blunsom. “Multilingual distributed representations without word alignment”. In *Proceedings of the Second International Conference on Learning Representations (ICLR 2014)*. 2014

122. Hickey, Thomas B. and Jenny A. Toves. „Managing Ambiguity in VIAF”. *D-Lib Magazine* Vol. 20, No. 7/8 (2014), preuzeto 2.2.2019, http://mirror.dlib.org/dlib/july14/hickey/07hickey.print.html

123. Hochstein, Juliane. „Ihr Bibliothekare habt doch jetzt…”. Ein Jahr “Gemeinsame Normdatei.” *Theke aktuell* 20, no. 1 (2013): 19-23.

124. Hodge, Gail. „Metapodaci na lak način”. *Glasnik Narodne biblioteke Srbije* 1 (2004): 157–180, preuzeto 14. 2. 2016, http://www.nb.rs/view_file.php?file_id=860

125. Hurley, Bernard, John Price-Wilkin, Merrilee Proffitt and Howard Besser. *The Making of America II Testbed Project: A Digital Library Service Model.* Washington: The Digital Library Federation, 1999, preuzeto 26. 2. 2016, http://www.clir.org/pubs/reports/pub87/pub87.pdf.

126. Hutchins, John. “Translation technology and the translator”. In *Proceedings of the Eleventh Conference of the Institute of Translation and Interpreting*, *8-10 May 1997,*

113-120. London: ITI, 1997, преузето 15.03.2017, http://hutchinsweb.me.uk/ITI-1997.pdf

127. Ikonomov, Nikola and Milena Dobreva. “The making of... digital book”. *NCD Review* 13 (2008): 1-8, preuzeto 27.02.2017, http://elib.mi.sanu.ac.rs/files/journals/ncd/13/ncd13001.pdf

128. The IMS Open Corpus Workbench (CWB): CQP Query Language Tutorial, 2016, preuzeto 26.3.2019, http://cwb.sourceforge.net/files/CQP_Tutorial/

129. An introduction to the DPLA Metadata Model, March 5, 2015, preuzeto 11.08.2016, https://dp.la/info/wp-content/uploads/2015/03/Intro_to_DPLA_metadata_model.pdf

130. “ISO 8879:1986. Information processing - Text and office systems - Standard Generalized Markup Language (SGML)”. *ISO,* преузето 28.01.2017, http://www.iso.org/iso/catalogue_detail.htm?csnumber=16387

131. „ISO 24613:2008. Language resource management - Lexical markup framework (LMF)”. *ISO*, преузето 14.2.2019, https://www.iso.org/standard/37327.html

132. „ISO 2709:2008. Information and documentation – Format for information exchange”. *ISO,* преузето 14.2.2019, www.iso.org/obp/ui/#iso:std:iso:2709:ed-4:v1:en

133. „ISO 30042:2008. Systems to manage terminology, knowledge and content - TermBase eXchange (TBX)”. *ISO,* преузето 14.2.2019, https://www.tbxinfo.net/

134. „ISO 16642:2017. Computer applications in terminology -- Terminological markup framework”. *ISO*, преузето 14.2.2019, https://www.iso.org/standard/56063.html

135. „ISO 639”. *ISO*, https://www.loc.gov/standards/iso639-2/langhome.html

136. „ISO 3166”. *ISO*, https://www.iso.org/iso/en/prods-services/iso3166ma/index.html

137. „ISO 8601”. *ISO*, https://www.w3.org/TR/1998/NOTE-datetime-19980827

138. Jäger, Gerhard and James Roger. “Formal language theory: refining the Chomsky hierarchy”. *Philosophical transactions of the royal society B* Vol. 367, No. 1598 (2012):

1956-1970, преузето 12.03.2016, http://rstb.royalsocietypublishing.org/content/royptb/367/1598/1956.full.pdf. DOI: 10.1098/rstb.2012.0077

139. Janik, Maciej, Ansgar Scherp, and Steffen Staab. “The semantic web: collective intelligence on the web”. *Informatik-Spektrum* Vol. 34, No. 5 (2011): 469-483, преузето 17.2.2019, http://www.ansgarscherp.net/publications/pdf/J07-JanikScherpStaab-TheSemanticWeb-Preprint.pdf

140. Kalchbrenner, Nal and Phil Blunsom. “Recurrent Continuous Translation Models”. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, *Seattle, Washington, USA, 18-21 October 2013*, 1700-1709. Washington: Association for Computational Linguistics, 2013, преузето 17.03.2017, http://anthology.aclweb.org/D/D13/D13-1176.pdf

141. Kemps-Snijders, Marc, Menzo Windhouwer, Peter Wittenburg, and Sue Ellen Wright. “ISOcat: Corralling data categories in the wild”. In *Proceedings of the international conference on language resource and evaluation (LREC’08)*, 887-891. European Language Resources Association, 2008, преузето 15.03.2019, http://www.lrec-conf.org/proceedings/lrec2008/pdf/222_paper.pdf

142. Klarin, Sofija. „Strukturalni metapodaci digitalnih objekata”. U *10. seminar Arhivi, knjižnice, muzeji: mogućnosti suradnje u okruženju globalne informacijske strukture: zbornik radova,* ur. Ivana Marinković Zenić, Mirna Willer, 41-53. Zagreb: Hrvatsko knjižničarsko društvo, 2007, preuzeto 12.2.2016, https://www.academia.edu/5317522/Strukturalni_metapodaci_digitalnih_objekata_Structural_metadata_for_digital_objects_Sofija_Klarin_2006_

143. Koehn, Philipp. “Europarl: A parallel corpus for statistical machine translation”. In *MT summit*, Vol. 5, 79-86. 2005, preuzeto 14.07.2015, http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf

144. Koehn, Philipp, Franz Josef Och, and Daniel Marcu. “Statistical phrase-based translation”. In *Proceedings of the 2003 Conference of the North American Chapter of*

the Association for Computational Linguistics on Human Language Technology-Volume 1, 48-54. Association for Computational Linguistics, 2003.

145. Koehn, Philipp, and Josh Schroeder. “Experiments in domain adaptation for statistical machine translation”. In *Proceedings of the second workshop on statistical machine translation*, *Prague, June 2007,* 224-227. Association for Computational Linguistics, 2007, преузето 17.03.2017, http://www.statmt.org/wmt07/WMT-2007.pdf#page=238

146. Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan et al. “Moses: Open source toolkit for statistical machine translation”. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177-180. Stroudsburg: Association for Computational Linguistics, 2007

147. Koehn, Philipp, Alexandra Birch and Ralf Steinberger. “462 machine translation systems for Europe”. In *Proceedings of MT Summit XII*, 65-72. 2009, preuzeto 11.05.2016, https://www.researchgate.net/profile/Philipp_Koehn/publication/228613050_462_machine_translation_systems_for_europe/links/09e4150db363e5ae18000000.pdf

148. Koivunen, Marja-Riitta, and Eric Miller. “W3C semantic web activity”. *Semantic Web Kick-Off in Finland* 2 (2001): 27-44., preuzeto 19.2.2019, https://www.w3.org/2001/12/semweb-fin/w3csw

149. Kostić, Aleksandar. “Electronic corpus of Serbian language from 12th to 18th century”. *Преглед НЦД* 24 (2014): 35-42, преузето 18.05.2016, http://elib.mi.sanu.ac.rs/files/journals/ncd/24/ncd24035.pdf

150. Korpus savremenog srpskog jezika na Matematičkom fakultetu Univerziteta u Beogradu. *Matemetički fakultet Univerziteta u Beogradu,* preuzeto 12.07.2016, http://korpus.matf.bg.ac.rs/prezentacija/istorija.html

151. Kovačević, Ljiljana, Vesna Injac i Dobrila Begenišić. „Bibliotekarski terminološki rečnik: englesko-srpski, srpsko-engleski“ [Library Terminological Dictionary: English-Serbian, Serbian-English]. Beograd: Narodna biblioteka Srbije, 2004

152. Krstev, Cvetana, Smiljana Jović-Puač i Duško Vitas. “Analiza podjezika uputstava za lekove na srpskohrvatskom i slovenačkom jeziku, deo I”. U *Zbornik radova sa IV naučnog skupa “Računarska obrada jezičkih podataka”, Portorož, 3-7 oktobar 1988*, eds. Damjan Bojadžijev, Petar Tancig, Duško Vitas, 249-255. Ljubljana: Institut “Jožef Štefan”, Društvo za uporabno jezikoslovje Slovenije, 1988

153. Krstev, Cvetana i Duško Vitas. “Konkordancije paralelizovanih tekstova”. U *Zbornik radova XXXVIII konferencije ETRAN, Niš, juni 1994*, ed. Slobodan Lazović, 229-230. Beograd: Društvo za elektroniku, telekomunikacije, računarstvo, automatiku i nuklearnu tehniku, 1994

154. Krstev, Cvetana, Duško Vitas and Tomaž Erjavec. “Morpho-Syntactic Descriptions in MULTEXT-East - the Case of Serbian”. In *Informatica* No. 28 (2004): 431-436

155. Krstev, Cvetana and Duško Vitas. “Corpus and Lexicon - Mutual Incompletness”. In *Proceedings of the Corpus Linguistics Conference, 14–17 July 2005, Birmingham*, eds. P. Danielsson & M. Wagenmakers, 14-27. 2005, преузето 26.01.2016, http://poincare.matf.bg.ac.rs/~cvetana/biblio/CVDV-CorpusLing05.pdf

156. Krstev, Cvetana. “Specifični koncepti Balkana u semantičkoj mreži Wordnet”. U *Zborniku radova "Susreti kultura"*, *Novi Sad, decembar 2004,* eds. Ljiljana Subotić et al, 275-285. Novi Sad: Univerzitet u Novom Sadu, Filozofski fakultet, 2006, преузето 26.5.2019, http://poincare.matf.bg.ac.rs/~cvetana/biblio/CK-SusretiKultura.pdf

157. Krstev, Cvetana, Duško Vitas and Agata Savary. “Prerequisites for a Comprehensive Dictionary of Serbian”. In *Proceedings of the 5th International Conference on NLP*, FinTAL 2006, Turku, Finland, August, 2006, eds. Tapio Salakoski, Filip Ginter, Sampo Pyysalo, Tapio Pahikkala. Serija *Lecture Notes in Artificial Intelligence: Subseries of Lecture Notes in Computer Science*, eds. J.G. Carbonell, J. Siekmann, 552-564. Heidelberg, Berlin: Springer, 2006a

158. Krstev, Cvetana, Ranka Stanković, Duško Vitas and Ivan Obradović. “WS4LR - a Worksation for Lexical Resources”. In *Proceedings of the Fifth Interantional Conference on Language Resources and Evaluation*, 1692-1697. Genoa, 2006b, преузето 17.03.2016, http://poincare.matf.bg.ac.rs/~cvetana/biblio/Krstev_467_new.pdf

159. Krstev, Cvetana, Svetla Koeva, Duško Vitas. “Towards the Global Wordnet”. In *Conference Abstracts of the First Interantional Conference of Digital Humanities Organisations (ADHO) Digital Humanties 2006*, *Paris-Sorbonne, 5-9 July 2006,* 114-117, 2006c

160. Krstev, Cvetana. *Processing of Serbian: automata, texts and electronic dictionaries.* Belgrade: Faculty of Philology, 2008

161. Krstev, Cvetana, Duško Vitas and Gordana Pavlović-Lažetić. “Resources and Methods in the Morphosyntactic Processing of Serbo-Croatian”. In *Formal Description of Slavic Languages: The Fifth Conference, Leipzig 2003*, (eds.) Zybatow, Gerhild et al., 3-17. Peter Lang: Frankfurt am Main, 2008, преузето 13.04.2016, http://poincare.matf.bg.ac.rs/~cvetana/biblio/FDSL5-final.pdf

162. Krstev, Cvetana and Duško Vitas. “An Effective Method for Developing a Comprehensive Morphological E-dictionary of Compounds.” In *Arena Romanistica,* eds. B. Lamiroy, E. Laporte, T. Kyriakopoulou, 204-212. Bergen: University of Bergen, Department of Foreign Languages, 2009, преузето 25.02.2014, http://poincare.matf.bg.ac.rs/~cvetana/biblio/Krstev-Vitas-LGC09.pdf

163. Krstev, Cvetana, Ranka Stanković, Ivan Obradović, Duško Vitas and Miloš Utvić. “Automatic Construction of a Morphological Dictionary of Multi-Word Units”. In *Proceedings of the 7th International Conference on NLP, IceTAL 2010, Reykjavik, Iceland, August 16-18, 2010*, eds. Hrafn Loftsson, Eiríkur Rögnvaldsson, Sigrún Helgadóttir. Lecture Notes in Computer Science 6233, 226-237. Berlin, Heidelberg: Springer, 2010.

164. Krstev, Cvetana and Duško Vitas. “Analigned English-Serbian corpus”. In *ELLSIIR Proceedings (English Language and Literature Studies: Image, Identity, Reality), Volume I, Belgrade, 4-6 December 2009*, eds. N. Tomović & J. Vujić, 495-508. Belgrad: Faculty of Philology, University of Belgrade, 2011, преузето http://poincare.matf.bg.ac.rs/~cvetana/biblio/AlignedCorpus-full-final.pdf

165. Krstev, Cvetana, Duško Vitas and Aleksandra Trtovac. “Orwell’s *1984* – the case of Serbian revisited”. In *Proceedings of 5th Language & Technology Conference*, ed.

Zygmunt Vetulani, 570-574. Poznań: Fundacja Uniwersytetuim. A. Mickiewicza, 2011, преузето 13.02.2016, http://poincare.matf.bg.ac.rs/~cvetana/biblio/LTC-105-Krstev.pdf

166. Krstev, Cvetana, Ivan Obradović, Miloš Utvić and Duško Vitas. "A System for Named Entity Recognition Based on Local Grammars". *Journal of Logic and Computation*, Vol. 24, Issue 2 (2014): 473-489, 2014, doi:10.1093/logcom/exs079

167. Krstev, Cvetana. *Kurs iz XML-a*, приступљено 11.07.2016a, http://poincare.matf.bg.ac.rs/~cvetana/kurs-xml/xml-sadr.html

168. Krstev, Cvetana. *Pronalaženje obrazaca u tekstu*. Kurs Leksičko prepoznavanje u obradi prirodnih jezika, preuzeto 12.07.2016b, http://poincare.matf.bg.ac.rs/~cvetana/Nastava/1516/nastava1516-new.html

169. Krstev, Cvetana, Anđelka Zečević, Duško Vitas, and Tita Kyriacopoulou. "NERosetta for the Named Entity Multi-lingual Space". In *Human Language Technology Challenges for Computer Science and Linguistics*, LNCS, 327-340. Springer International Publishing, 2016, DOI 10.1007/978-3-319-43808-5_25

170. Krstev, Cvetana, Branislava Sandrih, Ranka Stanković, and Miljana Mladenović. "Using English Baits to Catch Serbian Multi-Word Terminology". In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, *LREC 2018, Miyazaki, Japan, May 7-12, 2018*, eds. Nicoletta Calzolari et al., 2487-2494. European Language Resources Association, 2018.

171. Kučera, Henry. "Obituary for W. Nelson Francis". *Journal of English Linguistics* Vol. 30, No. 4(2002): 306-309

172. Laporte, Eric. *The RELEX Network.* 2003, преузето 13.04.2016, http://infolingu.univ-mlv.fr/english/Relex/Relex.html

173. Laporte, Eric, Duško Vitas and Cvetana Krstev. "Preparation and exploitation of Bilingual Texts". In *Lux Coreana* No. 1, 110-132. Han-Seine, 2006, преузето 20.04.2016, http://poincare.matf.bg.ac.rs/~cvetana/biblio/VKL.pdf

174. Leech, Geoffrey. "Corpora and theories of linguistic performance". In *Trends in Linguistics. Studies and Monographs*, *Vol. 65*, *Directions in corpus linguistics:*

proceedings of Nobel Symposium, ed. Jan Svartvik, 105-122. Berlin: Mouton de Gruyter, 1992

175. Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann et al. "DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia". *Semantic Web* 6, no. 2 (2015): 167-195.

176. Library of Congress. MODS Users Guidelines (Version 3), preuzeto 1. 3. 2016, http://www.loc.gov/standards/mods/userguide/

177. Linked Data - Connect Distributed Data across the Web: Frequently Asked Questions (FAQs), preuzeto 23.2.2019, http://linkeddata.org/faq

178. Lison, Pierre and Jörg Tiedemann. "OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles". In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016),* eds. Nicoletta Calzolari et al, 923-929. Portorož, 2016, preuzeto 20.07.2016, http://www.lrec-conf.org/proceedings/lrec2016/pdf/947_Paper.pdf

179. Lyons, John. "On competence and performance and related notions". In *Performance and competence in second language acquisition*, eds. Gillian Brown, Kirsten Malmkjaer, and John Williams, 11-32. Cambridge: Cambridge University Press, 1996

180. Ljubešic, Nikola, and Filip Klubicka. "{bs, hr, sr} WaC–web corpora of Bosnian, Croatian and Serbian". In *Proceedings of the 9th Web as Corpus Workshop* (WaC-9), 29-35. 2014.

181. McCallum, Sally. "An introduction to the Metadata Object Description Schema (MODS)". *Library Hi Tech* Vol. 22, No. 1 (2004): 82-88, preuzeto 1. 3.2016, http://polaris.gseis.ucla.edu/gleazer/260_readings/McCallum.pdf

182. McCrae, John, Guadalupe Aguado-de-Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia et al. "Interchanging lexical resources on the semantic web". *Language Resources and Evaluation* 46, no. 4 (2012): 701-719, preuzeto 14.03.2019, https://link.springer.com/content/pdf/10.1007%2Fs10579-012-9182-3.pdf

183. McEnery, Tony and Andrew Wilson. *Corpora and translation: uses and future prospects*. 1993, преузето 04.03.2016, http://ucrel.lancs.ac.uk/papers/techpaper/vol2.pdf

184. McEnery, Tony, and Andrew Wilson. *Corpus linguistics: An introduction*. Edinburgh: Edinburgh University Press, 2001.

185. McEnery, Tony, Richard Xiao and Yukio Tono. *Corpus-based language studies: an advanced resource book.* London: Routledge, 2006

186. McEnery, Tony and Richard Xiao. "Parallel and comparable corpora: What is happening". In *Incorporating Corpora. The Linguist and the Translator,* eds. Gunilla Anderman and Margaret Rogers, 18-31. Cleveldon: Multilingual Matters, 2008

187. McEnery, Tony and Andrew Hardie. *Corpus Linguistics: method, theory and practice*. Cambridge: Cambridge University Press, 2012

188. Mell, Peter, and Tim Grance. "The NIST definition of cloud computing". Gaithersburg: National Institute of Standards and Technology, 2011, преузето 16.10.2017, http://faculty.winthrop.edu/domanm/csci411/Handouts/NIST.pdf

189. Mendes, Pablo N., Max Jakob and Christian Bizer. "DBpedia: A Multilingual Cross-Domain Knowledge Base". In *LREC*, 1813-1817. Luxemburg: European Language Resources Association 2012, preuzeto 23.06.2016, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.680.5157&rep=rep1&type=pdf

190. Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine J. Miller. „Introduction to WordNet: An On-line Lexical Database". *International Journal of Lexicography* Vol. 3, Issue 4(1990): 235-244

191. Miles, Alistair, Brian Matthews and Michael Wilson. "SKOS core: simple knowledge organisation for the web". In *International Conference on Dublin Core and Metadata Applications*, 3-10. 2005, preuzeto 26.05.2017, http://dcpapers.dublincore.org/pubs/article/view/798/794

192. Miletic, Aleksandra, Cécile Fabre, and Dejan Stosic. "Construction du jeu d'étiquettes pour le parsing du serbe". *Actes de la conférence conjointe JEP-TALN-*

RECITAL 2016, Vol. 2, 1-12. 2015, preuzeto 04.02.2017, http://www.atala.org/taln_archives/ateliers/2015/TASLA/tasla-2015-long-001.pdf

193. Miller, Eric. “An Introduction to the Resource Description Framework”. *Bulletin of the Association for Information Science and Technology* 25, 1(1998): 15-19, preuzeto 08.05.2017, http://onlinelibrary.wiley.com/doi/10.1002/bult.105/ful

194. Miller, Eric, Uche Ogbuji, Victoria Mueller, and Kathy MacDougall. “Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services” (11 November 2012), preuzeto 26.05.2017, http://www.loc.gov/bibframe/pdf/marcld-report- 11-21-2012.pdf

195. Miller, Vaughne. “The EU's Acquis Cummunautaire”, SN/IA/5944: Library House of Commons, 2011

196. Milošević, Uroš, Vuk Mijović, and Sanja Vraneš. “Taking DBpedia Across Borders: Building the Serbian Chapter”. *ICIST 2014*, преузето 23.2.2019, https://www.researchgate.net/publication/261375671_Taking_DBpedia_Across_Borders_Building_the_Serbian_Chapter

197. Mitchell, Erik T. “Three case studies in linked open data”. *Library Technology Reports* 49, 5 (2013): 26-43, preuzeto 11.08.2016, https://journals.ala.org/ltr/article/view/4693/5587

198. Mosavi Miangah, Tayebeh. “Applications of corpora in translation”. *Translation Studies* 12 (2006): 43-56, преузето http://www.researchgate.net/profile/Tayebeh_Mosavi_Miangah/publication/271604433_Applications_of_corpora_in_translation/links/54cddc580cf298d6565e3ef8.pdf

199. Mühlberger, Günter. “H2020 Project READ (Recognition and Enrichment of Archival Documents) - 2016-2019”. Преузето 6.8.2018, http://www.academia.edu/22653102/H2020_Project_READ_Recognition_and_Enrichment_of_Archival_Documents_-_2016-2019

200. Navigli, Roberto and Simone Paolo Ponzetto. “BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic

network". *Artificial Intelligence* 193 (2012): 217–250, preuzeto 31.01.2017, http://wwwusers.di.uniroma1.it/~navigli/pubs/AIJ_2012_Navigli_Ponzetto.pdf

201. Neudecker, Clemens. "An Open Corpus for Named Entity Recognition in Historic Newspapers". In *LREC 2016 May*, 4348-4352. 2016.

202. Obitko, Marek. "Ontologies and Semantic Web". 2007, preuzeto 16.10.2017, http://obitko.com/tutorials/ontologies-semantic-web/semantic-web-architecture.html

203. Obradović, Ivan, Ranka Stanković i Miloš Utvić. „Integrisano okruženje za pripremu paralelizovanog korpusa". In *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen*, 563-578. Münster: LitVerlag, 2008

204. Och, Franz Josef and Hermann Ney. "A Systematic Comparison of Various Statistical Alignment Models". *Computational linguistics*, Vol. 29 No. 1(2003), 19–51.

205. Paskaleva, Elena, and Stoyan Mihov. "Second language acquisition from aligned corpora". In *Proceedings of International Conference "Language Technology and Language Teaching*, *Groningen, April 1997*, 43-52. Netherlands: Swets and Zeitingler, 1997

206. Paumier, Sébastien. *Unitex 3.0 User Manual*, 2011, преузето 05.04.2016, http://www.cis.uni-muenchen.de/people/lg3/ManuelUnitex.pdf

207. Pavlović-Lažetić, Gordana, Cvetana Krstev, Ivan Obradović and Duško Vitas, ed. *Natural language processing for Serbian.* Belgrade: University of Belgrad, Faculty of Mathematics, 2014, dostupno na: http://jerteh.rs/index.php/zbornici/

208. Pekárek, Aleš, and Marieke Willems. "The Europeana newspapers–a gateway to European newspapers online". In *Euro-Mediterranean Conference*, 654-659. Berlin, Heidelberg: Springer, 2012

209. Popović, Nataša i Milica Numović. „Upravljanje u oblaku – savremeni izazov u sistemima automatskog upravljanja". U *Infoteh-Jahorina, 16-18. mart 2016,* urednika Slobodan Milojković, 816-820. Istočno Sarajevo: Elektrotehnički fakultet, Univerzitet u Istočnom Sarajevu, 2016, preuzeto 16.10.2017, http://infoteh.etf.unssa.rs.ba/zbornik/2016/radovi/SUP-1/SUP-1-14.pdf

210. Prlja, Dragan, Mario Reljanović i Zvonimir Ivanović. *Internet pravo.* Beograd: Institut za uporedno pravo Beograd, 2012, преузето 21.02.2017, http://www.comparativelaw.info/ip.pdf

211. Proisl, Thomas and Peter Uhrig. "Efficient Dependency Graph Matching with the IMS Open Corpus Workbench". In *LREC 2012, Eighth International Conference on Language Resources and Evaluation, May 21-27, 2012, Istanbul, Turkey*, eds. Nicoletta Calzolari et al., 2750-2756. European Language Resources Association: 2012, преузето 07.04.2016, http://www.lrec-conf.org/proceedings/lrec2012/pdf/709_Paper.pdf

212. Radulovic, Filip, María Poveda-Villalón, Daniel Vila-Suero, Víctor Rodríguez-Doncel, Raúl García-Castro and Asunción Gómez-Pérez. "Guidelines for Linked Data generation and publication: An example inbuilding energy consumption". *Automation in Construction* 57 (2015): 178–187

213. Ramshaw, Lance A., and Mitchell P. Marcus. "Text chunking using transformation-based learning". In *Natural language processing using very large corpora,* 157-176. Dordrecht: Springer, 1999

214. „Regeln für den Schlagwortkatalog". Leipzig, Frankfurt am Main, Berlin: Deutsche Nationalbibliothek, 2010, преузето 2.2.2019, https://d-nb.info/1003001890/34

215. Reitz, Joan M. "ODLIS – Online Dictionary for Library and Information Science, Authority file", http://www.abc-clio.com/ODLIS/searchODLIS.aspx(приступљено 23.2.2019)

216. Reppen, Randi. *Using corpora in the language classroom*. New York: Cambridge University Press, 2014

217. "RDF Primer". *World Wide Web Consortium*, preuzeto 16.06.2016, https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/

218. Sarić, Ivana, Antonio Magdić i Mario Essert. „Sheme metapodataka značajne za knjižničarstvo s primjerom implementacije OPENURAL-a standarda". *Vjesnik bibliotekara Hrvatske* 54, 1/2 (2011): 134-157, preuzeto 24.02.2016, http://bib.irb.hr/datoteka/516390.vbh_54_1-2_saric_magdic_essert_sheme.pdf

219. Sánchez, J.A., Schofield, P., Depuydt, K., Gatos, B., Davis, R.M., Mühlberger, G.: tranScriptorium: an European Project on Handwritten Text Recognition. DocEng'13, Sept. 2013, Florence, Italy. 227–228 (2013)

220. Savary, Agata. "Computational Inflection of Multi-Word Units, a contrastive study of lexical approaches". *Linguistic Issues in Language Technology* 1-2 (2008): 1-53

221. Schiller, Anne; Teufel, Simone; Stöckert, Christine; Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS. (Kleines und großes Tagset). Universität Stuttgart, Institut für maschinelle Sprachverarbeitung; Universität Tübingen, Seminar für Sprachwissenschaft, preuzeto 1.4.2019, http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf

222. Schmid, Helmut. "Probabilistic Part-of-Speech Tagging Using Decision Trees". In *Proceedings of the International Conference on New Methods in Language Processing*: 44-49. Manchester, 1994, преузето 02.04.2016, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.1139&rep=rep1&type=pdf

223. Seaward, Louise and Maria Kallio. "Transkribus: Handwritten Text Recognition technology for historical documents". Preuzeto 5.8.2018, https://dh2017.adho.org/abstracts/649/649.pdf

224. Sekine, Satoshi, Kiyoshi Sudo, and Chikashi Nobata. "Extended Named Entity Hierarchy". In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands, Spain, May 29-31 2002, 1818-1824. European Language Resources Association, 2002, preuzeto 27.2.2019, http://www.lrec-conf.org/proceedings/lrec2002/pdf/120.pdf

225. Sekine, Satoshi, and Chikashi Nobata. "Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy". In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pp. 1977-1980. European Language Resources Association, 2004, preuzeto 27.2.2019, http://www.lrec-conf.org/proceedings/lrec2004/

226. Shadbolt, Nigel, Tim Berners-Lee and Wendy Hall. „The Semantic Web Revisited". In *IEEE Intelligent Systems* 21, 3 (2006): 96-101, preuzeto 16.06.2016, https://eprints.soton.ac.uk/262614/2/OLD_Semantic_Web_Revisted.pdf

227. Shukair, Gofran, Nikolaos Loutas, Vassilios Peristeras and Sebastian Sklarß. "Towards semantically interoperable metadata repositories: The Asset Description Metadata Schema". *Computers in Industry* vol. 64, iss. 1 (2013): 10–18

228. Sikos, Leslie. *Mastering structured data on the Semantic Web: From HTML5 microdata to linked open data*. Apress, 2015

229. Silberztein, Max. *Dictionnarie selectroniques et analyse automatique de textes. Le systeme INTEX*. Paris, 1993

230. Simeon, Ivana. "Paralelni korpusi I višejezični rječnici". *Filologija* br. 38-39 (2002): 209-215

231. Sinclair, John. "Corpus and text – basic principles". In *Developing Linguistic Corpora: A Guide to Good* Practice, ed. Martin Wynne, 1-16. Oxford: Oxbow Books, 2005

232. Slocum, Jonathan. "A survey of machine translation: its history, current status, and future prospects". *Computational linguistics* vol. 11, no. 1 (1985): 1-17, preuzeto 17.03.2017, http://wing.comp.nus.edu.sg/~antho/J/J85/J85-1001.pdf

233. SPARQL 1.1 Query Language. W3C, 2013, preuzeto 26.3.2019, https://www.w3.org/TR/sparql11-query/

234. "SRPS ISO 24616:2018". *ISO*, https://www.iss.rs/rs/standard/?natstandard_document_id=60906

235. Stanković, Ranka, Ivan Obradović, Cvetana Krstev and Duško Vitas. "Production of morphological dictionaries of multi-word units using a multipurpose tool". In *Proceedings of the Computational Linguistics-Applications Conference, October 17–19, 2011*, eds. K. Jassem, P. W. Fuglewicz, M. Piasecki and A. Przepiórkowski, 77 – 84. Jachranka: Polish Information Processing Society, 2011

236. Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac and Miloš Utvić. "A Tool for Enhanced Search of Multilingual Digital Libraries of E-journals". In *Proceedings of the 8th International Conference on Language Resources and Evaluation,*

LREC 2012, Istanbul, Turkey, 23--25 May 2012, eds. Nicoletta Calzolari et al., 1710-1717. Istanbul: European Language Resources Association, 2012, преузето 16.04.2016, http://www.lrec-conf.org/proceedings/lrec2012/pdf/375_Paper.pdf

237. Stanković, Ranka, Ivan Obradović and Miloš Utvić. "Developing termbases for expert terminology under the TBX standard". In *Natural language processing for Serbian: resources and applications*, eds. Gordana Pavlović-Lažetić et al., 12-26. Belgrade: University of Belgrade, Faculty of Mathematics, 2014

238. Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Olivera Kitanović. "Indexing of Textual Databases Based on Lexical Resources: - A Case Study for Serbian". In *Semantic Keyword-Based Search on Structured Data Sources - First COST Action IC1302 International KEYSTONE Conference, IKC 2015, Coimbra, Portugal, September 8-9, 2015. Revised Selected Papers*, LNCS 9398. Springer, 167-181, 2015. DOI 10.1007/978-3-319-27932-9_15

239. Stanković, Ranka, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac. "Rule-based Automatic Multi-Word Term Extraction and Lemmatization". In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, LREC 2016, Portorož, Slovenia, 23--28 May 2016, 2016, eds. Nicoletta Calzolari *et al.*, 507-514. European Language Resources Association, 2016, preuzeto 1.4.2019, https://pdfs.semanticscholar.org/b4bf/f89ec40e64e63df57a848ac5e83e30a474e6.pdf

240. Stanković, Ranka, Cvetana Krstev, Biljana Lazić and Dalibor Vorkapić. "A bilingual digital library for academic and entrepreneurial knowledge management". In *Proceeding of 10th International Forum on Knowledge Asset Dynamics — IFKAD 2015: Culture, Innovation and Entrepreneurship: connecting the knowledge dots*, *Bari, Italy, 10-12 June 2015*, eds. JC Spender, Giovanni Schiuma and Vito Albino, 1764-1777, preuzeto 04.02.2017, http://poincare.matf.bg.ac.rs/~cvetana/biblio/IFKAD_RS_CK_BL_DV-fin.pdf

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

Selected Papers, eds. A. Calì, D. Gorgan and M. Ugarte, LNCS 10151, 112-123. Springer, 2017

242. Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, Dániel Varga. "The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages". *CoRR* vol. abs/cs/0609058 (2006): n. pag, преузето 06.05.2016, https://arxiv.org/ftp/cs/papers/0609/0609058.pdf

243. Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. "BRAT: a Web-based Tool for NLP-Assisted Text Annotation". In *Proceedings of the Demonstrations at EACL-2012, Avignon, France*, 102-107. Association for Computational Linguistics, 2012

244. Sure, York and Rudi Studer. „Semantic Web technologies for digital libraries". *Library managemant* 26, 4/5(2005): 190-195, preuzeto 16.06.2016, https://eprints.soton.ac.uk/262614/2/OLD_Semantic_Web_Revisted.pdf

245. Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks". In *Advances in neural information processing systems*, 3104-3112. 2014

246. Svartvik, Jan, ed. *Trends in Linguistics. Studies and Monographs*, *Vol. 31*, *Directions in corpus linguistics: proceedings of Nobel Symposium.* Berlin: Mouton de Gruyter, 1991

247. Taylor, Charlotte. "What is corpus linguistics? What the data says". *ICAME Journal* No. 32(2008): 179-200, преузето 19.01.2016, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.184.3803&rep=rep1&type=pdf

248. TEI Consortium. „TEI P5: Guidelines for Electronic Text Encoding and Interchange". Version 3.0.0, 29.03.2016, preuzeto 01.06.2016, http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

249. TEI Guidelines, preuzeto 5. 4. 2016, www.tei-c.org/Guidelines/.

250. Text Encoding Initiative. TEI: History, preuzeto 25.2.2016, http://www.tei-c.org/About/history.xml

251. Teubert, Wolfgang. “Comparable or Parallel Corpora?”. *International Journal of Lexicografy* Vol. 9, No. 3 (1996): 238-264

252. Tharani, Karim. „Linked Data in Libraries: a Case Study of Harvesting and Sharing Bibliographic Metadata with BIBFRAME”. In *Information Technology and Libraries* 34, 1 (2015): 5-19, preuzeto 12.06.2016, http://ejournals.bc.edu/ojs/index.php/ital/article/view/5664/pdf

253. Tiedemann, Jörg. “Improved sentence alignment for movie subtitles”. In *Proceedings of RANLP, Borovets, Vol. 7*. 2007, preuzeto 21.07.2016, http://s3.amazonaws.com/tm-town-nlp-resources/ranlp07-subalign.pdf

254. “TMX 1.4b Specification. Open Standards for Container/Content Allowing Re-use (OSCAR) Recommendation”, 2005, preuzeto 24.10.2017, http://www.ttt.org/oscarstandards/tmx/tmx14b.html

255. Töny, Luzius. “Corpora als Ressourcen für die maschinelle Übersetzung”. Preuzeto 17.04.2016, http://www.swanrad.ch/downloads/mt_1.pdf

256. Trunk, Daniela. „Informationsseite zur GND”, преузето 2.2.2019, https://wiki.dnb.de/display/ILTIS/Informationsseite+zur+GND

257. Truong, Hong-Linh, and Schahram Dustdar. “On analyzing and specifying concerns for data as a service”. In *2009 IEEE Asia-Pacific Services Computing Conference (APSCC)*, pp. 87-94. IEEE, 2009

258. Tufiş, Dan, Svetla Koeva, Tomaž Erjavec, Maria Gavrilidou, Cvetana Krstev. “Building Language Resources and Translation Models for Machine Translation Focused on South Slavic and Balkan Languages”. In *Scientific results of the SEE-ERA.NET Pilot Joint Call* eds. Jana Machačová, Katarina Rohsmann, 37-48. Vienna: Centre for Social Innovation, 2009, preuzeto 15.03.2016, http://poincare.matf.bg.ac.rs/~cvetana/biblio/Tufis-KEGK-FASSBL2008.pdf

259. Tyers, Francis M., and Murat Serdar Alperen. “South-east European times: A parallel corpus of Balkan languages”. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, 49-53. 2010

260. Utvić, Miloš. „Konačni automati u regularnoj imenskoj derivaciji". Mag. teza, Matematički fakultet Univerziteta u Beogradu, 2008

261. Utvić, Miloš. *Izgradnja referentnog korpusa savremenog srpskog jezika: doktorska disertacija*. Dokt. disertacija, Filološki fakultet Univerziteta u Beogradu, 2013, preuzeto 14.2.2016, http://phaidrabg.bg.ac.rs/o:10061

262. Varga Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh and Viktor Trón. "Parallel corpora for medium density languages". In *Proceedings of RANLP'2005*, 590-596. Borovets: 2005

263. Vatant, Bernard. "Porting library vocabularies to the Semantic Web, and back: A win-win round trip". In *Proceedings of the IFLA World Library and Information Congress (IFLA'10), Gothenborg*. 2010, preuzeto 25.2.2019, https://www.ifla.org/past-wlic/2010/149-vatant-en.pdf

264. "VIAF. Connect authority data across cultures and languages to facilitate research". *OCLC,* preuzeto 2.2.2019, https://www.oclc.org/en/viaf.html

265. Vitas, Duško. „Prikaz jednog programskog sistema za automatsku obradu teksta". U *Zbornik II znanstvenega srečanja „Računalniška obdelava lingvističnih podatkov",* 457-465. Bled: Institut "Jožef Štefan", oktobar 1982 7-9. oktobar 1982

266. Vitas, Duško, Goran Nenadić and Cvetana Krstev [Electronic edition of Serbian translation of Plato's Republic aligned with 17 languages by Duško Vitas, Goran Nenadić, Cvetana Krstev]. "East meets West – A compendium of Multilingual Resources", eds. Tomaž Erjavec, Ann Lawson, Laurent Romary, TELRI Association. Mannheim: Institut für deutsche Sprache, 1998

267. Vitas, Duško and Cvetana Krstev. "A lexical approach to text alignment using Intex". In *Intex pour la linguistiqueet le traitement automatique des langues*, eds. Claude Muller, Jean Royaute, Max Silberztein, 249-263. Besancon: Presses Universitaires de Franche Comte, 2004, преузето 16.03.2016, http://poincare.matf.bg.ac.rs/~cvetana/biblio/krstev-vitas.pdf

268. Vitas, Duško. „Resursi i metode za obradu srpskog – stanje i perspektive". U *Srpska lingvistika / Serbische Linguistik*, Eine Bestandsaufnahme, Vol. 7 of Studies on

Language and Culture in Central and Eastern Europe (SLCCEE), eds. B. Golubović & C. Voß. 257–277. München: Verlag Otto Sagner, 2010

269. Vitas, Duško and Cvetana Krstev. “Derivational Morphology in an E-Dictionary of Serbian”. In *Proceedings of 2nd Language and Technology Conference, April 21-23, 2005, Poznań, Poland*, ed. Zygmunt Vetulani, 139-143. Poznań: Wydawnictwo Poznańskie Sp. z o.o., 2005

270. Vitas, Duško and Cvetana Krstev. “Literature and Aligned Texts”. In *Readings in Multilinguality* eds. Milena Slavcheva, Galia Angelova and Kiril Simov, 148-155. Sofia: Institute for Parallel Processing, Bulgarian Academy of Sciences, 2006, преузето 16.03.2016, http://poincare.matf.bg.ac.rs/~cvetana/biblio/CvDv-Paskaleva.pdf

271. Vitas, Duško and Cvetana Krstev. “Construction and Exploitation of X-Serbian Bitexts”. In *Multilingual Processing in Eastern and Southern EU Languages: Low-Resourced Technologies and Translation* eds. Cristina Vertan and Walther v. Hahn, 207-227. Cambridge: Cambridge Scholars Publishing, 2012a, преузето 13.03.2016, http://poincare.matf.bg.ac.rs/~cvetana/biblio/DvCv-CambridgeS-2012.pdf

272. Vitas, Duško and Cvetana Krstev. “Processing of Corpora of Serbian Using Electronic Dictionaries”. U *Prace Filologiczne*, Vol. LXIII. Warszawa, 2012b, 279-292. Преузето 13.04.2016, http://poincare.matf.bg.ac.rs/~cvetana/biblio/22_Vitas_Krstev.pdf

273. Vitas, Duško, Svetla Koeva, Cvetana Krstev and Ivan Obradović. „*Tour du monde* through the dictionaries”. In *Actes du 27eme Colloque International sur le Lexique et la Gammaire*, eds. M. Constant, T. Nakamura, M. De Gioia, S. Vecchiato, 249-256. Paris: Universite Paris-Est, Institut Gaspard-Monge, 2008. Преузето 13.03.2016, http://poincare.matf.bg.ac.rs/~cvetana/biblio/akvila-en-fin.pdf

274. Vitas, Duško, Ljubomir Popović, Cvetana Krstev, Anđelka Zečević. “How to Differentiate the Closely Related Standard Languages?”. In Proceedings of the Second International Conference*Computational Linguistics in Bulgaria (CLIB 2016)*, *September 9, 2016, Sofia, Bulgaria,* 559-574. Sofia: The Institute for Bulgarian Language *Prof. Lyubomir Andreychin*, Bulgarian Academy of Sciences, 2016

275. Volz, Raphael, Daniel Oberle, Steffen Staab and Boris Motik. „KAON SERVER - A Semantic Web Management System". In *Proceedings of the 12th World Wide Web, Alternate Tracks – Practiceand Experience, Hungary, Budapest*. 2003, preuzeto 11.08.2016, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.13.9286&rep=rep1&type=pdf

276. Vrandečić, Denny, and Markus Krötzsch. "Wikidata: a free collaborative knowledgebase". *Communications of the ACM* 57, no. 10 (2014): 78-85, преузето 15.2.2019, https://dl.acm.org/citation.cfm?id=2629489

277. Yimam, Seid Muhie, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann, „WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations". In *Proceedings of ACL-2013, demo session, Sofia, Bulgaria*, 1-6. Sofia: Association for Computational Linguistics, 2013

278. Zanettin, Federico. "Corpora in translation practice". In *Proceedings of the First International Workshop on Language Resources (LR) for Translation Work and Research*, 10-14. 2002

279. W3C Incubator Group. "Library Linked Data Incubator Group Final Report 25 October 2011", preuzeto 23.2.2019, https://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

280. "W3C Style: Cool URIs don't change". *W3C*, preuzeto 23.2.2019, https://www.w3.org/Provider/Style/URI

281. Wiebel, Stuart. "The Dublin Core: a simple content description model for electronic resources". *Bulletin of the American Society for Information Science* vol. 24, iss. 1(1997): 9–11, preuzeto 24. 2. 2016, http://onlinelibrary.wiley.com/doi/10.1002/bult.70/epdf

282. Workman, Michael, ed. *Semantic web: Implications for technologies and business practices*. Springer, 2016.

9 Додаци

9.1 Списак слика

Слика 1. Резултат претраге регуларним изразом „[Ll]xubav[a-z]*” у СрпФранКор-у 53
Слика 2 Резултат претраге регуларним изразом „amour” у СрпФранКор-у 53
Слика 3. Резултат претраге регуларним изразом „[Zz]dravlx[a-z]* radnika” у СрпЕнгКор-у 57
Слика 4. Резултат претраге регуларним изразом „occupational health” у СрпЕнгКор-у 58
Слика 5. Структура Библише 60
Слика 6. Окружење у Библиши за претрагу преко метаподатака 63
Слика 7 Резултати претраге преко кључне речи „biblioteka” 63
Слика 8. Окружење у Библиши за претрагу комплетног текста са резултатима за упит „biblioteka” . 65
Слика 9. Произведене конкорданце за упит „biblioteka” са опцијом “EN&SR” 66
Слика 10. Пример једног дела XML документа за дело Томаса Бернхарда „Моје награде” 90
Слика 11. Радни простор у систему COBISS за извоз метаподатака 119
Слика 12. Oxygen XML Editor – радно окружење 121
Слика 13. Вишеслојна „торта” семантичког веба 130
Слика 14. Пример RDF графа за ентитет „Томас Бернхард“ у бази Википодаци 133
Слика 15. Пример RDF записа у Turtle формату за ентитет „Thomas Bernhard” у бази GND 135
Слика 16. Први скупови података у LOD облаку 140
Слика 17. Облак „Отворени повезани подаци” 141
Слика 18. Облак „Отворени повезани подаци из области лингвистике” 142
Слика 19. Прва део записа „Томас Бернхард” у Википодацима 146
Слика 20. Примери изјава у Википодацима за Q44336 „Томас Бернхард” 147
Слика 21. Примери из листе референтних веб страна за ентитет „Томас Бернхард” 149
Слика 22. Упит у бази Википодаци „Дела немачких писаца рођених 1750-1990 која су преведена” са резултатима исписа на временској скали 150
Слика 23. Процес повезивања података у систему „Отворени повезани подаци” 151
Слика 24. Процес објављивања скупа повезаних података 153
Слика 25. Структура BIBFRAME модела 170
Слика 26. Сумећа алата за креирање записа у BIBFRAME модел 171
Слика 27. Сумећа алата за конверзију записа у BIBFRAME на основу идентификатора записа 171
Слика 28. Транскрибус - датотека за рад 180
Слика 29. Транскрибус – радно окружење за оптичко препознавање карактера (подешавање параметара) 181
Слика 30. Транскрибус - радно окружење за прегледање и корекцију текста после оптичког препознавања карактера 181
Слика 31. Транскрибус - радно окружење за анотацију структурних целина текста и форматирање текста 183

Слика 32. Транскрибус – радно окружење за избор излазног формата 184
Слика 33. Hunspell окружење за контролу и корекцију текстова на немачком језику 185
Слика 34. Едитор Notepad++: Томас Бернхард „Моје награде” 188
Слика 35. Мени Alignment – део програма ACIDE за аутоматско упаривање сегмената 190
Слика 36. Concordancier – део програма ACIDE који приказује упарене сегменте у виду паралелних конкорданци 191
Слика 37. Садржај три излазне XML датотеке, идентификатори сегмената из прве две се користе у трећој 191
Слика 38. Concordancier - сегмент изворног језика (немачки) су две реченице, а одговарајући сегмент циљног језика (српски) је једна реченица 193
Слика 39. Concordancier - сегмент изворног језика (немачки) је једна реченица, а одговарајући сегмент циљног језика (српски) је део реченице 193
Слика 40. Concordancier - сегмент изворног (немачки) и еквивалентан сегмент циљног језика (српски) се састоје од две или више реченица које нису у истом редоследу 194
Слика 41. Concordancier - сегмент изворног језика (немачки) је једна реченица, а одговарајући сегмент циљног језика (српски) не постоји 194
Слика 42. Пример упарених сегмената у роману „Излет у небо“ Гроздане Олујић 196
Слика 43. Пример једног упареног сегмента у роману „Излет у небо“ Гроздане Олујић 196
Слика 44. Мени TMX – део програма ACIDE за генерисање TMX документа 197
Слика 45. TMX формат паралелизованог текста 199
Слика 46. Табела интегрисаних јединица паралелизованог текста 200
Слика 47. Текстуални запис у такозваном “Vanila” формату 200
Слика 48. XML формат интегрисаних података 201
Слика 49. Генерисани HTML формат погодан за приказивање на вебу 201
Слика 50. Разлагање TMX документа - Split TMX 202
Слика 51. Резултат разлагања TMX документа – немачки део паралелног текста 202
Слика 52. Резултат разлагања TMX документа – српски део паралелног корпуса 203
Слика 53. Вертикализација текста 203
Слика 54. Број преводних парова у корпусу СрпНемКор према романима појединачно 204
Слика 55. Запис у Библиши за роман „Моје награде” аутора Томаса Бернхарда 207
Слика 56. Листа романа у потколекцији „Романи оригинално написани на немачком” у корпусу СрпНемКор 209
Слика 57. Приказ метаподатака у Библиши за роман „Моје награде / Meine Preise” Томаса Бернхарда са облацима текста 210
Слика 58. Роман „Моје награде / Meine Preise” Томаса Бернхарда у формату TMX 211
Слика 59. Лексичка јединица „мајка” у бази Терми у формату TBX са синонимима 216
Слика 60. Окружење у Библиши за претрагу преко метаподатака и резултати исписа за упит “Language: SR AND collection: SrpNemKor AND title: prozor” 217
Слика 61. Резултати семантичког проширења упита позивањем српског и енглеског Ворднета 218
Слика 62. Пример неких означених именованих ентитета XML ознакама у српском тексту романа „Парфем” 224

Слика 63. Визуализација именованих ентитета применом алата за анотацију и визуализацију WebAnno 226
Слика 64. Визуализација именованих ентитета применом веб алата BRAT 227
Слика 65. Stanford NER веб алат за анотацију именованих ентитета 229
Слика 66. Резултати анотације именованих ентитета применом Stanford NER веб алата на примеру текста „Парфем" на немачком језику 230
Слика 67. Визуализација именованих ентитета из табеле 13 применом веб алата BRAT 231
Слика 68. Библиша - везе ка базама VIAF, GND, Википодаци (Wikidata) и LCNAF за роман „Moje награде" и писца „Томаса Бернхарда" 239
Слика 69. Пример лексичких јединица „grenzenlos/unendlich" и „beskrajan/bezgraničan" у LMF/XML 243
Слика 70. Графички приказ преводних парова "grenzenlos"="beskrajan", "grenzenlos"="bezgraničan", "unendlich"="beskrajan" 247

9.2 Списак табела

Табела 1. Број нових речи у речницима DELAS опште лексеме и DELAS-PROP властитих имена према романима 186
Табела 2. Број речи у корпусу СрпНемКор према романима појединачно 205
Табела 3. Неки примери из немачко-српске листе преводних парова произведени на основу текстуалне колекције СрпНемКор 214
Табела 4. Примери синонима на немачком и српском језику 215
Табела 5. Примери генерисаних конкорданци поравнатих сегмената за упит "Language: SR and collection: SrpNemKor and keyword: majka" уз одабрану могућност приказа "DE&SR" 219
Табела 6. Генерисане конкорданце поравнатих сегмената за упит "Language: SR and collection: SrpNemKor and keyword: majka" уз одабрану могућност приказа "DE" 220
Табела 7. Генерисане конкорданце поравнатих сегмената за упит "Language: SR and collection: SrpNemKor and keyword: majka" уз одабрану могућност приказа "SR" 221
Табела 8. Примери генерисаних конкорданци поравнатих сегмената за упит "Language: SR and collection: SrpNemKor and keyword: bračni par" уз одабрану могућност приказа "DE&SR" 222
Табела 9. Генерисане конкорданце поравнатих сегмената за упит "Language: SR and collection: SrpNemKor and keyword: bračni par" уз одабрану могућност приказа "SR" 223
Табела 10. Листа ознака за именоване ентитете у српском језику 224
Табела 11. Примери именованих ентитета у IOB формату 226
Табела 12. Примери именовани ентитети у standoff формату 227
Табела 13. Примери именованих ентитета из текста "Parfum" у standoff формату 231
Табела 14. Примери анотације именованих ентитета применом модела DE.lft на роман „Parfum" 232
Табела 15. Примери анотације именованих ентитета применом модела DE.onb на роман „Parfum" 233

9.3 Списак скраћеница

9.3.1 Скраћенице на ћирилици

СрпКор - Корпус савременог српског језика
СрпФранКор – Српско-француски корпус
СрпЕнгКор – Српско-енглески корпус
УДК - Универзална децимална класификација
ДДК - Дјуијева децимална класификација

9.3.2 Скраћенице на латиници

AACR2 - Anglo-American Cataloguing Rules
ACIDE - Aligned Corpora Integrated Development Environment
ACH - Association of Computer in the Humanities
ACL - Association for Computational Linguistics
Acquis - Acquis Communautaire
ALLC - Association of Literary and Linguistic Computing
API - Application Programming Interface
ASCII - American Standard Code for Information Interchange
BBC - British Broadcasting Corporation,
BIBFRAME - Bibliographic Framework
BIBO - Bibliographic Ontology
BNC - British National Corpus
CDWA - Categories for the Description of Works of Art
CEE језици - Central and Eastern European Languages
CES - Corpus Encoding Standard
CIDOC CRM – International Committee for Documentation, Conceptual Reference Model
COBISS - Cooperative Online Bibliographic System and Services
CQP - Corpus Query Processor
DC - Dublin Core
DCAM - DCMI Abstract Model
DCMES - Dublin Core Metadata Element Set
DCMI - Dublin Core Metadata Initiative
DIEF - DBpedia Information Extraction Framework
DLF - DigitalLibraryFederation
DOI - Digital Object Indentifier
DPLA - Digital Public Library of America
DPLA MAP - DPLA Metadata Application Profile
DTD - Document Type Definition
EADH - European Association for Digital Humanities

EAGLES - Expert Advisory Group on Language Engineering Standards
EAD - Encoded Archival Description
ESE - Europeana Semantic Elements
EDM - Europeana Data Model
FOAF - Friend of Friend
FRBR - Functional Requirements for Bibliographic Records
GeolISS - Geologic Information System of Serbia
GND - Gemeinsame Normdatei
GNU GPL - GNU General Public License
HTML - Hypertext Markup Language
HTTP - Hyper Text Transfer Protocol
IMS OCWB - Institut für Maschinelle Sprachverarbeitung Open Corpus Workbench
Intera - Integrated European Language data Repository Area
IRI - International Resource Identification
ISBD - International Standard Bibliographic Description
ISBN - International Standard Book Number
ISO – International Organization for Standardization
ISSN - International Standard Serial Number
IZUM – Institut informacijskih znanosti Maribor
JRC-Acquis - JRC Collection of the Acquis Communautaire
KOS - Knowledge Organisation System
LADL - Laboratoire d'Automatique Documentaire et Linguistique
LCNAF - Library of Congress Name Authority File
LCSH - Library of Congress Subject Headings
LGPL - Lesser General Public License
LGPLLR - Lesser General Public License For Linguistic Resources
LIDO - Lightweight Information Describing Objects,
LISA - Localisation Industry Standards Association
LOD – Linked Open Data
LORIA - Laboratoire Lorrain de RechercheenInformatiqueetses Applications
LRE - Language Resources and Evaluation
LSC - Longman Spoken Corpus
MARC - Machine Readable Catalogue
MADS - Metadata Authority Description Standard
METS - Metadata Encoding and Transmission Standard
MoA II - Making of America II
MODS - Metadata Object Description Schema
MSD - Morphosyntactic Description
MUC - Message Understanding Conference
Multext-East - Multilingual Text Tools and Corpora for Eastern and Central European Languages
NCSA - National Center for Supercomputing Application
NISO - National Information Standards Organization
OAI-ORE - Open Archives Initiative - Object Reuse and Exchange
OAI-PMH - Open Archives Initiative – Protocol for Metadata Harvesting

OCLC - Online Computer Library Center
OCR - Optical Character Recognition
OIE - Open Information Extraction
OMR - Open Metadata Registry
OPAC - Online Public Access Catalog
OWL - Ontology Web Language
PoS - Part of Speech
POSIX - Portable Operating System Interface
PURL - Persistent Uniform Resource Locator
RAMEAU – Répertoire d'autorité matière encyclopédique et alphabétique unifié
RELAX NG - REgular LAnguage for XML Next Generation
RDA - Resource Description and Access
RDF - Resource Description Framework
SELFEH - Serbian-English Law Finance Education and Health
SETimes - South-East European Times
SGML - Standard Generalized Markup Language
SKOS - Simple Knowledge Organization System
SPARQL - Simple Protocol and RDF Query Language
TEI - Text Encoding Initiative
TELRI - Trans-European Language Resources Infrastructure
TMX - Translation Memory eXchange
TU - Translation Unit
TUV - Translation Unit Variant
UML - Universal Modelling Language
UNIMARC - Universal Machine Readable Catalogue
URI - Unifrom Resource Identifier
URL - Uniform Resource Locator
VIAF - Virtual International Authority File
WS4LR - Work Station for Language Resources
WWW - World Wide Web
W3C - World Wide Web Consortium
XML - eXtensible Markup Language
XSL - Extensible Stylesheet Language
YAGO - Yet Another Great Ontology

Прилози

Прилог 1 - Пример записа у формату COMARC/B / према стандарду ISO2709

ID=193724940 LN=0000305180 **M V4** 28.09.2012 NBS::ANAS **Updated:** 09.02.2017 UBSM::JELENAA **Copied:** 23.07.2013 UBSM::JEJA **First Copied:** 23.07.2013 **COBISS3:** 09.02.2017 UBSM::JELENAA

001	**a**c - ispravljeni zapis **b**a - tekstualna građa, štampana **c**m - monografska publikacija **d**0 - nema hierarhijskog odnosa **7**ba - latinica
010	**a**978-86-7958-064-1 **b**broš.
100	**c**2012 **h**srp - srpski **l**ba - latinica
101 1	**a**srp - srpski **c**ger - nemački
102	**a**srb - Srbija **b**cs - Centralna Srbija
200 0	**a**Moje nagrade **f**Tomas Bernhard **g**prevela s nemačkog Bojana Denić
210	**a**Beograd **c**LOM **d**2012 **e**Beograd **g**Caligraph
215	**a**115 str. **d**20 cm
225 1	**a**‡Edicija ‡Do-des-ka-den **v**‡knj. ‡5
300 0	**a**Prevod dela: Meine Preise / Thomas Bernhard
300 0	**a**Tiraž 700
300 0	**a**Str. 115-116: Pogovor / Periša Perišić.
500 0	**a**Meine Preise **m**srp
675	**a**821.112.2(436)-3 **b**821.112.2(436) **c**821.112.2(436) - Austrijska književnost **s**821.112.2(.)
700	**a**Bernhard **b**Thomas **f**1931-1989 **4**070 - autor
702 0	**a**Denić **b**Bojana **4**730 - prevodilac
702 0	**a**Peršić **b**Periša **4**080 - autor dodatnog teksta
900	**a**Бернхард **b**Томас **f**1931-1989
992	**b**a1307jđ
996	**d**lАБ**f1u**82**a**Bern **f528004457o**20130723 **q**2 - u obradi **t**20130723 **v**c - poklon **2**O1 **3**RSD 550,00 **3**EUR 5,00 **6**305383

Прилог 2 - Пример записа у формату MARC21 / према стандарду ISO2709

LDR 00850cam a2200253 i 4500
001 193724940
003 RS-BgCOB
005 20130723000000.0
008 120928s2012 rb |||||||||||||| ||srp c
020__$a9788679580641$qbroš.
040__$aNBS$bsrp$cSI-MaIIZ$dUBSM$eppiak
041_1$asrp$hger
080__$a821.112.2$x(436)
100_1$aBernhard, Thomas, $d1931-1989. $4aut
24010$aMeine Preise.$lsrp
24500$aMoje nagrade / $cTomas Bernhard ; prevela s nemačkog Bojana Denić.
260__$aBeograd : $bLOM, $c2012$e(Beograd : $fCaligraph)
300__$a115 str. ; $c20 cm.
490_0$aEdicija Do-des-ka-den ;$vknj. 5
500__$aPrevod dela: Meine Preise / Thomas Bernhard.
500__$aTiraž 700.

500__$aStr. 115-116: Pogovor / Periša Perišić.
700_1$aDenić, Bojana. $4trl
700_1$aPeršić, Periša. $4aui

Прилог 3 - Пример записа у формату Даблинско језгро / XML синтакса

```
<dc:collection xmlns:dc=http://purl.org/dc/elements/1.1/ xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance xmlns:dcterms=http://purl.org/dc/terms/>
    <dc:dc>
        <dc:title>Moje nagrade / Tomas Bernhard ; prevela s nemačkog Bojana Denić. </dc:title>
            <dcterms:alternative xml:lang="de">Meine Preise. </dcterms:alternative>
            <dc:creator>Bernhard, Thomas, 1931-1989. (aut)</dc:creator>
            <dc:contributor>Denić, Bojana. (trl)</dc:contributor>
            <dc:contributor>Peršić, Periša. (aui)</dc:contributor>
            <dc:language xsi:type="dcterms:ISO639-2">srp</dc:language>
            <dc:type>monograph</dc:type>
            <dc:type>text</dc:type>
            <dc:type xsi:type="dcterms:DCMIType">Text</dc:type>
            <dc:publisher>Beograd :  LOM,</dc:publisher>
            <dcterms:issued>2012</dcterms:issued>
            <dc:date xsi:type="dcterms:W3CDTF">2012</dc:date>
            <dc:relation>Series: Edicija Do-des-ka-den;knj. 5</dc:relation>
            <dc:format>115 str. ;  20 cm. </dc:format>
            <dc:description>Prevod dela: Meine Preise / Thomas Bernhard. </dc:description>
            <dc:description>Tiraž 700. </dc:description>
            <dc:description>Str. 115-116: Pogovor / Periša Perišić.</dc:description>
            <dc:subject xsi:type="dcterms:UDC">821.112.2(436)</dc:subject>
            <dc:identifier xsi:type="dcterms:URI">urn:ISBN:978-86-7958-064-1</dc:identifier>
            <dc:identifier>193724940</dc:identifier>
    </dc:dc>
</dc:collection>
```

Прилог 4 - Пример записа у формату METS / XML синтакса

```
<?xml version="1.0" encoding="UTF-8"?>
<mets:mets xmlns:mets="http://www.loc.gov/METS/" xmlns:mods="http://www.loc.gov/mods/v3"
xmlns:rts="http://cosimo.stanford.edu/sdr/metsrights/" xmlns:mix="http://www.loc.gov/mix/v10"
xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/mets.xsd
http://cosimo.stanford.edu/sdr/metsrights/ http://cosimo.stanford.edu/sdr/metsrights.xsd
http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-2.xsd http://www.loc.gov/mix/v10
http://www.loc.gov/standards/mix/mix10/mix10.xsd" OBJID="ark:/13030/hb4199p148" LABEL="Extending the
Lexicon by Exploiting Subregularities" PROFILE="http://www.loc.gov/mets/profiles/00000013.xml">
  <mets:metsHdr CREATEDATE="2017-10-13T11:18:40" ID="METS193724940">
    <mets:agent ROLE="CREATOR" TYPE="INDIVIDUAL">
      <mets:name>Jelena Andonovski</mets:name>
```

```
    </mets:agent>
    <mets:agent ROLE="CREATOR" TYPE="INDIVIDUAL">
      <mets:name>Cvetana Krstev</mets:name>
    </mets:agent>
  </mets:metsHdr>
  <mets:dmdSec ID="DMR1">
    <mets:mdRef
xlink:href="http://www.vbs.rs/scripts/cobiss?command=DISPLAY&base=70036&rid=193724940&fmt=11&lani=sc"
LOCTYPE="URL" MDTYPE="COMARC" LABEL="Catalog Record"/>
  </mets:dmdSec>
  <mets:dmdSec ID="DM1">
    <mets:mdWrap MDTYPE="DC">
      <dc:dc>
        <dc:title>Moje nagrade / Tomas Bernhard ; prevela s nemačkog Bojana Denić. </dc:title>
        <dcterms:alternative xml:lang="de">Meine Preise. </dcterms:alternative>
        <dc:creator>Bernhard, Thomas, 1931-1989. (aut)</dc:creator>
        <dc:contributor>Denić, Bojana. (trl)</dc:contributor>
        <dc:contributor>Peršić, Periša. (aui)</dc:contributor>
        <dc:language xsi:type="dcterms:ISO639-2">srp</dc:language>
        <dc:type>monograph</dc:type>
        <dc:type>text</dc:type>
        <dc:type xsi:type="dcterms:DCMIType">Text</dc:type>
        <dc:publisher>Beograd :  LOM,</dc:publisher>
        <dcterms:issued>2012</dcterms:issued>
        <dc:date xsi:type="dcterms:W3CDTF">2012</dc:date>
        <dc:relation>Series: Edicija Do-des-ka-den;knj. 5</dc:relation>
        <dc:format>115 str. ;  20 cm. </dc:format>
        <dc:description>Prevod dela: Meine Preise / Thomas Bernhard. </dc:description>
        <dc:description>Tiraž 700. </dc:description>
        <dc:description>Str. 115-116: Pogovor / Periša Perišić.</dc:description>
        <dc:subject xsi:type="dcterms:UDC">821.112.2(436)</dc:subject>
        <dc:identifier xsi:type="dcterms:URI">urn:ISBN:978-86-7958-064-1</dc:identifier>
        <dc:identifier>193724940</dc:identifier>
      </dc:dc>
    </mets:mdWrap>
  </mets:dmdSec>
  <mets:amdSec>
    <mets:techMD ID="ADM1">
      <mets:mdWrap MDTYPE="OTHER">
        <mets:xmlData>
          <mix:mix>
            <mix:BasicDigitalObjectInformation>
              <mix:FormatDesignation>
                <mix:formatName>TEXT/PDF</mix:formatName>
              </mix:FormatDesignation>
            </mix:BasicDigitalObjectInformation>
```

```
            </mix:mix>
          </mets:xmlData>
        </mets:mdWrap>
      </mets:techMD>
    </mets:amdSec>
  <METS:fileSec>
    <mets:fileGrp USE="application">
      <METS:fileGrp ID="GID1" USE="use_tmx">
        <METS:file ID="FID1" MIMETYPE="application/tmx" ADMID="ADM1">
          <METS:FLocat LOCTYPE="URL" xlink:href="http://jerteh.rs/biblisha/Default.aspx"/>
        </METS:file>
      </METS:fileGrp>
    </mets:fileGrp>
  </METS:fileSec>
  <METS:structMap TYPE="physical">
    <METS:div LABEL="Moje nagrade" DMDID="DM1" ADMID="ADM1" ADMID="RMD1" TYPE="pdf_file"
ID="div1">
      <METS:fptr FILEID="FID1"/>
    </METS:div>
  </METS:structMap>
</mets:mets>
```

Прилог 5 - Пример записа у формату MODS / XML синтакса

```
<modsCollection xmlns="http://www.loc.gov/mods/v3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xsi:schemaLocation="http://www.loc.gov/mods/v3 http://www.loc.gov/standards/mods/v3/mods-3-
3.xsd">
    <mods version="3.3">
        <titleInfo><title>Moje nagrade</title></titleInfo>
        <titleInfo type="uniform"><title>Meine Preise.  srp</title></titleInfo>
        <name type="personal">
            <namePart>Bernhard, Thomas</namePart>
            <namePart type="date">1931-1989</namePart>
            <role>
                <roleTerm authority="marcrelator" type="text">creator</roleTerm>
            </role>
            <role>
                <roleTerm authority="marcrelator" type="code">aut</roleTerm>
            </role>
        </name>
        <name type="personal">
            <namePart>Denić, Bojana.</namePart>
            <role>
                <roleTerm authority="marcrelator" type="code">trl</roleTerm>
            </role>
        </name>
        <name type="personal">
```

```
            <namePart>Perišić, Periša.</namePart>
            <role>
                <roleTerm authority="marcrelator" type="code">aui</roleTerm>
            </role>
        </name>
        <typeOfResource>text</typeOfResource>
        <originInfo>
            <place>
                <placeTerm authority="marccountry" type="code">rb</placeTerm>
            </place>
            <place>
                <placeTerm type="text">Beograd</placeTerm>
            </place>
            <publisher>LOM</publisher>
            <dateIssued>2012</dateIssued>
            <issuance>monographic</issuance>
        </originInfo>
        <language>
                <languageTerm authority="iso639-2b" type="code">srp</languageTerm>
        </language>
        <language objectPart="translation">
                <languageTerm authority="iso639-2b" type="code">ger</languageTerm>
      </language>
        <physicalDescription><extent>115 str. ;  20 cm. </extent></physicalDescription>
        <note type="statement of responsibility">Tomas Bernhard ; prevela s nemačkog Bojana Denić. </note>
        <note>Prevod dela: Meine Preise / Thomas Bernhard. </note>
        <note>Tiraž 700. </note>
        <note>Str. 115-116: Pogovor / Periša Perišić.</note>
        <relatedItem type="series">
                <titleInfo><title>Edicija Do-des-ka-den ; knj. 5</title></titleInfo>
        </relatedItem>
        <identifier type="isbn">978-86-7958-064-1</identifier>
        <recordInfo>
                <descriptionStandard>ppiak</descriptionStandard>
                <recordContentSource authority="marcorg">NBS</recordContentSource>
                <recordCreationDate encoding="marc">120928</recordCreationDate>
                <recordChangeDate encoding="iso8601">20130723000000.0</recordChangeDate>
                <recordIdentifier source="RS-BgCOB">193724940</recordIdentifier>
                <recordOrigin>Converted from MARCXML to MODS version 3.3 using MARC21slim2MODS3-3.xsl
                (Revision 1.50)</recordOrigin>
                <languageOfCataloging>
                        <languageTerm authority="iso639-2b" type="code">srp</languageTerm>
                </languageOfCataloging>
        </recordInfo>
    </mods>
</modsCollection>
```

Прилог 6 - Пример записа у формату MARC21 / XML синтакса

```
<?xml version="1.0" encoding="UTF-8"?>
<marc:collection xsi:schemaLocation="http://www.loc.gov/MARC21/slim
```

```
http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:fn="http://www.w3.org/2005/xpath-functions"
xmlns:xs="http://www.w3.org/2001/XMLSchema" xmlns:marc="http://www.loc.gov/MARC21/slim">
  <marc:record>
    <marc:leader>00850cam a2200253 i 4500</marc:leader>
    <marc:controlfield tag="001">193724940</marc:controlfield>
    <marc:controlfield tag="003">RS-BgCOB</marc:controlfield>
    <marc:controlfield tag="005">20130723000000.0</marc:controlfield>
    <marc:controlfield tag="008">120928s2012 rb |||||||||||||| ||srp c</marc:controlfield>
    <marc:datafield tag="020" ind2=" " ind1=" ">
      <marc:subfield code="a">9788679580641</marc:subfield>
      <marc:subfield code="q">broš.</marc:subfield>
    </marc:datafield>
    <marc:datafield tag="040" ind2=" " ind1=" ">
      <marc:subfield code="a">NBS</marc:subfield>
      <marc:subfield code="b">srp</marc:subfield>
      <marc:subfield code="c">SI-MaIIZ</marc:subfield>
      <marc:subfield code="d">UBSM</marc:subfield>
      <marc:subfield code="e">ppiak</marc:subfield>
    </marc:datafield>-<marc:datafield tag="041" ind2=" " ind1="1">
      <marc:subfield code="a">srp</marc:subfield>
      <marc:subfield code="h">ger</marc:subfield>
    </marc:datafield>
    <marc:datafield tag="080" ind2=" " ind1=" ">
      <marc:subfield code="a">821.112.2</marc:subfield>
      <marc:subfield code="x">(436)</marc:subfield>
    </marc:datafield>
    <marc:datafield tag="100" ind2=" " ind1="1">
      <marc:subfield code="a">Bernhard, Thomas, </marc:subfield>
      <marc:subfield code="d">1931-1989. </marc:subfield>
      <marc:subfield code="4">aut</marc:subfield>
    </marc:datafield>
    <marc:datafield tag="240" ind2="0" ind1="1">
      <marc:subfield code="a">Meine Preise. </marc:subfield>
      <marc:subfield code="l">srp</marc:subfield>
    </marc:datafield>
    <marc:datafield tag="245" ind2="0" ind1="0">
      <marc:subfield code="a">Moje nagrade / </marc:subfield>
      <marc:subfield code="c">Tomas Bernhard ; prevela s nemačkog Bojana Denić. </marc:subfield>
    </marc:datafield>
    <marc:datafield tag="260" ind2=" " ind1=" ">
      <marc:subfield code="a">Beograd : </marc:subfield>
      <marc:subfield code="b">LOM, </marc:subfield>
      <marc:subfield code="c">2012</marc:subfield>
      <marc:subfield code="e">(Beograd : </marc:subfield>
      <marc:subfield code="f">Caligraph) </marc:subfield>
```

```
    </marc:datafield>
    <marc:datafield tag="300" ind2=" " ind1=" ">
      <marc:subfield code="a">115 str. ; </marc:subfield>
      <marc:subfield code="c">20 cm. </marc:subfield>
    </marc:datafield>
    <marc:datafield tag="490" ind2=" " ind1="0">
      <marc:subfield code="a">Edicija Do-des-ka-den ;</marc:subfield>
      <marc:subfield code="v">knj. 5</marc:subfield>
    </marc:datafield>
    <marc:datafield tag="500" ind2=" " ind1=" ">
      <marc:subfield code="a">Prevod dela: Meine Preise / Thomas Bernhard. </marc:subfield>
    </marc:datafield>
    <marc:datafield tag="500" ind2=" " ind1=" ">
      <marc:subfield code="a">Tiraž 700. </marc:subfield>
    </marc:datafield>
    <marc:datafield tag="500" ind2=" " ind1=" ">
      <marc:subfield code="a">Str. 115-116: Pogovor / Periša Perišić.</marc:subfield>
    </marc:datafield>
    <marc:datafield tag="700" ind2=" " ind1="1">
      <marc:subfield code="a">Denić, Bojana. </marc:subfield>
      <marc:subfield code="4">trl</marc:subfield>
    </marc:datafield>
    <marc:datafield tag="700" ind2=" " ind1="1">
      <marc:subfield code="a">Peršić, Periša. </marc:subfield>
      <marc:subfield code="4">aui</marc:subfield>
    </marc:datafield>
  </marc:record>
</marc:collection>
```

Прилог 7 - Пример записа у формату COMARC / XML синтакса

```
<?xml version="1.0" encoding="UTF-8"?>
<collection>
  <record>
    <datafield ind2=" " ind1=" " tag="000">
      <subfield code="a">0013</subfield>
      <subfield code="b">2012092820130723</subfield>
      <subfield code="c">NBS::ANAS</subfield>
      <subfield code="d">UBSM::JEJA</subfield><subfield code="e"/>
      <subfield code="f">20130723</subfield>
      <subfield code="g">0000305180</subfield>
      <subfield code="h">UBSM::JEJA</subfield>
      <subfield code="i">20130723</subfield>
      <subfield code="p">UBSM::JEJA</subfield>
```

<subfield code="o">20130723</subfield>
<subfield code="x">193724940</subfield>
</datafield>
<datafield ind2=" " ind1=" " tag="001">
<subfield code="a">c</subfield>
<subfield code="b">a</subfield>
<subfield code="c">m</subfield>
<subfield code="d">0</subfield>
<subfield code="7">ba</subfield>
</datafield>
<datafield ind2=" " ind1=" " tag="010">
<subfield code="a">978-86-7958-064-1</subfield>
<subfield code="b">broš.</subfield>
</datafield>
<datafield ind2=" " ind1=" " tag="100">
<subfield code="c">2012</subfield>
<subfield code="h">srp</subfield>
<subfield code="l">ba</subfield>
</datafield>
<datafield ind2=" " ind1="1" tag="101">
<subfield code="a">srp</subfield>
<subfield code="c">ger</subfield>
</datafield>
<datafield ind2=" " ind1=" " tag="102">
<subfield code="a">srb</subfield>
<subfield code="b">cs</subfield>
</datafield>
<datafield ind2=" " ind1="0" tag="200">
<subfield code="a">Moje nagrade</subfield>
<subfield code="f">Tomas Bernhard</subfield>
<subfield code="g">prevela s nemačkog Bojana Denić</subfield>
</datafield>
<datafield ind2=" " ind1=" " tag="210">
<subfield code="a">Beograd</subfield>
<subfield code="c">LOM</subfield>
<subfield code="d">2012</subfield>
<subfield code="e">Beograd</subfield>
<subfield code="g">Caligraph</subfield>
</datafield>
<datafield ind2=" " ind1=" " tag="215">
<subfield code="a">115 str.</subfield>
<subfield code="d">20 cm</subfield>
</datafield>
<datafield ind2=" " ind1="1" tag="225">
<subfield code="a">Edicija Do-des-ka-den</subfield>
<subfield code="v">knj. 5</subfield>

```
</datafield>
<datafield ind2=" " ind1="0" tag="300">
  <subfield code="a">Prevod dela: Meine Preise / Thomas Bernhard</subfield>
</datafield>
<datafield ind2=" " ind1="0" tag="300">
  <subfield code="a">Tiraž 700</subfield>
</datafield>
<datafield ind2=" " ind1="0" tag="300">
  <subfield code="a">Str. 115-116: Pogovor / Periša Perišić.</subfield>
</datafield>
<datafield ind2="0" ind1="0" tag="500">
  <subfield code="a">Meine Preise</subfield>
  <subfield code="m">srp</subfield>
</datafield>
<datafield ind2=" " ind1=" " tag="675">
  <subfield code="a">821.112.2(436)-3</subfield>
  <subfield code="b">821.112.2(436)</subfield>
  <subfield code="c">821.112.2(436)</subfield>
  <subfield code="s">821.112.2(.)</subfield>
</datafield>
<datafield ind2="1" ind1=" " tag="700">
  <subfield code="a">Bernhard</subfield>
  <subfield code="b">Thomas</subfield>
  <subfield code="f">1931-1989</subfield>
  <subfield code="4">070</subfield>
</datafield>
<datafield ind2="1" ind1="0" tag="702">
  <subfield code="a">Denić</subfield>
  <subfield code="b">Bojana</subfield>
  <subfield code="4">730</subfield>
</datafield>
<datafield ind2="1" ind1="0" tag="702">
  <subfield code="a">Peršić</subfield>
  <subfield code="b">Periša</subfield>
  <subfield code="4">080</subfield>
</datafield>
<datafield ind2="4" ind1=" " tag="900">
  <subfield code="a">Бернхард</subfield>
  <subfield code="b">Томас</subfield>
  <subfield code="f">1931-1989</subfield>
</datafield>
<datafield ind2=" " ind1=" " tag="978">
  <subfield code="x">SR</subfield>
</datafield>
<datafield ind2=" " ind1=" " tag="992">
  <subfield code="b">a1307jđ</subfield>
```

```
    </datafield>
    <datafield ind2="6" ind1=" " tag="996">
      <subfield code="d">IAБ\f1\u82\aBern</subfield>
      <subfield code="f">528004457</subfield>
      <subfield code="o">20130723</subfield>
      <subfield code="q">2</subfield>
      <subfield code="t">20130723</subfield>
      <subfield code="v">c</subfield>
      <subfield code="2">O1</subfield>
      <subfield code="3">RSD 550,00</subfield>
      <subfield code="3">EUR 5,00</subfield>
      <subfield code="6">305383</subfield>
    </datafield>
  </record>
</collection>
```

Прилог 8 – Пример записа у формату TEI заглавље / XML синтакса

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_lite.rng"
schematypens="http://relaxng.org/ns/structure/1.0"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader type="text">
    <fileDesc>
      <titleStmt>
        <title type="main">Moje nagrade</title>
        <author>Jelena Andonovski</author>
        <editor>Cvetana Krstev</editor>
      </titleStmt>
      <editionStmt>
        <edition>Elektronsko izdanje</edition>
      </editionStmt>
      <extent>
        <seg type="format">text/XML</seg>
      </extent>
      <publicationStmt>
        <pubPlace>Beograd, Srbija</pubPlace>
        <publisher>Univerzitetska biblioteka "Svetozar Marković"</publisher>
        <publisher>Grupa za jezičke tehnologije</publisher>
        <date>2017</date>
      </publicationStmt>
      <seriesStmt>
        <title>Paralelni srpsko-nemački korpus književnih tekstova - SrpNemKor</title>
      </seriesStmt>
      <notesStmt>
        <note>Tekst koji se opisuje je skaniran u Unicerzitetskoj biblioteci u Beogradu. Prilikom digitalizacije
urađen je i OCR teksta. Tekst je sastavni deo paralelnog srpsko-nemačkog korpusa koji je napravljen prilikom rada
```

```
na doktorskoj disertaciji “Mreža otvorenih podataka i jezički resursi u procesu izgradnje srpsko-nemačkog literarnog korpusa”</note>
            <note>Delo Moje nagrade predstavlja srpski prevod dela Meine Preise austrijskog pisca Thomasa Bernharda.</note>
        </notesStmt>
        <sourceDesc>
          <bibl type=“monogr”>
            <title xml:lang=“srp” ref=“viaf:239728568”>Moje nagrade</title>
            <title xml:lang=“ger” ref=“viaf:239728568”>Meine Preise</title>
            <author ref=“viaf:12305044”>Tomas Bernhard, 1931-1989 </author>
            <editor role=“prevodilac”>Bojana Denić</editor>
            <editor role=“autor_dodatnog_teksta”>Periša Perišić</editor>
            <pubPlace>Beograd, Srbija</pubPlace>
            <publisher>LOM</publisher>
            <date>2012</date>
            <distributor>Caligraph</distributor>
            <extent>115 str.</extent>
            <extent>20 cm</extent>
            <idno type=“UDC”>821.112.2(436)-3</idno>
          </bibl>
        </sourceDesc>
      </fileDesc>
      <profileDesc>
        <langUsage>
          <language ident="srp">Serbian</language>
        </langUsage>
      </profileDesc>
      <revisionDesc>
        <change>
          <date>2017-10-11</date>
          <name>Jelena Andonovski</name>
        </change>
      </revisionDesc>
    </teiHeader>
```

Прилог 9 – Упоредни приказ метаподатака

	COMARC/B	COMARC/XML	MARC/XML	DC	MODS	MARC 21	METS	TEI заглавље
Подаци о запису који се системски додељују[214]	ДА	ДА	ДА	/	/	ДА	ДА	ДА [215]
ISBN	ДА	ДА	ДА	ДА	ДА	ДА	ДА	ДА
Језик и писмо публикације	ДА	ДА	ДА	Језик публикације	Језик публикације	Језик публикације	Језик публикације	ДА
Превод и оригинал	ДА	ДА	ДА	/	ДА	ДА	/	ДА
Држава издавања	ДА	ДА	/	/	/	/	/	ДА
Наслов	ДА	ДА	ДА	ДА	ДА	ДА	ДА	ДА
Година издавања	ДА	ДА	ДА	ДА	ДА	ДА	ДА	ДА
Податак о штампању	ДА	ДА	ДА	ДА	ДА	ДА	ДА	ДА
Број страна и димензије	ДА	ДА	ДА	ДА	ДА	ДА	ДА	ДА
Податак о збирци	ДА	ДА	ДА	ДА	ДА	ДА	ДА	ДА
Напомене о преводу дела,	ДА	ДА	ДА	ДА	ДА	ДА	ДА	Напомена о преводу дела.

[214]Креатор записа, датум креирања записа, датум последње промене у запису, податак о кориснику који је последњи пут мењао запис, податак о врсти грађе која се описује

[215] Бележе се подаци о креатору записа и датум последње измене

додатном тексту и тиражу								**Напомена о настанку дигиталног објекта.**
Оригинални наслов	**ДА**	**ДА**	**ДА**	**ДА**	**ДА**	**ДА**	**ДА**	**ДА**
УДК класификација	**ДА**	**ДА**	**ДА**	**ДА**	**/**	**ДА**	**ДА**	**ДА**
Подаци о аутору	**ДА**	**ДА**	**ДА**	**ДА**	**ДА**	**ДА**	**ДА**	**ДА**
Подаци о сарадницима	**ДА**	**ДА**	**ДА**	**ДА**	**ДА**	**ДА**	**ДА**	**ДА**
Упутница на фонетски облик	**ДА**	**ДА**	**/**	**/**	**/**	**/**	**/**	**/**
Подаци о холдингу	**ДА**	**ДА**	**/**	**/**	**/**	**/**	**/**	**/**
Носилац ауторског права	**/**	**/**	**/**	**/**	**/**	**/**	**ДА**	**/**
Техничке карактеристике дигиталног објеката	**/**	**/**	**/**	**/**	**/**	**/**	**ДА**	**ДА**

Прилог 10а - Пример записа за Томаса Бернхарда у бази GND / корисничко окружење

Link zu diesem Datensatz	http://d-nb.info/gnd/118509861	
Person	Bernhard, Thomas	
Geschlecht	männlich	
Andere Namen	Tuo ma si Bo en ha de Tuomasi-Boenhade Boenhade, Tuomasi Bernhârd, Tômâs Bernard, Tômâs Berunharuto, Tōmasu Perŭnharŭt'ŭ, T'omasŭ Mperncharnt, Tomas Birnhārt, Tūmaš Bernhardi, Tʹomas Bernhard, Tomas Bernchard, Tomas	Bernhard, Nicolaas Thomas To ma seu Be leun ha leu teu Tomaseu-Beleunhaleuteu Beleunhaleuteu, Tomaseu ベルンハルト, トーマス (Schriftcode: Jpan) 托马斯·伯恩哈德 (Schriftcode: Hans) 伯恩哈德, 托马斯 (Schriftcode: Hans) 베른하르트, 토마스 (Schriftcode: Kore) תומס, ברנהרד (Schriftcode: Hebr)
Quelle	M LCAuth	
Zeit	Lebensdaten: 1931-1989	
Land	Niederlande (XA-NL); Österreich (XA-AT)	
Geografischer Bezug	Geburtsort: Heerlen Sterbeort: Gmunden	
Beruf(e)	Schriftsteller Dramatiker	
Funktion(en)	sonstige Person (s) ; Interpret (i) ; Textverfasser (Text)	
Instrumente/Vokalstimmen	Sprechstimme(n) (Sprechst.)	
Weitere Angaben	Österreichischer Schriftsteller, Dichter, Dramatiker, geboren in den Niederlanden; Georg-Büchner-Preisträger 1970	
Beziehungen zu Personen	Freumbichler, Johannes (Großvater) Stavianicek, Hedwig (Lebensgefährtin)	

Systematik	12.2p Personen zu Literaturgeschichte (Schriftsteller)
Typ	Person (piz)
Autor von	1167 Publikationen 1. *Autobiographische Schriften* *Bernhard, Thomas. - Salzburg : Residenz Verlag, [2019], [1. Auflage]* 2. *[Bernhard]* *Alte Meister* *Bernhard, Thomas. - Berlin : Suhrkamp, 2018, Erste Auflage* 3. ...
Interpret von	1 Publikation 1. *Die Macht der Gewohnheit (Bernhard). Gesamtaufnahme der Uraufführung. Komödie* *[S.l.] : Deutsche Grammophon, [1977?]*
Beteiligt an	69 Publikationen 1. *Der Stimmenimitator* *Leipzig : Deutsche Nationalbibliothek, 2018* 2. *Schuldiger* *Ostermaier, Albert. - Mattighofen : Korrektur Verlag, 2015, Erste Auflage dieser Ausgabe* 3. ...
Thema in	487 Publikationen 1. *Bernhard-Handbuch* *Stuttgart : J.B. Metzler, [2018], [1. Auflage]* 2. *[Marten]* *Bernhards Baukasten* *Marten, Catherine. - Berlin : De Gruyter, [2018], [1. Auflage]* 3. ...
Maschinell verknüpft mit	42 Publikationen 1. *Die ‚Antiautobiografie'. Eine Inszenierung der Fragwürdigkeit des Ichs* *Reimer, Madlen. - Bamberg : Otto-Friedrich-Universität Bamberg, 2018* 2. *"Einerseits die passende Hose / andererseits Richard den Dritten im Kopf" : Theater und Mode(rne) bei Thomas Bernhard und Elfriede Jelinek* *Schwieren, Alexander. - Frankfurt am Main : Universitätsbibliothek Johann Christian Senckenberg, 2018* 3. ...

Прилог 10б - Пример записа за Томаса Бернхарда у бази GND / формат RDF/Turtle

```
@prefix schema: <http://schema.org/> .
@prefix gndo: <http://d-nb.info/standards/elementset/gnd#> .
@prefix lib: <http://purl.org/library/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix editeur: <https://ns.editeur.org/thema/> .
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix umbel: <http://umbel.org/umbel#> .
@prefix rdau: <http://rdaregistry.info/Elements/u/> .
@prefix sf: <http://www.opengis.net/ont/sf#> .
@prefix bflc: <http://id.loc.gov/ontologies/bflc/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix vivo: <http://vivoweb.org/ontology/core#> .
@prefix isbd: <http://iflastandards.info/ns/isbd/elements/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix mo: <http://purl.org/ontology/mo/> .
@prefix marcRole: <http://id.loc.gov/vocabulary/relators/> .
@prefix dnba: <http://d-nb.info/standards/elementset/agrelon#> .
@prefix dcmitype: <http://purl.org/dc/dcmitype/> .
@prefix dbp: <http://dbpedia.org/property/> .
@prefix dnbt: <http://d-nb.info/standards/elementset/dnb#> .
@prefix madsrdf: <http://www.loc.gov/mads/rdf/v1#> .
@prefix dnb_intern: <http://dnb.de/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix v: <http://www.w3.org/2006/vcard/ns#> .
@prefix wdrs: <http://www.w3.org/2007/05/powder-s#> .
@prefix ebu: <http://www.ebu.ch/metadata/ontologies/ebucore/ebucore#> .
@prefix bibo: <http://purl.org/ontology/bibo/> .
@prefix gbv: <http://purl.org/ontology/gbv/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
<http://d-nb.info/gnd/118509861> wdrs:describedby <http://d-nb.info/gnd/118509861/about> .
<http://d-nb.info/gnd/118509861/about> dcterms:license <http://creativecommons.org/publicdomain/zero/1.0/> ;
        dcterms:modified “2018-12-11T14:48:41.000”^^xsd:dateTime .
<http://d-nb.info/gnd/118509861> a gndo:DifferentiatedPerson ;
        gndo:gndIdentifier “118509861” ;
        foaf:page <https://de.wikipedia.org/wiki/Thomas_Bernhard> ;
        owl:sameAs <http://dbpedia.org/resource/Thomas_Bernhard> , <http://viaf.org/viaf/12305044> ;
        gndo:oldAuthorityNumber “(DE-588a)118509861” , “(DE-588a)135804701” , “(DE-588a)134623584” , “(DE-
101c)310313589” , “(DE-588c)4005792-6” ;
        gndo:variantNameForThePerson “Tuo ma si Bo en ha de” ;
        gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433560 .
_:node1d2kbdtnkx433560 gndo:personalName “Tuo ma si Bo en ha de” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “Tuomasi-Boenhade” ;
        gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433561 .
_:node1d2kbdtnkx433561 gndo:personalName “Tuomasi-Boenhade” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “Boenhade, Tuomasi” ;
```

gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433562 .
_:node1d2kbdtnkx433562 gndo:forename “Tuomasi” ;
gndo:surname “Boenhade” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “Bernhaȓd, Tơmaŝ” ;
gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433563 .
_:node1d2kbdtnkx433563 gndo:forename “Tơmaŝ” ;
gndo:surname “Bernhaȓd” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “Bernard, Tơmaŝ” ;
gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433564 .
_:node1d2kbdtnkx433564 gndo:forename “Tơmaŝ” ;
gndo:surname “Bernard” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “Berunharuto, Tōmasu” ;
gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433565 .
_:node1d2kbdtnkx433565 gndo:forename “Tōmasu” ;
gndo:surname “Berunharuto” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “Perun̆harut̆'u̇, T'omasu̇” ;
gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433566 .
_:node1d2kbdtnkx433566 gndo:forename “T'omasu̇” ;
gndo:surname “Perun̆harut̆'u̇” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “Mpernchарnt, Tomas” ;
gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433567 .
_:node1d2kbdtnkx433567 gndo:forename “Tomas” ;
gndo:surname “Mpernchарnt” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “Birnhaīt, Tūmaŝ” ;
gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433568 .
_:node1d2kbdtnkx433568 gndo:forename “Tūmaŝ” ;
gndo:surname “Birnhaīt” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “Bernhardi, Tomas” ;
gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433569 .
_:node1d2kbdtnkx433569 gndo:forename “Tomas” ;
gndo:surname “Bernhardi” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “Bernhard, Tomas” ;
gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433570 .
_:node1d2kbdtnkx433570 gndo:forename “Tomas” ;
gndo:surname “Bernhard” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “Bernchard, Tomas” ;
gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433571 .
_:node1d2kbdtnkx433571 gndo:forename “Tomas” ;
gndo:surname “Bernchard” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “Bernhard, Nicolaas Thomas” ;
gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433572 .
_:node1d2kbdtnkx433572 gndo:forename “Nicolaas Thomas” ;
gndo:surname “Bernhard” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “To ma seu Be leun ha leu teu” ;
gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433573 .
_:node1d2kbdtnkx433573 gndo:personalName “To ma seu Be leun ha leu teu” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “Tomaseu-Beleunhaleuteu” ;
gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433574 .
_:node1d2kbdtnkx433574 gndo:personalName “Tomaseu-Beleunhaleuteu” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “Beleunhaleuteu, Tomaseu” ;
gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433575 .
_:node1d2kbdtnkx433575 gndo:forename “Tomaseu” ;
gndo:surname “Beleunhaleuteu” .
<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “ベルンハルト, トーマス” ;
gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433576 .
_:node1d2kbdtnkx433576 gndo:forename “トーマス” ;

gndo:surname “ベルンハルト” .

<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “托马斯·伯恩哈德” ;

gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433577 .

_:node1d2kbdtnkx433577 gndo:personalName “托马斯·伯恩哈德” .

<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “伯恩哈德, 托马斯” ;

gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433578 .

_:node1d2kbdtnkx433578 gndo:forename “托马斯” ;

gndo:surname “伯恩哈德” .

<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “베른하르트, 토마스” ;

gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433579 .

_:node1d2kbdtnkx433579 gndo:forename “토마스” ;

gndo:surname “베른하르트” .

<http://d-nb.info/gnd/118509861> gndo:variantNameForThePerson “תומס, ברנהרד” ;

gndo:variantNameEntityForThePerson _:node1d2kbdtnkx433580 .

_:node1d2kbdtnkx433580 gndo:forename “תומס” ;

gndo:surname “ברנהרד” .

<http://d-nb.info/gnd/118509861> gndo:preferredNameForThePerson “Bernhard, Thomas” ;

gndo:preferredNameEntityForThePerson _:node1d2kbdtnkx433581 .

_:node1d2kbdtnkx433581 gndo:forename “Thomas” ;

gndo:surname “Bernhard” .

<http://d-nb.info/gnd/118509861>

gndo:familialRelationship <http://d-nb.info/gnd/118832883> ;

dnba:hasGrandParent <http://d-nb.info/gnd/118832883> ;

gndo:familialRelationship <http://d-nb.info/gnd/12335420X> ;

gndo:professionOrOccupation <http://d-nb.info/gnd/4053309-8> , <http://d-nb.info/gnd/4140241-8> ;

gndo:gndSubjectCategory <http://d-nb.info/standards/vocab/gnd/gnd-sc#12.2p> ;

gndo:geographicAreaCode <http://d-nb.info/standards/vocab/gnd/geographic-area-code#XA-NL> , <http://d-nb.info/standards/vocab/gnd/geographic-area-code#XA-AT> ;

gndo:biographicalOrHistoricalInformation “Österreichischer Schriftsteller, Dichter, Dramatiker, geboren in den Niederlanden; Georg-Büchner-Preisträger 1970”@de ;

gndo:placeOfBirth <http://d-nb.info/gnd/4240795-3> ;

gndo:placeOfDeath <http://d-nb.info/gnd/4021376-6> ;

owl:sameAs <http://www.filmportal.de/person/1F32BF48822046F0BEC23C162E9F7122> ;

gndo:gender <http://d-nb.info/standards/vocab/gnd/gender#male> ;

gndo:dateOfBirth “1931-02-09”^^xsd:date ;

gndo:dateOfDeath “1989-02-12”^^xsd:date .

Прилог 11 - Пример записа за Томаса Бернхарда у бази VIAF / HTML

Bernhard, Thomas, 1931-1989
Bernhard, Thomas
Thomas Bernhard österreichischer Schriftsteller
ברנהרד, תומס, 1931-1989
Bernhardt, Thomas (1931-1989)
Bernhard, Thomas, 1931-
برنهارد، توماس، 1931-1989
VIAF ID: 12305044 (Personal)
Permalink: http://viaf.org/viaf/12305044

- Preferred Forms
- 4xx's: Alternate Name Forms (99)
- 5xx's: Related Names (33)
- Works
- Selected Co-authors
- Countries and Regions of Publication (40)
- Publication Statistics
- Selected Publishers (15)
- About
- Record Views

MARC-21 record
VIAF Cluster in XML
RDF record
Just Links in JSON

- History of VIAF ID:12305044 (45)

Прилог 12а - Пример записа за Томаса Бернхарда у бази LCNAF / корисничко окружење

Bernhard, Thomas
URI(s)
http://id.loc.gov/authorities/names/n50007084
Instance Of
MADS/RDF PersonalName
MADS/RDF Authority
SKOS Concept
Scheme Membership(s)
Library of Congress Name Authority File
Collection Membership(s)
Names Collection - Authorized Headings
LC Names Collection - General Collection
Variants
Berŭnharŭt‘ŭ, T‘omasŭ
Bernhard, Nicolaas Thomas
Berncharnt, Tomas
ברנהרד, תומס
トーマス・ベルンハルト
Fabian, Thomas
Additional Information
http://id.loc.gov/rwo/agents/n50007084
Birth Date
(edtf) 1931-02-09
Death Date
(edtf) 1989-02-12
Birth Place
Heerlen (Netherlands)
Death Place
Gmunden (Austria)
Gender
male
Associated Language
ger
Occupation
Author
Exact Matching Concepts from Other Schemes
http://viaf.org/viaf/sourceID/LC%7Cn+50007084#skos:Concept
Closely Matching Concepts from Other Schemes
Bernhard, Thomas
Earlier Established Forms
Bernhard, Thomas, 1931-1989
Sources
found: His Die Rosen der Einöde, 1959.
found: His Wittgenstein's nephew, 1990, c1988:CIP t.p. (Thomas Bernhard) pub. info. (d. Jan. 1989)

found: Fialik, M. Der konservative Anarchist, c1991:t.p. (Thomas Bernhard) p. 17, etc. (former Intendant of the Burgtheater Wien)
found: T‘omasŭ Berŭnharŭt‘ŭ yŏn’gu, 1996:t.p. (T‘omasŭ Berŭnharŭt‘ŭ) p. 267, etc. (Nicolaas Thomas Bernhard; b. Feb. 9, 1931; d. Feb. 12, 1989)
found: Beton, 1996:t.p. (Tomas Berncharnt [in Greek])
found: Info. converted from 678, 2012-10-02(b. 1931)
found: Wikipedia, German, via WWW, Jan. 7, 2013(Nicolas Thomas Bernhard; Austrian author; born Feb. 9, 1931 in Heerlen, Netherlands; died Feb. 12, 1989 in Gmunden, Austria; published first in 1950 under the pseudonym Thomas Fabian)
LC Classification
PT2662.E7
Editorial Notes
[Machine-derived non-Latin script reference project.]
[Non-Latin script references not evaluated.]
Change Notes
1980-03-24: new
2016-01-28: revised
Alternate Formats
RDF/XML (MADS and SKOS)
N-Triples (MADS and SKOS)
JSON (MADS/RDF and SKOS/RDF)
MADS - RDF/XML
MADS - N-Triples
MADS/RDF - JSON
SKOS - RDF/XML
SKOS - N-Triples
SKOS - JSON
MADS/XML
MARC/XML

Прилог 12б - Пример записа за Томаса Бернхарда у бази LCNAF / формат RDF/XML

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <madsrdf:PersonalName rdf:about="http://id.loc.gov/authorities/names/n50007084"
xmlns:madsrdf="http://www.loc.gov/mads/rdf/v1#">
    <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#Authority"/>
    <madsrdf:authoritativeLabel xml:lang="en">Bernhard, Thomas</madsrdf:authoritativeLabel>
    <madsrdf:elementList rdf:parseType="Collection">
     <madsrdf:FullNameElement>
           <madsrdf:elementValue xml:lang="en">Bernhard, Thomas</madsrdf:elementValue>
     </madsrdf:FullNameElement>
    </madsrdf:elementList>
    <madsrdf:hasVariant>
     <madsrdf:PersonalName>
           <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#Variant"/>
           <madsrdf:variantLabel xml:lang="en">Berŭnharŭt‘ŭ, T‘omasŭ</madsrdf:variantLabel>
           <madsrdf:elementList rdf:parseType="Collection">
            <madsrdf:FullNameElement>
             <madsrdf:elementValue xml:lang="en">Berŭnharŭt‘ŭ, T‘omasŭ</madsrdf:elementValue>
            </madsrdf:FullNameElement>
           </madsrdf:elementList>
     </madsrdf:PersonalName>
    </madsrdf:hasVariant>
    <madsrdf:hasVariant>
     <madsrdf:PersonalName>
           <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#Variant"/>
           <madsrdf:variantLabel xml:lang="en">Bernhard, Nicolaas Thomas</madsrdf:variantLabel>
           <madsrdf:elementList rdf:parseType="Collection">
            <madsrdf:FullNameElement>
             <madsrdf:elementValue xml:lang="en">Bernhard, Nicolaas Thomas</madsrdf:elementValue>
            </madsrdf:FullNameElement>
           </madsrdf:elementList>
     </madsrdf:PersonalName>
    </madsrdf:hasVariant>
    <madsrdf:hasVariant>
     <madsrdf:PersonalName>
           <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#Variant"/>
           <madsrdf:variantLabel xml:lang="en">Berncharnt, Tomas</madsrdf:variantLabel>
           <madsrdf:elementList rdf:parseType="Collection">
            <madsrdf:FullNameElement>
             <madsrdf:elementValue xml:lang="en">Berncharnt, Tomas</madsrdf:elementValue>
            </madsrdf:FullNameElement>
           </madsrdf:elementList>
     </madsrdf:PersonalName>
    </madsrdf:hasVariant>
    <madsrdf:hasVariant>
     <madsrdf:PersonalName>
           <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#Variant"/>
           <madsrdf:variantLabel xml:lang="en">תומס, ברנהרד</madsrdf:variantLabel>
           <madsrdf:elementList rdf:parseType="Collection">
```

```
            <madsrdf:FullNameElement>
              <madsrdf:elementValue xml:lang="en">תומס, ברנהרד</madsrdf:elementValue>
            </madsrdf:FullNameElement>
          </madsrdf:elementList>
    </madsrdf:PersonalName>
   </madsrdf:hasVariant>
   <madsrdf:hasVariant>
    <madsrdf:PersonalName>
          <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#Variant"/>
          <madsrdf:variantLabel xml:lang="en">トーマス・ベルンハルト</madsrdf:variantLabel>
          <madsrdf:elementList rdf:parseType="Collection">
            <madsrdf:FullNameElement>
              <madsrdf:elementValue xml:lang="en">トーマス・ベルンハルト</madsrdf:elementValue>
            </madsrdf:FullNameElement>
          </madsrdf:elementList>
    </madsrdf:PersonalName>
   </madsrdf:hasVariant>
   <madsrdf:hasVariant>
    <madsrdf:PersonalName>
          <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#Variant"/>
          <madsrdf:variantLabel xml:lang="en">Fabian, Thomas</madsrdf:variantLabel>
          <madsrdf:elementList rdf:parseType="Collection">
            <madsrdf:FullNameElement>
              <madsrdf:elementValue xml:lang="en">Fabian, Thomas</madsrdf:elementValue>
            </madsrdf:FullNameElement>
          </madsrdf:elementList>
    </madsrdf:PersonalName>
   </madsrdf:hasVariant>
   <madsrdf:classification>PT2662.E7</madsrdf:classification>
   <madsrdf:hasCloseExternalAuthority>
    <madsrdf:Authority rdf:about="http://id.worldcat.org/fast/1920">
          <madsrdf:authoritativeLabel>Bernhard, Thomas</madsrdf:authoritativeLabel>
    </madsrdf:Authority>
   </madsrdf:hasCloseExternalAuthority>
   <madsrdf:hasEarlierEstablishedForm>
    <madsrdf:PersonalName>
          <rdf:type rdf:resource="http://www.loc.gov/mads/rdf/v1#Variant"/>
          <madsrdf:variantLabel xml:lang="en">Bernhard, Thomas, 1931-1989</madsrdf:variantLabel>
          <madsrdf:elementList rdf:parseType="Collection">
            <madsrdf:FullNameElement>
              <madsrdf:elementValue xml:lang="en">Bernhard, Thomas,</madsrdf:elementValue>
            </madsrdf:FullNameElement>
            <madsrdf:DateNameElement>
              <madsrdf:elementValue xml:lang="en">1931-1989</madsrdf:elementValue>
            </madsrdf:DateNameElement>
          </madsrdf:elementList>
    </madsrdf:PersonalName>
   </madsrdf:hasEarlierEstablishedForm>
   <madsrdf:isMemberOfMADSCollection
rdf:resource="http://id.loc.gov/authorities/names/collection_NamesAuthorizedHeadings"/>
   <madsrdf:isMemberOfMADSCollection rdf:resource="http://id.loc.gov/authorities/names/collection_LCNAF"/>
   <madsrdf:hasExactExternalAuthority rdf:resource="http://viaf.org/viaf/sourceID/LC%7Cn+50007084#skos:Concept"/>
   <madsrdf:identifiesRWO>
    <madsrdf:RWO rdf:about="http://id.loc.gov/rwo/agents/n50007084">
          <rdf:type rdf:resource="http://id.loc.gov/ontologies/bibframe/Person"/>
          <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
```

<rdfs:label xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">Bernhard, Thomas</rdfs:label>
<madsrdf:birthDate>
<skos:Concept xmlns:skos="http://www.w3.org/2004/02/skos/core#">
<rdfs:label xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">(edtf) 1931-02-09</rdfs:label>
</skos:Concept>
</madsrdf:birthDate>
<madsrdf:deathDate>
<skos:Concept xmlns:skos="http://www.w3.org/2004/02/skos/core#">
<rdfs:label xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">(edtf) 1989-02-12</rdfs:label>
</skos:Concept>
</madsrdf:deathDate>
<madsrdf:birthPlace>
<madsrdf:Geographic>
<rdfs:label xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">Heerlen (Netherlands)</rdfs:label>
</madsrdf:Geographic>
</madsrdf:birthPlace>
<madsrdf:deathPlace>
<madsrdf:Geographic>
<rdfs:label xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">Gmunden (Austria)</rdfs:label>
</madsrdf:Geographic>
</madsrdf:deathPlace>
<madsrdf:occupation>
<madsrdf:Occupation>
<rdfs:label xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">Author</rdfs:label>
</madsrdf:Occupation>
</madsrdf:occupation>
<madsrdf:gender>
<skos:Concept xmlns:skos="http://www.w3.org/2004/02/skos/core#">
<rdfs:label xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">male</rdfs:label>
</skos:Concept>
</madsrdf:gender>
<madsrdf:associatedLanguage>
<madsrdf:Language rdf:about="http://id.loc.gov/vocabulary/languages/ger">
<madsrdf:code>ger</madsrdf:code>
</madsrdf:Language>
</madsrdf:associatedLanguage>
</madsrdf:RWO>
</madsrdf:identifiesRWO>
<madsrdf:isMemberOfMADSScheme rdf:resource="http://id.loc.gov/authorities/names"/>
<madsrdf:editorialNote>[Machine-derived non-Latin script reference project.]</madsrdf:editorialNote>
<madsrdf:editorialNote>[Non-Latin script references not evaluated.]</madsrdf:editorialNote>
<identifiers:lccn xmlns:identifiers="http://id.loc.gov/vocabulary/identifiers/">n 50007084</identifiers:lccn>
<identifiers:local xmlns:identifiers="http://id.loc.gov/vocabulary/identifiers/">(OCoLC)oca00042624</identifiers:local>
<madsrdf:hasSource>
<madsrdf:Source>
<madsrdf:citation-source>His Die Rosen der Einöde, 1959.</madsrdf:citation-source>
<madsrdf:citation-status>found</madsrdf:citation-status>
</madsrdf:Source>
</madsrdf:hasSource>
<madsrdf:hasSource>
<madsrdf:Source>
<madsrdf:citation-source>His Wittgenstein's nephew, 1990, c1988:</madsrdf:citation-source>
<madsrdf:citation-note xml:lang="en">CIP t.p. (Thomas Bernhard) pub. info. (d. Jan. 1989)</madsrdf:citation-note>
<madsrdf:citation-status>found</madsrdf:citation-status>
</madsrdf:Source>
</madsrdf:hasSource>

```
   <madsrdf:hasSource>
    <madsrdf:Source>
          <madsrdf:citation-source>Fialik, M. Der konservative Anarchist, c1991:</madsrdf:citation-source>
          <madsrdf:citation-note xml:lang="en">t.p. (Thomas Bernhard) p. 17, etc. (former Intendant of the Burgtheater
Wien)</madsrdf:citation-note>
          <madsrdf:citation-status>found</madsrdf:citation-status>
    </madsrdf:Source>
   </madsrdf:hasSource>
   <madsrdf:hasSource>
    <madsrdf:Source>
          <madsrdf:citation-source>T'omasŭ Berŭnharŭt'ŭ yŏn'gu, 1996:</madsrdf:citation-source>
          <madsrdf:citation-note xml:lang="en">t.p. (T'omasŭ Berŭnharŭt'ŭ) p. 267, etc. (Nicolaas Thomas Bernhard; b. Feb. 9,
1931; d. Feb. 12, 1989)</madsrdf:citation-note>
          <madsrdf:citation-status>found</madsrdf:citation-status>
    </madsrdf:Source>
   </madsrdf:hasSource>
   <madsrdf:hasSource>
    <madsrdf:Source>
          <madsrdf:citation-source>Beton, 1996:</madsrdf:citation-source>
          <madsrdf:citation-note xml:lang="en">t.p. (Tomas Bernchamt [in Greek])</madsrdf:citation-note>
          <madsrdf:citation-status>found</madsrdf:citation-status>
    </madsrdf:Source>
   </madsrdf:hasSource>
   <madsrdf:hasSource>
    <madsrdf:Source>
          <madsrdf:citation-source>Info. converted from 678, 2012-10-02</madsrdf:citation-source>
          <madsrdf:citation-note xml:lang="en">(b. 1931)</madsrdf:citation-note>
          <madsrdf:citation-status>found</madsrdf:citation-status>
    </madsrdf:Source>
   </madsrdf:hasSource>
   <madsrdf:hasSource>
    <madsrdf:Source>
          <madsrdf:citation-source>Wikipedia, German, via WWW, Jan. 7, 2013</madsrdf:citation-source>
          <madsrdf:citation-note xml:lang="en">(Nicolas Thomas Bernhard; Austrian author; born Feb. 9, 1931 in Heerlen,
Netherlands; died Feb. 12, 1989 in Gmunden, Austria; published first in 1950 under the pseudonym Thomas
Fabian)</madsrdf:citation-note>
          <madsrdf:citation-status>found</madsrdf:citation-status>
    </madsrdf:Source>
   </madsrdf:hasSource>
   <madsrdf:adminMetadata>
    <ri:RecordInfo xmlns:ri="http://id.loc.gov/ontologies/RecordInfo#">
          <ri:recordChangeDate rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">1980-03-
24T00:00:00</ri:recordChangeDate>
          <ri:recordStatus rdf:datatype="http://www.w3.org/2001/XMLSchema#string">new</ri:recordStatus>
          <ri:recordContentSource rdf:resource="http://id.loc.gov/vocabulary/organizations/dlc"/>
          <ri:languageOfCataloging rdf:resource="http://id.loc.gov/vocabulary/iso639-2/eng"/>
    </ri:RecordInfo>
   </madsrdf:adminMetadata>
   <madsrdf:adminMetadata>
    <ri:RecordInfo xmlns:ri="http://id.loc.gov/ontologies/RecordInfo#">
          <ri:recordChangeDate rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2016-01-
28T17:24:01</ri:recordChangeDate>
          <ri:recordStatus rdf:datatype="http://www.w3.org/2001/XMLSchema#string">revised</ri:recordStatus>
          <ri:recordContentSource rdf:resource="http://id.loc.gov/vocabulary/organizations/dlc"/>
          <ri:languageOfCataloging rdf:resource="http://id.loc.gov/vocabulary/iso639-2/eng"/>
    </ri:RecordInfo>
```

```
   </madsrdf:adminMetadata>
   <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
   <skos:prefLabel xml:lang="en" xmlns:skos="http://www.w3.org/2004/02/skos/core#">Bernhard, Thomas</skos:prefLabel>
   <skosxl:altLabel xmlns:skosxl="http://www.w3.org/2008/05/skos-xl#">
    <rdf:Description>
          <rdf:type rdf:resource="http://www.w3.org/2008/05/skos-xl#Label"/>
          <skosxl:literalForm xml:lang="en">Berŭnharŭt‘ŭ, T‘omasŭ</skosxl:literalForm>
    </rdf:Description>
   </skosxl:altLabel>
   <skos:altLabel xmlns:skos="http://www.w3.org/2004/02/skos/core#">Berŭnharŭt‘ŭ, T‘omasŭ</skos:altLabel>
   <skosxl:altLabel xmlns:skosxl="http://www.w3.org/2008/05/skos-xl#">
    <rdf:Description>
          <rdf:type rdf:resource="http://www.w3.org/2008/05/skos-xl#Label"/>
          <skosxl:literalForm xml:lang="en">Bernhard, Nicolaas Thomas</skosxl:literalForm>
    </rdf:Description>
   </skosxl:altLabel>
   <skos:altLabel xmlns:skos="http://www.w3.org/2004/02/skos/core#">Bernhard, Nicolaas Thomas</skos:altLabel>
   <skosxl:altLabel xmlns:skosxl="http://www.w3.org/2008/05/skos-xl#">
    <rdf:Description>
          <rdf:type rdf:resource="http://www.w3.org/2008/05/skos-xl#Label"/>
          <skosxl:literalForm xml:lang="en">Berncharnt, Tomas</skosxl:literalForm>
    </rdf:Description>
   </skosxl:altLabel>
   <skos:altLabel xmlns:skos="http://www.w3.org/2004/02/skos/core#">Berncharnt, Tomas</skos:altLabel>
   <skosxl:altLabel xmlns:skosxl="http://www.w3.org/2008/05/skos-xl#">
    <rdf:Description>
          <rdf:type rdf:resource="http://www.w3.org/2008/05/skos-xl#Label"/>
          <skosxl:literalForm xml:lang="en">ברנהרד, תומס</skosxl:literalForm>
    </rdf:Description>
   </skosxl:altLabel>
   <skos:altLabel xmlns:skos="http://www.w3.org/2004/02/skos/core#">ברנהרד, תומס</skos:altLabel>
   <skosxl:altLabel xmlns:skosxl="http://www.w3.org/2008/05/skos-xl#">
    <rdf:Description>
          <rdf:type rdf:resource="http://www.w3.org/2008/05/skos-xl#Label"/>
          <skosxl:literalForm xml:lang="en">トーマス・ベルンハルト</skosxl:literalForm>
    </rdf:Description>
   </skosxl:altLabel>
   <skos:altLabel xmlns:skos="http://www.w3.org/2004/02/skos/core#">トーマス・ベルンハルト</skos:altLabel>
   <skosxl:altLabel xmlns:skosxl="http://www.w3.org/2008/05/skos-xl#">
    <rdf:Description>
          <rdf:type rdf:resource="http://www.w3.org/2008/05/skos-xl#Label"/>
          <skosxl:literalForm xml:lang="en">Fabian, Thomas</skosxl:literalForm>
    </rdf:Description>
   </skosxl:altLabel>
   <skos:altLabel xmlns:skos="http://www.w3.org/2004/02/skos/core#">Fabian, Thomas</skos:altLabel>
   <skos:exactMatch rdf:resource="http://viaf.org/viaf/sourceID/LC%7Cn+50007084#skos:Concept"
xmlns:skos="http://www.w3.org/2004/02/skos/core#"/>
   <skos:closeMatch xmlns:skos="http://www.w3.org/2004/02/skos/core#">
    <rdf:Description rdf:about="http://id.worldcat.org/fast/1920">
          <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
          <skos:prefLabel>Bernhard, Thomas</skos:prefLabel>
    </rdf:Description>
   </skos:closeMatch>
   <skos:editorial xmlns:skos="http://www.w3.org/2004/02/skos/core#">[Machine-derived non-Latin script reference
project.]</skos:editorial>
```

```
   <skos:editorial xmlns:skos="http://www.w3.org/2004/02/skos/core#">[Non-Latin script references not
evaluated.]</skos:editorial>
   <skos:inScheme rdf:resource="http://id.loc.gov/authorities/names" xmlns:skos="http://www.w3.org/2004/02/skos/core#"/>
   <skosxl:altLabel xml:lang="en" xmlns:skosxl="http://www.w3.org/2008/05/skos-xl#">Berŭnharŭt‘ŭ, T‘omasŭ</skosxl:altLabel>
   <skosxl:altLabel xml:lang="en" xmlns:skosxl="http://www.w3.org/2008/05/skos-xl#">Bernhard, Nicolaas
Thomas</skosxl:altLabel>
   <skosxl:altLabel xml:lang="en" xmlns:skosxl="http://www.w3.org/2008/05/skos-xl#">Berncharnt, Tomas</skosxl:altLabel>
   <skosxl:altLabel xml:lang="en" xmlns:skosxl="http://www.w3.org/2008/05/skos-xl#">ברנהרד, תומס</skosxl:altLabel>
   <skosxl:altLabel xml:lang="en" xmlns:skosxl="http://www.w3.org/2008/05/skos-
xl#">トーマス・ベルンハルト</skosxl:altLabel>
   <skosxl:altLabel xml:lang="en" xmlns:skosxl="http://www.w3.org/2008/05/skos-xl#">Fabian, Thomas</skosxl:altLabel>
   <skos:changeNote xmlns:skos="http://www.w3.org/2004/02/skos/core#">
    <cs:ChangeSet xmlns:cs="http://purl.org/vocab/changeset/schema#">
          <cs:subjectOfChange rdf:resource="http://id.loc.gov/authorities/names/n50007084"/>
          <cs:creatorName rdf:resource="http://id.loc.gov/vocabulary/organizations/dlc"/>
          <cs:createdDate rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">1980-03-
24T00:00:00</cs:createdDate>
          <cs:changeReason rdf:datatype="http://www.w3.org/2001/XMLSchema#string">new</cs:changeReason>
    </cs:ChangeSet>
   </skos:changeNote>
   <skos:changeNote xmlns:skos="http://www.w3.org/2004/02/skos/core#">
    <cs:ChangeSet xmlns:cs="http://purl.org/vocab/changeset/schema#">
          <cs:subjectOfChange rdf:resource="http://id.loc.gov/authorities/names/n50007084"/>
          <cs:creatorName rdf:resource="http://id.loc.gov/vocabulary/organizations/dlc"/>
          <cs:createdDate rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2016-01-
28T17:24:01</cs:createdDate>
          <cs:changeReason rdf:datatype="http://www.w3.org/2001/XMLSchema#string">revised</cs:changeReason>
    </cs:ChangeSet>
   </skos:changeNote>
  </madsrdf:PersonalName>
</rdf:RDF>
```


Прилог 13 - Пример записа за Томаса Бернхарда у Википодацима / корисничко окружење

Thomas Bernhard (Q44336)

Austrian writer
In more languages

Language	Label	Description	Also known as
English	Thomas Bernhard	Austrian writer	
српски	No label defined	No description defined	
srpski	No label defined	No description defined	

Albanian	No label defined	shkrimtar austriak	

Statements

instance of	human
	2 references
image	
	
	Bernhardhaus094o.JPG
	1,235 × 1,010; 681 KB
	1 reference
sex or gender	male
	5 references
country of citizenship	Austria
	1 reference
birth name	Nicolaas Thomas Bernhard (German)
	0 references
given name	
	Thomas
	1 reference
	Nicolaas
	0 references
family name	Bernhard
	1 reference
pseudonym	Thomas Fabian
	0 references
date of birth	9 February 1931
	5 references
place of birth	Heerlen
	2 references
date of death	12 February 1989
	8 references
place of death	Gmunden
	2 references
place of burial	Grinzinger Friedhof
	1 reference
partner	Marianne Hoppe
	0 references
languages spoken, written or signed	German
	1 reference
occupation	
	writer

2 references
novelist
1 reference
playwright
1 reference
screenwriter
0 references
poet
0 references
author
0 reference

genre	play 0 references prose 0 references
award received	Anton Wildgans Prize point in time 1967 1 reference Prix Médicis for foreign literature point in time 1988 0 references Georg Büchner Prize point in time 1970 1 reference Feltrinelli Prize point in time 1987 0 reference
educated at	Mozarteum University Salzburg 1 reference
member of political party	Austrian People's Party 0 references
signature	 Thomas Bernhard (signature).jpg 240 × 56; 2 KB 1 reference
notable work	Correction 1 reference Extinction 1 reference Der Untergeher 1 reference Woodcutters 1 reference Heldenplatz 0 references
official website	http://www.thomasbernhard.at/ 0 references http://ausloeschung.virtusens.de/ 0 references http://www.thomasbernhard.org/ 0 references
Commons Creator page	Thomas Bernhard 0 references
Commons category	Thomas Bernhard 0 references

Identifiers

VIAF ID	12305044	3 references

ISNI	0000 0001 2120 7957
	1 reference
MusicBrainz artist ID	ed2967cf-4140-4a3b-a760-2ecf9dfa73a9
	1 reference
Library of Congress authority ID	n50007084
	2 references
GND ID	118509861
	1 reference
NDL Auth ID	00433110
	1 reference
Perlentaucher ID	thomas-bernhard
	0 references
Freebase ID	/m/041flc
	1 reference
BnF ID	118915601
	1 reference
CANTIC-ID	a10049769
	1 reference
BIBSYS ID	90080307
	1 reference
IMDb ID	nm0076767
	1 reference
National Thesaurus for Author Names ID	068468806
	1 reference
DBNL author ID	bern074
	0 reference
NKCR AUT ID	jn19990000738
	0 references
BVMC person ID	40713
	0 references
PORT person ID	118601
	0 references
ČSFD person ID	128891
	0 references
Gran Enciclopèdia Catalana ID	0240699
	0 references
National Library of Greece ID	60912
	0 references
SNAC Ark ID	w61g16zc
	0 references
Cultureel Woordenboek identifier	literatuur-internationaal/thomas-bernhard
0 references	
Great Russian Encyclopedia Online ID	1861269
	0 references
Babelio author ID	2693
	0 references
NE.se ID	thomas-bernhard
	0 references

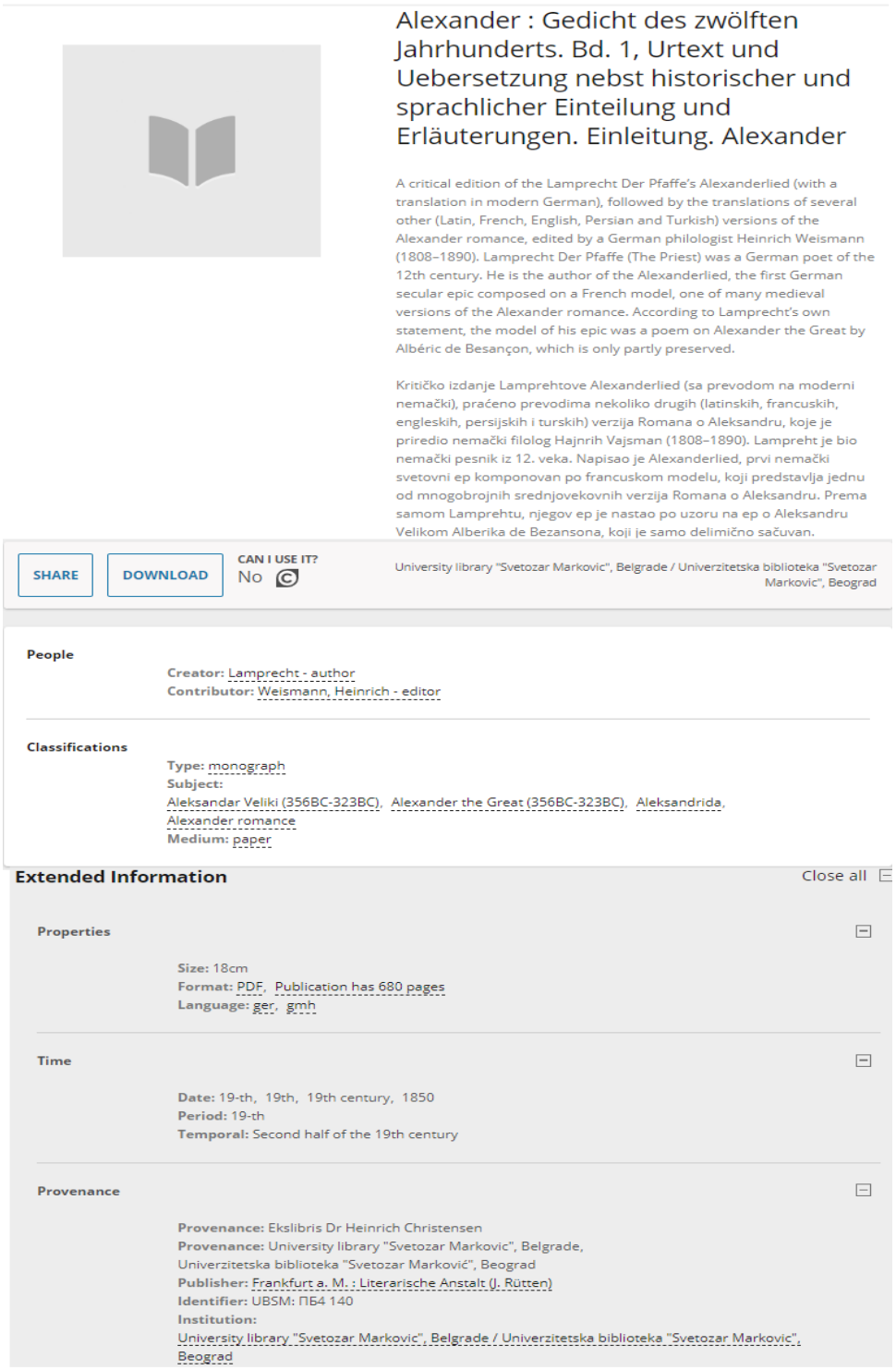
Munzinger IBA	00000012770
	0 references
KLG Kritisches Lexikon der Gegenwartsliteratur	16000000043
	0 references
SBN author ID	IT\ICCU\CFIV\012248
	0 references
Discogs artist ID	1371486
	1 reference
FAST ID	1920
	0 references
Open Library ID	OL4326320A
	0 references
BNE ID	XX843164
	0 references
Filmportal ID	1f32bf48822046f0bec23c162e9f7122
	1 reference
National Library of Israel ID	000019548
	0 references
	000605998
	0 references
	000610729
	0 references
	001671993
	0 references
Dialnet author ID	75355
	0 references
Encyclopædia Britannica Online ID	biography/Thomas-Bernhard
	0 references
PM20 folder ID	pe/001588
number of works	*unknown value*
number of works accessible online	0
0 references	
Enciclopédia Itaú Cultural ID	pessoa428625/thomas-bernhard
	0 references
SELIBR ID	178058
	0 references
Encyclopædia Universalis ID	thomas-bernhard
	0 references
PTBNP ID	27614
	0 references
gravsted.dk ID	thomasbernhard
	0 references
Flanders Arts Institute person ID	1877686
	0 references
CONOR ID	15336291
	0 references

SHARE Catalogue author ID	132758
	0 references
Poetry Foundation ID	thomas-bernhard
	0 references
GTAA ID	81913
	0 references
SUDOC authorities ID	028219260
1 reference	
Bibliothèque de la Pléiade ID	Thomas-Bernhard
	0 references
Libris-URI	gn244t381g818d8

	1 reference
Les Archives du Spectacle Person ID	727
	0 references
Adelphi author ID	131
	0 references
Bitraga author ID	3387
	0 references
WikiTree person ID	Bernhard-255
	0 references
University of Barcelona authority ID	a1372669
	0 references

Wikipedia (42 entries)

Прилог 14 - Запис у Европеани у EDM моделу / корисничко окружење



Alexander : Gedicht des zwölften Jahrhunderts. Bd. 1, Urtext und Uebersetzung nebst historischer und sprachlicher Einteilung und Erläuterungen. Einleitung. Alexander

A critical edition of the Lamprecht Der Pfaffe's Alexanderlied (with a translation in modern German), followed by the translations of several other (Latin, French, English, Persian and Turkish) versions of the Alexander romance, edited by a German philologist Heinrich Weismann (1808–1890). Lamprecht Der Pfaffe (The Priest) was a German poet of the 12th century. He is the author of the Alexanderlied, the first German secular epic composed on a French model, one of many medieval versions of the Alexander romance. According to Lamprecht's own statement, the model of his epic was a poem on Alexander the Great by Albéric de Besançon, which is only partly preserved.

Kritičko izdanje Lamprehtove Alexanderlied (sa prevodom na moderni nemački), praćeno prevodima nekoliko drugih (latinskih, francuskih, engleskih, persijskih i turskih) verzija Romana o Aleksandru, koje je priredio nemački filolog Hajnrih Vajsman (1808–1890). Lampreht je bio nemački pesnik iz 12. veka. Napisao je Alexanderlied, prvi nemački svetovni ep komponovan po francuskom modelu, koji predstavlja jednu od mnogobrojnih srednjovekovnih verzija Romana o Aleksandru. Prema samom Lamprehtu, njegov ep je nastao po uzoru na ep o Aleksandru Velikom Alberika de Bezansona, koji je samo delimično sačuvan.

SHARE DOWNLOAD CAN I USE IT? No

University library "Svetozar Markovic", Belgrade / Univerzitetska biblioteka "Svetozar Markovic", Beograd

People

Creator: Lamprecht - author
Contributor: Weismann, Heinrich - editor

Classifications

Type: monograph
Subject:
Aleksandar Veliki (356BC-323BC), Alexander the Great (356BC-323BC), Aleksandrida, Alexander romance
Medium: paper

Extended Information Close all

Properties

Size: 18cm
Format: PDF, Publication has 680 pages
Language: ger, gmh

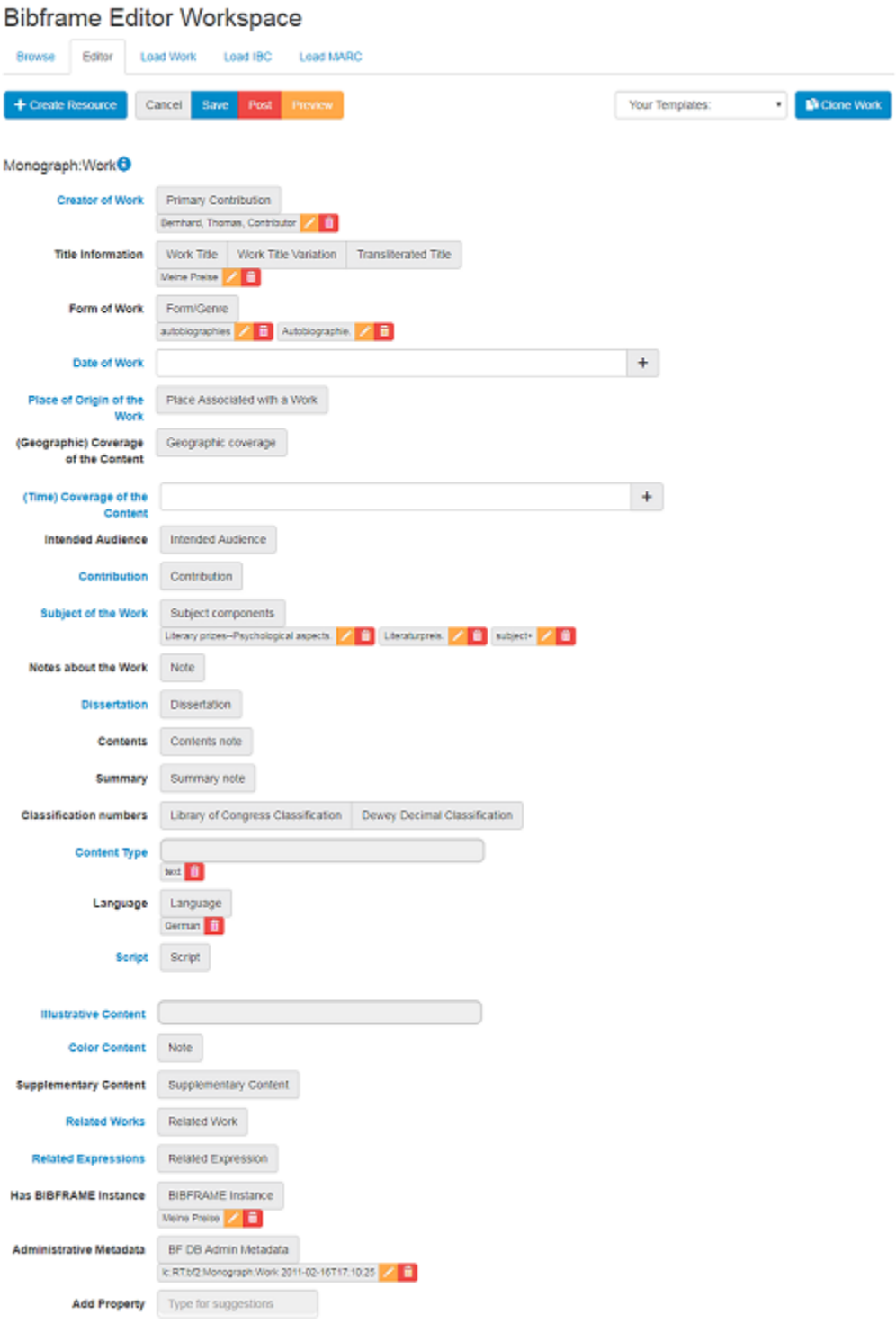
Time

Date: 19-th, 19th, 19th century, 1850
Period: 19-th
Temporal: Second half of the 19th century

Provenance

Provenance: Ekslibris Dr Heinrich Christensen
Provenance: University library "Svetozar Markovic", Belgrade, Univerzitetska biblioteka "Svetozar Marković", Beograd
Publisher: Frankfurt a. M. : Literarische Anstalt (J. Rütten)
Identifier: UBSM: П64 140
Institution:
University library "Svetozar Markovic", Belgrade / Univerzitetska biblioteka "Svetozar Markovic", Beograd

Прилог 15 - Пример записа за дело „Моје награде“ Томаса Бернхарда у моделу BIBFRAME / корисничко окружење



Биографија аутора

Јелена Андоновски рођена је 21. септембра 1987. године у Лесковцу. Дипломирала је на Катедри за библиотекарство и информатику 2010. године. На истој Катедри је новембра 2011. године одбранила мастер рад на тему *„Европеана – врата до европског културног наслеђа“*. Школске 2012/2013. уписала је докторске студије Језик, књижевност, култура на Филолошком факултету Универзитета у Београду, модул Култура.

Од децембра 2010. године ради у Универзитетској библиотеци „Светозар Марковић” у Београду. Звање вишег библиотекара стекла је 2016. године. Тренутно ради у Одељењу за обраду библиотечког материјала на пословима каталогизације и предметне и стручне класификације монографских публикација (стране и домаће књиге). Такође, део је тима за опис дигиталних колекција и објеката у погледу израде записа и доделе метаподатака у оквиру дигиталних збирки које креира, организује и одржава Универзитетска библиотека „Светозар Марковић” у оквиру домаћих и међународних пројеката дигитализације културне баштине. Учесник је и бројних домаћих и међународних конференција и радионица на којима је презентовала више радова.

Члан је Групе за језичке технологије Универзитета у Београду у оквиру које ради на припреми паралелних текстова и успостављању нових паралелних корпуса. Говори енглески и немачки, а служи се и шпанским језиком.

Изјаве о докторској дисертацији

Изјава о ауторству

Име и презиме аутора **Јелена С. Андоновски**

Број индекса **12024Д**

Изјављујем

да је докторска дисертација под насловом

Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора

У Београду, 12.07.2019

Андоновски Јелена

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора **Јелена Андоновски**

Број индекса **12024Д**

Студијски програм **Култура – Библиотекарство и информатика**

Наслов рада **Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса**

Ментор **проф. др Цветана Крстев**

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањена у **Дигиталном репозиторијуму Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис аутора

У Београду, 12.07.2019.

Андоновски Јелена

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Мрежа отворених података и језички ресурси у процесу изградње српско-немачког литерарног корпуса

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство (CC BY)

2. **Ауторство – некомерцијално (CC BY-NC)**

3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)

4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)

5. Ауторство – без прерада (CC BY-ND)

6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, 12.07.2019.

Предраг Јелић

1. Ауторство. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. **Ауторство – некомерцијално**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. Ауторство – некомерцијално – без прерада. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. Ауторство – некомерцијално – делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. Ауторство – без прерада. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. Ауторство – делити под истим условима. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.